Revista Brasileira de Cartografia
ISSN 1808-0936 | https://doi.org/10.14393/revbrascartogr
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto

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

Um Arcabouço de Meta-Aprendizado e Triangulação de Dados para Imputação de Valores Ausentes em Séries Temporais Climáticas

Vinícius Henrique Antunes Alves [1] e Marconi de Arruda Pereira [2]

[1] Universidade Federal de São João Del-Rei, Departamento de Tecnologias, Ouro Branco, Brasil. viniciushaalves97@gmail.com.
ORCID: https://orcid.org/ 0000-0003-4397-983X

[2] Universidade Federal de São João Del-Rei, Departamento de Tecnologias, Ouro Branco, Brasil. marconi@ufsj.edu.br.
ORCID: https://orcid.org/0000-0002-4292-4987

Recebido: 03.2021 | Aceito: 06.2021

Abstract: Machine learning and statistical methods can help model meteorological phenomena, especially in a context with many variables. However, it is not unusual that the measurement of those variables fails, generating data gaps and compromising data history analysis. The framework combines the predictions provided by three machine learning methods: decision trees, artificial neural networks and support vector machine, together with values calculated through five triangulation methods: arithmetic average, inverse distance weighted, optimized inverse distance weighted, optimized normal ratio and regional weight. Each machine learning algorithm generates eight regression models. One of the machine learning models makes predictions based only on the date. The remaining seven models make predictions based on one weather parameter (max. temperature, min. temperature, insolation, among others), in addition to the respective date. The triangulation methods use the climatic data from three neighboring cities to estimate the parameter of the target city. The generated dataset is, posteriorly, optimized by meta-learning algorithms. The results show that the additional information provided by the new machine learning models and the triangulation methods offered a significant increase in the accuracy of the imputed data. Moreover, the statistical analysis and coefficient of determination R^2 showed that the meta-learning model based on regression trees successfully combined the base-level outputs to generate outputs that best fill in the missing values of the time series studied in this paper.

Keywords: Meta-learning. Base-level learners. Machine learning. Triangulation. Time series.

Resumo: Aprendizado de máquina e métodos estatísticos podem ajudar a modelar fenômenos meteorológicos, principalmente num contexto com muitas variáveis. Porém, não raro, a medição destas variáveis pode sofrer falhas, gerando lacunas de dados e comprometendo a análise do histórico dos dados. Neste trabalho é proposto um arcabouço que combina as previsões fornecidas por três métodos de aprendizado de máquina: árvores de decisão, redes neurais artificiais e máquina de vetores suporte, juntamente com os valores calculados através de cinco métodos de triangulação: média aritmética, inverso da distância ponderada, inverso da distância ponderada otimizado e proporção normal otimizado. Cada algoritmo de aprendizado de máquina gera oito modelos de regressão. Um dos modelos de aprendizado de máquina gera previsões baseadas apenas na data e os sete modelos restantes geram previsões baseadas em um parâmetro climático (temperatura máxima, temperatura mínima, insolação, entre outros), além da respectiva data. Os métodos de triangulação usam dados climáticos de três cidades vizinhas para estimar o parâmetro da cidade-alvo. O conjunto de dados gerado é, posteriormente, otimizado por algoritmos de meta-aprendizado. Os resultados mostram que o acréscimo de informações fornecidas pelos novos modelos de aprendizado de máquina e os métodos de triangulação proporcionaram um aumento significativo na acurácia dos dados imputados. Além disso, a análise estatística e o coeficiente de determinação R^2 mostraram que o modelo de meta-aprendizado baseado em árvores de regressão combinou com sucesso os resultados do nível de base para gerar os resultados que melhor preenchem os valores faltantes das séries temporais estudadas neste artigo.

Palavras-chave: Meta-aprendizagem. Aprendizes base. Aprendizado de máquina. Triangulação. Série temporal.

1 INTRODUCTION

The weather is a relevant factor that impacts the management and planning of different areas, such as agriculture, energy generation and heavy civil construction. This relationship with the weather has motivated several searches in the area aiming to understand it (YANG et al., 2007). To predict the future condition of the weather, both machine learning and predictive analytics consider historical information to learn from past events, trying to recognize patterns. The material comes from stored data.

Concerning the elaboration of a study, it is crucial to verify data availability. Thus, using a complete and reliable data set, it is possible to generate studies with fewer errors (BAYMA; PEREIRA, 2018, 2017). On the other hand, inconsistencies and unsatisfactory data volumes cause a limited or even a false representation of the actual picture (GARCÍA et al., 2009).

Little and Rubin (2019) reviewed the theory about the different issues concerning the mechanisms that lead to missing data. They found out that there are three categories of incomplete data:

- Missing Completely at Random (MCAR), in which the missing values show dependency neither on the values of the parameter itself nor the values of any other parameters that compose the data
- Missing at Random (MAR), in which the missingness depends on the observed components
- Not Missing at Random (NMAR), in which the dependency relies on the missing values themselves

Despite having an extensive reservoir of climate data in Brazil, relevant institutions, such as the data division of CPTEC/INPE, do not have continuous information for all country regions (BARBOSA; CARVALHO, 2015). There are some periods without registration for different reasons, such as instrument failures, meteorological extremes, observation recording errors, and manual data entry procedures. This incomplete data may lead to the problems mentioned above.

When it comes to modeling the behavior of weather parameters, it is noticeable that building a model using only a single imputation approach (e.g., linear regression) becomes difficult and sometimes ineffective due to the system's complexity. 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; OSTFELD, 2008).

The meta-learning technique is applied when it is necessary to identify the best output from a set of previous predictions provided by multiple imputation techniques, called base-level learners. In addition, this approach offers a more significant variability on the first level predictions to eliminate biases that might occur when choosing a single method. Furthermore, the meta-learning model may identify the imputation technique that best fits a given set of parameters.

