Imputing Data That Are Missing at High Rates Using a Boosting Algorithm

Katherine Cauthen[1], Gregory Lambert[2], Jaideep Ray[3]
Sophia Lefantzi[3]
[1]Sandia National Laboratories, Albuquerque, NM 87158-5800
[2]Apple Inc., Cupertino, CA 95014
[3]Sandia National Laboratories, Livermore, CA 94550-0969

Abstract
Traditional multiple imputation approaches may perform poorly for datasets with high rates of missingness unless many *m* imputations are used. This paper implements an alternative machine learning-based approach to imputing data that are missing at high rates. We use boosting to create a strong learner from a weak learner fitted to a dataset missing many observations. This approach may be applied to a variety of types of learners (models). The approach is demonstrated by application to a spatiotemporal dataset for predicting dengue outbreaks in India from meteorological covariates. A Bayesian spatiotemporal CAR model is boosted to produce imputations, and the overall RMSE from a *k*-fold cross-validation is used to assess imputation accuracy.

Key Words: multiple imputation, machine-learning, boosting

1. Introduction

Missing data are common in disease surveillance studies, particularly in regions where public health infrastructure is limited, such as rural areas[1]. When data are missing at high rates, the challenge of imputing missing data accurately is amplified[2]. Machine learning approaches, such as boosting, may be advantageous in imputing data that are missing at high rates by leveraging the information gained by many weak learners to form one strong learner. We consider that one manifestation of a "weak" learner could be that it is trained on a dataset with many missing observations. An analogous approach has been shown to be effective for discrete data, so here we consider the continuous case[3]. The question we seek to answer is, does boosting improve imputation accuracy, and how does its efficacy change over rates of missingness?

1.1 Boosting

We chose to use Friedman's gradient boosting machine because it is one of the simplest implementations of boosting[4-7]. Assume we have a response, y, and we have a vector of predictors, $\boldsymbol{x}$. The goal is to estimate $\boldsymbol{x} \xrightarrow{f} \mathrm{y}$ with $\hat{f}(\boldsymbol{x})$ such that the expectation of the loss function $\Psi(y, f)$ is minimized:

$$\hat{f}(\boldsymbol{x}) = \underset{f(x)}{\overset{argmin}{}}\left(E_{y,\boldsymbol{x}}\Psi\left(y, \hat{f}(\boldsymbol{x})\right)\right) \quad .$$

[1] Corresponding author; kcauthe@sandia.gov

The procedure is as follows. Initialize $\hat{f}(x)$ to be constant. For t in $1, \ \dots, \ T$ do the following.

1. Compute the negative gradient
$$z_i = -\frac{\partial}{\partial f(\boldsymbol{x_i})} \Psi(y_i, f(\boldsymbol{x_i}))\Big|_{f(x_i)=\hat{f}(x_i)}$$
2. Fit model $g(\boldsymbol{x})$ that predicts z_i from $\boldsymbol{x_i}$
3. Choose a gradient descent step size as
$$\rho =_{\rho}^{\arg min} \sum_{i=1}^{N} \Psi(y_i, f(\boldsymbol{x_i}) + \rho g(\boldsymbol{x_i}))$$
4. Update the estimate of $f(x)$ as
$$\hat{f}(\boldsymbol{x}) \leftarrow \hat{f}(\boldsymbol{x}) + \rho g(\mathbf{x})$$

In this study, gradient boosting is equivalent to iteratively re-fitting the residuals of the model. We implemented this boosting procedure both in a simulation study and in an application to spatio-temporal data on dengue cases in India.

2. Methods

2.1 Simulation Study

2.1.1 Procedure

The goal of the simulation study was to determine if boosting resulted in more accurate imputations compared to a traditional, non-boosted approach. To do so we followed the procedure below.

1. Simulate data to have the following features:
$$x_{1i} = 0.05, \ 0.10, \ 0.15, \ \dots, \ 5.00$$
$$x_{2i} = \exp(0.25 * y_i) + \varepsilon_1$$
$$y_i = \sin(x) + 2\sin(x) + 6\sin(7x) + \ \varepsilon_2$$
$$\varepsilon_1 \sim N(0, \ 0.5)$$
$$\varepsilon_2 \sim N(0, \ 0.5)$$
$$i = 1, \dots, \ 100.$$
2. Set the rate of missingness to be either 5, 10, 20, 30, 40, 50, 60, or 70%.
3. Remove values of y_i according to a missing completely at random (MCAR) missingness mechanism. This means that each data point is equally likely to be missing, so there is no relationship between whether a data point is missing and any missing or observed values in the dataset.
4. Fit a model to the data without using boosting, make predictions (i.e. imputations) based on said model, and calculate the measure of performance.
5. Fit a model to the data using boosting, make predictions (i.e. imputations) based on said model, and calculate the measure of performance.
6. Repeat steps 3-5 k times to utilize a Monte Carlo procedure.
7. Calculate median performance across Monte Carlo samples
8. Repeat steps 2-7 for each rate of missingness.

