# A Meta-Learning Framework for Imputing Missing Values in Weather Time Series

**Vinícius H. A. Alves**[1], **Marconi A. Pereira**[1]

[1]Departamento de Tecnologias em Eng. Civil, Computação e Humanidades – DTECH
Universidade Federal de Sao João del-Rei - Campus Alto Paraopeba
MG 443, KM 7 – Ouro Branco – MG – Brazil

viniciushaalves97@gmail.com, marconi@ufsj.edu.br

***Abstract.*** *This paper describes an application of a meta-learning framework based on bagged trees. The proposed tool is used to estimate missing weather values in time series. The framework combines 8 different models of bagged trees that were optimized by a meta-learning algorithm. One of those 8 models was trained using only the date and each one of the remaining seven was calibrated with one weather parameter (max. temperature, min. temperature, insolation, among others), in addition to the respective date. The results show improvements in accuracy of the predicted values, achieving values such as* $R^2$ *= 0.94.*

## 1. Introduction

Climatic forecasting is very relevant, for instance, in agriculture planning, energy generation, natural disaster alerts, among others. Thus, it is necessary to learn from the past, considering the historical information, what is possible through stored data. When it comes to the elaboration of a study, it is important to verify the availability of data. Thus, using a complete and reliable database it is possible to generate studies with fewer errors [Bayma and Pereira 2017, Bayma and Pereira 2018]. Inconsistencies and unsatisfactory volume of data generate a limited or even a false representation of the real picture [García et al. 2009]. Rates of 1-5% of missingness are considered manageable. However, dealing with rates of 5-15% of missing values requires advanced methods, and over 15% may lead to significant interpretation losses [Acuna and Rodriguez 2004].

Despite having a large reservoir of climate data in Brazil, relevant institutions, such as the data division of CPTEC/INPE, do not have continuous information for all regions of the country [Barbosa and Carvalho 2015]. There are some periods of time without registration, for different reasons, which can lead to the problems mentioned above.

Predictive modeling is used to develop models capable of predicting missing values with great accuracy, but that is not an easy task. Thus, it has motivated several researches in the area [Yang et al. 2007]. When it comes to modeling the behavior of weather parameters, it is noticeable that trying to build a model using only a single imputation approach (e.g. linear regression) becomes difficult and sometimes ineffective. Therefore, finding different processes that best describe the problem or even conceiving multiple ways of dealing with it becomes a more appropriate measure, bringing with it greater precision [Solomatine and Ostfeld 2008].

This paper presents as a contribution a proposal of a framework for imputing missing data, using meta-learning algorithms. The tool uses bagged trees as both base learners and meta-learner. The base learners suggest the values to be imputed in the gaps, and then it is used a meta-learner to combine the previously suggested values to generate the most suitable outputs to fill the data gaps. This proposal increases the accuracy of the outputs when compared with other related works.

The framework is applied in 10 databases, each one composed of weather time series. These databases hold weather information from cities located in regions with different climatic configurations, distributed throughout the Brazilian territory. This approach aims to bring robustness to the framework, showing that it can deal with climatic diversity.

This text is organized as follows. Section 2 presents the literature review. Section 3 describes the data acquisition and preprocessing analysis. The Section 4 presents the regression method and the meta-learning layers. Section 5 describes the proposed framework. Section 6 details the framework validation. Section 7 presents the results and their analysis. Section 8 presents the conclusions.

## 2. Literature Review

A lot of studies propose approaches to fill the missing data values in time series. The most recent ones, generally, apply some computational intelligence tool. The most relevant works will be presented below, which served as a basis for what was developed in this study.

[Olcese et al. 2015] presented a method that uses artificial neural networks (ANNs) to predict missing aerosol optical depth (AOD) values at an AERONET station. ANNs with different topologies were trained with historical AOD values at two stations and air mass trajectories passing through both of them, generating 18 different datasets that were individually used to train 56 ANNs. It was used the coefficient of determination $R^2$ to compare measured and calculated AOD values to choose the best ones to be used to calculate the missing values. The model created was capable of imputing missing values with the average relative error equals to 25% (with 45% of the values having a relative error of less than 10%) and $R^2$ between 0.67 and 0.86 for the Iberian Peninsula and Eastern US, respectively.