This paper is an extended version of Alves and Pereira (2020), presented in XXI Brazilian Symposium on GeoInformatics (GEOINFO, 2020), in which we address the data sets with the missing values problem. We investigate the correlation among weather parameters and the relevance of considering a weather relationship among nearby cities. This approach uses a meta-learning framework based on machine learning and data triangulation to estimate missing values in weather time series. The current framework brings two additional machine learning techniques, support vector machine (SVM) and neural networks (NN), as both base-level learners and meta-learners, in addition to the regression trees. Furthermore, we used five triangulation techniques to provide more information about the days with missingness, evaluating the weather relationship among nearby cities. Finally, we applied the concept of meta-learning to improve regular learning algorithms and methods in the imputation values task.

This text is organized as follows. Section 2 presents the literature review. Section 3 describes the data acquisition and preprocessing analysis. Section 4 presents the regression and triangulation methods and the meta-learning layers. Section 5 describes the proposed framework. Section 6 details the framework validation and shows the result analysis. Finally, section 7 presents the conclusions.

2 RELATED WORKS

Computational intelligence tools are increasingly present in the most recent studies that propose approaches to fill the missing data values in time series. For example, Olcese et al. (2015) and Bayma and Pereira (2018) presented methods that use artificial neural networks (ANNs), linear regression, support vector machines and regression bagged trees to impute missing data on time-series data sets. In addition, these studies investigate which machine learning technique increases the imputation accuracy and which input arrangement can better picture the behavior of the output parameter.

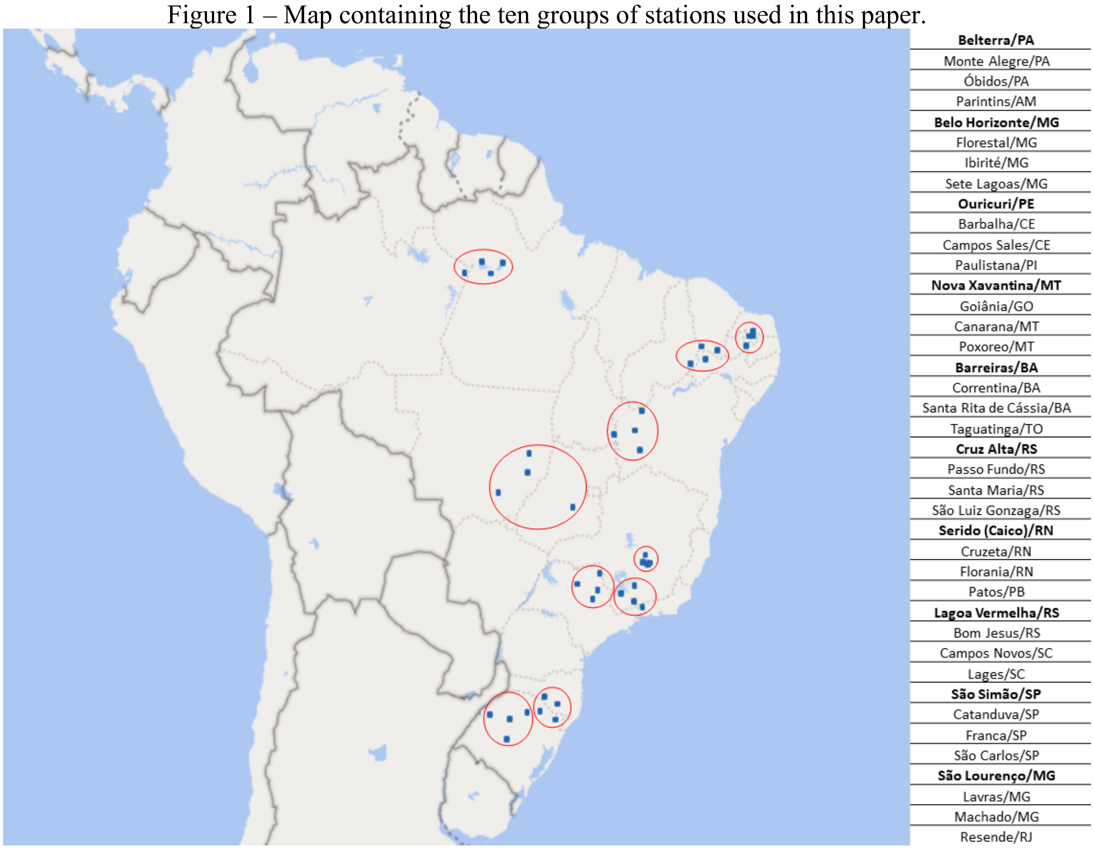
Despite being a relatively new research field mainly for meteorology purposes, the application of meta-learning is becoming more recurrent. This ensemble demonstrates robustness when it is crucial to consider multiples parameters with a too complex dependency for a single machine learning or statistical methods to model, as when identifying price trends for the stock market assets (ASSIS, 2019).

In the matter of defining which estimation method fits better the set of data, the use of statistical studies is a commonly adopted approach, as proposed by Jockers and Witten (2010) and Pirooznia et al. (2008). However, carrying out such a study demands elaborating the stages of the data treatment, model training and testing to the different data estimation methods. On the other hand, meta-learning is a powerful approach that allows matching several data estimation methods since it learns how to generate the most suitable output from the set of estimations available.

Furthermore, it is reasonable to expect that distinct models will stand out when estimating missing values of different parameters. However, using a meta-learning model to recognize the best imputation value among the set of predictions ensures that, no matter which base-level learning model fits the data best, it has the same or even better performance.

Our proposal is based on the concept of meta-learning, initially described by Maudsley (1980) as "*the process by which learners become aware of and increasingly in control of habits of perception, inquiry, learning and growth that they have internalized*" (Maudsley, 1980). Applying that concept for machine learning consists of algorithms that learn how best to combine predictions from estimation techniques in ensemble learning.

Figure 1 – Map containing the ten groups of stations used in this paper.



Source: The Authors (2021).

In comparison to these related studies, our main contributions are: the use of the correlation among weather parameter and the relationship among the weather of neighboring cities; the investigation of an automatic approach, since a thorough time series analysis by humans is often not feasible in practical applications that process multiple time series in minimal time; and the application of the concept of meta-learning combined with triangulation methods to improve regular learning algorithms in the imputation values task.

