Comparison of Machine Learning Techniques for the Estimation of Climate Missing Data in the State of Minas Gerais, Brazil

Lucas O. Bayma[1], **Marconi A. Pereira**[1]

[1] 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 paper describes a study of a combination of several machine learning techniques to determine missing climatic values. This study produced a computational framework, formed by four different methods: linear regression, neural networks, support vector machines and regression bagged trees. A statistical study is conducted to compare these four methods. The study statistically demonstrated that the regression bagged trees 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.*

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.

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 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].

Over time, several tools have been applied in order to identify these missing values [Gilat and Subramaniam 2009]. Several approaches have been proposed and improved in

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

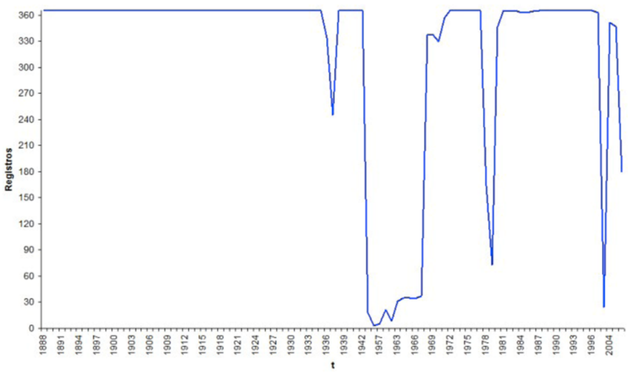


Figure 1. Existence of precipitation data between 1888 and 2006, Station of Light (Estação da Luz). Source: [Barbosa and Carvalho 2015]

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 boosting [Dudoit et al. 2002].

Within this perspective, this paper presents a framework for the study of several tools and techniques of machine learning for the imputation of missing data in time series in order to better predict the tendency data. The framework implements linear regression, neural networks, support vector machines and regression tree models, and is applied in the Minas Gerais state, Brazil.

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 paper 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

In machine learning there is a sub-area that aims to study techniques and models for the identification of missing data. [Dudoit et al. 2002] compare the performance of different discrimination methods for the classification of tumors based on gene expression data. The methods include nearest-neighbor classifiers, linear discriminant analysis, classification trees and also new approaches, such as bagging and boosting. The given methods were used for imputation of missing data of cancer genes. The results showed that diagonal linear discriminant analysis (DLDA) and nearest-neighbor obtained better results, with aggregated tree predictors had performance intermediate. The work used a framework to compare different methods, but it was used in time-series 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 prediction of new values. This work created a framework to predict rate of penetration during drilling using trees, bagged trees and random forest, with support of statistical comparison. Althought bagged trees and Random Forest methods increased substantially the accuracy of predictions, only bagged tree had the combination of computational efficiency and accuracy.

[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 compared different models, such as neural network, support vector machines (SVM) or bagged trees.

The present work aims to study of 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, some gaps in the previous works were considered by the current paper, such as the study of correlation between the time and the missing climate values.

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 by the following parameters: precipitation, maximum temperature, minimum temperature, insolation, evaporation rate, average relative humidity, average compensated temperature, and average wind speed time-series. Since the meteorological stations were built in different dates, the time-series

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

datasets also have different start dates. Each station collects automatically climatic data during the day and save them at midday (composed by data collected during the morning) and midnight (with the average data of the day). Due the highly noise data from midday values, just midnight values were considered in the study. For space restrictions, 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 paper.

3.2. Data Preprocessing Analysis

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.

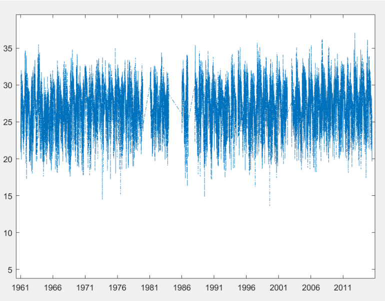


Figure 2. Maximum temperature data series of Belo Horizonte station.