Plots of the response versus each of the predictors are given in Figures 1 and 2. The response is periodically related to x_1 with some random error and logarithmically related to x_2 with independent random error as well.

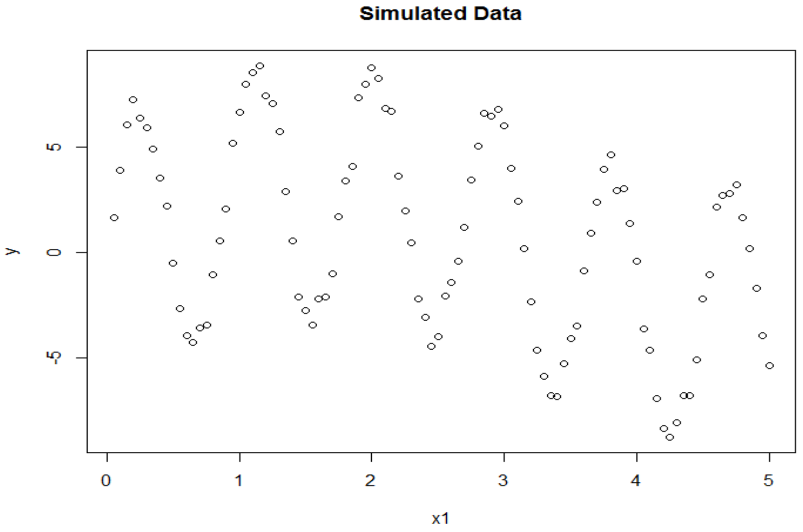


Figure 1. Plot of the simulated periodic relationship between the response and *x1*.

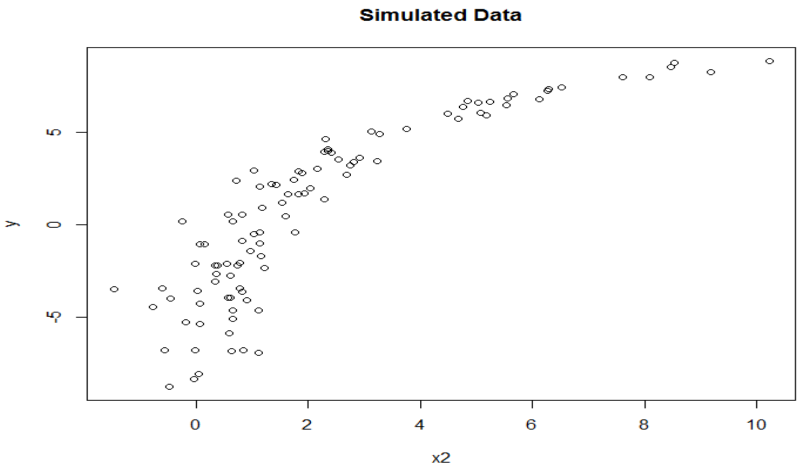


Figure 2. Plot of the simulated logarithmic relationship between the response and *x2*.

2.1.2 Boosting and performance assessment

The boosting algorithm will be specified to use a least-squares loss function where the learner was a first order autoregressive model with a p-spline for x_2. We used 100 boosting iterations and a learning rate of 0.05. The boosting was done using the `mboost` package in R.[8] We used *k* = 10,000 Monte Carlo samples, and performance was evaluated using the median relative root mean square error (rRMSE) over Monte Carlo samples, as specified below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m}(y_i \ - \hat{y}_i \ )^2}{m}}$$

$$rRMSE = 100 * \frac{RMSE}{\bar{y}}$$

where y_i is the observed value of y, $\hat{y}_i$ is the predicted value of y, $\bar{y}$ is the mean observed value of y, and m represents the number of missing values.