3 DATA ACQUISITION AND PREPROCESSING

3.1 Data acquisition

The Brazilian Institute of Meteorology (INMET) concentrates daily data from more than 400 meteorological stations distributed throughout the country on its digital collection. Thus, it provides hourly, daily and monthly data on its website. In this research, we used daily data from 40 different meteorological stations. The parameters used were: date; rainfall (R); maximum temperature (MaT); minimum temperature (MiT); insolation (I); evaporation rate (ER); average relative humidity (ARH); average compensated temperature (ACT); and average wind speed (AWS) time-series.

Figure 1 shows a list of the stations used in this paper with the respective latitudes, longitudes, and altitudes. The target city is marked in bold. After the target city, three neighboring cities of the first are shown.

3.2 Data reduction

As presented in Bayma and Pereira (2018, 2017), the learning methods show a better performance using a window of 5 years of data from the time series. For instance, to fill a month gap, a maximum of the last five years of data can be used to train the learning methods. Therefore, we selected different intervals of five years to apply the framework, picturing different cycles.

As discussed in (ACUNA and RODRIGUEZ 2004), 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. Thus, we admitted a limit of 15% of missing values providing more intervals without much distortion of the actual picture to the study.

3.3 Data normalization

Han et al. (2011) describe that, in algorithms such as neural networks or those that use distance measures such as the classification of the nearest neighbor or clustering, normalization gives the process a significant speed in the learning phase. Hence, normalizing the parameters becomes a crucial measure to avoid issues when processing data with different lower and upper limits. When all the information is represented within the same boundaries, for example, [-1,1] or [0,1], the machine learning algorithms present better performance (HAN et al., 2011).

Al Shalabi et al. (2006) made a comparative study between three widely used normalization methods, Min-Max normalization, Z-score normalization, and decimal scale normalization. The authors concluded that the Min-Max method provided, in the experiments carried out, greater accuracy in the machine learning process. Hence, in the present work, we adopted Min-Max normalization with an interval that ranges from 0.2 to 1, according to Eq. (1):

$$A_N = \frac{A - A_{min}}{A_{max} - A_{min}} * (L_{max} - L_{min}) + L_{min} \quad (1)$$

where A_N is the normalized value, A is the original value, A_{min} is the minimum value among the original training dataset values, A_{max} is the maximum value among the original training dataset values, L_{min} and L_{max} are the lower and the upper limit of the normalized data, respectively.

4 THEORETICAL FOUNDATIONS

4.1 Regression methods

In this section, the regression methods used will be presented. These algorithms may be employed in the process of data imputation and forecasting. In this article, we used the techniques only in the imputation process.

4.1.1 REGRESSION TREE AND BAGGED TREES

Regression trees are acyclic graphs in which the nodes represent tests to compare the features with a constant or a range of values. The test determines whether a given value is less than or greater than a predefined constant, which generates a binary division, or whether this value is below, within, or above a range, developing a division into three nodes. It is applied successively with different constants or intervals. Thus, each leaf node represents an average value among all the training set values.

This work used 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 version used in this work consists of 100 bootstrap replications of the climate time series dataset. We chose to grow each tree using at least five observations per leaf.

4.1.2 ARTIFICIAL NEURAL NETWORKS

An artificial neural network, also called ANN, is an arrangement of several processing units. According to Silverman and Dracup (2000), numerical values go from the input nodes to the hidden layer. Each unit, also called a node or neuron, has connections to the next level node, and each of those connections receives an individual weighting factor. Thus, those values are multiplied by the weights and, at the next layer, each neuron adds up all weighted inputs. Furthermore, an activation function is applied to the neurons to determine whether they will or will not be activated depending on their significance to the result.

The network consists of the input layer representing the matrix created by the day, month, year and weather parameter for the base-level learners or the previously estimated values for the meta-learner. In contrast, the output layer means the vector formed by the feature to be analyzed. We parameterized the number of hidden layers to three. 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 yield good results for the trained data, yet it can generate an overfitted model. The neural network training function used in this study was BFGS Quasi-Newton. The function used to measure the network's performance was the sum of absolute error.

4.1.3 SUPPORT VECTOR MACHINE

The support vector machine is considered a binary classifier of the nonlinear type with linear machine learning in a resource space induced by the kernel. The strategy is to separate the different entries into two classes generating a boundary that distinguishes both; that is called a hyperplane (CRISTIANINI et al., 2000). The kernel function is one of the essential points for this classifier, as it reduces its computational complexity. A significant advantage of this algorithm relies on the fact that it is possible to generate responses through the average values obtained by a small subset of the training data.

Although hyperplane is a concept of classifiers, SVM can also perform the function of numerical prediction through the regression that generalization provides. Thus, using the binary classification methodology, a model is generated, and it is expressed by a set of support vector machines that use the kernel function (WITTEN et al., 2005). We used the gaussian kernel function with a kernel s scale of 20.654, box constraint of 122.14, and epsilon equals 0.006107.

4.2 Triangulation techniques

As Stake (1995) observed, the triangulation technique determines the position of a ship in the ocean throughout the positions of three stars in the sky. In the context of this paper (data analysis), we used triangulation to determine the value of unknown data from known data from adjacent cities. Five triangulation methods were selected to compose the framework: arithmetic average (AA), inverse distance weighted (IDW), optimized inverse distance weighted (OIDW), optimized normal ratio (ONR) and regional weight (RW).

4.2.1 ARITHMETIC AVERAGE

Arithmetic Average (AA) is a commonly used triangulation method used due to its simplicity. The missing value is obtained by arithmetically averaging the known values, as Eq. (2) shows.

$$M = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{2}$$

where M is the missing value, Y_i is the feature of the neighboring station i, and n is the number of used neighboring stations in triangulation.

4.2.2 INVERSE DISTANCE WEIGHTED AND OPTIMIZED INVERSE DISTANCE WEIGHTED

The inverse distance weighted (IDW) method is widely used due to its simplicity (HUBBARD, 1994). This method uses weights for each feature used in the calculation. The weights are based on the inverse distance between the two stations, as Eq. (3) shows.

$$M = \frac{\sum_{i=1}^{n} \frac{Y_i}{d_i}}{\sum_{i=1}^{n} \frac{1}{d_i}} \tag{3}$$

where d_i is the distance between the neighboring station and the target station, that distance was calculated using the Haversine method (Robusto 1957).

