Revista Brasileira de Meteorologia, **v. 38,** e38220030, 2023 rbmet.org.br
DOI: http://dx.doi.org/10.1590/0102-778638220030

Article

Application of a Computational Hybrid Model to Estimate and Filling Gaps for Meteorological Time Series

Eluã Ramos Coutinho[1], Jonni Guiller Ferreira Madeira[2], Robson Mariano da Silva[3], Elizabeth Mendes de Oliveira[2], Angel Ramon Sanchez Delgado[3]

[1]Departamento de Engenharia Civil, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil.
[2]Departamento de Matemática, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Angra dos Reis, RJ, Brazil.
[3]Departamento de Matemática, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, Brazil.

Received: 29 March 2022 - Accepted: 9 August 2022

Abstract

The present study applies computational intelligence techniques in the development of a hybrid model composed of Artificial Neural Networks (ANNs) and Genetic Algorithms (GAs) (MLP-GA) to estimate and fill in the gaps in the monthly variables of evaporation, maximum temperature and relative humidity to six regions in the state of Rio de Janeiro (RJ), Brazil. The results were evaluated using statistical techniques and compared with results obtained by the Multiple Linear Regression (RLM), Multilayer Perceptron (MLP) and Radial Basis Function (RBF) models and also compared with the data recorded by the weather stations. The correlation coefficient (*r*) between the evaporation estimates generated by MLP-GA with the recorded data showed a high relationship, remaining between 0.82 to 0.97. The average percentage error (*MPE*) ranged from 6.01% to 9.67%, indicating a accuracy between 90% to 94%. For the maximum temperature generated by MLP-GA the correlation with the recorded data remained between 0.97 to 0.99. It also presented the *MPE* between 0.95% to 1.57%, maintaining the accuracy of the estimated data between 98% to 99%. The correlation coefficient (*r*) between the relative humidity estimates generated with the MLP-GA remained between 0.89 a 0.97, the *MPE* between 1.15% to 1.89%, which guaranteed a rate higher than 98% of correctness in its estimates. Such results demonstrated gains in relation to the other applied models and allowed the accomplishment of the filling of most of the missing values.

Keywords: fault filling, Artificial Neural Networks, Genetic Algorithms.

Aplicação de um Modelo Computacional Híbrido para Estimar e Preencher Falhas em Séries Temporais Meteorológicas

Resumo

O presente estudo aplica técnicas de inteligência computacional no desenvolvimento de um modelo híbrido composto por Redes Neurais Artificiais (RNAs) e Algoritmos Genéticos (AGs) (MLP-GA) para estimar e preencher lacunas nas variáveis mensais de evaporação, temperatura máxima e umidade relativa em seis regiões do estado do Rio de Janeiro (RJ), Brasil. Os resultados foram avaliados por meio de técnicas estatísticas e comparados com os resultados obtidos pelos modelos de Regressão Linear Múltipla (RLM), Perceptron de Multicamadas (MLP) e Redes de Função de Base Radial (RBF), além de serem comparados com os dados registrados pelas estações meteorológicas. O coeficiente de correlação (*r*) entre as estimativas de evaporação geradas pelo MLP-GA com os dados registrados mostrou uma relação elevada, permanecendo entre 0,82 e 0,97. O erro percentual médio (*MPE*) variou de 6,01% a 9,67%, indicando uma precisão entre 90% e 94%. Para a temperatura máxima gerada pelo MLP-GA, a correlação com os dados registrados permaneceu entre 0,97 e 0,99. Apresentou também o *MPE* entre 0,95% e 1,57%, mantendo a precisão dos dados esti-

Autor de correspondência: Eluã Ramos Coutinho, eluaramos@hotmail.com.

mados entre 98% e 99%. O coeficiente de correlação (*r*) entre as estimativas de umidade relativa geradas pelo MLP-GA permaneceu entre 0,89 e 0,97, com *MPE* entre 1,15% e 1,89%, garantindo uma taxa superior a 98% de acerto em suas estimativas. Tais resultados demonstraram ganhos em relação aos outros modelos aplicados e permitiram o preenchimento da maioria dos valores ausentes.

Palavras-chave: preenchimento de falhas, Redes Neurais Artificiais, Algoritmos Genéticos.

1. Introduction

Monitoring natural phenomena by observing meteorological information helps to understand the climatic characteristics of a region (Pappas *et al.*, 2014). These are extremely important for decision making in areas such as agriculture, energy, transport, ecology, safety and health (Brito *et al.*, 2016; Coutinho *et al.*, 2018). Therefore, continuous and reliable time series are necessary. However, problems such as imprecise functioning of monitoring equipment, extreme weather conditions, absence of observers, human errors and other factors make the availability of continuous series a rarity, with failures or lack of records one of the problems commonly occurring in meteorological data series (Tardivo and Berti, 2014; Woldesenbet *et al.*, 2016; Dembélé *et al.*, 2019; Vega-Garcia *et al.*, 2019).

Failures are a common problem, this situation can make research unfeasible, hinder the use of information and even make it impossible to understand the climate of a region, limiting the understanding of the spatial or temporal variability of various meteorological and hydrological processes (Wanderley *et al.*, 2014). Thus, the solution to this problem is the reconstruction of time series by estimating and filling in the missing information (Tardivo and Berti, 2014; Anjomshoaa and Salmanzadeh, 2018).

Over the years, several methods have been used to estimate and fill in missing data, such as averages, spatial interpolation, inverse distance weighting, linear regression, logistic regression, multiple regression, kriging, remote sensing, among others (Teegavarapu and Chandramouli, 2005; Pappas *et al.*, 2014; Tardivo and Berti, 2014). More details on these methodologies can be found in the studies of Wanderley *et al.* (2012), Samanta *et al.* (2012), Eccel *et al.* (2012), Clack (2016), Brito *et al.* (2016), Woldesenbet *et al.* (2016), Anjomshoaa and Salmanzadeh (2018), Brubacher *et al.* (2020); Giovanella *et al.* (2021) and in Liu *et al.* (2017).

These methodologies are generally applied to estimate and fill meteorological data, which may require information from other locations and often only consider the proximity between the locations, discarding the climatic differences that may occur due to the relief and the altitude. Thus, the methodology can present low quality estimates. Another negative aspect is the difficulty of representing extreme situations with non-linear trends (Silva *et al.*, 2018; Dembélé *et al.*, 2019; Aieb *et al.*, 2019; Ren *et al.*, 2019). These factors have directly influenced the use of techniques of Artificial Intelligence that are characterized by the attempt to reproduce human knowledge or natural biological processes, adapting to different situations and managing to extract characteristics and infer responses from a set of data. (Haykin, 2001; Russell and Norvig, 2013).

There are several studies with this theme in the literature: Teegavarapu and Chandramouli (2005) applied the techniques of artificial neural networks (ANN) and the Krigagem model to estimate absent precipitation data from 20 pluviometric stations in the state of Kentucky in the United States of America; Coulibaly and Evora (2007) compared 6 types of ANNs to fill in missing records of daily precipitation and temperature from 15 Gatineau weather stations in northeastern Canada; Kim and Pachepsky (2010) applied ANN to reconstruct daily precipitation data from 39 meteorological stations in the Chesepeake Bay watershed in the USA; Yozgatligil *et al.* (2013) compared different techniques, including Multilayer Perceptron ANN (MLP) to fill in missing values for total monthly precipitation time series and average monthly temperature for stations belonging to 7 regions of Turkey; Ford and Quiring (2014) evaluated the performance of ANNs, weighting for inverse distance, mean, kriging and spatial regression for daily filling of soil moisture in Oklahoma in the USA; Wanderley *et al.* (2014) applied an ANN to fill monthly rainfall data gaps in the state of Alagoas, Brazil; Canchala-Nastar *et al.* (2019) used ANN to fill in missing precipitation data from 45 stations located in southwest Colombia; Gunawardena *et al.* (2022) compare multivariate linear regression model and artificial neural networks to predict and fill gaps in meteorological data in southeastern France; Brubacher *et al.* (2020) applied multiple linear regression and artificial neural networks to fill historical series of daily rainfall in Rio Grande do Sul; Aschauerand Marty (2021) compare the inverse distance weighted, elastic network regression and random forest for time series forecasting of depth of snow in Switzerland; Vega-Garcia *et al.* (2019) used a feed-forward ANN to estimate and populate precipitation data from 5 stations located in the Ebro river basin in Spain.

However, even if models based on artificial intelligence present satisfactory results, the definition of estimating variables and the individual choice of hyperparameters for each model require time and in-depth knowledge of the technologies applied. For this reason, the present study aims to apply a hybrid methodology composed by the ANN junction of Multilayer Perceptron and Genetic Algorithms (MLP-GA) to create a model with autonomous training and adjustment characteristics, which can estimate and reconstruct meteorological data. In addition, to prove the efficiency of the proposed model, its results are compared statistically with those presented by MLP, Radial Basis Function (RBF) and Multiple Linear Regression (RLM) models.

Section 2 describes the study areas, data sets, preprocessing, proposed model to estimate and fill gaps, the models used for comparison and the methods applied to evaluate the performance of each model. Section 3 presents the results obtained by the models for each region. Section 4 discusses and compares the results with others found in the literature and section 5 presents the conclusion of the study.

2. Material and Methods

2.1. Data and study site

The data used in this study were provided by the National Meteorological Institute of Brazil (INMET). The information is monthly averages of evaporation, maximum temperature and relative humidity, recorded during the period from 05/31/2002 to 12/31/2012, adding 128 data for each variable of each season.

This series of meteorological information belongs to six stations located in the municipalities of Campos dos Goytacazes (CG) (21.74° S; 41.33° W and 11.20 m), Cordeiro (CO) (22.02° S; 42.36° W and 505.92 m), Itaperuna (IT) (21.20° S; 41.90° W and 123.59 m), Rio de Janeiro (RJ) (22.89° S; 43.18° W and 11.10 m), Paty do Alferes (PA) (-22.35° S; -43.41° W and 507 m), and Resende (RE) (-22.45° S; -44.44° W and 439.89 m), located in Rio de Janeiro state, Brazil (Fig. 1).

