Identifying Finest Machine Learning Algorithm for Climate Data Imputation in the State of Minas Gerais, Brazil

Lucas O. Bayma, Marconi A. Pereira

Departamento de Tecnologia e Eng. Civil, Computação e Humanidades – DTECH
Universidade Federal de São João Del Rei - Campus Alto Paraopeba
MG 443, KM 7 Ouro Branco – MG – Brazil lucasobayma@gmail.com, marconi@ufsj.edu.br

Abstract. Climate prediction is a relevant activity for humanity and, for the success of the climate forecast, a good historical database is necessary. However, because of several factors, large historical data gaps are found at different meteorological stations, and studies to determine such missing weather values are still scarce. This work describes a study of a combination of several machine learning techniques to determine missing climatic values. This study extends our previous work, producing a computational framework, formed by three different methods: neural networks, regression bagged trees and random forest. Deep data analysis and a statistical study is conducted to compare these three methods. The study statistically demonstrated that the random forest technique was successful in obtaining missing climatic values for the state of Minas Gerais and can be widely used by the responsible agencies to improve their historical databases, consequently, their climate forecasts.

Categories and Subject Descriptors: G.1 [**Numerical Analysis**]: Interpolation; G.3 [**Probability and Statistics**]: Imputation; I.6 [**Simulation and Modeling**]: Miscellaneous

1. INTRODUCTION

An important task to better study and predict weather is the storage of historical data. The governments and industries that are affected by the weather must store time series of climate data. This historical data can feed forecast models, increasing the accuracy of the forecast. The measurement of time series allows the identification of cycles and patterns repeated over time, in such a way that, if properly combined with the current observational data, they can help in the task of predicting and validating future data.

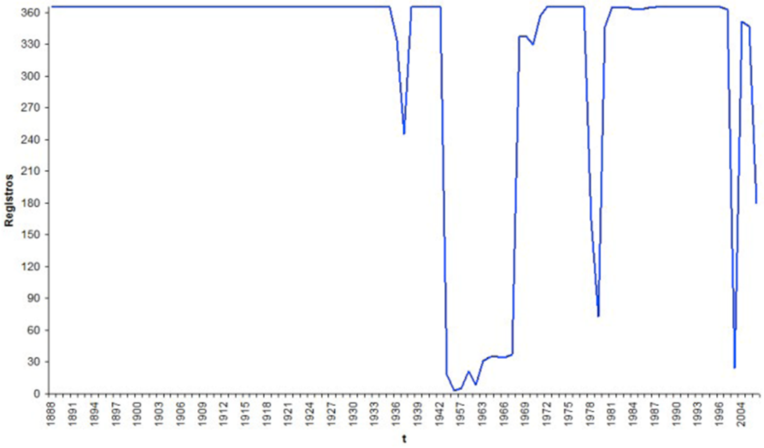


Fig. 1: Existence of precipitation data between 1888 and 2006, Estação da Luz, São Paulo city.

The database division of CPTEC/INPE[1] has an important role in the collection and storage of climate data. Particularly, there is a large body of observational data [Barbosa and Carvalho 2015] such as precipitation (since 1880). On the other hand, the historical series of these data are not always continuous and there may be momentary interruptions caused by different reasons.

Figure 1 ([Barbosa and Carvalho 2015]) shows a set of data measured at the Estação da Luz, in São Paulo city, between the years 1888 and 2006. A significant interruption was noted in the 1940s, 1950s and 1960s. These missing data are relevant for the historical series and can be inferred from other context-related attributes [Lakshminarayan et al. 1999].

Mechanisms of missing data relies into three categories [Enders 2010]: Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR).

(1) *Missing Completely At Random (MCAR).* It occurs when the probability of an instance having a missing value on some variable X is independent of the variable itself and of the values of any other variables in the dataset. In other words, no systematic differences exist between participants with missing data and those with complete data. Typical examples of MCAR, are when the missing elements in the data matrix X are located completely at random, or when some laboratory values are missing because a batch of lab samples was processed improperly. Possible reasons for MCAR include manual data entry procedure, incorrect measurements, equipment error, changes in experimental design, etc.

(2) *Missing At Random (MAR).* It occurs when the missing data is systematically related to the observed but not with the unobserved data. Whether an observation is missing has nothing to do with the missing values, but it does have to do with the values of an individual's observed variables. For example, if the temperature data gaps occur on different days throughout the week, it is MAR. If the temperature data is missing only on weekends, it is a MNAR variable.

(3) *Missing Not At Random (MNAR).* It occurs when the probability of an instance having a missing value on some variable X depends on the variable itself, that is, the missingness is related to events or factors which are not measured by the researcher. MNAR is called "non-ignorable" because the missing data mechanism itself has to be modeled as you deal with the missing data. This is the case where the lower temperatures from a meteorological station are missing.

Over time, several tools have been applied in order to identify these missing values [Gilat and Subramaniam 2009], studying the imputation methods in different mechanisms of missing data [Sefidian and Daneshpour 2019]. Several approaches have been proposed and improved in this context, such as algorithms based on artificial neural networks [Luengo et al. 2010; Olcese et al. 2015; Singh 2016], decision trees [Valdiviezo and Van Aelst 2015], support vector machines [García-Laencina et al. 2015; Sapankevych and Sankar 2009], and recent machine learning approaches, such as bagged trees [Hegde et al. 2015] and random forest [Tang and Ishwaran 2017]. Studies of how the prediction of new values can be improved, when the identification of missing values is compared with different rates of missing data [Sefidian and Daneshpour 2019].

Within this perspective, this article presents a framework for the study of several tools and techniques of machine learning for the imputation of *MCAR* missing data with missingness of 25% in climatic time series in order to better predict the tendency data. The framework was made to allow a cross-validation between the models. This validation is important for verifying the effectiveness of missing data imputation for predicting new values.

This work is an extension of [Bayma and Pereira 2017], and proposes a more complex framework. The previous framework was composed of multiple linear regression (MLR), support vector machine (SVM) neural networks (NN) and bagged regression trees (BRT). The current work keeps the algorithms that performed better in the identification of the missing data task, neural network (**NN**),

[1] http://www.cptec.inpe.br/

bagged regression trees (**BT**), adding random forest algorithm (**RF**) to the framework. Since more works within missing data imputation area have been released, e.g. [Tang and Ishwaran 2017; Yang and Hu 2018; Sefidian and Daneshpour 2019; Jordanov et al. 2018], it is necessary to update the work and include the random forest algorithm to both missing data imputation and new data prediction. To better understand how the missing data imputation is correlated to the final model performance, this article have a data analysis more robust, using a box and whisker plot for data visualization and a heat map, to support the previous statistical study, which is conduced by the end of it.

This work is structured as follows. Section 2 presents the related works. The Datasets are described in section 3, along with preliminary data processing and analysis. Section 4 discusses the regression methods presented in the proposed framework. The quality measurement of the imputed data is discussed in section 5, along with the study design for the comparison. The results of the comparison are presented in section 6. Finally, the conclusions are presented in section 7.

2. RELATED WORK

The imputation of missing data has great importance in many areas of industry and governmental researches. The inappropriate use of missing data can lead to a biased result, therefore in machine learning there is a growing sub-area that aims to study techniques and models for the identification of missing data and a lot of recent research has focused on improvements in data imputation. [Sefidian and Daneshpour 2019] makes a study in how the identification of data is different for MAR and MNAR missing values. Using a well know dataset from the researches, they studied the impact of different missingness in the data. Their proposed imputation method had better results when dealing with higher rates of missing data.