Khosravi et al. (2015) proposed a modification to IDW, including the monthly average of the parameter and altitude to calculate the variable's weight, as Eq. (4) shows.

$$M = \frac{\sum_{i=1}^{n} \frac{Y_i}{d_i} * \frac{A}{A_i} * \frac{logH}{logH_i}}{\sum_{i=1}^{n} \frac{1}{d_i}} \tag{4}$$

where A and A_i are the monthly parameter averages of the target station and the neighboring station, respectively. H and H_i are the altitudes of the target station and the neighboring station, respectively.

4.2.3 REGIONAL WEIGHT

Regional weight (RW) is one of the classes of methods that uses a weight parameter. As described in (PAULHUS and KOHLER, 1952), the weight to each value used in the calculation is based on the monthly average, as Eq. (5) shows.

$$M = \frac{1}{n}\sum_{i=1}^{n} \frac{A}{A_i} Y_i \tag{5}$$

4.2.4 OPTIMIZED NORMAL RATIO

Optimized normal ratio (ONR) is one of the classes of methods that uses a weight parameter. Young (1992) proposed the weight represents the correlation between the target station and the neighboring station, as Eq. (6) shows.

$$M = \frac{\sum_{i=1}^{n} Y_i * \left({r_i}^{2^{\frac{p_i-2}{1-r^2}}} \right)}{\sum_{i=1}^{n} {r_i}^{2^{\frac{p_i-2}{1-r^2}}}} \tag{6}$$

where r_i represents the correlation between the target station and the neighboring station, and p_i is the number of days the correlation coefficient is based.

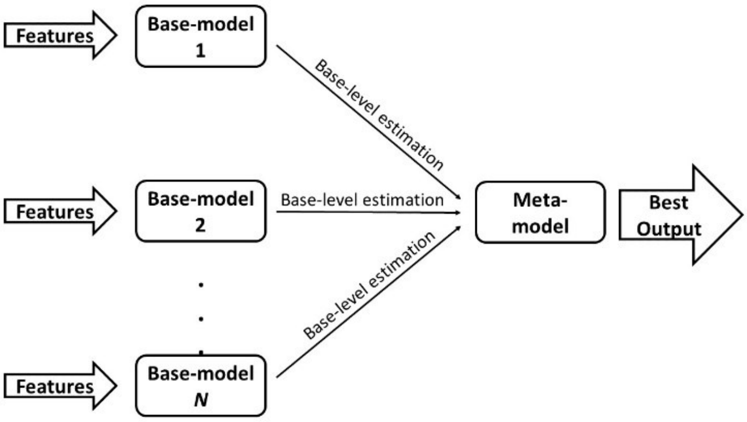
4.3 Meta-learning

The multi-classification technique may be described as a knowledge combination of an ensemble of classifiers that seek more accurate decisions (KUNCHEVA, 2014). Some multi-classifiers are voting, ranking, the mixture of experts and meta-classifiers. The last one is based on accumulating knowledge from the multiple classification methods under a learning system (BRAZDIL et al., 2008). In the context of this paper, we used regressors instead of classifiers. Hence, they are given the term meta-learners.

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

Stacking generalization is adopted as a solution to combine multiple model's outputs. Stacking is an approach that uses several generalizers to estimate values individually with their biases through a particular learning set and then filter out those biases (WOLPERT, 1992). This strategy is applied in such a way to return an output that may not be better than the output of the best imputation method. Still, it diminishes or gets rid of the possibility of choosing an inadequate one. Data of the target city generates values.

Figure 2 – A scheme of the generated framework.



Source: The Authors (2021).

Meta-learning algorithms are essentially composed of two levels: level 0 (also called base-level) and level 1. The first level takes several learning algorithms {L_1, L_2,..., L_N} and trains them using the dataset *D* composed of the features under consideration (i.e., base-learning data) to produce a set of models {m_1, m_2,..., m_N}.

The second level takes a meta-dataset *D'* composed of the predictions of each base-level model, for that instance, to train a new learning algorithm L_{meta} that builds a meta-model m_{meta} that fits the predictions of the base-level learners to target classes.

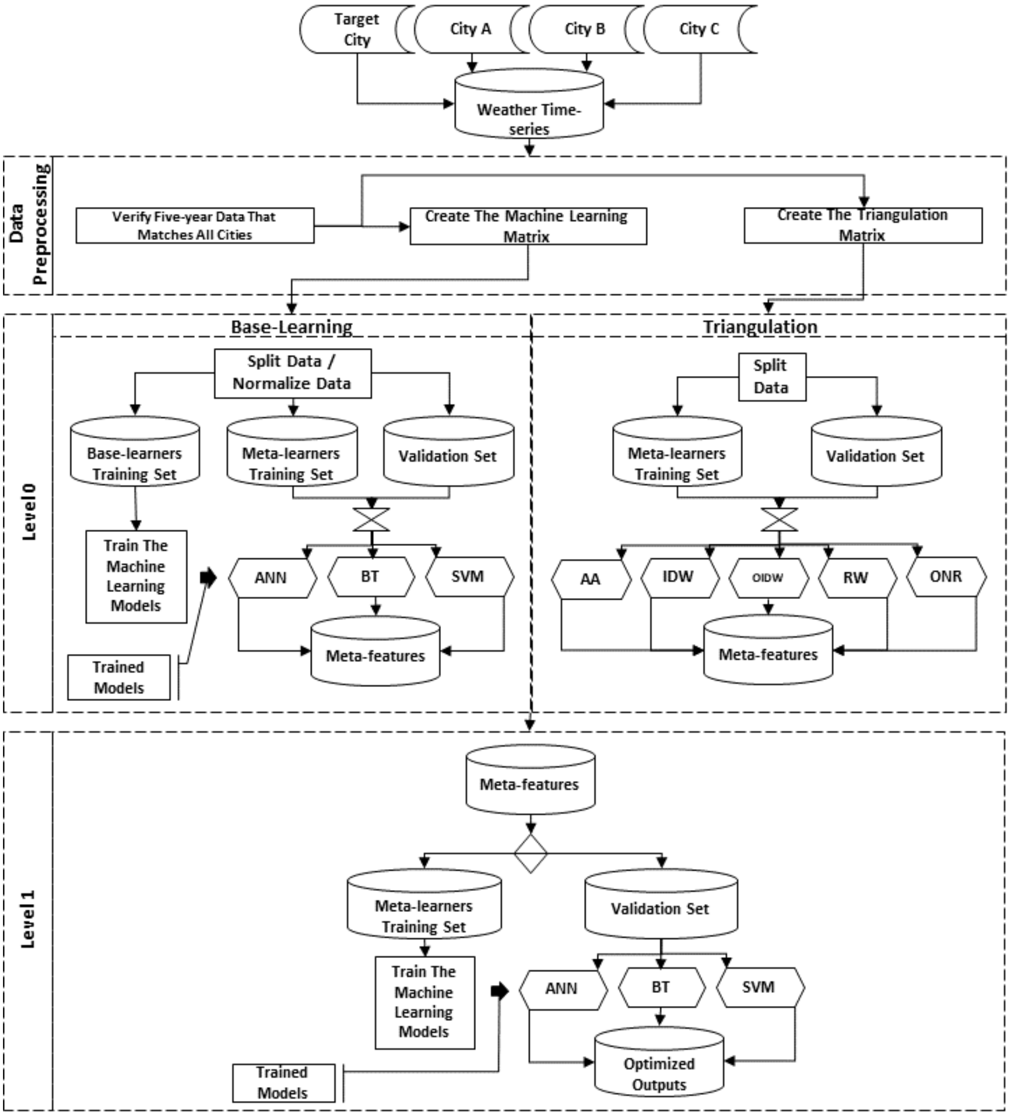
Finally, the framework consists of *N* base-level models and one meta-model, which estimates new instances throughout two calculation stages.