The Rio de Janeiro state is located in the southeastern region of Brazil and borders the states of Espírito Santo, Minas Gerais, São Paulo and the Atlantic Ocean (Brito *et al.*, 2016). The state is characterized by the second largest metropolis and for being the largest oil producer in the country. It has an approximate population of 17,264,943 and its territorial extension is 43,750,423 km^2 (IBGE, 2020). It also presents a great climatic diversity due to the relief, altitude and its proximity to the Atlantic

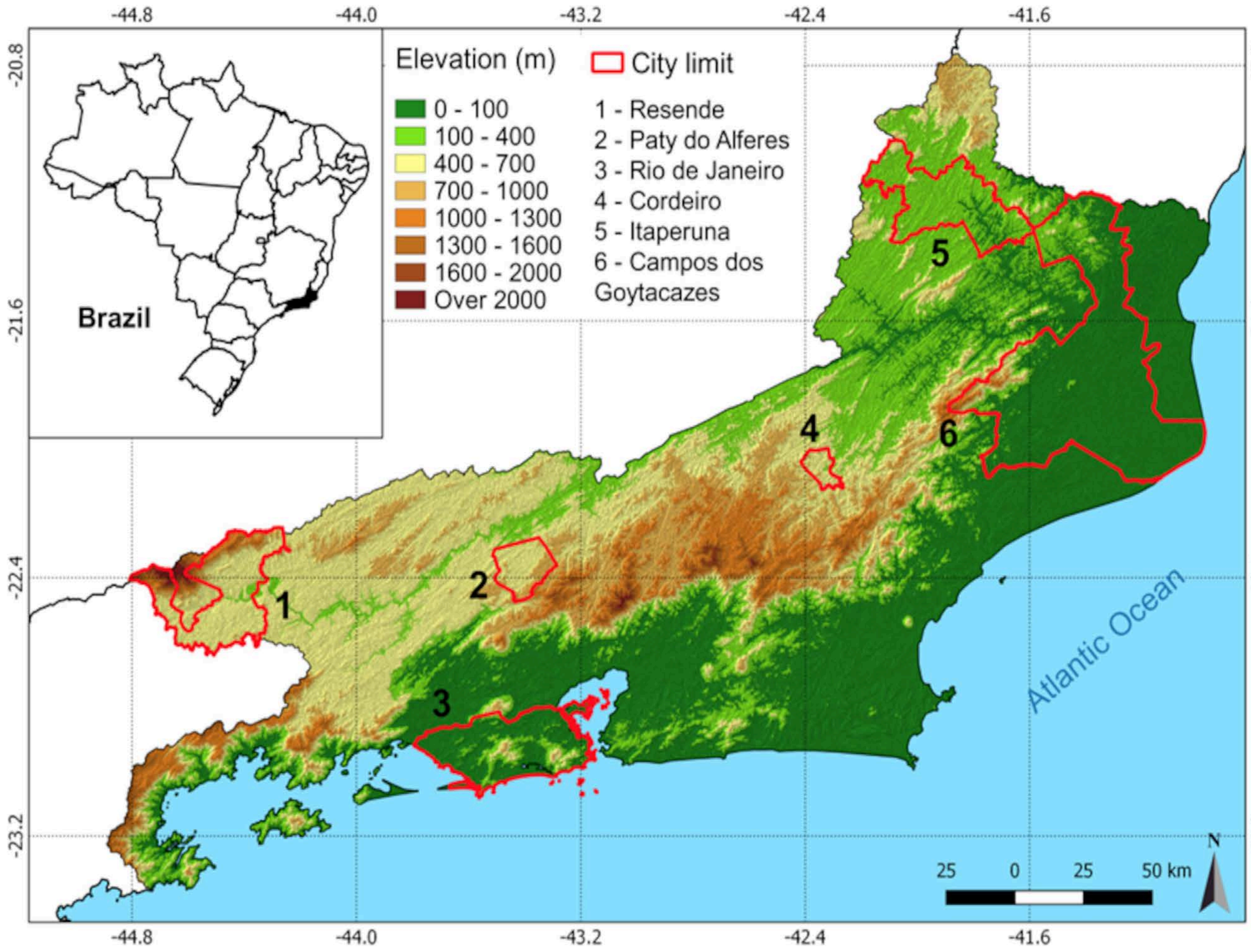

Figure 1 - Map of the state of Rio de Janeiro with the regions set used in the study.

Ocean, and for that reason it is possible to observe the predominance of tropical semi-humid, tropical altitude and topical climates (Bastos and Napoleão, 2011).

Thus, the locations used in this study were determined in order to represent the climatic diversity found in the different regions of Rio de Janeiro state.

2.2. Proposed model for filling gaps

The fill proposal in the occurred failures in the six meteorological stations was to compare the results of a hybrid MLP-GA model with the different techniques commonly used and thus apply the model that presented the best result in filling each variable in its respective region. For this, the following methodology consisted of carrying out several steps, ranging from the preparation of information to the evaluation of the results obtained by the methods (Fig. 2).

In the identification failure and inconsistencies stage, it was possible to verify the percentage of missing values for each meteorological variable in each location (Table 1).

After this stage, the missing data were removed from all stations as described in the methodology addressed in the study by Coutinho *et al.* (2018), which is based on the missing data of a variable for a given period in a given station. The main objective of this methodology is to create a homogeneous training and model testing set, where data from the same period from other stations should be taken by creating a set with the same records. For example, if station x in the set of stations did not have the evaporation record or the monthly average of the maximum temperature or relative humidity for the period of 30/04/2008, the same must be removed from all other stations. This process guarantees a homogeneous data set, making all stations have the same record numbers, totaling about 108 records for evaporation, 90 for maximum temperature and 64 for relative humidity.

Table 1 - Number of occurred failures in the period from 05/31/2002 to 12/31/2012.

	Missing data (%)		
Stations	Evaporation	Maximum temperature	Relative humidity
CG	2.34	5.47	3.91
CO	2.34	3.13	17.97
IT	0.00	3.10	5.47
RJ	2.34	5.47	5.47
PA	0.78	5.47	8.59
RE	7.81	13.28	21.09

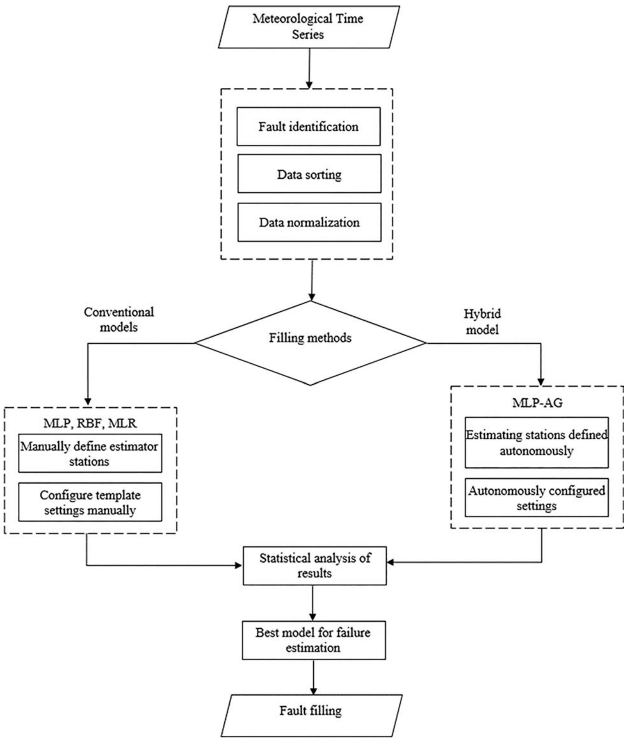
Meteorological Time Series
Fault identification
Data sorting
Data normalization
Conventional models
Filling methods
Hybrid model
MLP, RBF, MLR
Manually define estimator stations
Configure template settings manually
MLP-AG
Estimating stations defined autonomously
Autonomously configured settings
Statistical analysis of results
Best model for failure estimation
Fault filling

Figure 2 - Model applied to fill in failures.

With the removal of the missing information, the data set was normalized, changing the values actual scale to an interval between zero and Eq. (1). Such transformation was intended to encode all attributes at similar intervals, making all data have the same importance. This facilitates the adjustment of training algorithms and also the presentation of better results (Coutinho *et al.*, 2016).

$$x_j^{norm} = \frac{x_j - x^{min}}{x^{max} - x^{min}} \tag{1}$$

where x_j^{norm}: normalized variable; x_j: variable in position j; x^{min}: minimum value observed between variables; x^{max}: maximum observed value between variables.

To compare the values observed and estimated by the techniques, the information set for each variable was divided and submitted to the Multiple Linear Regression models, Multilayer Perceptron Networks (MLP), Radial Base Function Networks (RBF) and the Hybrid model composed by Genetic Algorithms and by ANN MLP (MLP-GA) in two parts: 75% to training/adjustment and 25% to validation.

The validation stage consisted of submitting the estimator data set to the models to estimate each of the data for the variables of evaporation, maximum temperature or relative humidity. efficiency was assessed using statistical techniques applied to the results obtained. Once the model's ability to predict the submitted variable was confirmed, the set of data belonging to the stations determined as estimators removed in the screening process was used to fill in the real gaps. Thus, if the Rio de Janeiro station does not have the maximum temperature measurement for the period of 30/04/2006, but the other remaining stations have it, then these data are submitted to the models to estimate the data that will be filled in the Rio de Janeiro station.

2.3. Filling methods

2.3.1. Multiple linear regression

Multiple linear regression is a technique that analyzes or relates a dependent variable to several independent variables (Fonseca *et al.*, 2012). The relationship between a dependent variable Y and other independent variables (X_1, X_2, X_3) is formulated by the following linear model Eq. (2) (Sousa *et al.*, 2007):

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \qquad (2)$$

The resolution of this problem is linked to the estimation of the values of the parameters α, β_1, $\beta_{2,}$ β_k which can be performed by the method of least squares, which aims to determine values of α e β, minimizing the sum of the squared errors (more information about the application and resolution of this model can be found at Sousa *et al.*, 2007; Lyra *et al.*, 2011; Coutinho *et al.*, 2016; Coutinho *et al.*, 2018; Brubacher *et al.*, 2020; Dias and Soares, 2021).

2.3.2. Multilayer perceptrons (MLP)

The MLP is a supervised neural network that belongs to the feed-foward class. Its structure is completely connected, consisting of an input layer, one or more hidden layers and an output layer (Haykin, 2001; Coulibaly and Evora, 2007; Russell and Norvig, 2013) (Fig. 3). The flexibility of application of the model and its ability to present favorable results make it widely applicable in the complex problems resolution such as pattern recognition, classification, forecasting, image processing and reconstruction of missing information (Shah and Ghazali, 2011; Anochi and Campos Velho, 2015).

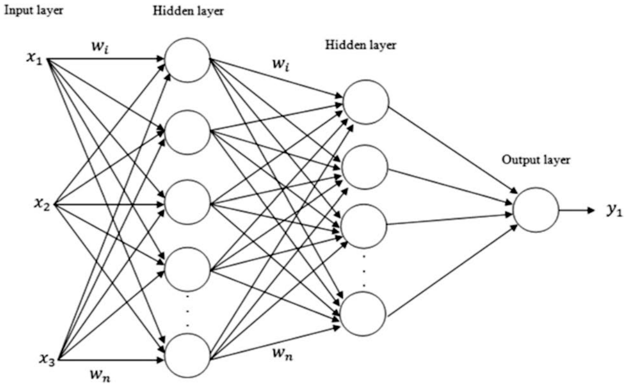

Figure 3 - ANN MLP architecture applied to estimate the maximum temperature, relative humidity and evaporation data.

The operation of this model consists of extracting characteristics from a known data set, which occurs through the adjusting weights process (w_i) by mapping and inputs data set (x_i) and outputs (y_i). This adjustment is performed by the Back-propagation algorithm or by its Quasi-Newton Back-propagation, Resilient Back-propagation, Levenberg-Marquardt Back-propagation variations (more information about the application, training types and transfer functions can be found at Haykin 2001; Braga *et al.*, 2012; Coutinho *et al.*, 2016; Coutinho *et al.*, 2018; Milidonis *et al.*, 2021).

2.3.3. Radial basis function networks (RBF)

The RBF ANN is activated by the function of the distance between its input vectors, centers, intermediate or hidden layer. The method uses radial base functions and aims to group the input data into clusters and transform a set of non-linearly separable input patterns in a set of linearly separable outputs. The output layer has the function of classifying the patterns received from the previous layer through the linear combination of the functions outputs (Braga *et al.*, 2012; Haykin (2001); Coutinho *et al.*, 2016). Fig. 4 demonstrates the basic architecture of an RBF-type ANN.