[Bayma and Pereira 2018] compared the effectiveness of four imputation methods, used to fill data gaps in the historical time series from databases of the Brazilian Institute of Meteorology (INMET)[1]. The used methods were linear regression, ANNs, support vector machines and regression bagged trees. To compare the performance of each model, a part of the data was artificially removed so that the imputation methods could identify the missing values. In order to emphasize the importance of data imputation, the study also performed prediction of future data, considering the bases with and without the imputed data. A total of 20 models were generated by combining the four regression models and five different inputs that represented one scenario without imputation and four scenarios that represent the imputation of each method. In addition, the $k$-folds cross-validation method was implemented for all machine learning techniques to perform

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

a statistical test. The study concludes that, when the database was filled with the imputed data, there was an improvement in the forecast of new climatic values. This improvement was more significant with the use of bagged trees, both for imputation and for forecasting future data.

Another relevant approach was proposed in [Assis 2019]. In this work, a framework based on meta-learning methods was presented to identify price trends for the stock market assets. The implemented tool was based on the WEKA[2] API through which 7 regressors were combined to predict values and trends: ANNs, support vector machines, decision trees, random forest, Bayesian networks, minimum sequential optimization and genetic programming. The results showed an accuracy with up to 57% and financial results with gains of up to 100% of the capital value initially invested. The proposed framework can be used both to identify future values as well as to perform imputation to past values.

Meta-learning is a relatively new methodology, but its application is becoming more recurrent. The present study applies the concept of meta-learning to improve regular learning algorithms in the imputation values task.

## 3. Data Acquisition and Preprocessing Analysis

### 3.1. Data Acquisition

The Brazilian Institute of Meteorology (INMET) has more than 400 meteorological stations spread across the country and provides hourly, daily and monthly data on its website, in addition to several other resources that go beyond the interests of this work. The data acquisition for each city studied was made through the INMET website. In this research, daily data from 10 different meteorological stations were used. The parameters used were: date, rainfall, maximum temperature, minimum temperature, insolation, evaporation rate, average relative humidity, average compensated temperature, and average wind speed time-series.

Table 1 shows a summary of the used time series data. The second and third columns present the start date and end date of each analyzed city. The last column presents the total number of days used from each database.

| Cities | Start Date | End Date | Number of days |
|---|---|---|---|
| **Barreiras** | 01/01/1961 | 31/12/2019 | 21548 |
| **Belo Horizonte** | 01/01/1961 | 31/12/2019 | 21548 |
| **Cruz Alta** | 01/01/1961 | 31/12/2019 | 21548 |
| **Cuiabá** | 01/01/1961 | 31/12/2019 | 21548 |
| **Curitiba** | 01/01/1961 | 31/12/2019 | 21548 |
| **Diamantino** | 01/01/1961 | 31/12/2019 | 21548 |
| **Ouricuri** | 01/10/1975 | 31/12/2019 | 16162 |
| **Rio Branco** | 01/06/1969 | 31/12/2019 | 18475 |
| **São Felix do Xingu** | 01/09/1972 | 31/12/2019 | 17287 |
| **São Paulo** | 01/01/1961 | 31/12/2019 | 21548 |

**Table 1. Analyzed periods and total number of days.**

[2]https://www.cs.waikato.ac.nz/ml/weka/

### 3.2. Data Preprocessing Analysis

Pearson product-moment correlation, "R", and the $p$-value represent dimensionless measures of the covariance between two variables, which is a scale that ranges from –1 to +1 [Wackerly et al. 2014]. The closer to those limits the correlation value is, the stronger is the association between the variables compared (their linear dependence). It is 0 whether there is no correlation between them. Moreover, it is possible to evaluate that relationship through the $p$-value, which the closer to 0 the $p$-value is, the stronger is the correlation between the variables compared.