Figure 2 presents a scheme of the proposed framework.

5 THE META-LEARNING FRAMEWORK TO FILL MISSING VALUES

The proposed framework consists of 8 base-level learning models (level 0) that match each input, five triangulation methods and three meta-learning models (level 1). The framework removes the parameter that represents the one that is being imputed along with 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. Thus, both inputs and the output represent records of the same day. Moreover, we adopted stacking generalization to combine the outputs from the eight machine learning models and the five triangulation methods to generate the most suitable values to fill in the data gaps.

Figure 3 – A scheme representative of the data flow and the framework.


Target City
City A
City B
City C
Weather Time-series
Data Preprocessing
Verify Five-year Data That Matches All Cities
Create The Machine Learning Matrix
Create The Triangulation Matrix
Base-Learning
Triangulation
Level 0
Split Data / Normalize Data
Base-learners Training Set
Meta-learners Training Set
Validation Set
Train The Machine Learning Models
Trained Models
ANN
BT
SVM
Meta-features
Split Data
Meta-learners Training Set
Validation Set
AA
IDW
OIDW
RW
ONR
Meta-features
Level 1
Meta-features
Meta-learners Training Set
Validation Set
Train The Machine Learning Models
Trained Models
ANN
BT
SVM
Optimized Outputs

Source: The Authors (2021).

We deal with the imputations of missing values from different time series. To encompass the complexity of each climatic parameter and to picture the dependence among each input and the output, it would demand a study or multiple tests to define the best parameterizations for each machine learning model. Moreover, it would take a specific parametrization for each model created. Thus, instead, we chose to set each machine learning method with the parametrizations cited in subsections 4.1.1, 4.1.2 and 4.1.3 to perform a simplified approach.

We built the framework in three stages, as Figure 3 shows. First, we make the data acquisition and preprocessing to acquire data from a text file and process it. This step is necessary to structure the data into tables and remove the information that corresponds to the days with missing values. Then, we reduce the data to 5 years in which the four data sets have the amount of data of the same period that matches the requirement of 10% of completeness. Moreover, we split the data collected by each of the selected meteorological stations into three datasets. One of the datasets is used to train the base-level learners, one is used to generate the meta-dataset to train the meta-learning models and the last one is used to test the framework. Furthermore, we ensured that no data used in training was also part of the validation amount. Finally, we normalize the machine learning datasets.

We chose to use a 40/40/20 portion for the base-level learners training dataset, meta-learners training dataset, and validation dataset, respectively. To the best of our knowledge, a fragmentation like 80/20 for training and validating machine learning models, respectively, is commonly adopted when one makes statistical studies Dannenberg et al. (1997). Moreover, as suggested by (ASSIS, 2019), we used half of the training to train the base-level learners and a half to train the meta-learners.

Algorithm 1 shows the second and third stages represented in Figure 3. The second stage consists of the base-level learners training using the base-level learners training set (line 8). The base-level learners are responsible for 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. In addition to the base-level learners, we used triangulation methods to calculate the missing values based on the corresponding parameter values from 3 neighboring cities. For example, to calculate the missing precipitation value of the target city, the precipitation values from the three neighboring cities are used. The outputs generated by the base-level learners and the triangulation methods are called the level 0 imputation values. This ensemble allows that, for a given day, there are 29 different imputing values available for the same missing parameter.

Subsequently, in the third stage, level 1, the meta-learners are trained using the meta-learner training set (line 17). Then, that set of inputs is applied to the models generated in the learning stage (line 9) and triangulation methods (line 15) to estimate the level 0 imputation values. Therefore, the meta-learning data have as many attributes as base-level learning models, triangulation methods, and dates. The final purpose is to learn from the group of imputation methods to generate models capable of combining those level 0 outputs to calculate an optimized one that is more accurate.

Algorithm 1 – Meta-Learning.

Input: machineLearningData, triangulationData, triangulationMethods, learners, metaLearners, indl, indm, indp

Output: Predictions

```
1:   start
2:   for i ← 1 to metaLearners.size()
3:       for j ← 1 to learners.size()
4:           for k ← 1 to machineLearningData.getQuantParam()
5:               L ← machineLearningData.get(indl, k);
6:               M ← machineLearningData.get(indm, k);
7:               P ← machineLearningData.get(indp, k);
8:               model ← train(learners[j], L);
9:               M'[i, k] ← fit(model, M);
10:              P'[i, k] ← fit(model, P);
11:          end
```

```
            end
        T1 ← triangulationData.get(indm);
        T2 ← triangulationData.get(indp);
        M'.append(date, triangulationMethods(T1));
        P'.append(date, triangulationMethods(T2));
        metaModel ← train(metaLearners[i], M');
        predictions[i] ← fit(metaModel,P');
        end
        return predictions
        end
```

Source: The Authors (2021).