2.2 Application

2.2.1 Data

The observed data consisted of two predictor variables, temperature and precipitation, and one response variable, dengue case count proxy. In India, the monsoon season sweeps from the southwest to the northeast, and shortly after its arrival mosquitoes breed and they spread dengue fever. As such, there is variation in dengue outbreak onset times across India, and meteorological covariates are known to be predictive of dengue counts[9-23]. Temperature and precipitation measurements were obtained from MERRA (Modern Era Retrospective-analysis for Research and Applications) datasets.[24] The dengue case count proxy was obtained from HealthMap, an automated web-scraping tool for worldwide disease surveillance. We assume that the density of news reports regarding dengue in a given region, and uses this as a proxy for dengue cases.[25]

The dataset was comprised of monthly observations from August of 2011 – December of 2013 for each of 15 sub-selected Indian states, for a total of 435 observations that were spatially and temporally correlated. We constructed a neighborhood matrix of the states, where neighborhood was defined as sharing a border with one another. Across the entire dataset, 60.23% of the observations were missing, and we considered them to be missing at random (MAR). To assess the efficacy of using a boosting approach to imputing the missing data, we compared single imputations based on a spatio-temporal conditional autoregressive (STCAR) model to a boosted, single imputations based on the same model. We fitted the STCAR model using a Bayesian approach in the `CARBayesST` package in R.[26]

2.2.2 Bayesian STCAR model

The study region is a set of $k = 1, \dots, 15$ states $S = \{S_1, \dots, S_{15}\}$, with data recorded for each state for $t = 1, \dots, 29$ months. The dengue case count proxy is denoted by $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_{29})_{15x29}$ where $\mathbf{Y}_t = (Y_{1t}, \dots, Y_{15t})$ is the $15\ x\ 1$ column vector of observations for all 15 states for time period t. A vector of two known covariates, temperature and precipitation, for state k at time t is denoted by $\mathrm{x}_{kt} = (x_{kt1}, x_{kt2})$. The generalized linear mixed model that we fit is of the form

$$Y_{kt}|\mu_{kt} \sim f(y_{kt}|\mu_{kt}, v^2) \text{ for } k = 1, \dots, K, \quad t = 1, \dots, N,$$
$$\mu_{kt} = \mathrm{x}_{kt}^T \boldsymbol{\beta} + \Psi_{kt},$$
$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_\beta, \Sigma_\beta).$$

The vector of covariate regression parameters are given by $\boldsymbol{\beta} = (\beta_1, \beta_2)$ with a multivariate Gaussian prior with mean $\boldsymbol{\mu}_\beta$ and diagonal variance matrix Σ_β. A latent component for state k and time t is represented by Ψ_{kt}, and it includes a set of spatio-temporally autocorrelated random effects. Since we assume a Gaussian distribution of the dengue case counts, we specify $Y_{kt} \sim \mathrm{N}(\mu_{kt}, v^2)$ and $\mu_{kt} = \mathbf{x}_{kt}^{\mathrm{T}} \boldsymbol{\beta} + \Psi_{kt}$, where v^2 is the observation variance with an $\mathrm{Inverse} - \mathrm{Gamma}(1, 0.01)$ prior.

We chose to model the spatio-temporal structure with a multivariate first order autoregressive process with a spatially correlated precision matrix. This formulation was selected because it models a spatial response surface that is allowed to vary over time, as specified below.

$$\Psi_{kt} = \phi_{kt},$$
$$\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1} \sim \mathrm{N}(\rho_T\phi_{t-1}, \tau^2\mathbf{Q}(\mathbf{W}, \rho_S)^{-1}) \qquad t = 2, \dots, 29,$$
$$\boldsymbol{\phi}_1 \sim \mathrm{N}(\mathbf{0}, \tau^2\mathbf{Q}(\mathbf{W}, \rho_S)^{-1}),$$
$$\tau^2 \sim \mathrm{Inverse} - \mathrm{Gamma}(1, 0.01),$$
$$\rho_S\rho_T \sim \mathrm{Uniform}(0,1).$$

Here, $\boldsymbol{\phi}_t = (\phi_{1t}, \dots, \phi_{15t})$ is the vector of random effects for time t, and they are allowed to evolve over time according to a multivariate first order autoregressive process that has a temporal autoregressive parameter ρ_T. This means that the temporal autocorrelation is implemented by the mean $\rho_T\boldsymbol{\phi}_{t-1}$, and the spatial autocorrelation is implemented by the variance $\tau^2\boldsymbol{Q}(\boldsymbol{W}, \rho_S)^{-1}$. The precision matrix $\boldsymbol{Q}(\boldsymbol{W}, \rho_S)$is specified as

$$\mathbf{Q}(\mathbf{W}, \rho_S) = \rho_S[\mathrm{diag}(\mathbf{W1}) - \mathbf{W}] + (1 - \rho_S)\mathbf{I},$$

where **I** is the 15x15 identity matrix. The random effects have mean zero, the temporal and spatial autocorrelation parameters have uninformative priors, and τ^2 has a conjugate prior.