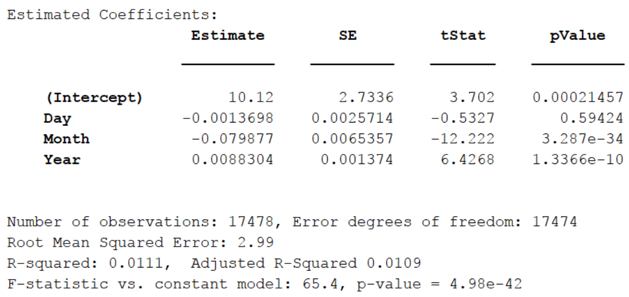
As the climate undergoes great changes throughout the year, it was necessary to evaluate which components of the date variable provided the greatest changes in climate data. Due to this, the date information has been separated between day, month and year, since each of them retains different information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

Figure 3 shows the relationship between day, month and year values. We can see 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].

In Figure 4 is possible to visualize how the variables influence the response. It is possible to visualize 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

The framework was composed by four machine learning regression models: linear regression, neural network, support vector machine and bagged regression trees. Regres-



```
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
```

Figure 3. Pearson correlation method.

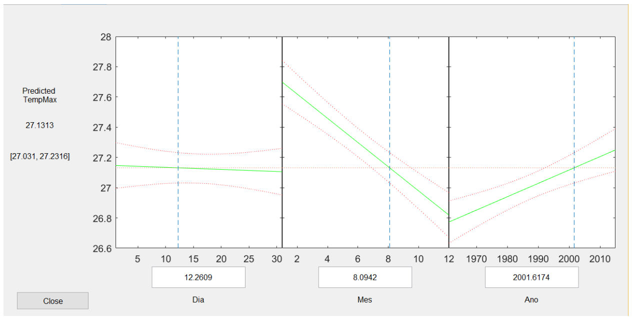


Figure 4. Prediction slice plots.

sion 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 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 following will be detailed each used algorithm for the input of the missing data.

4.1. Linear Regression

Linear regression is one of the simplest methods in statistics and machine learning techniques and when the attributes are numeric, is a natural technique to consider. Given a data set $X = \{y_i, x_{i1}, ..., x_{ip}\}_{i=1}^{n}$, of n units, a linear regression tries to map the output y_i onto a continuous expected result function $y_i = \theta_0 + \theta_0 x_{i1} + ... + \theta_p x_{ip}$. Often written as a matrix form $Y_{n,1} = X_{n,m}\theta_{m,1}$, where Y is the array of n dependent variables, X is the matrix of m arrays of n independents variables and θ is a $m+1$ dimensional parameter vector, called weight. θ_0 is the offset term [Witten et al. 2011]. The weight θ can be found by measuring the cost function $J(\theta_0, \theta_1) = 1/2m + \sum_{i=1}^{m}(h_\theta(x_i) - y_i)^2$ until it reaches the lowest value, where h_θ is the hypothesis function of the linear model. The cost function is otherwise called *Mean Squared Error* and it represents, graphically, the smallest distance between the independent variables and the regression line [Seber and Lee 2012].

In the proposed framework, $X_{3,n}$ was the matrix of date variable. The lines represent the 3 inputs: day, month and year. The columns represent the size of the dataset collected, varying according to the meteorological stations data storage. The *Cost Function* was able to find the most suitable curve that represents the missing climate data (Figure 5).

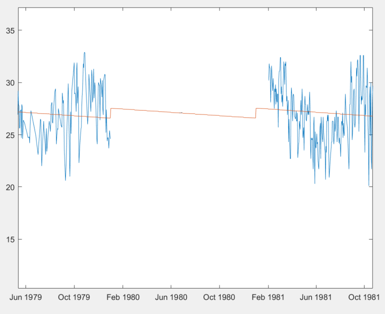


Figure 5. Imputed data (in orange line) using linear regression model with multiple variables.

4.2. 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. 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].

Several neural networks with different hidden layers were tested to find the layer value that predicts the data with the minimum error. The number of hidden layers found, which made the model computationally feasible to perform the calculations and with minimum error, was 10. 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 6.

4.3. Support Vector Machine

The SVM is known as a non-probabilistic binary linear classifier, since it, given different inputs, selects which of two classes the inputs belong to, finding a frontier of separation between these two classes, known as a hyperplane [Cristianini and Shawe-Taylor 2000]. The main characteristic of SVM algorithms is the kernel function, used to reduce the computational complexity. Kernel functions are any functions *K(x,y)* if it can be written as $K(x, y) = \Phi(x) \cdot \Phi(y)$, where Φ is a function that maps an instance into a feature space [Schölkopf et al. 1999].