2.3.4. Configuration of the MLP-GA hybrid model

Unlike the previous models, where ANN characteristics such as number of entries, transfer functions, training methods, learning rates and others are defined manually, in the MLP-GA hybrid model, genetic algorithms (GA) were used to define the characteristics autonomously, using its evolutionary aspect to arrive at the best combination. Fig. 5 demonstrates the use of GA in conjunction with ANN MLP.

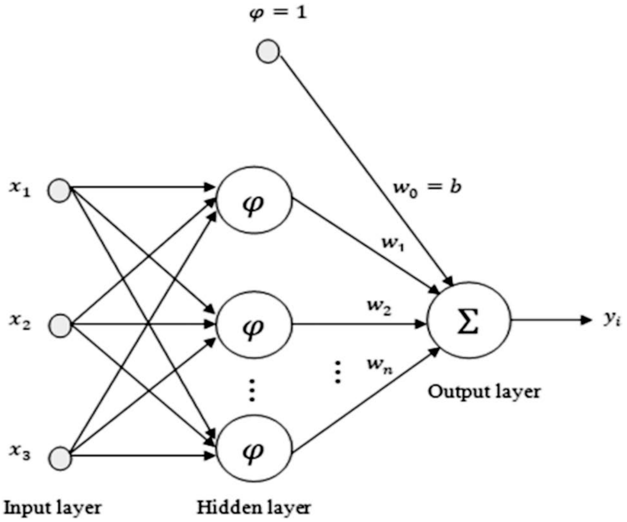

Figure 4 - Basic structure of a RBF ANN.

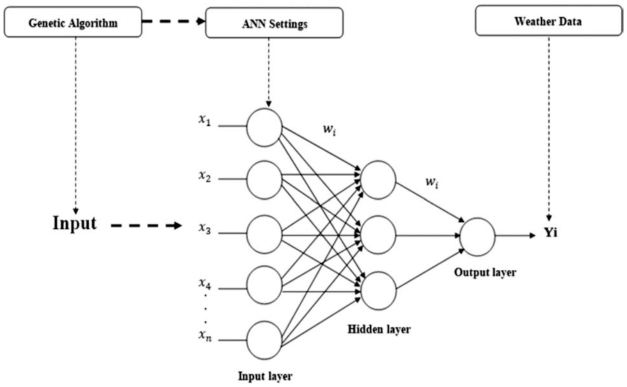

Figure 5 - The MLP-GA structure model.

In this model, GA was responsible for making several combinations, evaluating the training and validation data set, transfer the first and second intermediate layers functions, training algorithm, learning rate and momentum rate at each moment. To determine these characteristics, each chromosome or individual in a population was encoded in binary in a structure with 14 bits (Fig. 6). This representation allows the description of various features and has been adopted by other researchers pesquisadores (Ghareeb e Saadany, 2013; Haidar e Verma, 2016; Ventura *et al.*, 2019).

The initial 5 bits belonging to (a) are associated with the variables belonging to each of the stations that can be used to fill a fault. For example, if this model was being executed to fill a gap in the relative humidity variable belonging to the Campos dos Goytacazes station, then these five bits indicate whether or not to use the relative humidity data for the regions of Paty do Alferes, Cordeiro, Itaperuna, Resende and Rio de Janeiro. This is achieved through the binary configuration, in which each gene that has a value of 1 identifies that the data set belonging to a station x is active and should be used, and the value 0 identifies that it is inactive and should be ignored.

The part of the chromosome belonging to (b) and (c) identifies which function will be applied to the first and second intermediate layers. These bits allow the choice between the step activation functions (satlin) (1 1) Eq. (3), Linear (purelin) (0 0) Eq. (4), Hyperbolic Tangent (tansig) (0 1) Eq. (5) and Sigmoid (logsig) (1 0) Eq. (6):

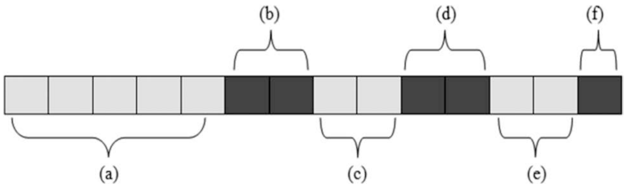

Figure 6 - Example of a chromosome adopted to configure the ANN.

$$\mathrm{satlin}(u_i) = \begin{cases} 1 & \text{se } u_i \geq 0 \\ 0 & \text{se } u_i < 0 \end{cases} \tag{3}$$

$$purelin(u_i) = u_i \tag{4}$$

$$f(u_i) = tgh\left(\frac{u_i}{2}\right) = \frac{1-\exp(-u_i)}{1+\exp(-u_i)} \tag{5}$$

$$f(u_i) = \frac{1}{(1+\exp^{(-u_i)})} \tag{6}$$

The space belonging to (d) has the responsibility to make the choice between the training types, being able to choose the Back-propagation-traingd (0 1) algorithm or its variations, Quasi-Newton Back-propagation - trainbfg (1 0), Resilient Back-propagation - RProp - trainrp (0 0), Levenberg-Marquardt Back-propagation - trainlm (1 1).

The learning rate and momentum adopted are obtained through the bits of (e). The learning rate can range from 0.01 to 0.86, and the momentum rate is obtained by multiplying over the learning rate values. Multipliers can vary between 0.8, 0.15, 0.12 and 0.10 depending on the bits choice (1 0), (0 0), (1 1) and (0 1).

The last bit identified by (f) has the creating purpose of an aid factor (λ) to adjust the network, which increases the training data set by inserting a generic set obtained by averaging the data used in the training.

Other important factors of this methodology, such as the configuration of the genetic algorithm, were fixed in 40 individuals, maximum generation numbers equal to 60, use of elitism to propagate about 15 individuals with high evaluations to the next generation, random drawing of 5 individuals with evaluations low to propagate to the next generation, in order to avoid that genetic characteristics contained only in an low fitness individual are lost, creating a population without diversity with similar individuals.

In addition to these parameters, GA also used the tournament method with its size set at 5, both for choosing the first parent and for choosing the second parent, and a two-point crossover operator to generate new individuals.

The GA objective function has been described by minimize the absolute global error. This function was developed based on the creation of an ANN with the parameters defined by an individual to be evaluated. After training this network, its suitability is assessed by estimating the validation data set. The minimization of the value originated by the sum of the differences between the estimated and the expected values for the validation data set is the objective function.

2.4. Definition of model characteristics and stations selection used in data estimation

One of the major problems in using data from other locations to estimate values and fill in the faults of a station x is how to choose the estimator stations. There are several ways to define them, which may be by closer regions, statistical methods such as stepwise linear regression, correlation coefficient between your data, or others (Coutinho et. al, 2018).

However, Serrano *et al.* (2010) points out that there is no general criterion for selecting the appropriate stations, and for this reason, possible combinations were tested and analyzed in this study, where it was decided to use for the MLR, MLP and RBF models three stations chosen in part by the proximity criteria of the regions with the season to be filled. This methodology was adopted in order to compare the manual choice without statistical support with the choices defined autonomously by the MLP-GA hybrid model. From Table 2 it is possible to check the input data and also the settings adopted by all models (MLR, MLP, RBF and MLP-GA).

2.5. Performance evaluation

To assess the models' ability to estimate the variables: evaporation, maximum temperature and relative air humidity, statistical measures were used, such as Pearson's correlation coefficient (r) Eq. (7) which assesses the degree of association between the estimated and observed data, ranging between -1 and 1, with 1 being a perfect correlation, the mean absolute error (MAE) Eq. (8) that evalu-

Table 2 - Characteristics defined manually and automatically used to estimate meteorological data. Legend: Neurons number in the first layer (N1), Neurons number in the second layer (N2), First layer functions (F1), Second layer functions (F2),Training Type (TR), Unused data (X).

Characteristics defined manually

Data	*Stations*	Estimators					MLP			MLR	RBF	
		1°	2°	3°	N1	N2	F1	F2	TR	X	N1	F1
Evaporation, Maximum Temperature, Relative Humidity	CG	CO	IT	RJ	30	15	logsig	tansig	trainbfg	X	30	Gauss
	CO	RJ	IT	CG	30	15	logsig	tansig	trainbfg	X	30	Gauss
	IT	RJ	CO	CG	30	15	logsig	tansig	trainbfg	X	30	Gauss
	RJ	PA	RE	CO	30	15	logsig	tansig	trainbfg	X	30	Gauss
	PA	RJ	RE	CO	30	15	logsig	tansig	trainbfg	X	30	Gauss
	RE	RJ	PA	CO	30	15	logsig	tansig	trainbfg	X	30	Gauss

Characteristics defined in the model MLP-GA

Data	Stations	Estimators					λ	N1	N2	F1	F2	TR
		1°	2°	3°	4°	5°						
Evaporation	CG	CO	IT	PA	X	X	S	30	15	satlin	logsig	traingd
	CO	CG	IT	RE	RJ	X	S	30	15	satlin	satlin	trainbfg
	IT	CO	CG	PA	RE	RJ	S	30	15	purelin	purelin	trainrp
	RJ	CG	PA	X	X	X	N	30	15	logsig	satlin	trainbfg
	PA	CG	IT	RJ	X	X	N	30	15	purelin	purelin	trainrp
	RE	CO	IT	RJ	X	X	N	30	15	purelin	logsig	trainbfg
Maximum temperature	CG	CO	RJ	X	X	X	S	30	15	tansig	satlin	trainbfg
	CO	CG	IT	PA	X	X	S	30	15	satlin	satlin	trainrp
	IT	CG	CO	RJ	X	X	N	30	15	logsig	satlin	trainbfg
	RJ	CG	IT	X	X	X	S	30	15	purelin	logsig	trainlm
	PA	CG	CO	IT	RE	X	N	30	15	purelin	satlin	traingd
	RE	PA	X	X	X	X	N	30	15	purelin	purelin	trainrp
Relative humidity	CG	IT	RJ	PA	X	X	S	30	15	purelin	satlin	trainbfg
	CO	IT	RJ	X	X	X	N	30	15	purelin	logsig	trainrp
	IT	CG	CO	PA	X	X	S	30	15	purelin	purelin	trainbfg
	RJ	CG	CO	IT	PA	X	S	30	15	purelin	tansig	traingd
	PA	CO	IT	RE	X	X	N	30	15	Purelin	satlin	traingd
	RE	CG	CO	PA	X	X	S	30	15	Tansig	logsig	trainbfg

ates the absolute difference between the real and estimated values, root mean square error (*RMSE*) Eq. (9) which measures the root mean square of errors between actual and estimated values, mean absolute percent error (*MPE*) Eq. (10) which presents the average difference between the real and estimated values in percentage, concordance index (*D*) Eq. (11) which measures the predicted values accuracy in relation to estimated values and the confidence index (*C*) Eq. (12) which allows the joint analysis of the precision and accuracy of the results obtained. (Fonseca *et al.*, 2012; Pezzopane *et al.*, 2012; Bruce and Bruce, 2019; Korstanje, 2021; Auffarth, 2021; Aschauerand Marty, 2021; Giovanella *et al.*, 2021; Fine *et al.*, 2022).

$$r = \frac{\dfrac{\sum_{j=1}^{N}(x_j - \bar{x})(o_j - \bar{O})}{N}}{\sqrt{\dfrac{\sum_{j=1}^{N}(x_j - \bar{x})^2}{N}}.\sqrt{\dfrac{\sum_{j=1}^{N}(o_j - \bar{O})^2}{N}}} \tag{7}$$

$$MAE = \frac{\sum_{j=1}^{n}\left|O_j - x_j\right|}{n} \tag{8}$$

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n}(O_j - x_j)^2}{n}} \tag{9}$$

$$MPE = \frac{\sum_{j=1}^{n}\dfrac{\left|O_j - x_j\right|}{O_j}}{n}.100 \tag{10}$$

The confidence index (*C*) Eq. (12) is calculated by the product of the correlation coefficient (*r*) and by the agreement index (*D*) Eq. (11). Their values range from zero (0) for no agreement to one (1) for perfect agreement (Pezzopane *et al.*, 2012). Table 3 shows the performance evaluation criteria.