6 EXPERIMENTS AND RESULTS

As seen previously, this work presents the application of a meta-learning framework to estimate missing weather values in climatic time series. Furthermore, this section will present the validation test used in the experiments and the results. We carried out a study under ten groups of data sets, each one composed of four meteorological stations, to validate the methodology. One of the cities is called the target city and the other three cities are the neighboring ones, which data are used to infer the missing data.

The coefficient of determination R^2 (Eq. 7) was used to determine how well the models can reproduce the actual records (HOMMA; SALTELLI, 1996). This coefficient compares the difference between the calculated value and the real value, weighting the result with the difference between the average and the actual value, as Eq. 7 shows. Thus, the closer to 1 the coefficient of determination R^2 is, the better the model calculates the dependent variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y_i} - y_i)^2}{\sum_{i=1}^{N}(\bar{y} - y_i)^2} \tag{7}$$

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

We created artificial gaps in the dataset to simulate the real scenario where the lack of data occurs randomly as a second validation. A total of 20% of the data set was randomly chosen to validate the trained models.

Furthermore, we created artificial gaps in the inputs by removing some input information to simulate different scenarios with different combinations of parameter missingness. As we built this methodology to take advantage of the information available, we analyzed the following scenarios: one weather parameter available, two weather parameters available, three weather parameters available, and so on, until seven weather parameters available. Then, we combined all those scenarios with the triangulation values. The 0 constant replaced the gaps, as suggested by Han et al. (2011), one it is a value out of the bounds of all variables used in this study.

The algorithm was applied in data sets from cities with different climatic characteristics to work as a further validation method. For each data set, the methodology adopted was performed 30 times to generate sufficient material to make a statistical analysis. Furthermore, the analysis done was proposed in Carrano et al. (2011), which demands to build an empirical probability distribution function for the mean values of R^2 through bootstraps of the data produced by each method (PEREIRA et al., 2019). Then, it compares those functions using the method of analysis of variance – ANOVA - (FISHER, 1919) and Tukey's multiple comparison test. Finally, the models are arranged from the best to the worst. In other words, a model with mean values better than the other ones always lies in the first position. Moreover, we analyze whether the p-values between two models is lower than 0.05 to conclude that the methods are statistically different at such a confidence level of 95%.

Table 1 summarizes the models analyzed according to the different inputs created to simulate a real scenario with missing values. The models from 6 to 29 represent the base-level learners based on machine learning algorithms. It is noticeable that despite being eight models from each machine learning algorithm, the table shows nine inputs. Since the algorithm estimates the values from a given feature, this information must not be part of the set of input. For example, when imputing evaporation rate values, the models from 7 to 13 run with the inputs: date + insolation or date + rainfall or date + average compensated temperature or date + maximum temperature or date + minimum temperature or date + average relative humidity or date + average wind speed. The same configuration is observed to models 15-21 (SVM) and 23-29 (ANNs).

Table 1 – Codes used for each model in the presentation of the results.

Model Code	Method	Input
1	Arithmetic Average	Data from Three Neighboring Cities
2	Inverse Distance Weighted	Data from Three Neighboring Cities
3	Optimized Inverse Distance Weighted	Data from Three Neighboring Cities
4	Optimized Normal Ratio	Data from Three Neighboring Cities
5	Regional Weight	Data from Three Neighboring Cities
6	Bagged Trees	Date
7-13		Date + 1 weather parameter
14	Support Vector Machine	Date
15-21		Date + 1 weather parameter
22	Artificial Neural Networks	Date
23-29		Date + weather parameter
30	Bagged Trees	Six Missing Parameters
31		Five Missing Parameters
32		Four Missing Parameters
33		Three Missing Parameters
34		Two Missing Parameters
35		One Missing Parameter
36		No Missing Parameters
37	Artificial Neural Networks	Six Missing Parameters
38		Five Missing Parameters
39		Four Missing Parameters
40		Three Missing Parameters
41		Two Missing Parameters
42		One Missing Parameter
43		No Missing Parameters
44	Support Vector Machine	Six Missing Parameters
45		Five Missing Parameters
46		Four Missing Parameters
47		Three Missing Parameters
48		Two Missing Parameters
49		One Missing Parameter
50		No Missing Parameters

6.1 Summarized results

As discussed previously, we applied the framework to 10 datasets, which produced a large amount of data. Hence, for the sake of brevity, space restriction and to preserve the readability of the text, only the most relevant results are presented in this paper.

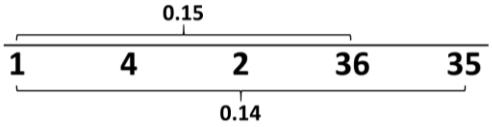
In Table 2 are summarized the results of the whole set of datasets used in this study. The first column contains the weather parameters under which we developed our study to impute missing values from their time series. The second and third columns contain the base-level models and meta-learning models that stood out the most on the ten datasets on the level-0 estimations and level-1 estimations. Finally, the last column represents the best model from the whole set of models. Thus, the meta-learning technique achieved its objective to recognize which base-level learning estimates the missing values with more accuracy, allowing the estimation process's optimization. Furthermore, we observe that different base-learning models present better performance for each weather parameter than the other level 0 models. Nevertheless, the robustness of our approach enabled meta-learning models with the same configuration to recognize the models with the best outputs when estimating different features, emphasizing its adaptability.

Table 2 – Summarization of the best models when estimating weather parameters.