[Wei et al. 2018] proposed a missing value imputation approach for a real missing data problem, medical research using metabolomics data. The study compared eight imputation methods, both statistical (median, zero, half minimun) and machine learning methods (random forest, k-nearest neighbors, quantile regression imputation of left-censored data) for different types of missing values. Normalized root mean squared error (NRMSE) and NRMSE-based sum of ranks (SOR) were applied to compare the imputation methods. Their results showed that random forest performed better for MAR and MCAR missing data, while quantile regression imputation of left-censored data (QRILC) had better results for MNAR data.

[Saar-Tsechansky and Provost 2007] proposed a comparison of different classification models to handle missing values. The authors compared reduced-feature models, regression trees, reduced-feature ensemble, bagged trees and a hybrid approach that combines reduced-feature and regression trees models. Concluding that reduced-feature ensemble has better performance than bagged trees, although reduced-feature modeling is significantly more expensive in terms of computation or storage, and the hybrid approach was similar to the bagged trees. [Hegde et al. 2015] showed that bagged trees and random forest are the state of the art in the prediction of new values. This work created a framework to predict the rate of penetration during drilling using trees, bagged trees and random forest, with support of statistical comparison. Although bagged trees and Random Forest methods increased substantially the accuracy of predictions, only bagged tree had the combination of computational efficiency and accuracy.

[Demirhan and Renwick 2018] proposed a comparison between 36 imputation methods for solar irradiance time series of several cities in Australia. Since solar irradiance prediction is crucial, having a solid database with identified missing values is also very important. The experiments were run in a Monte Carlo setting and the missingness was generated randomly. The imputation methods were focused in statistical models, such as linear and Stineman interpolations, weighted moving average. They also studied the imputation models in different missing frequencies: minutely, hourly, daily, and weekly. The accuracy of the algorithms was compared with root mean square error (RMSE) and relative root mean square error (rRMSE). The results showed that for different frequencies, different methods performed better: Stineman interpolation and Kalfman filtering performed well with minutely and hourly series, while weighted moving average gave better imputations for daily and weekly time series.

[Olcese et al. 2015] proposed a study using neural networks (NN) as a machine learning tool to identify missing values, using historical values at two stations, air mass trajectories passing through both of them and NN calculations to process all the information. This work made a comparison of several neural networks with different topologies, number of hidden layers and methods of propagation of the error and used the coefficient of determination r^2 to compare measured and calculated values. The result is a model capable of generating missing values and a great tool to predict values in several conditions. The result of the work was a model capable of generating missing values, with a 10% error in relation to the real data.

[Xiao et al. 2015] proposed a framework for consistent estimation of multiple land-surface parameters from time-series surface reflectance data. The framework was built combining pre-processing methods, such as Kalman filter and a two-layer canopy reflectance model (ACRM). The work showed that the proposed framework was successful to input missing and noisy data. Although this work used time-series data, it did not compare different models, such as neural network, support vector machines (SVM) or bagged trees.

[Tang and Ishwaran 2017] proposed a deep study in the use of random forest technique to impute missing data. Different ways to grow the random forest and impute data were tested, and different sizes of missing data were used, to correlate the best approach and the optimum size of missing data so that the imputation model can perform better. They also worked with a solid statistical study, to identify the better model. The paper showed that for big data gaps, random forest algorithms did noticeably better than other algorithms, such as k-nearest neighbors (KNN), since random forest is more adaptive.

[Jordanov et al. 2018] proposed a framework to compare state-of-art algorithms, such as neural networks, support vector machine, random forest, to identify missing values in a classification problem of radar signal. The framework also compared the imputation values in a final verification, concluding that the classifiers' accuracy improved when using identified missing values, compared to dataset with missing data. The framework, which used data samples with up to 60% of missingness, showed that the classifiers' improvement was higher when the missingness was higher.

The present work aims to study several tools to estimate new climatic data. Although the related works presented before had great results in the study of methods to identify missing values of different sources, there are some such as the study of correlation between the time and the missing climate values and a statistical study between different imputation models. This article also extends the previous work [Bayma and Pereira 2017], updating to the most recent models the previous framework and increasing the final data and statistical analysis.

3. DATASETS AND DATA PREPROCESSING ANALYSIS

3.1 Datasets

There are 48 meteorological automatic stations in the state of Minas Gerais, Brazil, whose data are available at the National Institute of Meteorology (INMET) website[2]. For this research, time-series daily data were used from 11 meteorological stations distributed around the state. The datasets used were composed of the following parameters: precipitation, maximum temperature, minimum temperature, insolation, evaporation rate, average relative humidity, average compensated temperature,

[2] http://www.inmet.gov.br/

and average wind speed time-series. Since the meteorological stations were built at different dates, the time-series datasets also have different start dates. Each station automatically collects climatic data during the day and saves them at midday (composed of data collected during the morning) and midnight (with the average data of the day). Due to the highly noisy data from midday values, just midnight values were considered in the study.

3.2 Data Analysis and Preprocessing

In order to better present the results and facilitates the comprehension, from this point of the article, all information generated based on the Belo Horizonte station will be detailed. Data from the other stations will be summarized at the end of this article.

The first approach was to analyze the time-series dataset to acquire better understanding of the correlation between the variables, in order to improve the study. The maximum temperature of the Belo Horizonte station can be seen in the Figure 2, showing that there is a large gap of missing data between 1980 and 1981, 1983 to 1986, 1987 to 1988, among other minor gaps. Such missing values represent about 13% of the total amount of values.

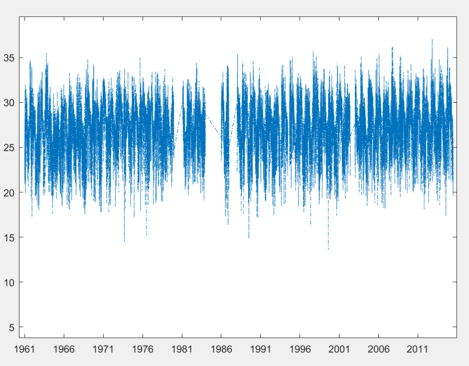


Fig. 2: Maximum temperature data series of Belo Horizonte station.

As mentioned before, the mechanisms of missing data belong to three categories [Enders 2010]: missing at random (*MAR*), missing completely at random (*MCAR*), and missing not at random (*MNAR*). When analyzing the missingness of the meteorological stations in study, all the 12 stations have a significant rate of missing data. In comparison with each variable collected by the stations, Figure 3 shows the missingness of the most used variables, maximum and minimum temperature, precipitation and the total missingness of each station.

Crucial information can be extracted from Figure 3. First, the total missingness of each station is relatively similar to the temperatures and precipitation missingness. This can be explained by the fact that the missing data is related to bad functioning or technical problems of the meteorological stations and, therefore, is expected to see similar rates of missing data within each station. This confirms that study's scenario fits into *MCAR* case. Secondly, the missingness of the stations is around 25%, with lower value of 14% (max. temperature of Montes Claros station) and higher value of 44% (precipitation of Lambari station). Thus, the focus of the work is to identify missing values for a database of 25% missingness.