2.2.3 Boosting and performance assessment

Since the `mboost` package is not flexible enough to accommodate a STCAR model for a base learner we wrote our own boosting algorithm. The boosting algorithm specified was identical to that used in the simulation study except for the base learner, which was specified as described above. Performance was evaluated using the root mean square error according to the following non-exhaustive, leave-3-out cross-validation procedure.

1. For each of states *k* = 1, … , 14
    a) Hold out Oct 2013 (*t*= 1), Nov 2013 (*t* = 2), and Dec 2013 (*t* = 3)
    b) Get imputations by given method
    c) Find RMSE

    $$RMSE_k = \sqrt{\frac{\sum_{t=1}^{3}(y_{tk} - \widehat{y_{tk}})^2}{N_k}}$$

    where y_{tk} is the observed dengue incidence for month *t* and state *k*, $\widehat{y_{tk}}$ is the imputed dengue incidence for month *t* and state *k*, and N_k is the number of hold-out observations y_{ik} that are not missing for state *k*.
2. Find overall RMSE. This is the accuracy metric.

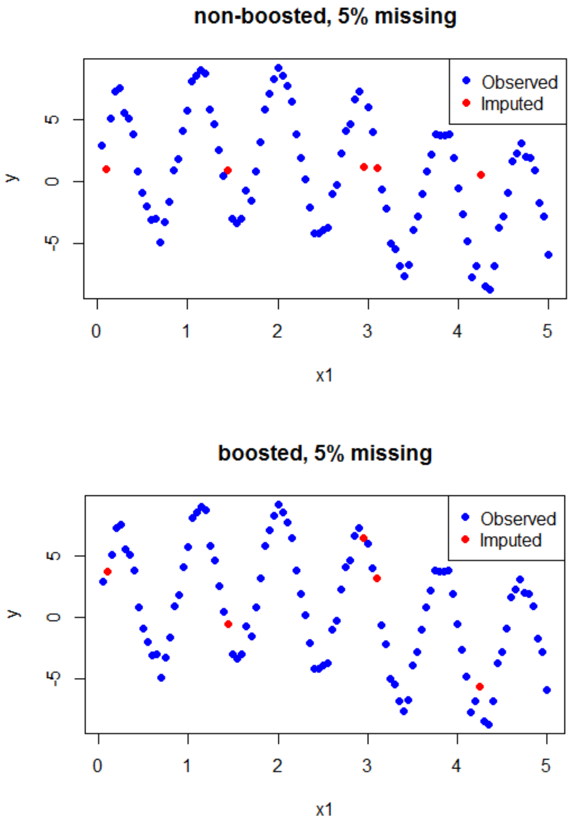
$$RMSE_{overall} = \sqrt{\frac{RMSE_1^2 + RMSE_2^2 + \cdots + RMSE_{15}^2}{14}}$$

Note that only 14 states were used in the validation procedure, since the state of Bengal was missing data on all three hold-out months.

3. Results

3.1 Simulation Study

For the simulation study we compared the imputation accuracy of a traditional, non-boosted approach and a boosted approach across varying rates of missingness, using a first order autoregressive model with a p-spline for the second covariate. We repeated this procedure for 10,000 Monte Carlo samples and calculated the performance metric, median rRMSE over the samples. Figure 3 shows the observed and imputed values of the non-boosted and boosted settings for one sample at a 5% rate of missingness. Figure 4 shows a similar graph for a 70% rate of missingness. For this sample, the imputed values follow the periodic pattern in the data better in the boosted setting compared to the non-boosted setting for both the 5% and 70% missing cases.


non-boosted, 5% missing

boosted, 5% missing

Figure 3. Observed and imputed values for 5% missing data in the non-boosted and boosted settings.