Table 3 - Criteria for evaluating and analyzing the performance of models based on the confidence index.

IC value	Performance
> 0.85	Optimal
0.76 a 0.85	Verygood
0.66 a 0.75	Good
0.61 a 0.65	Intermediate
0.51 a 0.60	Tolerable
0.41 a 0.50	Bad
≤ 0.40	Terrible

$$D = 1 - \frac{\sum_{j=1}^{n}\left(o_j - x_j\right)^2}{\sum_{j=1}^{n}\left(\left|x_j - \bar{O}\right| + \left|o_j - \bar{O}\right|\right)^2} \tag{11}$$

$$C = (r.D) \tag{12}$$

where: *n* or *N* represents the number of data used, O_j the observed value, x_j the value estimated by the techniques employed, $\bar{O}$ the average of the observed data and $\bar{x}$ the estimated data average.

In addition to these methods, statistical measures of average (M), maximum (MAX), minimum (MIN) and standard deviation (SD) were also used to compare the estimated meteorological data with the stations.

3. Results

3.1. Estimation results and filling in evaporation data

From Table 4 it is possible to see that the models based on artificial intelligence ANNs MLP, RBF and MLP-GA presented results superior to those presented by the MLR model. Analyzing the results highlighted by the measures of (*r*), *RMSE*, *MAE*, (*D*) and (*C*). Another important aspect is that in five of the six locations studied, the MLP-GA model was superior in the evaporation estimates. This fact can be confirmed by analyzing the results of each error measure obtained by the MLP-GA model, which were lower and presented *MPE* between 6.01% to 9.67%, characterizing that the data estimated by this model have an accuracy above 90%.

Comparing the results presented for the CG region, it is observed that the MLR, MLP and RBF models also demonstrate high indexes of (*r*) with the real data. However, the measurements of *RMSE*, *MAE* and *MPE* indicate that the MLP-GA model showed less variations in its estimates, characterizing greater precision.

For CO region, the MLP-GA model also presented the highest indexes of (*r*), having similarly demonstrated high values of (*D*) and (*C*), which characterizes the performance of this model in optimum. Comparing the *RMSE* obtained by the MLP-GA model with the measurements obtained by the other models, it appears that the *RMSE* obtained by the MLP-GA model is approximately 13% less than the *RMSE* obtained by the MLP, 30% less than that obtained by the RBF and 23% less than the error presented by the MLR.

For IT region, it is also proved that the Evaporation estimated by MLP-GA reached the highest values of (*r*), (*D*) and (*C*), which demonstrates a high relationship with the actual data. Another important point is that the *RMSE*, *MAE* and *MPE* error measurements generated with the MLP-GA model were lower than the errors shown by the

Table 4 - The Real data analysis and results of evaporation estimates for different regions of the state of Rio de Janeiro, Brazil. Indices: mean (M), maximum (MAX), minimum (MIN), standard deviation (SD), correlation coefficient (*r*), root of the mean square error (*RMSE*), mean absolute error (*MAE*), mean percentage error (*MPE*), agreement index (*D*), confidence index (*C*). Models: multiple linear regression (MLR), Multilayer Perceptron (MLP), Radial Basis Function (RBF), hybrid model genetic algorithm + Multilayer Perceptron (MLP-GA).

Actual data	M	MAX	MIN	SD
Evaporation (CG)	104.78	159.20	65.30	26.36
Evaporation (CO)	47.78	78.10	32.10	13.11
Evaporation (IT)	117.12	188.40	60.60	36.01
Evaporation (PA)	81.49	138.90	46.90	22.78
Evaporation(RE)	132.00	212.30	67.60	36.72
Evaporation (RJ)	99.22	142.10	70.00	18.82

Estimated data (CG)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Evaporation	MLR	98.83	144.57	68.31	20.19	0.82	16.10	12.40	11.90	0.87	0.71
	MLP	104.24	143.49	69.91	25.81	0.85	14.20	10.00	9.80	0.92	0.78
	RBF	99.74	135.77	69.23	20.79	0.79	16.54	12.69	12.10	0.86	0.68
	MLP-GA	106.78	146.62	74.53	21.62	0.88	12.45	9.39	9.67	0.93	0.82
Estimated data (CO)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Evaporation	MLR	45.66	67.17	29.16	10.82	0.94	5.03	4.10	8.32	0.95	0.90
	MLP	45.86	65.44	28.77	11.62	0.95	4.42	3.21	6.50	0.97	0.92
	RBF	45.69	65.38	30.66	10.25	0.93	5.58	4.04	7.93	0.94	0.87
	MLP-GA	47.97	72.25	32.22	11.77	0.96	3.86	3.07	6.65	0.97	0.93
Estimated data (IT)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Evaporation	MLR	117.21	176.85	81.78	27.53	0.95	13.11	11.10	13.11	0.95	0.90
	MLP	117.12	173.52	74.33	31.78	0.95	11.38	9.24	9.42	0.97	0.92
	RBF	115.58	181.53	75.43	32.02	0.94	12.00	10.35	10.01	0.97	0.91
	MLP-GA	114.71	194.42	65.71	33.70	0.97	8.73	6.87	6.01	0.98	0.96
Estimated data (PA)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Evaporation	MLR	91.08	134.97	60.86	18.44	0.96	11.83	10.76	15.18	0.92	0.88
	MLP	86.56	147.88	57.10	24.33	0.94	9.38	7.28	9.17	0.96	0.90
	RBF	87.33	134.29	55.10	20.02	0.91	10.78	9.12	11.90	0.93	0.85
	MLP-GA	82.51	127.53	48.52	21.17	0.95	6.97	5.34	6.54	0.97	0.93
Estimated data (RE)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Evaporation	MLR	107.24	180.07	64.62	29.05	0.92	28.73	25.32	18.44	0.82	0.76
	MLP	131.43	198.88	88.76	36.93	0.85	19.67	14.46	11.13	0.92	0.78
	RBF	114.00	196.66	81.77	31.18	0.85	25.98	21.39	15.54	0.85	0.72
	MLP-GA	126.49	192.94	81.77	30.53	0.88	17.86	12.22	9.25	0.92	0.82
Estimated data (RJ)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Evaporation	MLR	114.10	145.38	93.24	13.66	0.71	19.71	17.83	19.12	0.65	0.46
	MLP	100.13	139.79	68.89	15.61	0.84	9.96	7.77	7.89	0.91	0.76
	RBF	98.77	135.07	76.17	14.68	0.65	14.24	8.49	8.05	0.77	0.50
	MLP-GA	105.33	144.77	71.08	21.11	0.82	13.43	8.84	9.11	0.87	0.72

other models, remaining at 8.73 mm for *RMSE*, 6.87 mm for *MAE* and 6.01% for *MPE* (Table 4).

In PA region, the MLP-GA model also demonstrated a high aptitude in estimating Evaporation, indicating that the estimated data are highly associated with the actual values. Comparing the *MAE* values obtained by MLP-GA with those obtained by the other models, it is possible to observe that the *MAE* presented by MLP-GA is 26% lower than the *MAE* obtained by MLP, 41% less than that achieved by RBF and 50% lower than that obtained by the MLR.

In the RE region estimates, the MLP-GA model presented an *MPE* of 9.25%, and demonstrated high rates of (*r*), (*D*) and (*C*) with the real data, characterizing that the estimated data have more than 90% accuracy.

In RJ region, the model that proved to be superior was the MLP, obtaining the highest coefficient (*r*) between the estimated and registered data, the lowest *MAE* (7.77 mm), the lowest *RMSE* (9.96), and *MPE* equal to 7.89%, which characterizes that each data estimated by the MLP model had a hit rate of 92.11%.

After comparing the results, it was decided to use the MLP-GA to fill the gaps in the regions of CG, CO, IT, PA and RE, and the MLP model to fill the gaps in RJ. Fig. 7 presents the real data, the estimates presented by the best models, the dispersion of the values and the data filled in.

3.2. Estimation results and filling in maximum temperature data

The results indicatethatall models had low error rates and high (*r*) values in the maximum temperature estimate (Table 5). However, despite the low variation between the errors obtained, it is possible to verify that the MLP-GA model proved to be superior, presenting more accurate estimates than the other models.

For the CG region, it is noted that the values of *MAE* and *MPE* obtain a low variation for the dissipated models, keeping between 0.32 and 0.37 for *MAE* and 1.09 and 1.25 for *MPE*, which indicates that the estimated values may