[Bayma and Pereira 2017] applied the Pearson correlation method [Pearson 1900] to analyze the relationship between date and maximum temperature. They found out that the $p$-value of the variables month and year are less than the significance level of 0.05, which means that they are strong correlated. Then, they created an approach that considers just the day, month and year in the imputation process, but the month and year have a greater relevance in models than the day.

Aiming to characterize the correlation among the weather parameters, the correlation coefficients test was performed on the 10 cities' databases. Table 2 shows the average of the results. The main diagonal is set to 1, since it means the correlation between the parameter with itself. The other cells represent the $p$-value among the variables identified in each row and column. The $p$-values that are less than 0.05, indicate that the couple of variables has a statistically significant correlation [Bolboaca and Jäntschi 2006].

| | **R** | **MaT** | **MiT** | **I** | **ER** | **ACT** | **ARH** | **AWS** |
|---|---|---|---|---|---|---|---|---|
| **R** | 1 | 0.00 | 0.09 | 0.00 | 0.00 | 0.02 | 0.00 | 0.17 |
| **MaT** | 0.00 | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.09 |
| **MiT** | 0.09 | 0.00 | 1 | 0.05 | 0.03 | 0.00 | 0.00 | 0.10 |
| **I** | 0.00 | 0.00 | 0.05 | 1 | 0.00 | 0.00 | 0.00 | 0.09 |
| **ER** | 0.00 | 0.00 | 0.03 | 0.00 | 1 | 0.00 | 0.00 | 0.07 |
| **ACT** | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 1 | 0.00 | 0.02 |
| **ARH** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1 | 0.00 |
| **AWS** | 0.17 | 0.09 | 0.10 | 0.09 | 0.07 | 0.02 | 0.00 | 1 |

R - Rainfall – MaT - Maximum Temperature – MiT - Minimum Temperature –
I - Insolation – ER - Evaporation Rate – ACT - Average Compensated Temperature –
ARH - Average Relative Humidity – AWS - Average Wind Speed

**Table 2. Average $p$-values of the 10 cities.**

The Table 3 shows the distribution of the gaps, detailing the percentages by database parameters. In the first column, it is presented the cities studied in this work. The remaining columns indicate the percentage of records in which 0, 1, 2, 3, 4, 5, 6, 7 or 8 parameters are available. The second column ("0") indicates the percentage of records in which there is a lack of values for the all 8 parameters of the database, i.e. there is not any weather value in the record. The third column ("1") indicates the percentage of records where 1 of the 8 weather parameters is available, and so on. The last column indicates the percentage of complete records, that is, none of the eight parameters is missing.

This study presents a better performance in the cases in which there are from 1 to 7 weather parameters, since the framework seeks to take advantage of the existing

parameters to infer the others, especially those that have a high correlation value with the existing parameters in the record.

| Cities | Quantity of weather parameters | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| **Barreiras** | 12.572% | 0.009% | 0.074% | 1.615% | 2.409% | 12.711% | 2.418% | 21.362% | 46.830% |
| **Belo Horizonte** | 7.444% | 0.005% | 0.005% | 0.023% | 0.088% | 0.469% | 0.334% | 3.021% | 88.611% |
| **Cruz Alta** | 12.256% | 0.065% | 0.028% | 0.715% | 5.694% | 1.638% | 1.063% | 18.401% | 60.140% |
| **Cuiabá** | 3.007% | 0.023% | 0.107% | 0.691% | 1.703% | 11.259% | 11.880% | 17.241% | 54.089% |
| **Curitiba** | 0.650% | 0.005% | 0.014% | 0.975% | 2.836% | 0.464% | 0.575% | 8.869% | 85.614% |
| **Diamantino** | 8.275% | 0.023% | 0.037% | 0.733% | 2.497% | 4.799% | 20.763% | 39.748% | 23.125% |
| **Ouricuri** | 22.658% | 0.037% | 0.012% | 0.099% | 0.526% | 12.096% | 2.128% | 7.945% | 54.498% |
| **Rio Branco** | 0.698% | 2.425% | 0.081% | 4.141% | 0.774% | 1.846% | 6.490% | 23.648% | 59.897% |
| **São Felix do Xingu** | 19.194% | 2.493% | 0.289% | 0.978% | 0.445% | 6.207% | 5.733% | 17.181% | 47.481% |
| **São Paulo** | 1.402% | 0.181% | 0.065% | 0.111% | 0.051% | 0.900% | 0.367% | 25.752% | 71.171% |