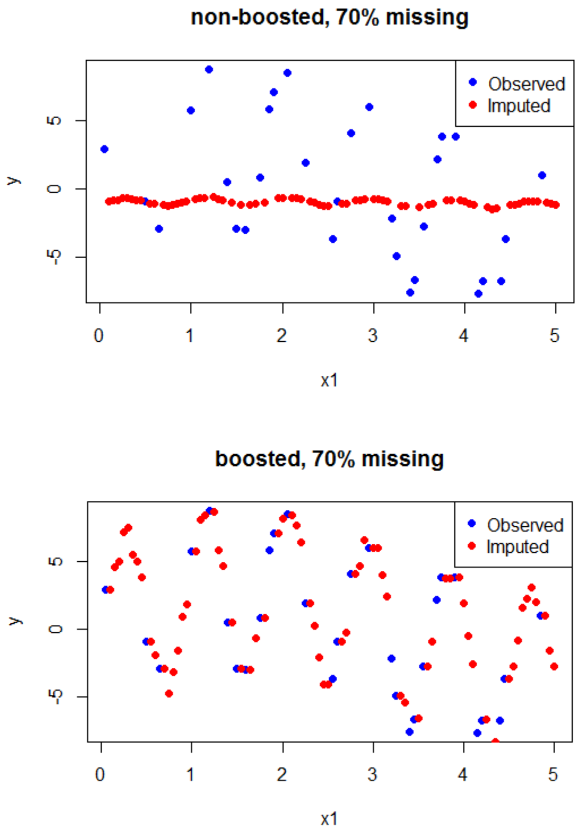


Figure 4. Observed and imputed values for 70% missing data in the non-boosted and boosted settings.

The same pattern holds across all Monte Carlo samples (Figure 5). The boosted imputations result in errors that are smaller compared to the non-boosted imputations, even at very high rates of missingness.

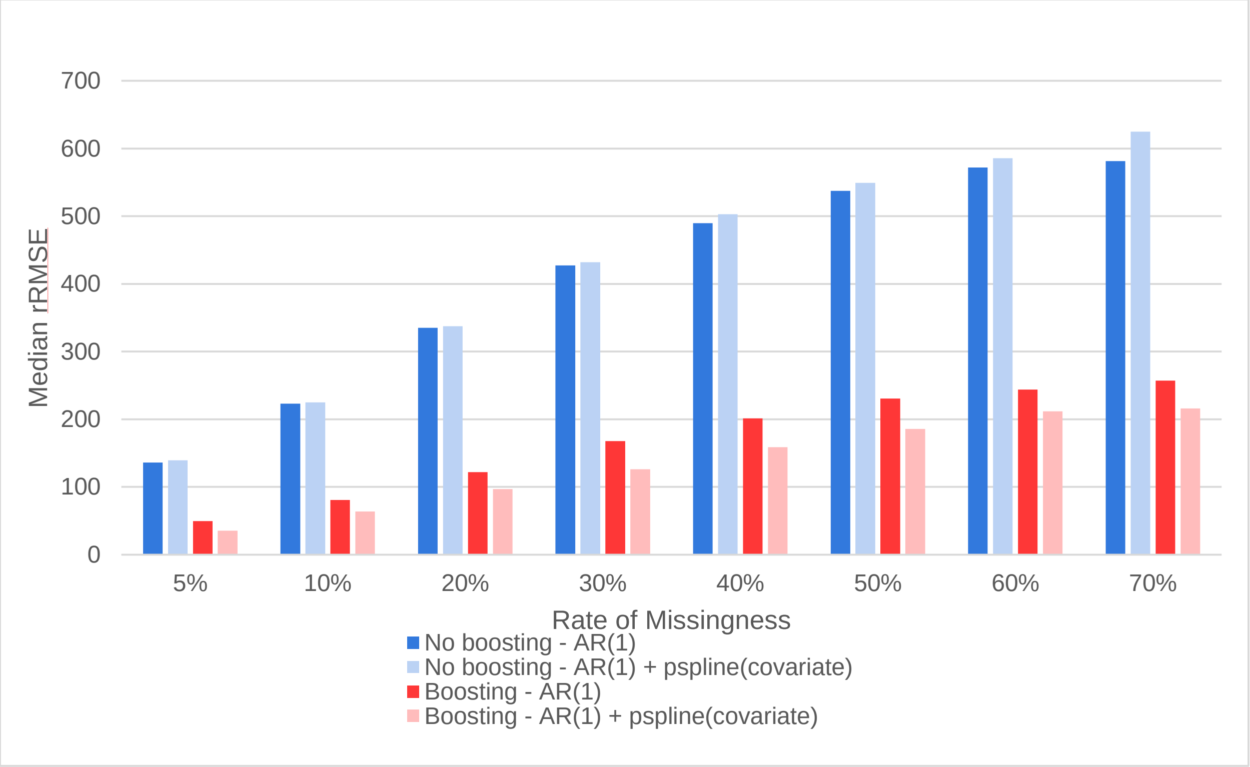


Figure 5. Results of the simulation study. Median rRMSE for the boosted and non-boosted settings across rates of missingness.

3.2 Application

For the dengue application, imputation accuracy was compared between a traditional, non-boosted approach and a boosted approach, where the base learner was an STCAR model. The resultant RMSEs from the cross-validation procedure are given in Figure 6. Overall, the boosted setting performs better than the non-boosted setting by a factor of nearly two. We observe variability in the relative performance of the non-boosted and boosted settings across states, but we were unable to identify the pattern responsible for this variability. It does not appear that these differences in relative performance are related to amount of missingness or to extremity of the values we were trying to predict.