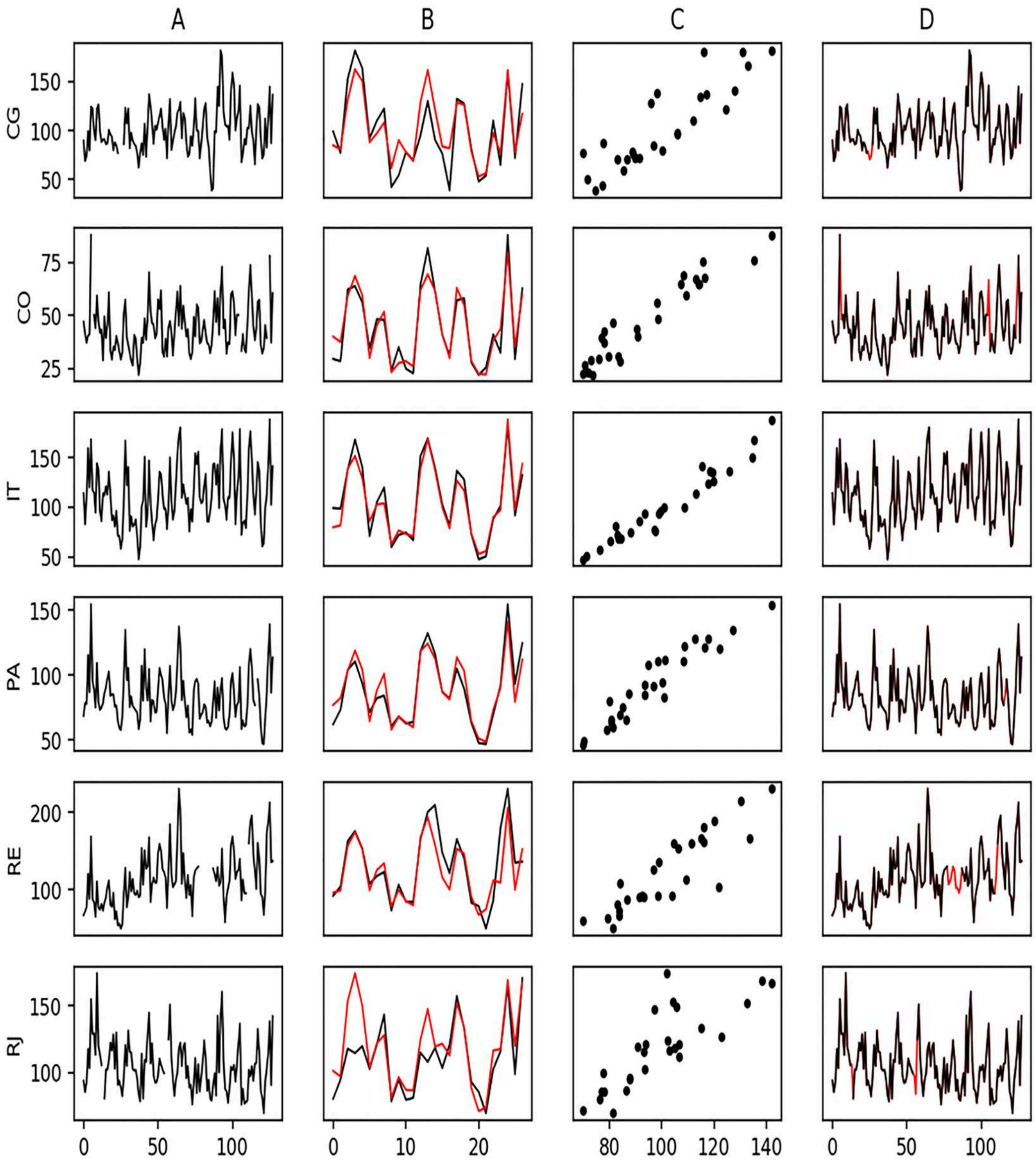

Figure 7 - Actual evaporation data (A), evaporation estimate results by models (B), real x estimated data dispersion (*C*) and filled in by models (*D*) for different regions of the state of Rio de Janeiro, Brazil. Regions: Campos dos Goytacazes (CG), Cordeiro (CO), Itaperuna (IT), Paty do Alferes (PA), Resende (RE), Rio de Janeiro (RJ).

Table 5 - Analysis of real data and results of maximum temperature estimates for different regions of the state of Rio de Janeiro, Brazil. Indices: mean (M), maximum (MAX), minimum (MIN), standard deviation (SD), correlation coefficient (*r*), root of the mean square error (*RMSE*), mean absolute error (*MAE*), mean percentage error (*MPE*), agreement index (*D*), confidence index (*C*). Models: multiple linear regression (MLR), Multilayer Perceptron (MLP), Radial Basis Function (RBF), hybrid model genetic algorithm + Multilayer Perceptron (MLP-GA).

Actual data	M	MAX	MIN	SD
Maximum temperature (CG)	29.86	34.32	26.96	2.11
Maximum temperature (CO)	27.05	31.59	23.21	2.07
Maximum temperature (IT)	30.08	35.21	26.45	2.16
Maximum temperature (PA)	27.76	32.96	24.00	2.22
Maximum temperature (RE)	28.15	32.70	23.80	2.34
Maximum temperature (RJ)	30.26	35.31	26.96	2.48

Estimated data (CG)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Maximum temperature	MLR	29.65	34.18	26.34	2.08	0.98	0.48	0.37	1.24	0.99	0.97
	MLP	29.73	34.08	26.71	1.99	0.98	0.46	0.36	1.19	0.99	0.97
	RBF	29.66	34.55	26.35	2.18	0.98	0.46	0.37	1.25	0.99	0.97
	MLP-GA	29.58	34.06	26.34	2.16	0.99	0.44	0.32	1.09	0.99	0.98
Estimated data (CO)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Maximum temperature	MLR	27.19	31.89	23.95	2.04	0.98	0.43	0.38	1.43	0.99	0.97
	MLP	27.17	32.02	23.54	2.08	0.98	0.41	0.34	1.25	0.99	0.97
	RBF	27.23	31.79	23.89	2.10	0.97	0.52	0.41	1.53	0.98	0.96
	MLP-GA	27.29	31.58	23.82	2.03	0.99	0.38	0.31	1.15	0.99	0.98
Estimated data (IT)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Maximum temperature	MLR	29.98	34.59	26.13	2.10	0.97	0.49	0.41	1.35	0.99	0.96
	MLP	30.19	35.22	26.24	2.21	0.98	0.47	0.36	1.21	0.99	0.97
	RBF	30.07	35.38	26.61	2.11	0.97	0.49	0.40	1.32	0.98	0.95
	MLP-GA	30.06	35.26	26.29	2.25	0.98	0.44	0.32	1.07	0.99	0.97
Estimated data (PA)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Maximum temperature	MLR	27.45	32.03	23.67	2.10	0.99	0.47	0.36	1.29	0.99	0.98
	MLP	27.66	32.72	24.17	2.09	0.99	0.38	0.32	1.17	0.99	0.98
	RBF	27.70	31.74	24.27	2.10	0.98	0.43	0.31	1.12	0.99	0.97
	MLP-GA	27.64	32.06	23.73	2.15	0.99	0.38	0.27	0.95	0.99	0.98
Estimated data (RE)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Maximum temperature	MLR	27.30	32.49	23.37	2.33	0.95	1.10	0.89	3.16	0.94	0.90
	MLP	27.65	32.26	23.16	2.37	0.97	0.72	0.56	2.00	0.98	0.95
	RBF	27.44	32.61	23.21	2.37	0.96	0.95	0.74	2.65	0.96	0.92
	MLP-GA	28.19	33.48	24.37	2.25	0.97	0.57	0.43	1.55	0.98	0.95
Estimated data (RJ)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Maximum temperature	MLR	30.76	36.09	26.54	2.39	0.91	1.14	0.90	3.05	0.94	0.86
	MLP	30.54	35.49	27.65	2.39	0.95	0.79	0.59	2.01	0.97	0.92
	RBF	30.55	35.32	28.03	2.46	0.94	0.89	0.69	2.33	0.96	0.91
	MLP-GA	30.20	35.66	27.62	2.43	0.97	0.58	0.47	1.57	0.99	0.96

have an error between 0.32 °C and 0.37 °C and that the estimated data are 99% accurate.

In the CO region it is also possible to prove that the maximum temperature estimated by the models is highly accurate. However, comparing the results obtained by the MLP-GA with the other models, it appears that the *RMSE* presented by the MLP-GA model is 13% less than that achieved by the MLR model, 8% less than that obtained by MLPe 37% less than the obtained by the RBF model, which demonstrates less variation in the estimated data and greater precision of the values obtained by the MLP-GA.

For IT region, it is observed that the *MPE* indices presented in the maximum temperature estimate ranged from 1.07% to 1.35%, which indicates that the models obtained an average precision above 98%. In addition to this fact, it is noticed that the MLP-GA model presented the highest indexes of (*r*), (*D*) and (*C*), which indicates that the model was more accurate.

For the PA region, the results exposed by the model in the maximum temperature estimate indicate that the (*r*) values are highly correlated with the actual data, and that the quality of the estimates is classified as optimal. Analyzing the *RMSE*, *MAE* and *MPE* values obtained by MLP-GA, it is observed that the *RMSE* was 0.38 °C, the *MAE* was 0.27 °C and the *MPE* was 0.95%, which means that the estimated values have more than 99% of accuracy with actual values.

In RE region, the results obtained by MLP-GA expose a great difference between the models. From the *MAE* values, it can be seen that the value obtained by the MLP-GA model is more than 100% less than that obtained by MLR, 30% less than that achieved by MLP and 72% less than the *MAE* achieved by RBF.

In RJ region, as well as in RE, it is possible to notice a greater variation between the estimates, where it appears that the MLP-GA model presented more accurate results than the other models. The values of *MAE* and *MPE* reached by MLP-GA indicate that the average error of this model is 0.47 °C in each estimate, and that the estimated data have an accuracy greater than 98%.

After comparing the results, it was decided to use the MLP-GA to fill the gaps in the maximum temperature variable in all six regions. Due to the lack of data from the same period in some of the regions applied in the estimate, it was not possible to fill 100% of the actual failures that occurred, being possible to fill 75% in IT, 71% in PA and 89% in RE, real data, those estimated by the best model, the dispersion of values and the data filled in data are represented in Fig. 8.

3.3. Estimation results and filling in relative humidity data

The results of the relative humidity estimate highlight that the MLP-GA model presented greater precision in its estimates when compared with the other models. Analyzing the results shown in Table 6, it can be seen that the (*r*), *MAE* and EMP indexes achieved by the MLP-GA model ranged from 0.89 to 0.97, 0.90 to 1.49 and from 1.15% to 1.90%, highlighting that the accuracy of this model was greater than 98%.

In the CG region, it is observed through the (*C*) indexes achieved by the models that only the MLP and MLP-GA models achieved a performance classified as very good. However, despite presenting a small difference, it is still possible to state that the MLP-GA model presents more accurate results than the MLP.

In CO region, the indexes of (*r*) and (*C*) indicate that the values estimated by MLP-GA demonstrate a high relationship with the actual relative humidity data. It is also confirmed that the value of *RMSE*, *MAE* and *MPE* were 1.31, 1.03 and 1.28%, being below the values of *RMSE*, *MAE* and *MPE* obtained by the other models (Table 6).

For the IT region, the *RMSE* index achieved with the MLP-GA model was 1.73, which is 41% less than the RBF, 26% less than the MLR and 19% less than the error obtained by the MLP model. It is also noticed that the indexes achieved by the MLR and MLP models were considerably close, reaching the same *MAE*, (*D*) and (*C*) values.

For PA region, the *MPE* exposed by the MLP-GA model in the relative humidity estimate indicates about 98.85% accuracy with the actual data. Analyzing the *MAE* and (*C*) values obtained by MLP-GA, it is noted that its error was only 0.90 and that the confidence of the estimated values was classified as excellent.

In the RE region, the MLP-GA model obtained, as in all the analyzed places, high indexes of (*r*), (*D*) and (*C*), indicating that the measured data have a high association with the real values and that the information has a high level of agreement and trust.

In the RJ region, the *MAE* and *MPE* values achieved by the MLP-GA indicate that the average error of this model is 1.36 in each estimate, and that the estimated data have an accuracy greater than 98%. Comparing the *MAE* values obtained by all models, we can verify that MLP-GA is 56% more accurate than MLR, 18% more than MLP and 49% more than RBF.

After comparing the results, it was decided to use the MLP-GA to fill in the gaps in the relative humidity variable of all six regions. Due to the lack of data from the same period in some of the regions applied in the estimate, it was not possible to fill 100% of the actual failures that occurred, making it possible to fill 96% in CO, 54.55% in PA, 63% in RE and 72% in RJ. Fig. 9 shows the real data, those estimated by the best model, the dispersion of the values and the completed data.

4. Discussion

Comparing the methodologies adopted for gaps in the meteorological data reconstruction used in this study, it was observed that the hybrid approach MLP-AG with characteristics of autonomous training demonstrated greater precision through the statistical measures used. It was also possible to verify that despite the model performing a large number of possible combinations in search of a solution, it presented a low execution time (Table 7).