**Table 3. Percentage between the days with missingness and the total number of days of each city's database.**

### 3.3. The Construction of Train and Test Datasets

As presented in [Bayma and Pereira 2017, Bayma and Pereira 2018], the learning methods present a better performance using a window of 5 years of data from the time series. For instance, to fill a gap in a month, a maximum of the last 5 years of data should be used to train the learning methods. Therefore, different intervals of 5 years were selected to apply the framework. A limit of 5% of missing values was admitted to provide more intervals without much distortion of the real picture to the study.

The data collected by each one of the 10 selected meteorological stations was splitted into three datasets: learners training set, meta-learning training set and validation set. It was ensured that no data used in the training was also part of the validation amount. Around 40% of the data was used to train the base learners, 40% was used to train the meta-learner and 20% for validating the final outputs.

## 4. Theoretical Foundations

### 4.1. Machine Learning

[Bishop 2006] describes the machine learning algorithm as being the task to represent a database as belonging to a function $Y(\bar{x})$, where a vector of independent variables $\bar{x}$ is taken as input and generates the output $Y$ as a function of $\bar{x}$. The function $Y(\bar{x})$ is determined during the training stage, also called the learning phase, which uses a part of the available data for the calculation. Once trained, new entries can have their set of images determined through this type of algorithm. This ability to determine correct outputs for unprecedented input values is called "generalization".

In the present work, the machine learning method used to correlate the weather parameters and to infer the missing values is the bagged trees [Witten et al. 2005]. This method is presented below with its respective configuration.

#### 4.1.1. Regression Tree and Bagged Trees

According to [Witten et al. 2005], decision trees employ the "divide to conquer" approach. The name "tree" comes from the relationship of learning nodes with the branches

and leaves of a real tree. Each node represents a test of the attributes for decision making. Typically, the test consists of comparing the attribute with a constant or a range of values. Each leaf node represents an average value among all the values of the training set to which the leaf applies to.

The difference between classification trees and regression trees refers to the content of their results. While the first one seeks to find classes among the data, the second one seeks numerical results according to the training set. In this work, the interest is in finding numerical values for weather parameters imputation, so the research turns to the regression trees. As the attributes are numeric, the test usually consists of determining whether a given value is less than or greater than a predefined constant, which generates each a binary division or whether this value is below, within or above a range, which generates a division into three nodes. The test is applied successively with different constants or intervals. Figure 1 represents a regression tree.

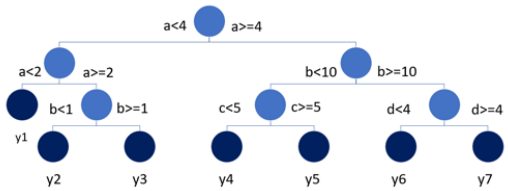


**Figure 1. Scheme of a generic regression tree.**

This work uses the concept of "Bootstrap Aggregating", sometimes known by the acronym "Bagging", so that the grouping of regression trees occurs, which tends to minimize the effects of overfitting [Witten et al. 2005]. The generated models had a maximum number of divisions of the branch node equals to 3 per tree, which characterizes trees that are not very deep.

### 4.2. Meta-Learning

The multi-classifiers may be described as a knowledge's combination of an ensemble of classifiers seeking for more accurate decisions [Kuncheva 2014]. Some multi-classifiers are: voting, ranking, mixture of experts and meta-classifiers. The last one is based on learning about the base classifiers to obtain a knowledge about which one may be the most efficiently applied [Brazdil et al. 2008]. In the context of this paper it was used regressors instead of classifiers, hence, they are given the term meta-learners.

[Kuncheva 2014] emphasizes that the meta-learning process implies in an increase in complexity. However, the authors still mention that combining an ensemble of base learners with less complex approaches becomes more straightforward than finding parameters' combination that best describes the problem's complexity.