As the climate undergoes great changes throughout the year, it was necessary to evaluate which components of the date variable provided the greatest changes in climate data. Due to this, the date information has been separated between day, month and year, since each of them retains different information about the climate data, such as maximum temperature. The Pearson correlation method [Pearson 1992] was used to verify correlation between the variables *date* and *maximum temperature*.

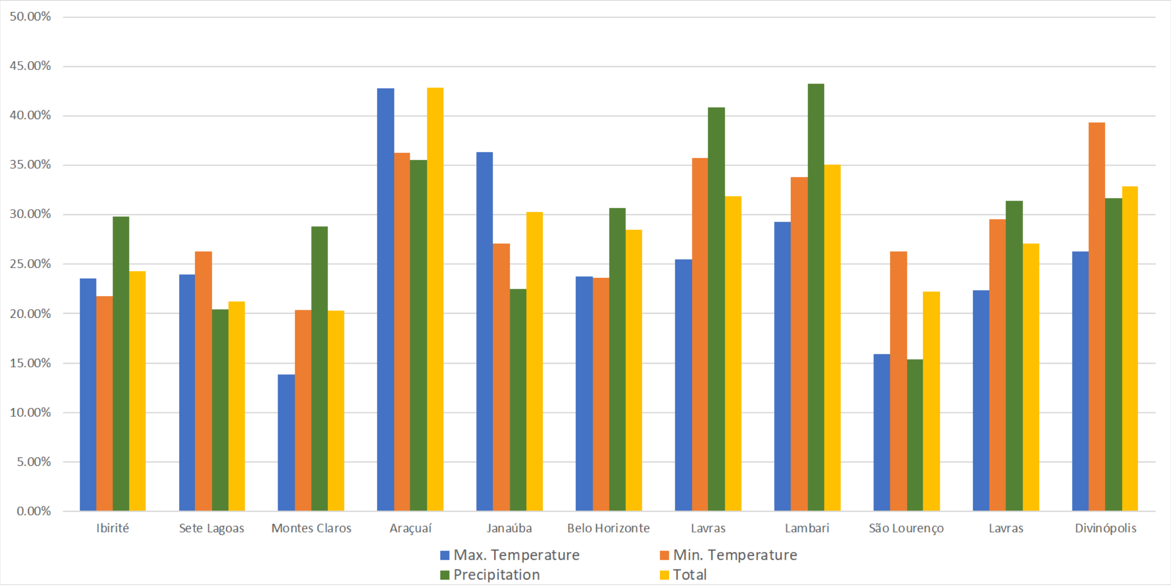
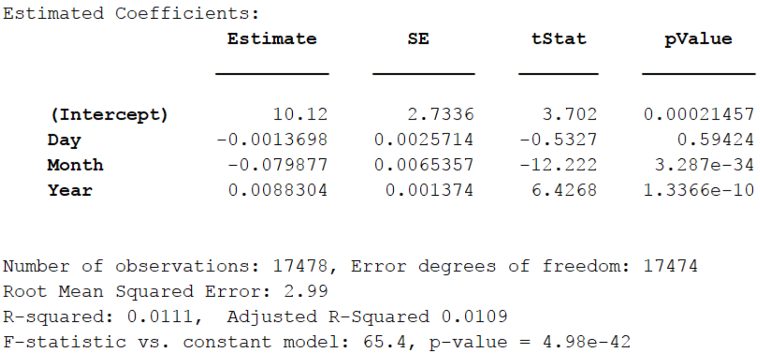


Fig. 3: Meteorological stations missingness

Figure 4 shows the relationship between day, month and year values. It is visible that the p-value of the month and year are extremely small, showing that both variables are statistically significant to generate the maximum temperature response [Carrano et al. 2011]. The p-value of the day is considered high (above 0.05), showing that this variable has no great influence on the response [Wasserstein and Lazar 2016].



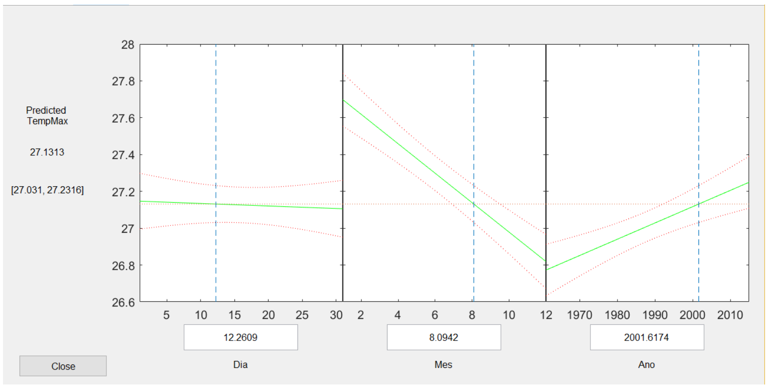
```
Estimated Coefficients:
                      Estimate        SE          tStat       pValue
                     __________    _________    ________    __________

    (Intercept)           10.12       2.7336       3.702    0.00021457
    Day              -0.0013698    0.0025714     -0.5327       0.59424
    Month             -0.079877    0.0065357     -12.222     3.287e-34
    Year              0.0088304     0.001374      6.4268    1.3366e-10


Number of observations: 17478, Error degrees of freedom: 17474
Root Mean Squared Error: 2.99
R-squared: 0.0111,  Adjusted R-Squared 0.0109
F-statistic vs. constant model: 65.4, p-value = 4.98e-42
```

Fig. 4: Pearson correlation method.

In Figure 5 it is possible to visualize how the variables influence the response. It is possible to notice that the month and day contribute inversely to the temperature. While the year contributes directly to the maximum temperature of Belo Horizonte, showing that, since 1961, the temperature has been increasing in the capital of Minas Gerais during the studied period. Therefore, it was proven that the date variable has highly correlation with the climate data and it was used as input into regression models.

4. THE PROPOSED FRAMEWORK

Regression models involve the following variables: unknown parameters, denoted as β, the independent variables, X, and the dependent variables Y. A regression model relates Y to a function of X and


Predicted TempMax
27.1313
[27.031, 27.2316]

Dia: 12.2609
Mes: 8.0942
Ano: 2001.6174

Fig. 5: Prediction slice plots.

β as $Y \approx f(X, \beta)$. The approximation is usually formalized as $(E \mid X) = f(X, \beta)$. The form of the function f is based on the machine learning technique. For all regression models, the result is a solution for unknown parameters β that will, for example, minimize the distance between the measured and predicted values of the dependent variable Y, also based on the machine learning technique [Draper and Smith 2014].

The framework architecture is shown in Figure 6. It can be partitioned into four parts: data preprocessing, missing data imputation, model building to predict new data, and model evaluation and comparison. The previous framework [Bayma and Pereira 2017] was composed of four machine learning regression models: linear regression, neural network, support vector machine and bagged regression trees. This works extends the previous by adding a new regression model, random forest. Since neural network and bagged regression models performed better in the past framework, this work will focus only on these three, neural networks (**NN**), bagged trees (**BT**), and random forest (**RF**), adding more details about their construction. The following will be detailed the three used algorithm for the input of the missing data.