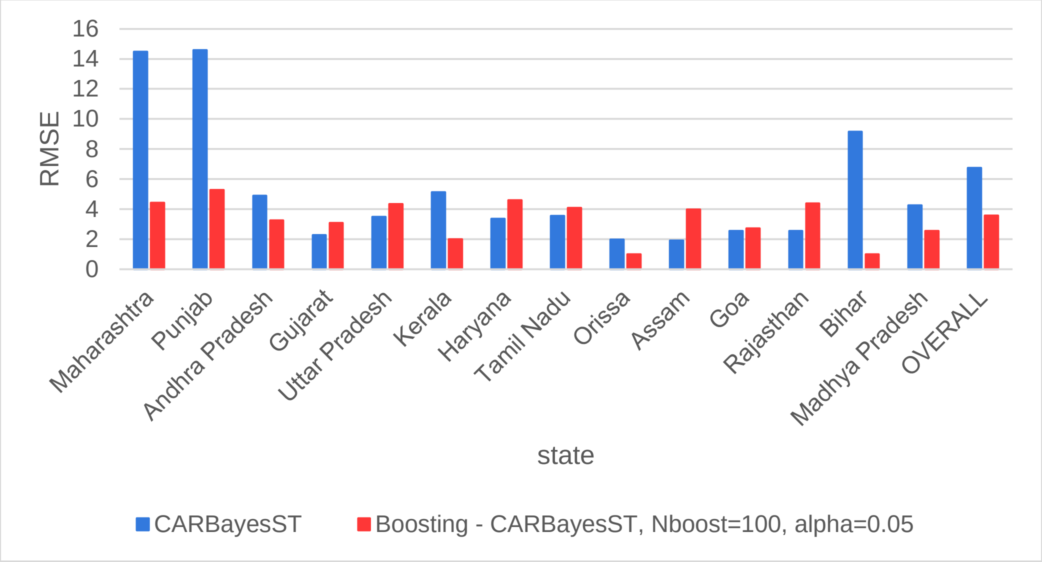


Figure 6. Results of the dengue application. RMSE for the boosted and non-boosted settings across states and overall.

A comparison between the observed data set with missing values and the resultant fully-imputed data set using boosting are shown in Figures 7 and 8. The seasonal trends are expected given the seasonal nature of dengue outbreaks over the course of the year. Likewise, the time series for the states are not fully aligned with one another due to the geographically-varying monsoon onset times across India.

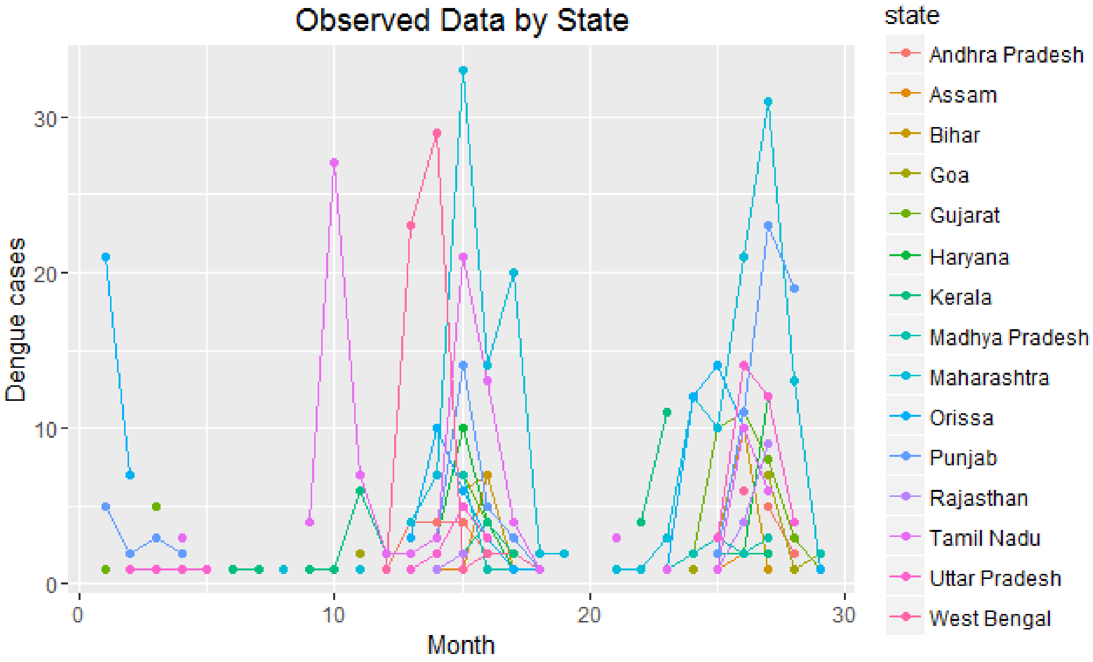


Figure 7. Observed dengue case counts by state.