## 5. The Meta-Learning Framework to Fill Missing Values

The proposed framework consists of 8 base learners (level 0), which suggest the values to be imputed in the gaps, and then a meta-learner (level 1) combines the previously suggested values to generate the most suitable outputs to fill the data gaps.

Figure 2 represents a scheme of the layers of the meta-learning, where the left hand side shows the $N$ base learners in which is applied $N$ different inputs, being one

input for each learner, and the right hand side shows a meta-learner that receives as input the outputs of the previous learners generating the optimized output.

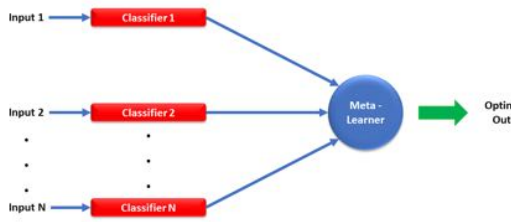


**Figure 2. A scheme of the generated framework.**

The framework is divided into two stages, where each one is represented by a block in the scheme in Figure 3. Those stages are: the learning stage, the meta-learning stage.

In the first stage, level 0, the base learners are trained using the base learners training set. The base learners are in charge of generating models capable of calculating the missing value from a given day based on the date and one of the weather parameters of the same day. Each model generated by each one of the 8 bagged trees must be fed with the inputs used in the learning stage. This ensemble allows that, for a given day, there are 8 different predictions available for the same missing parameter.

There are 8 base learners (level 0) that match each input. To generate those inputs, the framework removes the parameter that represents the one that is being imputed along the iteration, remaining 8 out of 9 types of inputs: (1) date, (2) date + rainfall, (3) date + maximum temperature, (4) date + minimum temperature, (5) date + insolation, (6) date + evaporation rate, (7) date + average relative humidity, (8) date + average compensated temperature and (9) date + average wind speed. Both inputs and the output represent records of the same day. In each iteration, a different weather parameter is imputed.

Subsequently, in the second stage, level 1, the meta-learner is trained using the meta-learner training set. That set of inputs are applied to the models generated in the learning stage to generate level 0 imputation. The level 0 outputs of the 8 models, in addition to the date of the day they refer to, become the inputs to feed the meta-learner. It seeks to learn from the group of base learners' knowledge to generate a model capable of combining those level 0 outputs to calculate an optimized one that is more accurate.

In the end, there are 9 trained models, being one of them in charge of giving the very best output. The trained ensemble is, then, validated.

## 6. Validation Methods

Aiming to measure the quality of the imputed data, the coefficient of determination $R^2$ was used to determine how well the models can reproduce the actual outputs [Homma and Saltelli 1996]. This coefficient compares the difference between the calculated value and the actual value, weighting the result with the difference between the average and the actual value. The closer to 1 the coefficient of determination $R^2$ is, the better the model calculates the dependent variable.