4.1 Neural Networks

The neural networks model implements a structure analogous to a neuronal cell. These cells can be linked as a network, using different layers, simulating the communication between the neurons [Ripley 2007]. In this work, the input layer represents the climatic data matrix, created from the data of day, month and year of operation of the station. While the output layer represents the output vector formed by the data to be analyzed. The number of hidden layers is parameterizable. A few hidden layers can generate a simplistic neural network model, unable to encompass the complexity of prediction. On the other hand, many hidden layers can generate good results for the trained data, however it can generate an overfitted model. The neural network training method used in this study was Bayesian backpropagation [Ripley 2007].

In order to determine the best network configuration to identify missing data, a 10 fold cross-validation was performed with different hidden layer numbers, using the climate data with 25% of missingness randomly generated. The average of the mean square error (MSE) and the average of the time required are presented in Figure 7.

From Figure 7, it is possible to identify that the error of the algorithm decreases from 1 hidden layer to 15. After that, the reduction of the error is imperceptive and even worse (higher error with 60 hidden layers). This can be explained with a possible overfitting. The time of the test was also checked, showing that it has a considerable increase in process time with 15 hidden layers or more. To illustrate the comparison, the 100 hidden layers network had an error similar to 10 hidden layers,

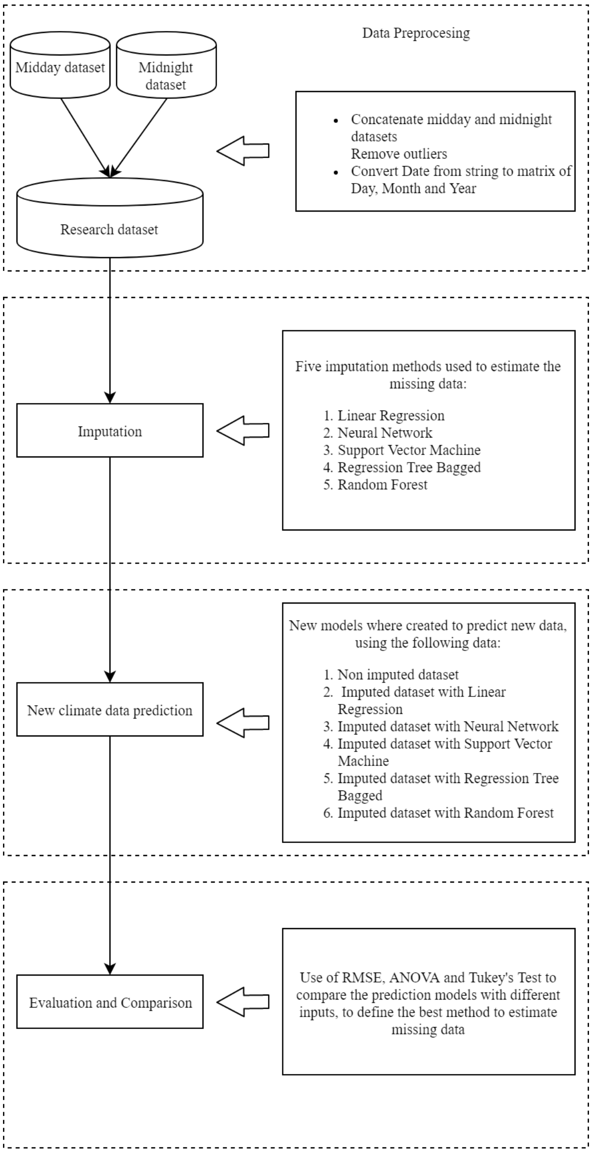

Data Preprocesing
Midday dataset
Midnight dataset
Research dataset
Concatenate midday and midnight datasets
Remove outliers
Convert Date from string to matrix of Day, Month and Year
Imputation
Five imputation methods used to estimate the missing data:
1. Linear Regression
2. Neural Network
3. Support Vector Machine
4. Regression Tree Bagged
5. Random Forest
New climate data prediction
New models where created to predict new data, using the following data:
1. Non imputed dataset
2. Imputed dataset with Linear Regression
3. Imputed dataset with Neural Network
4. Imputed dataset with Support Vector Machine
5. Imputed dataset with Regression Tree Bagged
6. Imputed dataset with Random Forest
Evaluation and Comparison
Use of RMSE, ANOVA and Tukey's Test to compare the prediction models with different inputs, to define the best method to estimate missing data

Fig. 6: Framework diagram.

but with process time 20 times higher. Finally, from this comparison between several neural network configurations, the number of hidden layers found was 10, which made the model computationally feasible to perform the calculations and with minimum error. The network model was assembled to estimate all missing data weather from the stations studied. With the neural network model it was possible to find values to replace the missing values with most similarity to the real values, as indicated in Figure 8.

4.2 Regression Tree, Bagged Trees and Random Forest

Regression Tree is a variation of the Classification Tree, designed to approximate real-valued functions. Classification trees are constructed by repeated splits of subsets (nodes) of the input values $X_¸$ into two descendant subsets, starting with X itself. Each terminal subset is assigned as belonging to a class, and the resulting partition of X corresponds to the classifier, called the leaf node [Breiman et al.

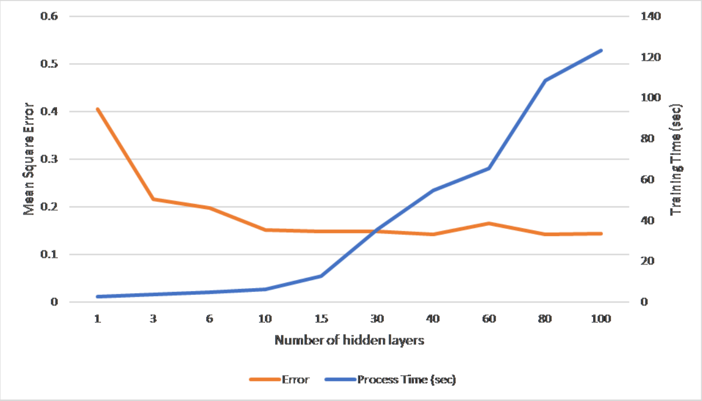


Fig. 7: MSE and process time of different neural network configurations

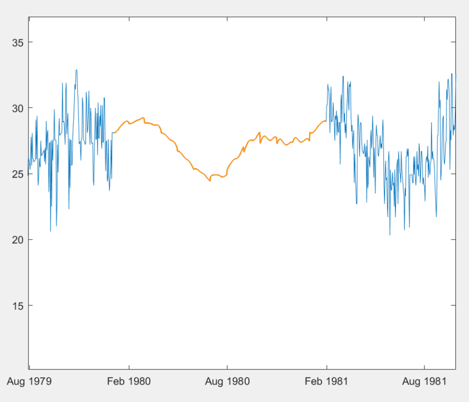


Fig. 8: Imputed data (in orange line) using neural networks regression model.

1984]. When the decision tree is used to predict numerical values, rather than predicting categories, the tree is called a regression tree. The leaves of a regression tree represent the expected mean values of the response.

[Breiman 1998] showed that gains in accuracy could be obtained by *aggregating predictors* from perturbed version of the learning set. Bagging can improve performance of good unstable methods by replicating the original learning set $\mathcal{L}$ with small changes, k times. Predictors are built for each k perturbed dataset and aggregated. *Classification and Regression Trees* (CART) and neural networks are unstable, whereas k-nearest neighbor methods are stable [Breiman et al. 1996]. Since neural nets progress slowly and replications require many days of computing, just bagged regression trees were used in this work.