Imputed Parameter	Best Model - Level 0	Best Model - Level 1	Best Model
Evaporation Rate	5	36	36
Insolation	1	36	36
Rainfall	5	36	36
Average Compensated Temperature	10	50	50
Maximum Temperature	2	36	36
Minimum Temperature	1,2,10	50	50
Average Relative Humidity	1,10	36	36
Average Wind Speed	5	36	36

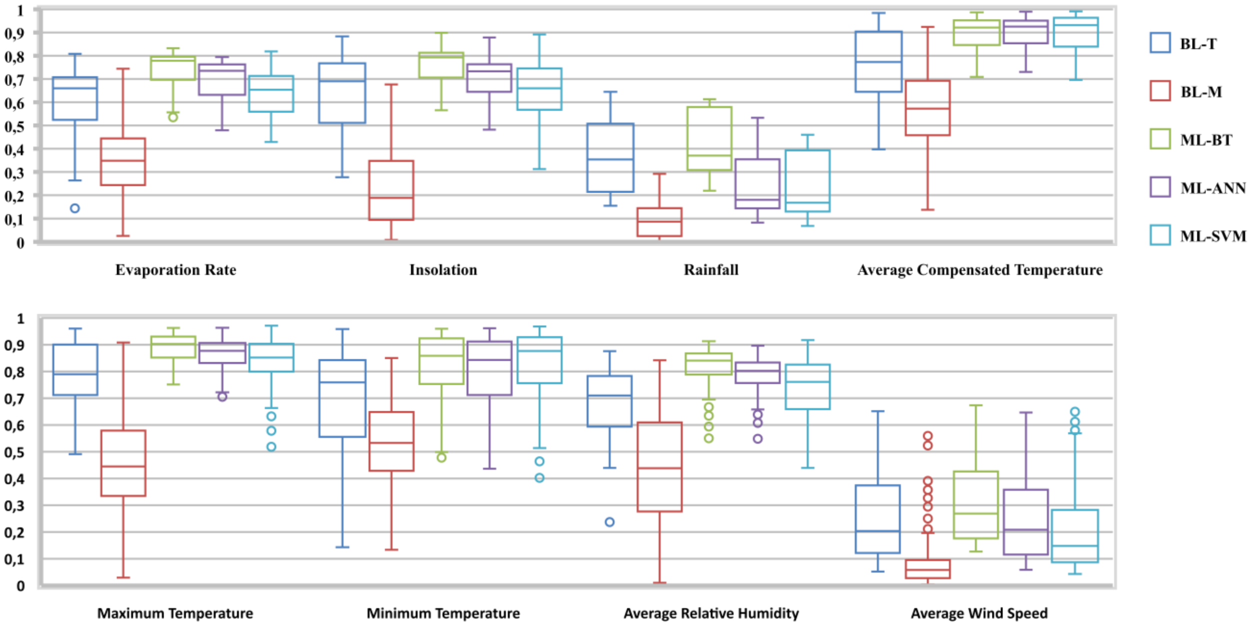
Figure 4 shows the five models with the best performances when estimating rainfall missing values from Belo Horizonte city. In the representation, when two or more models are overlined together, the *p*-value between them is greater than 0.05. In other words, it means that there is no statistical evidence to set these models apart from each other. The brackets indicate the *p*-values related to each comparison. Thus, the *p*-values between model 1 and models 36 and 35 show that the triangulation and the meta-learning models are tied.

Figure 4 – Order of the models according to *p*-value generated, using the statistical test – Belo Horizonte's rainfall estimators.



Source: The Authors (2021).

Figure 5 – Boxplots representing the median of the coefficients of determination R² of the 30 executions of the ten groups for each imputed weather parameter.



Source: The Authors (2021).

Figure 5 shows the boxplots representing the median of the coefficients of determination R² of the 30 executions of the ten groups for each weather parameter imputed. BL-T and BL-M represent the base-learning models based on triangulation methods and machine learning methods, respectively. ML-BT, ML-ANN and ML-SVM represent the meta-learning models based on bagged trees, artificial neural networks and support vector machine, respectively. It is noticeable that the meta-learning model based on bagged trees is always one of the models that present the highest medians of R^2 values for the whole set of imputed parameters, which is expected from this algorithm that captures nonlinear dependencies among the inputs and the output.

Another aspect that Figure 5 shows is that the boxplots representing the meta-learning models' performance present a small interquartile range revealing its precision in estimating the optimized outputs regardless of the different data sets used in this work. Furthermore, all models achieve lower values of R^2 when imputing missing values on rainfall and average wind speed time series compared to the other parameters.

7 CONCLUSION

In this paper, we proposed an alternative solution to the problem of imputing missing values in a MCAR weather time series. Instead of seeking an optimal input combination and the best model to fit each weather parameter, we showed that using a meta-learning model to optimize more straightforward approaches yields satisfactory results even under a weak correlation between parameters. As an outcome, our proposal brings the following advantages:

- It automates the process of choosing the best model out of a set of estimators. We argue that such an automatization can lead to advantages in the field of imputing parameters from different nature when for each weather parameter, a different model stands out from the others and when it allows less complex approaches at the base-learning level; and
- It includes triangulation methods to improve the imputation process. Those methods bring to our analysis the spatial relationship between the variables in addition to the temporal relationship.

Furthermore, we showed that high values of R^2 were achieved even when there were fewer weather parameters to improve the imputation, which does not occur when only the correlation between the weather parameter is explored.

Moreover, we comparatively tested base-learning models and meta-learning models in this work over well-known regression problems. As a result, we found out that a meta-learning model compared to the best base-learning model has either similar or better performance.

Meta-learning is a relatively new field of research. Thus, this article also contributes to the study of this field applied to the meteorological context and the improvement of results.

Acknowledgments

The authors thank the INMET and database sector of CPTEC/INPE 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.

Authors' Contribution

Both authors participated in the stages of conceptualization, formal analysis, investigation, methodology, project administration, resources, validation and data visualization. In addition, author Vinícius Henrique Antunes Alves contributed to the curation of data, software and the initial draft of the work. The author Marconi de Arruda Pereira contributed to the supervision, revision and editing of the work.

Interest conflicts

The authors declare that there are no conflicts of interest.

References

ACUNA, E.; RODRIGUEZ, C. The treatment of missing values and its effect on classifier accuracy. **Classification, clustering, and data mining applications**. p.639–647, 2004. Springer.

AL SHALABI, L.; SHAABAN, Z.; KASASBEH, B. Data mining: A preprocessing engine. **Journal of Computer Science**, v. 2, n. 9, p. 735–739, 2006.

ALVES, V. H. A.; PEREIRA, M. A. **A Meta-Learning Framework for Imputing Missing Values in Weather Time Series**. , p. 12, 2020.