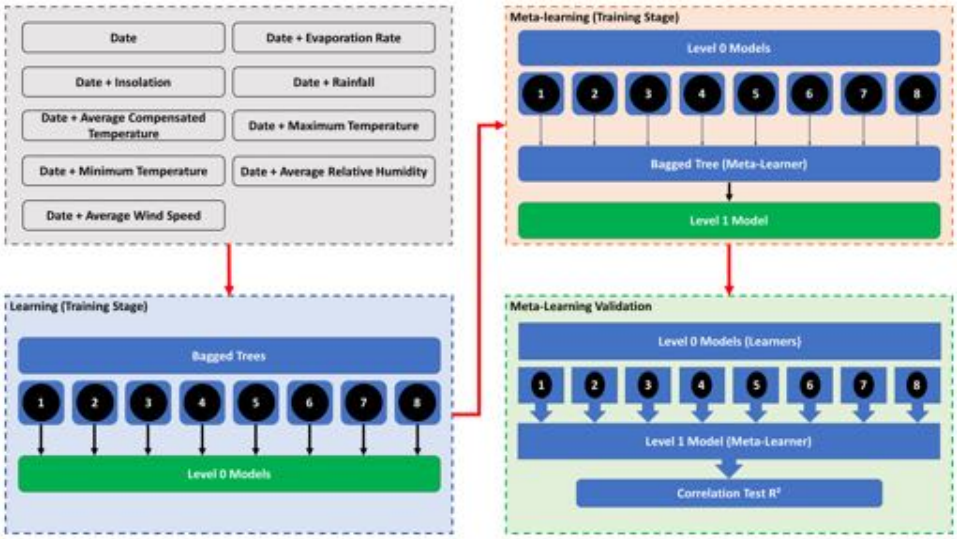


**Figure 3. Scheme of the framework.**

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\overline{y} - y_i)^2} \tag{1}$$

where $y_i$ is the $i$-th actual value, $\hat{y}_i$ is the $i$-th calculated value and $\overline{y}$ is the average of the $N$ actual values.

To perform the validation test, artificial gaps were created in the dataset. In order to simulate the real scenario where the lack of data occurs randomly, a total of 20% of the database was randomly chosen to validate the trained models. The predicted outputs were compared to the actual values using the coefficient determination $R^2$ test.

To simulate different scenarios with different combinations of parameters missingness, artificial gaps in the inputs were created by removing some parameter in the inputs. It creates 8 different scenarios: no weather parameter available; one weather parameter available; two weather parameters available; three weather parameters available; four weather parameters available; five weather parameters available; six weather parameters available; seven weather parameters available. The gaps were replaced by a constant which is a value that is completely out of the bounds of all variable used in this study to simulate $-\infty$, as suggested by [Han et al. 2011]. It was chosen the constant $-9999$.

Working as a second validation method, the algorithm was applied in databases from cities with different climatic characteristics, being each couple of meteorological stations located in each one of the 5 Brazilian regions: north, northeast, midwest, south, southeast. For each database, the methodology adopted was performed 30 times to generate sufficient material to make statistic analysis.

## 7. Results And Analysis

Due to space restrictions, since there are plenty of results to be analyzed in this study, as the framework is applied to 10 cities using 8 different inputs, all the results pre-

sented below refer only to the Belo Horizonte station. The results from the other stations will be summarized and available as appendix at the following address: `https://ufsj.edu.br/marconi/geoinfo2020_-_paper_1.php`. As soon as this article is published, the source code will also be made available at that same address.

Figure 4 shows the coefficient of determination $R^2$ between the measured and the calculated for the base learner (BL) that considers fewer variables and the meta-learning in 8 distinct scenarios: (0) when there is not any weather parameter and only the date is used to generate all the outputs of the base learners; (1) when there is only one parameter available to increase the calculation; (2) when there are two parameters available to increase the calculation; (3) when there are three parameters available to increase the calculation; (4) when there are four parameters available to increase the calculation; (5) when there are five parameters available to increase the calculation; (6) when there are six parameters available to increase the calculation; (7) when there are seven parameters available to increase the calculation. Except for the average wind speed and rain fall, the average of the coefficient of determination of the parameters increases, demonstrating the effectiveness of considering more variables than only the date when trying to calculate the missing values. There are some parameters that present improvement of more than 30% when it is the only information that is missing, for instance, insolation (49%), average compensated temperature (36%), maximum temperature (53%) and average relative humidity (38%).

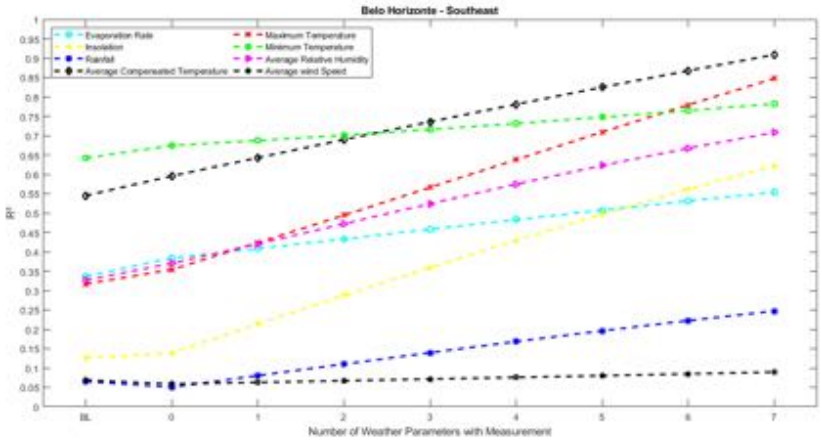


**Figure 4. The coefficient of determination $R^2$ of the base learner that considers fewer variables and the meta-lerning applied in 8 different scenarios when imputing each weather parameter - Belo Horizonte station.**