In the proposed framework, 100 bootstrap replications of the climate time-series dataset were used, in order to extract the missing data from the stations under study. The bagged trees model did much better than the previous models, since it was able to work with data that had a great temporal variation and, at the same time, it was not overloaded and could estimate with very low error the missing values, as shown in Figure 9.

Random forests are an improvement over bagged trees. Bagged trees are greedy, so they choose which variable to split on using a greedy algorithm that minimizes error. Even if the bagged trees have 100 bootstrap replications, the regression trees can have a lot of structural similarities and have high correlation in their predictions. Also, random forest algorithm selects a random subset of predictors

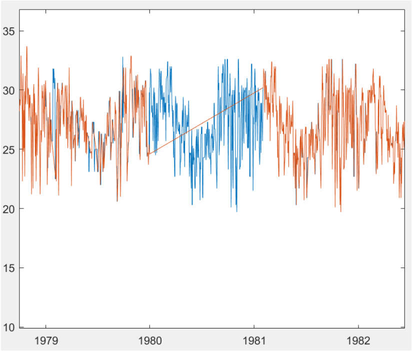


Fig. 9: Temperature data estimated (in blue line) with bagged trees model.

to use at each split, so that the resulting predictions from all of the subtrees have less correlation [Breiman et al. 1984].

In the framework, a cross validation between different random forest models was created, in order to tune the best parameters for the imputation model. Since the dataset has three input variables, $m = 2$ had better results in the cross validation, where m is the number of randomly selected features that can be searched at a split point. Due computational power, a random forest of 30 regression trees was used in the proposed framework. The imputation result, as shown in Figure 10, was very similar when compared with the regression bagged trees, even though the random forest used less bootstrap trees.

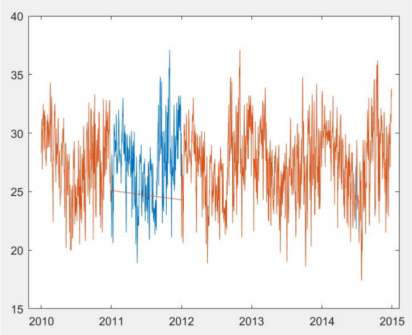


Fig. 10: Temperature data estimated (in blue line) with random forest model.

5. VALIDATION OF MISSING DATA ESTIMATION METHODS

In the absence of a method that compares the efficiency of imputed missing data when trying to predict new climate data, a cross-validation method was designed to handle this comparison. To compare the regression models, the method implements a k-folds cross-validation among all the machine learning techniques used in this research, using non-imputed data and data imputed by previous methods. It was selected 70% of the dataset to train the models, and 30% of the dataset to validate the models, ensuring that no training data were reused in the validation phase, avoiding overfitted prediction models. Thus, 12 models were created: 3 different methods, each method with 4 different imputation approaches: no imputation (NI); imputed data using: neural network (NN), bagged trees (BT) and random forest (RF). In addition, the data of the studied stations were reduced to 5 years, taking 1 year to simulate the missing data which corresponds to 25% of the dataset of each station.

For the quality measurement of the imputed data was used the normalized root mean square error - NRMSE (Equation 1). NRMSE is a parameter validation, that can be used when it is necessary to compare the performance of a model with other predictive models and it is being used in meteorology to see how effectively a mathematical model predicts the behavior of the atmosphere [Hyndman and Koehler 2006]. Given the *mean square error* (MSE) $\sum_{i=1}^{n}((X_{obs,i} - X_{model,i})^2/n$, where $X_{obs,i}$ is the vector of observed values corresponding to the inputs, and $X_{model,i}$ is the vector of i predictions, the RMSE of a model with respect to the estimated variable $\mathrm{X}_{\mathrm{model}}$ is defined by the square root of the MSE, normalized by the reach of the observed data $(X_{obs,max} - X_{obs,min})$, which is the difference between the maximum $X_{obs,max}$ and minimum $X_{obs,min}$ values of the vector of observed values.

$$\text{NRMSE} = \frac{\sqrt{\sum_{i=1}^{n} \frac{(X_{obs,i} - X_{model,i})^2}{n}}}{(X_{obs,max} - X_{obs,min})} \tag{1}$$

The 12-folds cross-validation method is executed 30 times in order to generate an array of NRMSE values, to be studied statistically. The method used to perform the statistical analysis was proposed in [Carrano et al. 2011], which consists of making a bootstrap of the data sent from each method to build a probabilistic distribution function of the mean of the NRMSE values. Such functions are compared using ANOVA [Fisher 1919] and Tukey's multiple comparison test. This test returns an ordered sequence of the validated models, using permutation. In addition to the ANOVA and Tukey's tests, the models were ordered according to a statistical analysis based on the p-value of 5%, to evaluate if one model is superior to another. If the analysis indicates that model A is higher than model B with p-value less than 5%, we consider that A is ahead of B; Otherwise we say that the models are tied.

6. RESULTS

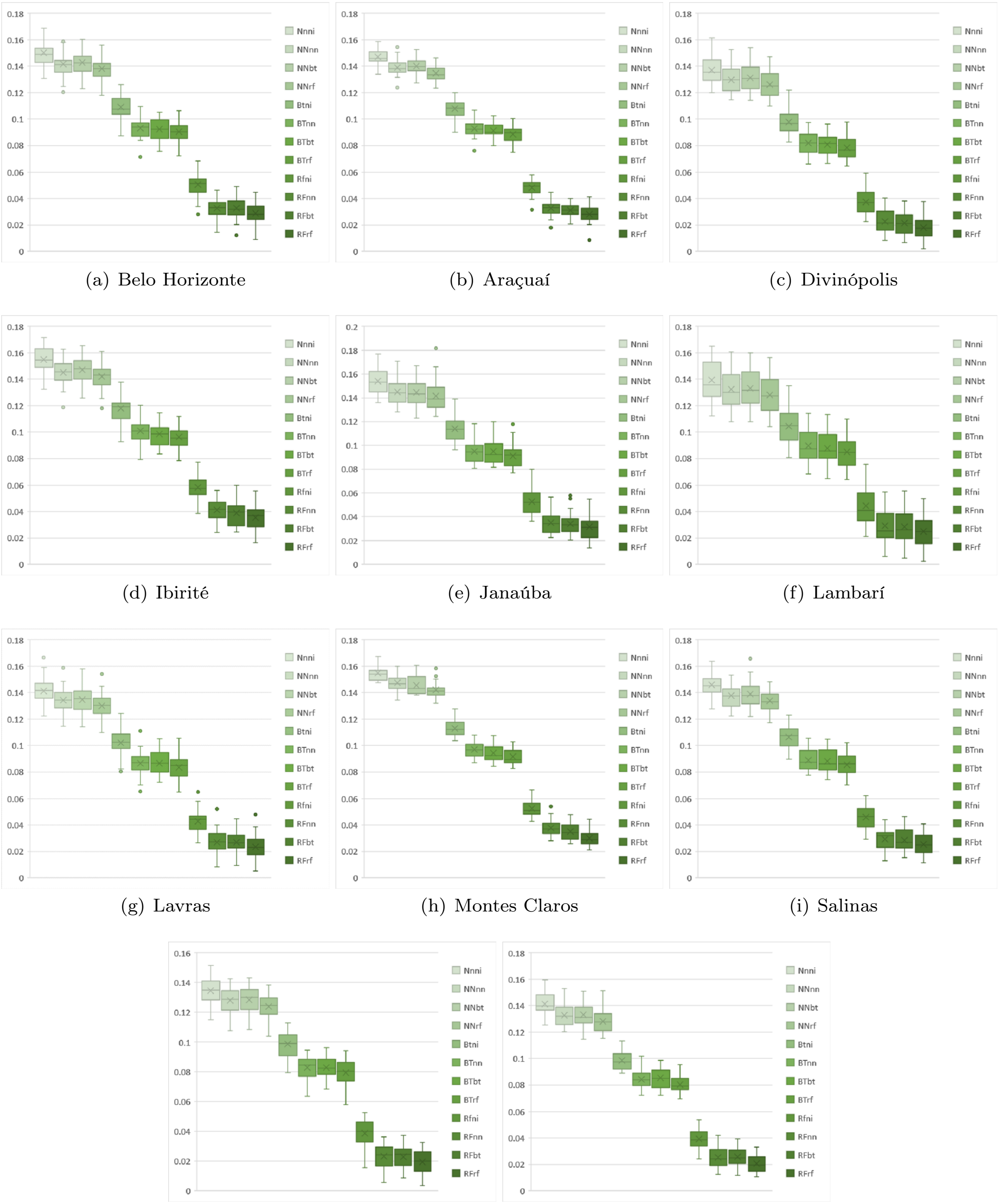
Figure 11 shows the box and whisker plot created from the 12-folds cross-validation NRMSE arrays, for each meteorological station. Each box plot represents one model and is created from the array of NRMSE values, generated from the combination between prediction model and the imputation algorithm. The bottom and top of the box are the first and third quartiles and the red band inside the box is the median. The ends of the whisker (black line) represent the minimum and maximum of all of the data, and the red plus symbols are the outliers.