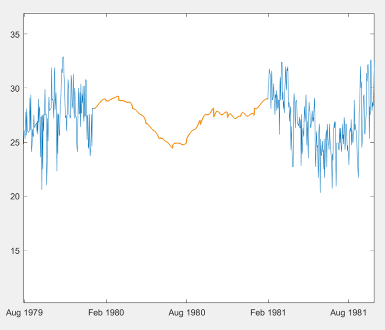


Figure 6. Imputed data (in orange line) using neural networks regression model.

The concept of hyperplane is only applied to classification. However, support vector machine was also developed to work with numerical prediction. Using the binary classification methodology, a model is produced that can usually be expressed in terms of some support vector machines and can be applied, using kernel functions [Witten et al. 2011]. For each model, the 10 folds cross-validation were performed to find the most suitable kernel function. The loss function for each sample was analyzed to test which model obtained the best result. The Gaussian kernel function model obtained better response, with loss function equal to 7.35 versus 8.96 of the loss function of the kernel model with linear function. With the SVM model, it was possible to find values to replace the time-series missing data (Figure 7), similar to those obtained using NN model.

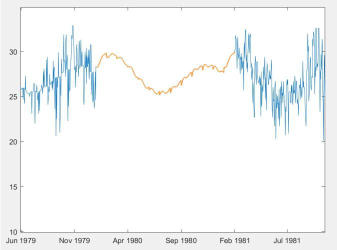


Figure 7. Estimated data (in orange line) using the support vector machine model.

4.4. Regression Tree and Bagged Trees

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. 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 much slower 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 8.

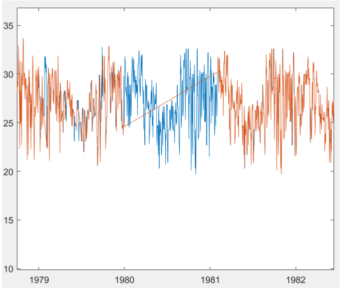


Figure 8. Temperature data estimated (in blue line) with bagged trees model.

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 model 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. 20 models were created: 4 different methods, each method with 5 different imputation approaches: (1) no imputation; imputed data using: (2) linear regression, (3) neural networks, (4) support vector machine and (5) bagged trees. 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 X_{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.

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

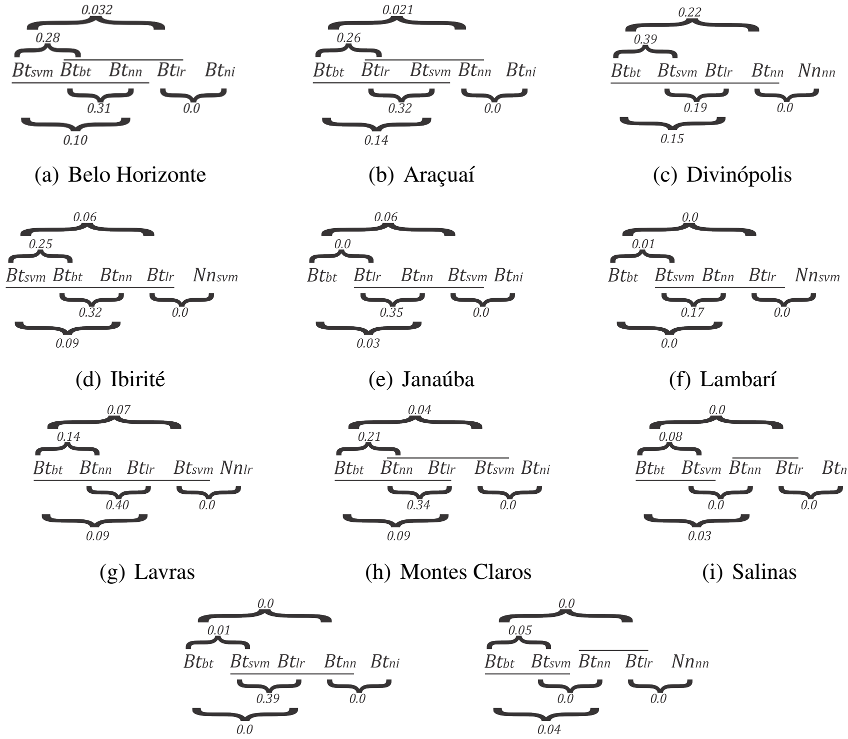
The 20-folds cross-validation method is executed 30 times in order to generate an array of NRMSE values, to be studied statistically. The method used to perform the statistical analysis was proposed in [Carrano et al. 2011], which consists of making a bootstrap of the data sent from each method to build 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 9 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 9(a) that, although the predicted model of bagged trees with values imputed by the SVM model (Bt_{svm}) is in front of the sequence, it has *p*-value greater than 5% in relation to the Bt_{bt} models (values imputed with bagged tree method) and Bt_{nn} (values imputed using the neural network method). Only in relation to the Bt_{lr} model (values imputed with linear regression method) that the Bt_{svm} model stands out, with a *p*-value of 3.1%. This demonstrates that the Bt_{svm}, Bt_{bt}, and Bt_{nn} models are statistically similar and are tied. The Bt_{lr} model has a *p*-value of 0% in relation to Bt_{ni} (prediction with values not estimated), and it can be concluded that, statistically, the prediction model of new values obtained better results with the imputed data than without imputation of data. The tied models are grouped by dashed lines, that is, at the Belo Horizonte station, the Bt_{svm}, Bt_{bt} and Bt_{nn} models are tied, while the Bt_{bt} and Bt_{nn} and Bt_{lr} models are also statistically similar, while the Bt_{ni} 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., Bt_{svm} and Bt_{bt} are tied as Bt_{bt} and Bt_{lr} are tied, but Bt_{svm} and Bt_{lr} are not tied. For more details about statistics comparison see [Carrano et al. 2011].