Figure 5 shows how the coefficients of determination of the base learner (BL) that only uses the date and the meta-learning in the different scenarios are distributed when calculating the average compensated temperature. Note that the boxplots with big areas (2, 3, 4 and 5) occurs because there is no differentiation among which weather parameters were available to generate the outputs, in other words, when parameters with lower correlation are used to estimate the missing data, it may lead to damages to the calculation and when the ones with high correlation are used, it may increase the results.

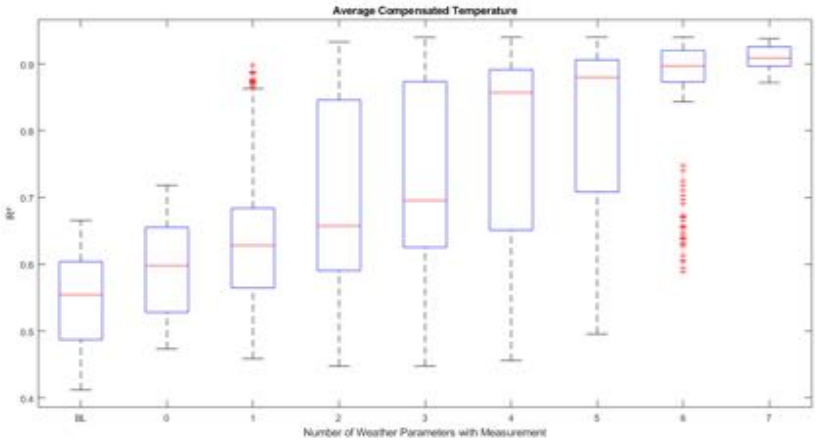


**Figure 5. The coefficient of determination $R^2$ of average compensated temperature - Belo Horizonte station.**

Figure 6 shows, in percentage, the comparison between the meta-learning and the base learner that only uses the date as input. It represents the probability that the result generated by the meta-learner is better than the approach that considers fewer variables, as presented by [Bayma and Pereira 2017, Bayma and Pereira 2018, Assis 2019]. It is possible to see that the meta-learning is affected by the different weather scenarios analyzed and the presence of parameters with low correlation. However, except for the average wind speed and rain fall, the meta-learning shows better results than the base learner with fewer inputs in at least seventy percent of the executions. In the best scenario, the meta-learning reaches better results 100% of the time, except for the average wind speed.

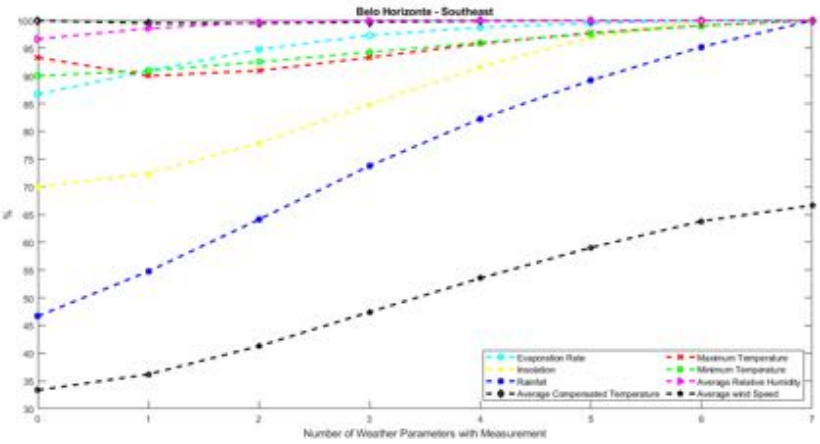


**Figure 6. Comparison between the meta-learning (proposed approach) and learning process [Bayma and Pereira 2017]. The $x$ axis present the number of weather parameters used in the meta-learning. The $y$ axis presents the percentage of improvement of meta-learning compared to leaning process - Belo Horizonte station.**

Through Figure 6 it is possible to conclude that, for example: except for average wind speed, whenever there are 7 weather parameters available, the meta-learner's output is better than the base learner's prediction; and except for average wind speed, when there are 3 weather parameters available, for more than $70\%$ of the inputs the meta-learner's predictions are better than the base learner's predictions.