Even considering that to achieve such results the MLP-AG model had a larger set of information, it's possible to notice that in several places it used the same number of variables or a lower number, such as to estimate eva-

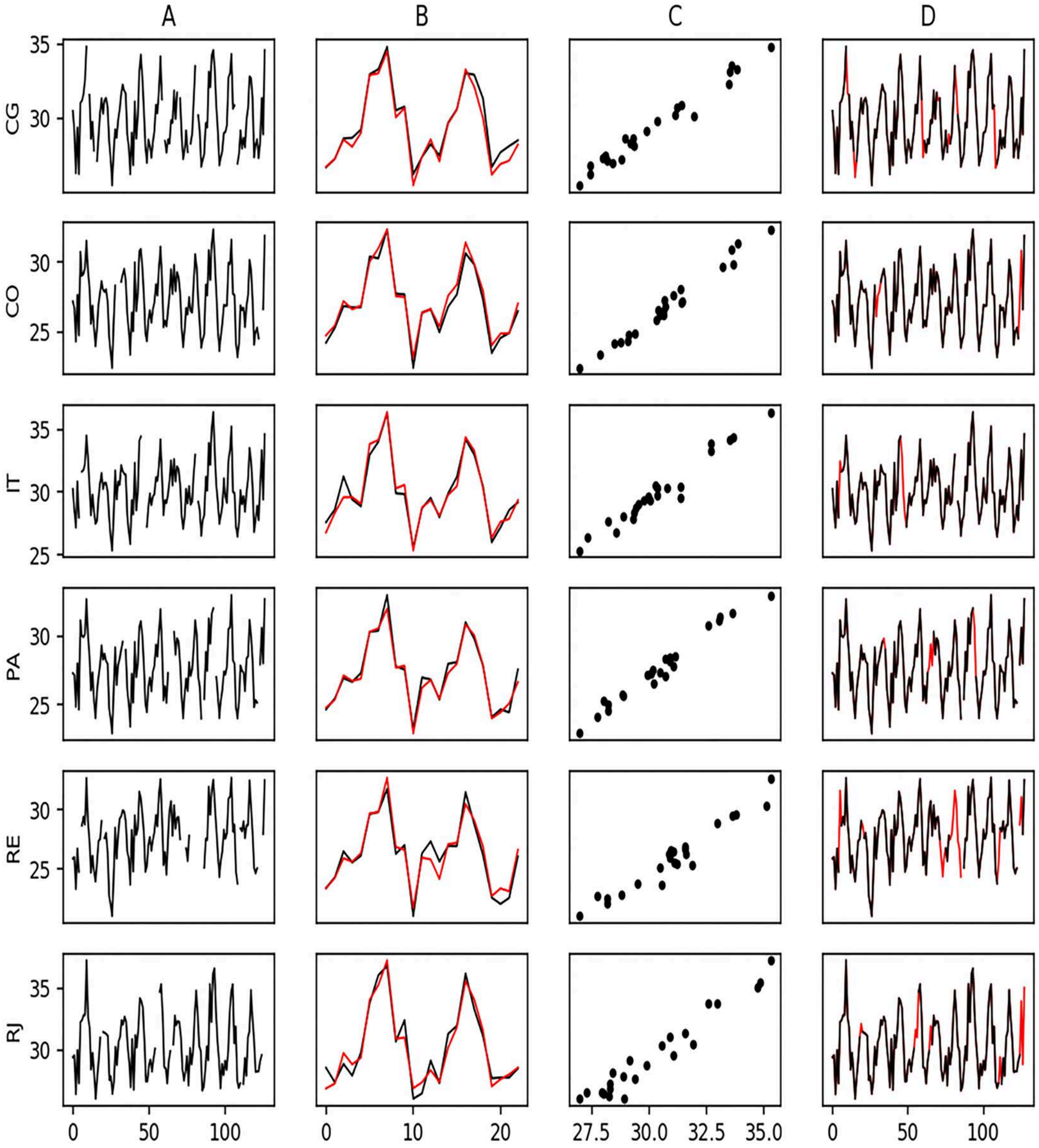

Figure 8 - Actual maximum temperature data (A), results of the maximum temperature estimate by the MLP-GA model (B), dispersion x estimated real data (*C*) and data filled by the model (*D*) for different regions of the state of Rio de January, Brazil. Regions: Campos dos Goytacazes (CG), Cordeiro (CO), Itaperuna (IT), Paty do Alferes (PA), Resende (RE), Rio de Janeiro (RJ).

poration in RJ, the maximum temperature in CG, RJ and RE and relative humidity in CO. In addition, other factors where differences between manual and autonomous training approaches can be noted are related to the training settings adopted, such as the transfer functions of the hidden layers and the types of training of each approach.

This fact justifies the idea that the choice of topology and characteristics of an ANN for a given problem can be seen as a problem of combinatorial optimization, defined in the space of possible architectures (Rocha *et al.*, 2008). Thus, a possible explanation for the best resource fulness of the MLP-GA model is that the greater number of tests and adjustments performed by the same model helped him in trying to achieve an optimal combination, while the manual training approach may have restricted the reach of the model. Therefore, we can infer that the MLP-GA model used the evolutionary process to adapt to the studied locations, which allowed for low error rates with small variations between regions (Table 8).

Within this context, factors such as the quality and quantity of data and various other information directly influence the performance of the models, causing them to have low or high error rates. Even so, it is possible to verify that the indexes achieved by MLP-GA are largely similar to those obtained by several methodologies found in the literature (Table 9).

Regarding the (*r*), *RMSE*, *MAE* and *MPE* indices obtained by MLP-GA for the evaporation estimate, it

Table 6 - The real data analysis and results of relative humidity estimates for different regions of the state of Rio de Janeiro, Brazil. Indices: mean (M), maximum (MAX), minimum (MIN), standard deviation (SD), correlation coefficient (*r*), root of the mean square error (*RMSE*), mean absolute error (*MAE*), mean percentage error (*MPE*), agreement index (*D*), confidence index (*C*). Models: multiple linear regression (MLR), Multilayer Perceptron (MLP), Radial Basis Function (RBF), hybrid model genetic algorithm + Multilayer Perceptron (MLP-GA).

Actual data	M	MAX	MIN	SD
Relative humidity (CG)	76.02	80.70	69.37	2.84
Relative humidity (CO)	79.66	86.73	71.12	4.98
Relative humidity (IT)	72.10	82.40	61.88	6.14
Relative humidity (PA)	79.33	83.62	74.12	3.54
Relative humidity (RE)	77.54	84.08	68.16	4.38
Relative humidity (RJ)	72.62	78.98	62.75	4.01

Estimated data (CG)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Relative humidity	MLR	76.29	81.85	69.58	3.28	0.80	1.94	1.46	1.92	0.88	0.71
	MLP	76.44	80.68	69.32	3.11	0.87	1.56	1.07	1.41	0.92	0.80
	RBF	76.45	82.09	70.24	3.09	0.81	1.86	1.40	1.84	0.88	0.71
	MLP-GA	76.29	82.90	69.88	3.07	0.89	1.39	1.03	1.35	0.94	0.84
Estimated data (CO)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Relative humidity	MLR	78.74	87.31	69.36	5.06	0.95	1.75	1.42	1.78	0.97	0.92
	MLP	78.95	87.47	70.53	4.86	0.94	1.74	1.24	1.54	0.97	0.91
	RBF	79.31	88.24	70.19	5.14	0.93	1.85	1.55	1.94	0.96	0.90
	MLP-GA	79.23	86.15	70.86	4.98	0.97	1.31	1.03	1.28	0.98	0.95
Estimated data (IT)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Relative humidity	MLR	73.21	81.01	63.56	5.06	0.96	2.19	0.96	2.61	0.96	0.92
	MLP	72.73	78.86	63.97	5.20	0.95	2.06	0.96	2.29	0.96	0.92
	RBF	73.01	78.21	64.64	4.63	0.94	2.45	0.94	2.91	0.94	0.89
	MLP-GA	72.45	80.97	61.85	5.56	0.96	1.73	0.98	1.89	0.98	0.94
Estimated data (PA)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Relative humidity	MLR	79.74	88.37	68.43	5.65	0.93	2.67	2.20	2.80	0.91	0.84
	MLP	79.64	85.54	69.69	4.80	0.91	2.12	1.67	2.12	0.93	0.84
	RBF	78.99	85.50	71.22	4.43	0.91	1.89	1.54	1.95	0.94	0.86
	MLP-GA	79.43	83.41	72.74	3.63	0.94	1.17	0.90	1.15	0.97	0.92
Estimated data (RE)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Relative humidity	MLR	76.38	85.32	64.98	6.00	0.85	3.33	2.50	3.24	0.88	0.75
	MLP	76.12	80.02	66.30	4.17	0.90	2.34	1.82	2.30	0.92	0.83
	RBF	76.61	82.99	67.59	4.79	0.89	2.31	1.79	2.30	0.93	0.83
	MLP-GA	76.57	80.32	68.56	3.86	0.90	2.09	1.49	1.87	0.93	0.84
Estimated Data (RJ)	Models	M	MAX	MIN	SD	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Relative humidity	MLR	72.13	74.11	69.61	1.45	0.83	2.88	2.13	2.98	0.69	0.57
	MLP	72.39	76.33	66.94	2.98	0.84	2.14	1.61	2.28	0.89	0.75
	RBF	72.20	74.40	68.99	1.66	0.81	2.77	2.03	2.86	0.73	0.59
	MLP-GA	72.80	77.14	66.48	3.16	0.90	1.79	1.36	1.90	0.93	0.83

appears that these ranged from 0.82 to 0.97, 3.86 to 17.86, 3.07mm to 12.22mm and 6.01% to 9.67% indicating that the model presented the average precision between the regions to which it was applied from 90.23% to 93.99%.

To estimate the maximum temperature, the (*r*), *RMSE*, *MAE* and *MPE* indices obtained by MLP-GA varied from 0.97 to 0.99, 0.38 to 0.58, 0.27 °C to 0.47 °C and 0.95% to 1.57%. These values indicate that the model presented an average precision between the regions, to which it was applied from 98.5% to 99%. Values with high precision indexes were also found by Coulibaly and Evora (2007) in the watershed of Gatineau, in northeastern