Figure 9 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 grouped trees (bagged trees). The prediction of new climate values had better performance with estimated missing values using bagged trees methods, as is shown in nine of eleven stations (Araçuaí, Divinópolis, Janaúba, Lambarí, Lavras, Montes Claros, Salinas, São Lourenço and Sete Lagoas). [Dudoit et al. 2002] and [Saar-Tsechansky and Provost 2007] also showed better results using bagged trees method to estimate missing values.

The final observation that Figure 9 provides is that when comparing the meteo-


0.032
0.28
Btsvm Btbt Btnn Btlr Btni
0.31
0.0
0.10
0.021
0.26
Btbt Btlr Btsvm Btnn Btni
0.32
0.0
0.14
0.22
0.39
Btbt Btsvm Btlr Btnn Nnnn
0.19
0.0
0.15
0.06
0.25
Btsvm Btbt Btnn Btlr Nnsvm
0.32
0.0
0.09
0.06
0.0
Btbt Btlr Btnn Btsvm Btni
0.35
0.0
0.03
0.0
0.01
Btbt Btsvm Btnn Btlr Nnsvm
0.17
0.0
0.0
0.07
0.14
Btbt Btnn Btlr Btsvm Nnlr
0.40
0.0
0.09
0.04
0.21
Btbt Btnn Btlr Btsvm Btni
0.34
0.0
0.09
0.0
0.08
Btbt Btsvm Btnn Btlr Btni
0.0
0.0
0.03
0.0
0.01
Btbt Btsvm Btlr Btnn Btni
0.39
0.0
0.0
0.0
0.05
Btbt Btsvm Btnn Btlr Nnnn
0.0
0.0
0.04

(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

Figure 9. Sequence of the most significant models and their *p*-values of the studied meteorological stations

rological station results, in none of them, the model that use data without previous estimation had similar or better results to the other models. Therefore we conclude that pre estimation of climatic missing values had improved models to predict new values.

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 short 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 paper showed that predictions of new climatic data have an increase in accuracy when the input data, that has 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. The final statistical analyzes were important to show the discrepancy between the forecast models with imputed data and the models without imputed data. Particularly, call attention the forecast model of regression bagged trees with imputation, which presented good performance in all data series.

The missing data imputation models 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, 15th edition.

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.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

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

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87.

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.

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).

Schölkopf, B., Burges, C. J., and Smola, A. J. (1999). *Advances in kernel methods: support vector learning*. MIT press.

Seber, G. A. and Lee, A. J. (2012). *Linear regression analysis*, volume 936. John Wiley & Sons.

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.

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.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2011). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

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.