From Figure 11 it is possible to identify that the prediction models random forest (models 9 to 12) and bagged trees (models 5 to 8) stand out from the neural network algorithm (models 1 to 4).

When comparing the imputation algorithms, prediction models with data without imputation (models number 1, 5 and 9) had worse results than the models with any imputation, with higher median of NRMSE, especially for the prediction models with bagged trees and random forest algorithm. The same analysis can be applied with other imputation methods. Models that used random forest imputation (number 4, 8 and 12) had lower median of NRMSE, compared with the other imputation algorithms. A heat map is also a good analysis to understand the difference between the models NRMSE.

Figure 12 shows the heat map of different meteorological stations. From the k-folds cross-validation results, random forest had better results when discovering new climate data, with NRMSE lower than 0.05. Besides that, the heat map also shows that random forest imputation method had better results for all the eleven stations, while the other imputation methods have floated between themselves (e.g. stations Ibirité and Montes Claros).

Analyzing all eleven meteorological stations heat maps in Figure 12, random forest model for prediction of new data had an improvement of 90%, compared to the second best algorithm, bagged



(a) Belo Horizonte (b) Araçuaí (c) Divinópolis

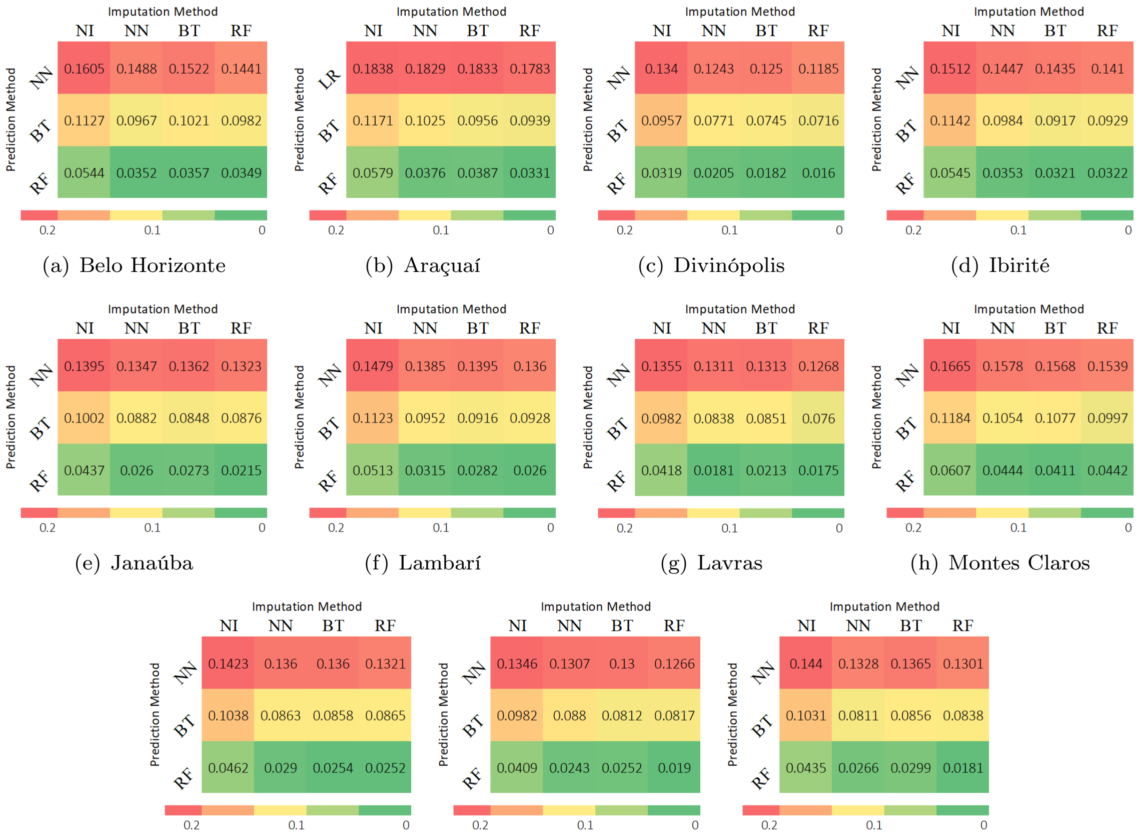
(d) Ibirité (e) Janaúba (f) Lambarí

(g) Lavras (h) Montes Claros (i) Salinas

(j) São Lourenço (k) Sete Lagoas

Fig. 11: Box and whisker plot for NRMSE of each station from the study. 1: $\mathrm{Nn}_{\mathrm{ni}}$; 2: $\mathrm{Nn}_{\mathrm{nn}}$; 3: $\mathrm{Nn}_{\mathrm{bt}}$; 4: $\mathrm{Nn}_{\mathrm{rf}}$; 5: $\mathrm{Bt}_{\mathrm{ni}}$; 6: $\mathrm{Bt}_{\mathrm{nn}}$; 7: $\mathrm{Bt}_{\mathrm{bt}}$; 8: $\mathrm{Bt}_{\mathrm{rf}}$; 9: $\mathrm{Rf}_{\mathrm{ni}}$; 10: $\mathrm{Rf}_{\mathrm{nn}}$; 11: $\mathrm{Rf}_{\mathrm{bt}}$; 12: $\mathrm{Rf}_{\mathrm{rf}}$;

trees. Besides that, when the column of imputation method is analyzed, imputation data with random forest had an improvement of about 60% to 90% (depending on the station), compared to no imputation dataset. Although this result sounds promising, a deep statistical analysis was still necessary to understand if these improvements are really significant. For that we used ANOVA and Tukey's