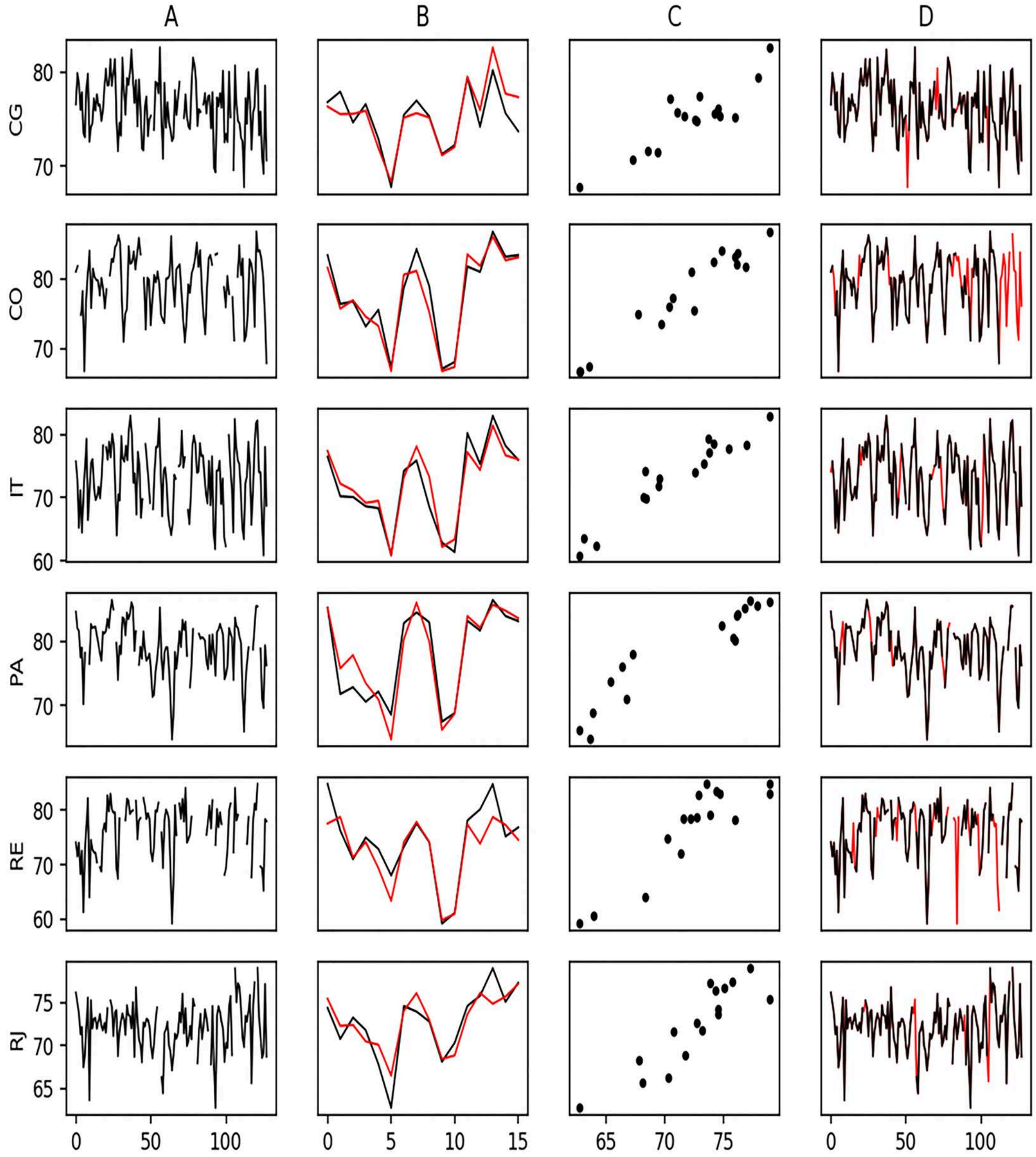

Figure 9 - Actual relative humidity data (A), results of the relative humidity estimate by the MLP-GA model (B), dispersion of real x estimated data (*C*) and filled in by the model (*D*) for different regions of the state of Rio de January, Brazil. Regions: Campos dos Goytacazes (CG), Cordeiro (CO), Itaperuna (IT), Paty do Alferes (PA), Resende (RE), Rio de Janeiro (RJ).

Canada, where they compared six types of artificial neural networks, among them MLP and RBF to estimate the maximum temperature. And by Kotsiantis *et al.* (2008) who obtained (*r*) indexes of 0.91 and *RMSE* of 3.07 through the application of different regression models.

In order to estimate the relative humidity, the MLP-GA model reached indexes of (*r*) between 0.89 and 0.97, *RMSE* between 1.17 and 2.09, *MAE* between 0.90 and 1.49 and *MPE* between 1.15% and 1.90%. Low error rates were also found by Altan and Ustundag (2012), Coutinho *et al.* (2018) and Anjomshoaa and Salmanzadeh (2018), obtained by applying the techniques of Regression, Wavelet Transform, RNA MLP and linear interpolation and cubic splines. Table 8 shows the variation of the results found in the literature.

Although the results found in the literature demonstrate high rates of accuracy in the estimates of the variables used in each study, it is still possible to observe that the MLP-GA model demonstrates more accurate precision. Among other factors that we can highlight, most of the approaches found in the literature are strictly supervised and depend directly on manual choices or statistical methods to define several parameters of each model, among them the predictors of each variable. These approaches differ from the MLP-GA that determines its characteristics autonomously, which can become extre-

Table 7 - Runtime and generation of the individual that converged to the expected solution.

	Evaporation		Maximum temperature		Relative humidity	
Stations	Generation	Time in seconds	Generation	Time in seconds	Generation	Time in seconds
CG	3	160.83	2	73.44	40	832.23
CO	14	281.26	3	186.71	14	419.80
IT	8	248.36	5	253.45	40	716.65
PA	13	353.78	2	134.72	5	200.36
RE	8	280.14	20	684.84	40	935.77
RJ	40	529.62	29	387.58	40	614.62

Table 8 - The results variation presented by the MLP-GA model in the estimation of meteorological data.

Dados	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	(*C*)
Evaporation	0.82 to 0.97	3.86 to 17.86	3.07 to 12.22	6.01 to 9.67	0.87 to 0.98	0.72 to 0.96
Maximum temperature	0.97 to 0.99	0.38 to 0.58	0.27 to 0.47	0.95 to 1.57	0.98 to 0.99	0.95 to 0.98
Relative humidity	0.89 to 0.97	1.17 to2.09	0.90 to 1.49	1.15 to 1.90	0.93 to 0.98	0.83 to 0.95

Table 9 - Results found in the literature on the estimation of meteorological data.

Dados	(*r*)	*RMSE*	*MAE*	*MPE* (%)	(*D*)	Referências
Mean temperature	0.93	2.46	X	X	X	Kotsiantis *et al.* (2006)
Maximum daily temperature	0.99	X	0.74	X	X	Coulibaly and Evora (2007)
Minimum daily temperature	0.94	2.18	X	X	X	Kotsiantis et al (2008)
Maximum daily temperature	0.91	3.07	X	X	X	
Maximum temperature	X	X	0.64	X	X	Altan and Ustundag (2012)
Relative Humidity	X	X	3.33	X	X	
Minimum daily temperature	X	0.70	X	X	X	Woldesenbet *et al.* (2016)
Maximum daily temperature	X	0.90	X	X	X	
Minimum daily temperature	0.90	X	X	X	X	Thevakaran and Sonnadara (2017)
Maximum daily temperature	0.90	X	X	X	X	
Maximum temperature	0.97	0.41	0.32	1.05	0.99	Coutinho *et al.* (2018)
Relative Humidity	0.94	1.95	1.47	1.85	0.96	
Mean temperature	X	0.79	0.59	X	X	Anjomshoaa and Salmanzadeh (2018)
Relative Humidity	X	X	2.25	X	X	
Maximum temperature	X	X	5.26	X	0.82	Beguería *et al.* (2019)
Minimum daily temperature	0.93	1.10	0.80	X	X	Shabalala *et al.* (2019)
Maximum daily temperature	0.92	1.20	0.55	X	X	

mely advantageous with the increase in the number of locations and variables to be filled, where the MLP-GA can streamline the process by finding acceptable combinations to estimate the values of each region.

5. Conclusion

The results analysis achieved by the models in the estimation and filling of the meteorological data used allowed to conclude that the model that best adapted to the studied locations was o MLP-GA. having obtained in the majority of the regions the lowest values of *RMSE*, *MAE* and *MPE*, and also the highest correlation indexes (*r*) and reliability (*C*) with the real data.

However, it also appears that the MLR, MLP and RBF models also showed satisfactory results in most of the locations, demonstrating high correlation indexes (*r*) and reliability with the real data. However, the results exposed by the ANN MLP highlight it as the second best alternative in estimating the studied variables.

Another important fact observed was that the use of GA maximized the adaptability of ANN MLP, which made the model present more accurate results. However, the failures occurrence in the predictor variables determined by the model impaired part of the series filling of maximum temperature and relative humidity, which were not performed in 100%.

In fact, more research is needed to explore the real gains from the MLP-GA model in estimating weather data. However, the results allow us to confirm that the MLP-GA model is a viable alternative to estimate and fill gaps in the meteorological data used.

Acknowledgements

The Coordination for the Improvement of Higher Education Personnel (CAPES) for providing the research grant. to the National Institute of Meteorology (INMET) for making meteorological data available to the Federal Center for Technological Education Celso Suckow da Fonseca (CEFET-ANGRA). The Federal Rural University of Rio de Janeiro (UFRRJ-SEROPÉDICA) and by the Postgraduate Program in Civil Engineering at the Federal University of Rio de Janeiro - UFRJ - COPPE.

References

AIEB, A.; MADANI, K.; SCARPA, M.; BONACORSO, B.; LEFSIH, K. A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed. Algeria. **Heliyon**, v. 5, n. 2, p. 1247-1274, 2019. doi

ALTAN, T.N.; USTUNDAG, B.B. Reconstruction of Missing Meteorological Data Using Wavelet Transform. **IEEE - First International Conference on Agro- Geoinformatics (Agro-Geoinformatics)**, p. 1-7, 2012. doi

ANJOMSHOAA, A.; SALMANZADEH, M. Filling missing meteorological data in heating and cooling seasons separately. **International Journal of Climatology**, v. 39, n. 2, p. 701-710, 2018. doi

ANOCHI, J.A.; CAMPOS VELHO, H.F. Meteorological data mining for climate precipitation prediction using neural networks. **Journal of Computational Interdisciplinary Sciences**, v. 2, n. 2, p. 71-78, 2015. doi

ASCHAUER, J.; MARTY, C. Evaluating methods for reconstructing large gaps in historic snow depth time series. **Geoscientific Instrumentation Methods and Data Systems**, v. 10, n. 2, p. 297-312, 2021. doi

AUFFARTH, B. **Machine Learning for Time-Series with Python, Forecast, Predict, and Detect Anomalies with State-of-the-Art Machine Learning Methods**. United Kingdom: Packt Publishing Ltd, 2021.

BASTOS, J.; NAPOLEÃO, P. **O Estado do Ambiente: Indicadores Ambientais do Rio de Janeiro de 2010**. INEA, 2011. Disponível em http://200.20.53.3:8081/cs/groups/public/documents/document/zwew/mde1/~edisp/inea0015448.pdf, acesso em 15/12/2019.

BEGUERÍA, S.; TOMAS-BURGUERA, M.; SERRANO-NOTIVOLI, R.; PEÑA-ÂNGULO, D.; VICENTE-SERRANO, S.M.; GONZÁLEZ-HIDALGO, J.C. Gap Filling of monthly temperature data and its effect on climatic variability and trends. **Journal Of Climate**, v. 32, n. 22, p. 7797-7821, 2019. doi

BRAGA, A.P.; CARVALHO, A.C.P.L.F.; LUDEMIR, T.B. **Redes Neurais Artificiais - Teoria e Aplicações**. Ed. 2. Rio de Janeiro: LTC - Livros Técnicos e Científicos Ltda, 2012.

BRITO, T.T.; OLIVEIRA-JÚNIOR, J.F.; LYRA, G.B.; GOIS. G.; ZERI, M. Multivariateanalysisappliedtomonthlyrainfall over Rio de Janeiro state.Brazil. **Meteorology and Atmospheric Physics**, v. 129, p. 469-478, 2017. doi

BRUBACHER, J.P.; OLIVEIRA, G.G.; GUASSELLI, L.A. Preenchimento de falhas em séries temporais de precipitação diária no Rio Grande do Sul. **Revista Brasileira de Meteorologia**, v. 35, n. 2, p. 335-344, 2020. doi

BRUCE, P.; BRUCE, A. **Estatística Prática para Cientistas de Dados**. Ed. 1. Rio de Janeiro: Alta Books, 2019.