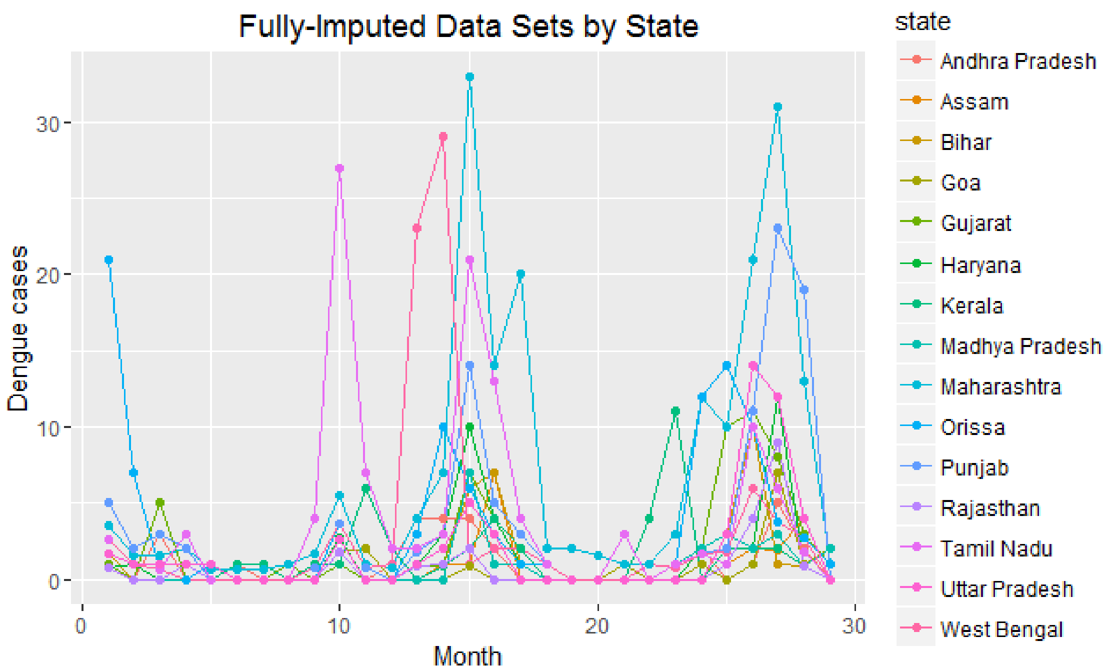


Figure 8. Observed and imputed dengue case counts by state.

4. Conclusions

We sought to determine if a boosting approach to imputing missing data could out-perform the traditional, single imputation method. To do so, we implemented a simulation study that compared boosted and non-boosted settings across varying rates of missingness for periodic data. We also tested the boosted imputation methodology in an application for geospatio-temporal dengue fever data in India. Boosted models result in improved imputation compared to non-boosted models, both in the simulation study and the application. The performance metrics were about 50% better in the boosted setting. These findings held even when the rate of missingness was very high.

The boosted imputation methodology is a promising one for future applications. It could be applied to a dataset with virtually any structure, and it does not necessarily require parametric assumptions, depending upon the learner chosen. Additionally, this methodology can use information from all cases, not just those that are complete.

One limitation of the boosting imputation methodology is that the time required to obtain boosted imputations is linearly related to the number of boosting iterations specified. If a base learner model takes a substantial amount of time to fit, then boosting it could potentially be memory and time consuming. More work is required to ascertain the robustness of these methods under various data conditions such as missingness mechanism, data structure, distribution, and missingness in multiple variables. Future efforts should focus on assessment of the robustness of the method and potential improvements that might be made by using adaptive boosting algorithms.