ASSIS, C. A. S. **Predição de Tendências em Séries Financeiras Utilizando Meta-Classificadores**, abr. 2019. PhD Thesis, Centro Federal de Educação Tecnológica de Minas Gerais.

BARBOSA, M.; CARVALHO, M. Sistemas de Armazenamento de Dados Observados do CPTEC/INPE. **Instituto Nacional de Pesquisas Espaciais**, 2015.

BAYMA, L. O.; PEREIRA, M. A. Identifying Finest Machine Learning Algorithm for Climate Data Imputation in the State of Minas Gerais, Brazil. **Journal of Information and Data Management**, v. 9, n. 3, p. 259–259, 2018.

BAYMA, L. O.; PEREIRA, M. DE A. Comparison of machine learning techniques for the estimation of climate missing mata in the State of Minas Gerais, Brazil. GEOINFO. **Anais...** . p.283–294, 2017.

BRAZDIL, P.; CARRIER, C. G.; SOARES, C.; VILALTA, R. **Metalearning: Applications to data mining**. Springer Science & Business Media, 2008.

CARRANO, E. G.; WANNER, E. F.; TAKAHASHI, R. H. C. A Multicriteria Statistical Based Comparison Methodology for Evaluating Evolutionary Algorithms. **IEEE Transactions on Evolutionary Computation**, v. 15, n. 6, p. 848–870, 2011.

CRISTIANINI, N.; SHAWE-TAYLOR, J.; OTHERS. **An introduction to support vector machines and other kernel-based learning methods**. Cambridge university press, 2000.

DANNENBERG, R. B.; THOM, B.; WATSON, D. **A Machine Learning Approach to Musical Style Recognition**. , p. 4, 1997.

FISHER, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. **Transactions of the Royal Society of Edinburgh**, v. 52, n. 2, p. 399–433, 1919.

GARCÍA, S.; FERNÁNDEZ, A.; LUENGO, J.; HERRERA, F. A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. **Soft Computing**, v. 13, n. 10, p. 959, 2009. Springer.

GRUBER, T.; WILLBERG, M. Signal and error assessment of GOCE-based high resolution gravity field models. **Journal of Geodetic Science**, v. 9, n. 1, p. 71–86, 2019.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. Elsevier, 2011.

HOMMA, T.; SALTELLI, A. Importance measures in global sensitivity analysis of nonlinear models. **Reliability Engineering & System Safety**, v. 52, n. 1, p. 1–17, 1996. Elsevier.

HUBBARD, K. Spatial variability of daily weather variables in the high plains of the USA. **Agricultural and Forest Meteorology**, v. 68, n. 1–2, p. 29–41, 1994. Elsevier.

JOCKERS, M. L.; WITTEN, D. M. A comparative study of machine learning methods for authorship attribution. **Literary and Linguistic Computing**, v. 25, n. 2, p. 215–223, 2010.

KHOSRAVI, G.; NAFARZADEGAN, A. R.; NOHEGAR, A.; FATHIZADEH, H.; MALEKIAN, A. A modified distance-weighted approach for filling annual precipitation gaps: application to different climates of Iran. **Theoretical and Applied Climatology**, v. 119, n. 1, p. 33–42, 2015. Springer.

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

LITTLE, R. J.; RUBIN, D. B. **Statistical analysis with missing data**. John Wiley & Sons, 2019.

MAUDSLEY, D. B. A THEORY OF META-LEARNING AND PRINCIPLES OF FACILITATION: AN ORGANISMIC PERSPECTIVE. , 1980.

OLCESE, L. E.; PALANCAR, G. G.; TOSELLI, B. M. A method to estimate missing AERONET AOD values based on artificial neural networks. **Atmospheric Environment**, v. 113, p. 140–150, 2015. Elsevier.

PAULHUS, J. L.; KOHLER, M. A. Interpolation of missing precipitation records. **Monthly Weather Review**, v. 80, n. 8, p. 129–133, 1952.

PEREIRA, M. DE A.; CARRANO, E. G.; DAVIS JÚNIOR, C. A.; DE VASCONCELOS, J. A. A comparative study of optimization models in genetic programming-based rule extraction problems. **Soft Computing**, v. 23, n. 4, p. 1179–1197, 2019.

PIROOZNIA, M.; YANG, J. Y.; YANG, M. Q.; DENG, Y. A comparative study of different machine learning methods on microarray gene expression data. **BMC Genomics**, v. 9, n. Suppl 1, p. S13, 2008.

ROBUSTO, C. C. The cosine-haversine formula. **The American Mathematical Monthly**, v. 64, n. 1, p. 38–40, 1957. JSTOR.

SILVERMAN, D.; DRACUP, J. A. Artificial Neural Networks and Long-Range Precipitation Prediction in California. **JOURNAL OF APPLIED METEOROLOGY**, v. 39, p. 10, 2000.

SOLOMATINE, D. P.; OSTFELD, A. Data-driven modelling: some past experiences and new approaches. **Journal of hydroinformatics**, v. 10, n. 1, p. 3–22, 2008. IWA Publishing.

STAKE, R. E. **The art of case study research**. Thousand Oaks: Sage Publications, 1995.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Practical machine learning tools and techniques**. Elsevier, 2005.

WOLPERT, D. H. Stacked generalization. **Neural Networks**, v. 5, n. 2, p. 241–259, 1992.

YANG, Y.; LIN, H.; GUO, Z.; JIANG, J. A data mining approach for heavy rainfall forecasting based on satellite image sequence analysis. **Computers & geosciences**, v. 33, n. 1, p. 20–30, 2007. Elsevier.

YOUNG, K. C. A three-way model for interpolating for monthly precipitation values. **Monthly Weather Review**, v. 120, n. 11, p. 2561–2569, 1992.

Author’s biography

Vinícius Henrique Antunes Alves, was born in 1997 in Itaúna, MG. He is currently graduating from the Mechatronics Engineering course at the Federal University of São João del-Rei, Ouro Branco, MG. He has experience in Geoinformatics in the field of machine learning. He has an article published at the Brazilian Symposium on Geoinformatics, 21 (GEOINFO) entitled The Meta-Learning Framework for Imputing Missing Values in Weather Time Series.