(a) Belo Horizonte (b) Araçuaí (c) Divinópolis (d) Ibirité

(e) Janaúba (f) Lambarí (g) Lavras (h) Montes Claros

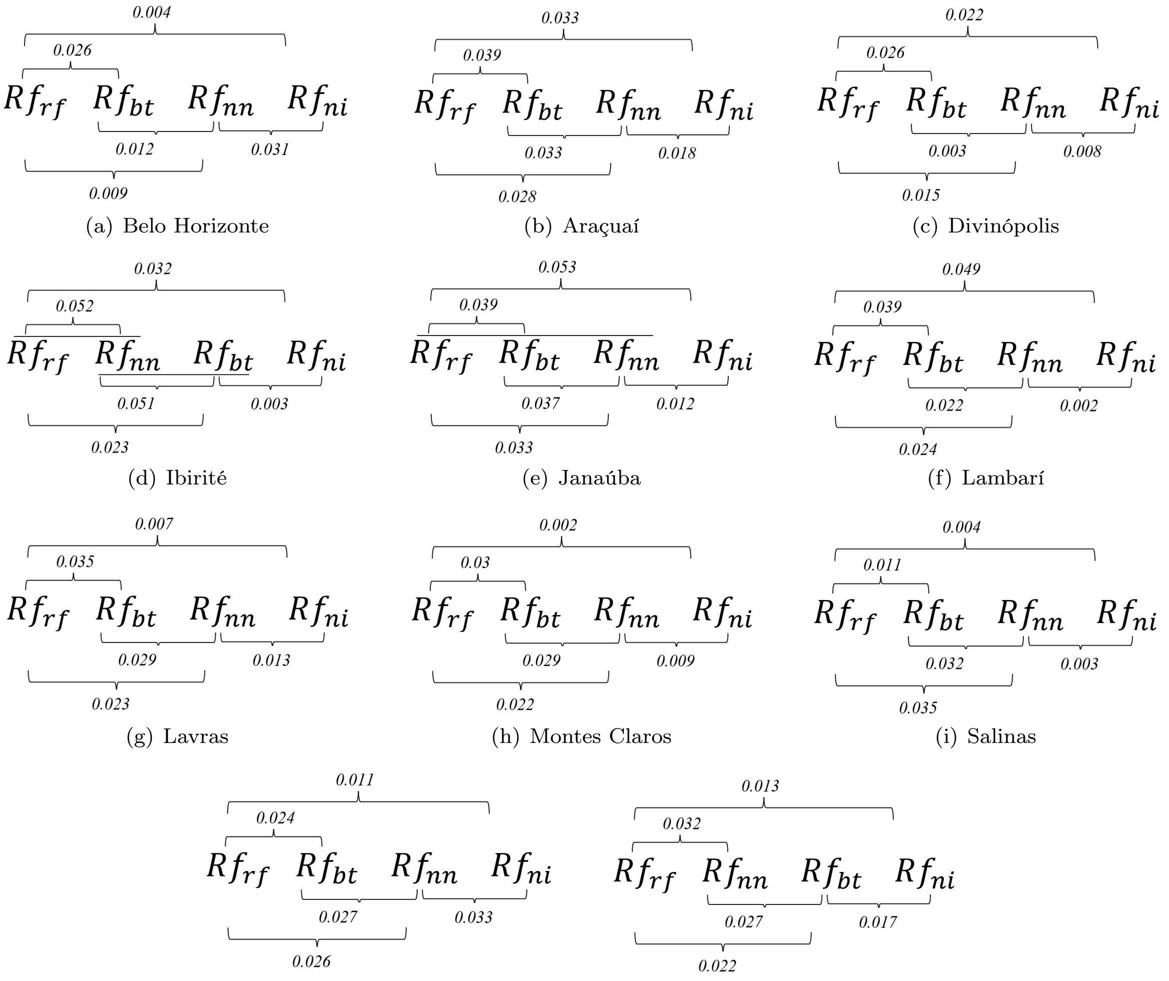
(i) Salinas (j) São Lourenço (k) Sete Lagoas

Fig. 12: Heat map of NRMSE from the study meteorological stations, showing the prediction methods with different types of data imputation. Strong green for better results and strong red for worst results.

tests.

Figure 13 shows the comparison between models for each meteorological station, ordered by ANOVA and Tukey's tests. The p-values show the significance between the models. It is possible to notice in Figure 13(d) that, although the predicted model of random forest with values imputed by the random forest model ($\mathrm{Rf}_{\mathrm{Rf}}$) is in front of the sequence, it has p-value greater than 0.05 in relation to the $\mathrm{Rf}_{\mathrm{nn}}$ models (values imputed with neural networks method). Only in relation to the $\mathrm{Rf}_{\mathrm{bt}}$ model (values imputed with bagged tree method) that the $\mathrm{Rf}_{\mathrm{rf}}$ model stands out, with a p-value of 0.023. This demonstrates that the $\mathrm{Rf}_{\mathrm{rf}}$ and $\mathrm{Rf}_{\mathrm{nn}}$ models are statistically similar and are tied. The tied models are grouped by dashed lines, that is, at 13(f), the $\mathrm{Rf}_{\mathrm{bt}}$ and $\mathrm{Rf}_{\mathrm{svm}}$ models are tied, while the $\mathrm{Rf}_{\mathrm{ni}}$ (values imputed without imputation) model is not relevant in comparison to any of the other data forecast models. As may be noted, the statistical comparison is not transitive, e.g. 13(d), $\mathrm{Rf}_{\mathrm{rf}}$ and $\mathrm{Rf}_{\mathrm{rf}}$ are tied as $\mathrm{Rf}_{\mathrm{nn}}$ and $\mathrm{Rf}_{\mathrm{bt}}$ are tied, but $\mathrm{Rf}_{\mathrm{rf}}$ and $\mathrm{Rf}_{\mathrm{bt}}$ are not tied. For more details about statistics comparison see [Carrano et al. 2011].

Figure 13 also shows the analysis obtained for the remaining 10 other stations. It is possible to notice that in all the stations analyzed in this work, the models with the highest performance in the prediction of new values were the models of random forest (Rf). The prediction of new climate values had better performance with estimated missing values using random forest methods, as is shown in all eleven stations. And in ten out of eleven, random forest imputation method stands out from the others imputation methods, showing the improvement in predicting new data with imputed values


0.004
0.026
Rf_{rf} Rf_{bt} Rf_{nn} Rf_{ni}
0.012
0.031
0.009
0.033
0.039
Rf_{rf} Rf_{bt} Rf_{nn} Rf_{ni}
0.033
0.018
0.028
0.022
0.026
Rf_{rf} Rf_{bt} Rf_{nn} Rf_{ni}
0.003
0.008
0.015
0.032
0.052
Rf_{rf} Rf_{nn} Rf_{bt} Rf_{ni}
0.051
0.003
0.023
0.053
0.039
Rf_{rf} Rf_{bt} Rf_{nn} Rf_{ni}
0.037
0.012
0.033
0.049
0.039
Rf_{rf} Rf_{bt} Rf_{nn} Rf_{ni}
0.022
0.002
0.024
0.007
0.035
Rf_{rf} Rf_{bt} Rf_{nn} Rf_{ni}
0.029
0.013
0.023
0.002
0.03
Rf_{rf} Rf_{bt} Rf_{nn} Rf_{ni}
0.029
0.009
0.022
0.004
0.011
Rf_{rf} Rf_{bt} Rf_{nn} Rf_{ni}
0.032
0.003
0.035
0.011
0.024
Rf_{rf} Rf_{bt} Rf_{nn} Rf_{ni}
0.027
0.033
0.026
0.013
0.032
Rf_{rf} Rf_{nn} Rf_{bt} Rf_{ni}
0.027
0.017
0.022