Acknowledgements

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

1. World Health Organization. (2006). Communicable disease surveillance and response
2. Dong, Y., Peng, C.Y.J. (2013). Principled missing data methods for researchers. *SpringerPlus 2*, 222.
3. Wang, C.Y., Feng, Z. (2010). Boosting with missing predictors. *Biostatistics 11*, 2, 195-212.
4. Freund, Y., Schapire, R.E. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences, 55,* 11, 119-139.
5. Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*, 5, 1189-1232.
6. Friedman, J.H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis, 38*, 4, 367-378.
7. Friedman, J.H., Hastie, T., Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics, 28*, 2, 337-374.
8. Hothorn, T., Buehlmann, T., Kneib, M., Schmid, M., Hofner, B. mboost: Model-based boosting, Rpackage version 2.6-0. 2016. http://CRAN.R-project.org/package=mboost
9. Aggarwal, P., Gupta, P., Kandpal, S.D., Kakati, B., & Gupta, D. (2014). Post-monsoon season surveillance a must for curtailing annual dengue epidemic in rural India. *National Journal of Community Medicine, 5*, 1, 153-155.
10. Chakravarti, A., & Kumaria, R. (2005). Eco-epidemiological analysis of dengue infection during an outbreak of dengue fever, India. *Virology Journal, 2.*
11. Chandy, S., Ramanathan, K., Manoharan, A., Mathai, D., & Baruah, K. (2013). Assessing effect of climate on the incidence of dengue in Tamil Nadu. *Indian Journal of Medical Microbiology, 31,* 3.
12. Dhiman, R. C., Pahwa, S., Dhillon, G. P., & Dash, A. P. (2010). Climate change and threat of vector-borne diseases in India: are we prepared? *Parasitology Research, 106,* 4, 763-73.
13. Durani, K., Dund, J., Shingala, H., & Sinha, M. (2014). Epidemiological trend analysis of dengue virus infection in western part of Gujarat. *Indian Journal of Research, 3*, 6, 146-148.
14. Jeelani, S. & Sabesan, S. (2013). Aedes vector population dynamics and occurrence of dengue fever in relation to climate variables in Puducherry, South India. *International Journal of Current Microbiology and Applied Sceinces, 2*, 12, 313-322.
15. Joshi, V., Mathur, M. L., Dixit, A. K., & Singhi, M. (1996). Entomological studies in a dengue endemic area, Jalore, Rajasthan. *The Indian Journal of Medical Research, 104,* 161-5.
16. Kaup, S. & Sankarankutty, J. (2014). Seroprevalence and seasonal trend of dengue virus infection at a teaching hospital in Tumkur, India. *Scholars Journal of Applied Medical Sciences, 2*, 3A, 922-926.

17. Pandey, N., Nagar, R., Gupta, S., Omprakash, Kahn, D., Singh, D., Mishra, G., Prakash, S., Singh, K.P., Singh, M., & Jain, A. (2012). Trend of dengue virus infection at Lucknow, north India (2008-2010): A hospital-based study. *Indian Journal of Medical Research, 136*, 5, 862-867.
18. Pruthvi, D., Shashikala, P., & Shenoy, V. (2012). Evaluation of platelet count in dengue fever along with seasonal variation of dengue infection. *Journal of Blood Disorders and Transfusion, 3*, 4.
19. Raheel, U., Faheem, M., Riaz, M. N., Kanwal, N., Javed, F., Zaidi, N., & Qadri, I. (2011). Dengue fever in the Indian Subcontinent: an overview. *Journal of Infection in Developing Countries, 5,* 4, 239-47.
20. Ram, S., Khurana, S., Kaushal, V., Gupta, R., & Khurana, S. B. (1998). Original Articles - Incidence of dengue fever in relation to climatic factors in Ludhiana, Punjab. *The Indian Journal of Medical Research, 108,* 4, 128.
21. Rao, M.R.K. & Padhy, R.N. (2014). Prevalence of dengue viral infection (DI) in and around of Angul district of Odisha, India: A comprehensive eco-epidemiological study. *International Organization of Scientific Research - Journal of Pharmacy and Biological Sciences, 9,* 3, 55-64.
22. Ratho, R. K., Mishra, B., Kaur, J., Kakkar, N., & Sharma, K. (2005). An outbreak of dengue fever in periurban slums of Chandigarh, India, with special reference to entomological and climatic factors. *Indian Journal of Medical Sciences, 59,* 12, 518-26.
23. Sankari, T., Hoti, S., Singh, T., Shanmugavel, J. (2012). Outbreak of dengue virus serotype-2 (DENV-2) of Cambodian origin in Manipur, India - Association with meteorological factors. *Indian Journal Of Medical Research, 136*, 4, 649-655.
24. Modern Era Retrospective-analysis for Research and Applications (MERRA). Goddard Earth Sciences Data and Information Sciences Center, http://disc.sci.gsfc.nasa.gov/mdisc/data-holdings
25. HealthMap. Dengue. http://www.healthmap.org/dengue/
26. Lee, D., Rushworth, A., Napier, G. CARBayesST: Spatio-temporal generalised linear mixed models for areal unit data. R package version 2.4. 2016. https://CRAN.R-project.org/package=CARBayesST