CANCHALA-NASTAR, T.; CARVAJAL-ESCOBAR, Y.; ALFONSO-MORALES, W.; CERÓN, W.L.; CAICEDO, E. Estimation of missing data of monthly rainfall in southwestern Colombia using artificial neural networks. **Journal Data in Brief**, v. 26, p. 104517, 2019. doi

CLACK, C.T.M. Modeling solar irradiance and solar PV power output to create a resource assessment using linear multiple multivariate regression. **Journal Of Applied Meteorology And Climatology**, v. 56, n. 1, p. 109-125, 2017. doi

COULIBALY, P.; EVORA, N.D. Comparison of neural network methods for infilling missing daily weather records. **Journal of Hydrology**, v. 341, n. 2, p. 27-41, 2007. doi

COUTINHO, E.R.; SILVA, R.M.; DELGADO, A.R.S. Utilização de Técnicas de Inteligência Computacional na Predição de Dados Meteorológicos. **Revista Brasileira de Meteorologia**, v. 31, n. 1, p. 24-36, 2016. doi

COUTINHO, E.R.; SILVA, R.M.; MADEIRA, J.G.F.; COUTINHO, P.R.O.S.; BOLOY, R.A.M.; DELGADO, A.R.S. Application of Artificial Neural Networks (ANNs) in the gap filling of meteorological time series. **Revista Brasileira de Meteorologia**, v. 33, n. 2, p. 317-328, 2018. doi

DEMBÉLÉ, M.; ORIANI, F.; TUMBULTO, J.; MARIÉTHOZ, G.; SCHAEFLI, B. Gap-filling of daily streamflow time series using Direct Sampling in various hydroclimatic settings. **Journal of Hydrology**, v. 569, p. 573-586, 2019. doi

DIAS, A.S.; SOARES, W.A. Uso de metodologias de preenchimento de falhas para estimativas de dados de precipitação. **Research Society and Development**, v. 10, n. 5, p. e57610515383-e57610515383, 2021. doi

ECCEL, E.; CAU, P.; RANZI, R. Data reconstruction and homogenization for reducing uncertainties in high-resolution climate analysis in Alpine regions. **Theoretical and Applied Climatology**, v. 110, p. 345-358, 2012. doi

FINE, L.; RICHARD, A.; TANNY, J.; PRADALIER, C.; ROSA, R.; ROZENSTEIN, O. Introducing state-of-the-art deep learning technique for gap-filling of eddy covariance crop evapotranspiration data. **Water**, v. 14, n. 5, p. 763, 2022. doi

FONSECA, J.S.; MARTINS, G.A.; TOLEDO, G.L. **Estatística Aplicada**. Ed. 2, São Paulo: Atlas, 2012.

FORD, T.W.; QUIRING, S.M. Comparison and application of multiple methods for temporal interpolation of daily soil moisture. **International Journal of Climatology**, v. 34, n. 8, p. 2604-2621, 2014. doi

GHAREEB, W.T.; SAADANY, E.F.EL. 2013. A hybrid genetic radial basis function network with fuzzy corrector for short term load forecasting. **IEEE Electrical Power & Energy Conferenc (EPEC)**, p. 1-5, 2013. doi

GIOVANELLA, T.H.; OLIVEIRA, F.C.; MARCHI, V.A.A.; TLUSZCZ, J. Desempenho de métodos de preenchimento de falhas em dados de evapotranspiração de referência para região Oeste Paraná. **Revista Brasileira de Meteorologia**, v. 36, n. 3, p. 415-422, 2021. doi

GUNAWARDENA, N.; DURAND, P.; HEDDE, T.; DUPUY, F.; PARDYJAK, E. Data filling of micrometeorological variables in complex terrain for high-resolution nowcasting. **Atmosphere**, v. 13, n. 3, p. 408, 2022. doi

HAIDAR, A.; VERMA, B. A Genetic algorithm based feature selection approach for rainfall forecasting in sugarcane areas. **IEEE Symposium Series on Computational Intelligence (SSCI)**, p. 1-8, 2016. doi

HAYKIN, S. **Redes Neurais Princípios e Pratica**. Ed. 2, Porto Alegre: Bookmam, 2001.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **População Estimada do Rio de Janeiro 2019**. Disponível em https://cidades.ibge.gov.br/brasil/rj/panorama, acesso em janeiro de 2020.

KIM, J.W.; PACHEPSKY, Y.A. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. **Journal of Hydrology**, v. 394, n. 3-4, p. 305-314, 2010. doi

KOTSIANTIS, S.; KOSTOULAS, A.; LYKOUDIS, S.; ARGIRIOU, A.; MENAGIAS, K. Filling missing temperature values in weather data banks. **2nd IEEE International Conference on Intelligent Environments**, v. 1, p. 327-334, 2006. doi

KOTSIANTIS, S.; KOSTOULAS, A.; LYKOUDIS, S.; ARGIRIOU, A.; MENAGIAS, K. Using data mining techniques for estimating minimum. maximum and average daily temperature values. **International Journal of Mathematical. Physical and Engineering Sciences**, v. 1, n. 1, p. 16-20, 2008.

KORSTANJE, J. **Advanced Forecasting With Python, With State-of-the-Art-Models Including LSTMs, Facebook's Prophet, and Amazon's DeepAR**, United States: Apress, 2021.

LIU, S.; SU, H.; TIAN, J.; ZHANG, R.; WANG, W.; WU, Y. Evaluating four remote sensing methods for estimating surface air temperature on a regional scale. **Journal of Applied Meteorology and Climatology**, v. 56, n. 3, p. 803-814, 2017. doi

LYRA, G.B.; SOUZA, M.O.; VIOLA, D.N. Modelos lineares aplicados à estimativa da concentração do material particulado (PM_{10}) na cidade do Rio de Janeiro, RJ. **Revista Brasileirade Meteorologia**, v. 26, n. 3, p. 392-400, 2011. doi

MILIDONIS, K.; BLANCO, M.J.; GRIGORIEV, V.; PANAGIOTOU. C.F.; BONANOS, A.M.; CONSTANTINOU, M.; PYE, J.; ASSELINEAU, C.A. Review of application of AI techniques to solar tower systems. **Solar Energy**, v. 224, p. 500-515, 2021. doi

PAPPAS, C.; PAPALEXIOU, S.M.; KOUTSOYIANNIS, D. A quick gap filling of missing hydrometeorological data. **Journal of Geophysical Research: Atmospheres**, v. 119, n. 15, p. 9290-9300, 2014. doi

PEZZOPANE, J.E.M.; CASTRO, F.S.; PEZZOPANE, J.R.M.; CECÍLIO, R.A. **Agrometeorologia Aplicações para o Espírito Santo**. Alegre: CAUFES, 2012.

REN, H.; CROMWELL, E.; KRAVITZ, B.; CHEN, X. Using Deep Learning to Fill Spatio-Temporal Data Gaps in Hydrological Monitoring Networks. **Hydrologyand Earth System Sciences**, v. 196, p. 1-20, 2019.

ROCHA, M.; CORTEZ, P.; NEVES, J.M. **Análise Inteligente de Dados Algoritmo e Implementação em Java**. Lisboa: FCA – Editora de Informática, 2008.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial**. Ed. 3. Rio de Janeiro: Elsevier, 2013.

SAMANTA, S.; PAL, D. K.; LOHAR, D.; PAL, B. Interpolation of climate variables and temperature modeling. **Theoretical and Applied Climatology**, v. 107, p. 35-45, 2012. doi

SERRANO, S.M.V.; BEGUERÍA, S.; MORENO, J.I.L.; VERA, M.A.G.; STEPANEK, P. A complete daily precipitation database for Northeast Spain: Reconstruction, quality control and homogeneity. **International Journal of Climatology**, v. 30, n. 8, p. 1146-1163, 2010. doi

SILVA, M.T.; GILL, E.W.; HUANG, W. An improved estimation and gap-filling technique for sea surface wind speeds using NARX neural networks. **Journal of Atmospheric and Oceanic Technology**, v. 35, n. 7, p. 1521-1532, 2018. doi

SHABALALA, Z.P.; MOELETSI, M.E.; TONGWANE, M.I.; MAZIBUKO, S.M. Evaluation of infilling methods for time series of daily temperature data: Case study of Limpopo Province, South Africa. **Climate**, v. 7, n. 7, p. 86, 2019. doi

SHAH, H.; GHAZALI, R. Prediction of earthquake magnitude by an improved ABC-MLP. **IEEE- Developments in E-Systems Engineering**, p. 312-317, 2011. doi

SOUSA, N.M.N.; DANTAS, R.T.; LIMEIRA, R.C. Influência de variáveis meteorológicas sobre a incidência do dengue, Meningite e pneumônia em João Pessoa-PB. **Revista Brasileira de Meteorologia**, v. 22, n. 2, p. 183-192, 2007. https://doi.org/10.1590/S0102-77862007000200004

TARDIVO, G.; BERTI, A. The selection of predictors in a regression-based method for gap filling in daily temperature datasets. **International Journal of Climatology**, v. 34, n. 4, p. 1311-1317, 2014. doi

TEEGAVARAPU, R.S.V.; CHANDRAMOULI, V. Improved weighting methods. deterministic and stochastic data-driven models for estimation of missing precipitation records. **Journal of Hydrology**, v. 312, n. 1-4, p. 191-206, 2005. doi

THEVAKARAN, A.; SONNADARA, D.U.J. Estimating missing daily temperature extremes in Jaffna, Sri Lanka. **Theoretical and Applied Climatology**, v. 132, n. 1-2, p. 145-152, 2018. doi

VEGA-GARCIA, C.; DECUYPER, M.; ALCÁZAR, J. Applying cascade-correlation neural networks to in-fill gaps in

Mediterranean daily flow data series. **Water-Open Access Journal**, v. 11, n. 8, 1691, 2019. doi

VENTURA, T.M.; MARTINS, C.A.; FIGUEIREDO, J.M.; OLIVEIRA, A.G.; MONTANHER, J.R.P. MANNGA: A robust method for gap filling meteorological data. **Revista Brasileira de Meteorologia**, v. 34, n. 2, p. 315-323, 2019. doi

WANDERLEY, H.S.; AMORIM, R.F.C.; CARVALHO, F.O. Variabilidade espacial e preenchimento de falhas de dados pluviométricos para o Estado de Alagoas. **Revista Brasileira de Meteorologia**, v. 27, n. 3, p. 347-354, 2012. doi

WANDERLEY, H.S.; AMORIM, R.F.C.; CARVALHO, F.O. Interpolação espacial de dados médios mensais pluviométricos com redes neurais artificiais. **Revista Brasileira de Meteorologia**, v. 29, n. 3, p. 389-396, 2014. doi

WOLDESENBET, T.A.; ELAGIB, N.A.; RIBBE, L.; HEINRICH, J. Gap filling and homogenization of climatological datasets in the headwater region of the Upper Blue Nile Basin. Ethiopia. **International Journal Of Climatology**, v. 37, n. 4, p. 2122-2140, 2016. doi

YOZGATLIGIL, C.; ASLAN, S.; IYIGUN, C.; BATMAZ. I. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. **Theoretical and Applied Climatology**, v. 112, p. 143-167, 2013. doi