## 8. Conclusions

Computational resources, such as machine learning, plays an important role in modeling physical phenomena through less complex analysis that consider reduced numbers of variables that affects the system. Due to that, this resource can be used in meteorology to look for meteorological events models.

In this work, it is demonstrated through the analysis of coefficient of determination $R^2$, that meta-learning can increases the accuracy in imputing missing values in weather time series. Even though the meta-learner's output may not be better than the best level 0 model's output for any type of input, it diminishes or get rid of the possibility of choosing an inadequate single model.

It is noticeable that the more information is available, the better the results will be. Nevertheless, the results demonstrate that the meta-learner can recognize which parameters or which combinations of inputs can generate the most suitable values to fill the data gaps.

The low values of $R^2$ characteristic of rainfall and average wind speed may be due to their complex behavior and the low correlation with the other parameter (in the case of the average wind speed).

Despite the complexity of the climatic dynamics of the different regions impacts the recognition of patterns of different parameters, this approach takes advantage of the available information, which provides a better representation of the real picture of a given date and, consequently, of certain parameters. Moreover, the present paper modeled the complex weather parameters behaviors through less complex approaches, getting rid of the hard work of finding out the very best combination of independent variables to infer the dependent variable missing values.

## Acknowledgments

The authors thank FAPEMIG for financial support and the INMET and database sector of CPTEC/INPE[3] for the availability of data and technical report. We are especially grateful to Carlos Alberto da Silva Assis (*in memoriam*) for all the tips and guidance.

## References

Acuna, E. and Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In *Classification, clustering, and data mining applications*, pages 639–647. Springer.

Assis, C. A. S. (2019). *Predição de Tendências em Séries Financeiras Utilizando Meta-Classificadores*. PhD thesis, Centro Federal de Educação Tecnológica de Minas Gerais.

[3]http://www.inpe.br/

Barbosa, M. and Carvalho, M. (2015). Sistemas de armazenamento de dados observados do cptec/inpe. *Instituto Nacional de Pesquisas Espaciais*.

Bayma, L. O. and Pereira, M. A. (2018). Identifying finest machine learning algorithm for climate data imputation in the state of minas gerais, brazil. *Journal of Information and Data Management*, 9(3):259–259.

Bayma, L. O. and Pereira, M. d. A. (2017). Comparison of machine learning techniques for the estimation of climate missing mata in the state of minas gerais, brazil. In *GEOINFO*, pages 283–294.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Bolboaca, S.-D. and Jäntschi, L. (2006). Pearson versus spearman, kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences*, 5(9):179–200.

Brazdil, P., Carrier, C. G., Soares, C., and Vilalta, R. (2008). *Metalearning: Applications to data mining*. Springer Science & Business Media.

García, S., Fernández, A., Luengo, J., and Herrera, F. (2009). A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing*, 13(10):959.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Homma, T. and Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17.

Kuncheva, L. I. (2014). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

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. (1900). X. 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. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Solomatine, D. P. and Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *Journal of hydroinformatics*, 10(1):3–22.

Wackerly, D., Mendenhall, W., and Scheaffer, R. L. (2014). *Mathematical statistics with applications*. Cengage Learning.

Witten, I. H., Frank, E., and Hall, M. A. (2005). *Practical machine learning tools and techniques*. Elsevier.

Yang, Y., Lin, H., Guo, Z., and Jiang, J. (2007). A data mining approach for heavy rainfall forecasting based on satellite image sequence analysis. *Computers & geosciences*, 33(1):20–30.