(a) Belo Horizonte (b) Araçuaí (c) Divinópolis

(d) Ibirité (e) Janaúba (f) Lambarí

(g) Lavras (h) Montes Claros (i) Salinas

(j) São Lourenço (k) Sete Lagoas

Fig. 13: Sequence of the most significant models and their p-values of the studied meteorological stations

from random forest algorithm. [Tang and Ishwaran 2017] also showed better results using random forest algorithm to estimate missing values.

The final observation that Figure 13 provides is that when comparing the meteorological station results, in none of them, the model that uses data without previous estimation was statistically similar and tied with the other imputation methods and they were not relevant in comparison to any of the other data forecast models. Therefore, we conclude that estimation of climatic missing values had improved models to predict new values, especially with random forest algorithm.

7. CONCLUSIONS

Climate prediction is a relevant activity for humanity, since its beginnings. The various companies and public agencies have equipment capable of performing climate measurements as well as acting in the arduous task of predicting the climate for the near future. Time-series climate data have a great relevance in this task, since they can feed predictive models, and the lack of them can result in worse predictions. This article showed that predictions of new climatic data have an increase in accuracy when the input data with a considerable amount of missing values, is previous filled with data through machine learning techniques.

With the analysis of the imputed data and the final forecast of new values, it was possible to conclude that the imputed data allowed the forecast of new data to have a better performance. When there is a large amount of missing temporal data over a long period of time, it becomes difficult for machine learning models to deal with this lack of data. Box and whisker plots were useful to easily understand the difference between the errors whether using imputation data or not, and also identify the gains of the random forest algorithm for data imputation. The heat map data analysis showed a numerical improvement in any forecast model when using previous random forest algorithm for data imputation. The final statistical analyses were important to show the discrepancy between the forecast models with imputed data and the models without imputed data. Particularly, noteworthy were the forecast model of random forest and imputation with random forest algorithm, which presented good performance in all data series.

The missing data imputation models, with special attention to random forest algorithm, created in this article, can be widely used by diverse responsible companies and public agencies for improving their historical databases, hence their predictions. In a future work, a previous spatial analysis can be used within the framework, such as data triangulation between meteorological stations, in order to improve the forecast models.

Acknowledgment

We thank FAPEMIG and PROPE/UFSJ for financial support and the INMET and database sector of CPTEC/INPE for the availability of data and technical reports.

REFERENCES

Barbosa, M. and Carvalho, M. (2015). *Sistemas de Armazenamento de Dados Observados do CPTEC/INP*. Instituto Nacional de Pesquisas Espaciais.

Bayma, L. O. and Pereira, M. A. (2017). Comparison of machine learning techniques for the estimation of climate missing data in the state of minas gerais, brazil. *Proceedings of the XVIII Brazilian Symposium on Geoinformatics*, pages 283–294.

Breiman, L. (1998). Using convex pseudo-data to increase prediction accuracy. *breast (Wis)*, 699(9):2.

Breiman, L. et al. (1996). Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6):2350–2383.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees.* CRC press.

Carrano, E. G., Wanner, E. F., and Takahashi, R. H. (2011). A multicriteria statistical based comparison methodology for evaluating evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, 15(6):848–870.

Demirhan, H. and Renwick, Z. (2018). Missing value imputation for short to mid-term horizontal solar irradiance data. *Applied Energy*, 225:998–1012.

Draper, N. R. and Smith, H. (2014). *Applied regression analysis.* John Wiley & Sons.

Enders, C. K. (2010). *Applied missing data analysis.* Guilford press.

Fisher, R. A. (1919). Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Transactions of the royal society of Edinburgh*, 52(02):399–433.

García-Laencina, P. J., Abreu, P. H., Abreu, M. H., and Afonoso, N. (2015). Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Computers in biology and medicine*, 59:125–133.

Gilat, A. and Subramaniam, V. (2009). *Métodos numéricos para engenheiros e cientistas: uma introdução com aplicações usando o MATLAB.* Bookman Editora.

Hegde, C., Wallace, S., Gray, K., et al. (2015). Using trees, bagging, and random forests to predict rate of penetration during drilling. In *SPE Middle East Intelligent Oil and Gas Conference and Exhibition.* Society of Petroleum Engineers.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688.

Jordanov, I., Petrov, N., and Petrozziello, A. (2018). Classifiers accuracy improvement based on missing data imputation. *Journal of Artificial Intelligence and Soft Computing Research*, 8(1):31–48.

Lakshminarayan, K., Harp, S. A., and Samad, T. (1999). Imputation of missing data in industrial databases. *Applied intelligence*, 11(3):259–275.

Luengo, J., García, S., and Herrera, F. (2010). A study on the use of imputation methods for experimentation with radial basis function network classifiers handling missing attribute values: The good synergy between rbfns and eventcovering method. *Neural Networks*, 23(3):406–418.

Olcese, L. E., Palancar, G. G., and Toselli, B. M. (2015). A method to estimate missing aeronet aod values based on artificial neural networks. *Atmospheric Environment*, 113:140–150.

Pearson, K. (1992). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In *Breakthroughs in Statistics*, pages 11–28. Springer.

Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge university press.

Saar-Tsechansky, M. and Provost, F. (2007). Handling missing values when applying classification models. *Journal of machine learning research*, 8(Jul):1623–1657.

Sapankevych, N. I. and Sankar, R. (2009). Time series prediction using support vector machines: a survey. *IEEE Computational Intelligence Magazine*, 4(2).

Sefidian, A. M. and Daneshpour, N. (2019). Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model. *Expert Systems with Applications*, 115:68–94.

Singh, P. (2016). Neuro-fuzzy hybridized model for seasonal rainfall forecasting: A case study in stock index forecasting. In *Hybrid Soft Computing Approaches*, pages 361–385. Springer.

Tang, F. and Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377.

Valdiviezo, H. C. and Van Aelst, S. (2015). Tree-based prediction on incomplete data using imputation or surrogate decisions. *Information Sciences*, 311:163–181.

Wasserstein, R. L. and Lazar, N. A. (2016). The asa's statement on p-values: context, process, and purpose.

Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., and Ni, Y. (2018). Missing value imputation approach for mass spectrometry-based metabolomics data. *Scientific reports*, 8(1):663.

Xiao, Z., Liang, S., Wang, J., Xie, D., Song, J., and Fensholt, R. (2015). A framework for consistent estimation of leaf area index, fraction of absorbed photosynthetically active radiation, and surface albedo from modis time-series data. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6):3178–3197.

Yang, J. and Hu, M. (2018). Filling the missing data gaps of daily modis aod using spatiotemporal interpolation. *Science of the Total Environment*, 633:677–683.