Earth Science Informatics (2021) 14:37–50
https://doi.org/10.1007/s12145-020-00528-8

RESEARCH ARTICLE

Improving flood forecasting through feature selection by a genetic algorithm – experiments based on real data from an Amazon rainforest river

Alen Costa Vieira[1] · Gabriel Garcia[2] · Rosa E. C. Pabón[3] · Luciano P. Cota[3] · Paulo de Souza[4] · Jó Ueyama[5] · Gustavo Pessin[3]

Received: 11 June 2019 / Accepted: 15 September 2020 / Published online: 3 October 2020

Abstract

This paper addresses the problem of feature selection aiming to improve a flood forecasting model. The proposed model is carried out through a case study that uses 18 different time series of thirty-five years of hydrological data, forecasting the level of the Xingu River, in the Amazon rainforest in Brazil. We employ a Genetic Algorithm for the task of feature selection and exploit several different genetic parameters seeking to improve the accuracy of the prediction. The features selected by the Genetic Algorithm are used as input of a Linear Regression model that performs the forecasting. A statistical analysis verifies that the final model can predict the river level with high accuracy, which obtains a coefficient of determination equal to 0.988. Hence, the proposed Genetic Algorithm showed to be successful in selecting the most relevant features.

Keywords River level forecasting · Genetic algorithm · Feature selection · Linear regression

Introduction

A report by the Brazilian Natural Disaster agency showed that 12.04% of natural disasters in Brazil from 1991 to 2012 (UFSC 2013) were caused by floods and that these floods directly affected more than 14 million people. On a global scale, the large number of natural disaster incidents has made Brazil one of the countries with the highest number of flood in the world (EM-DAT 2016). Moreover, it is widely expected that global climate change may aggravate this problem and have serious effects on the global hydrological cycles, causing changes in the precipitation regime and a rise in sea levels (IPCC 2013).

River flood forecasting systems can assist in warning people and communities to leave hazardous areas prior to disasters and a reliable prediction of river levels can help to reduce the social, economic and environmental damage caused by flooding.

Based on recent work we developed on this matter (Ueyama et al. 2017) we conclude that accurate flood forecasting should incorporate two key stages: (i) defining a predictive model, such as Neural Networks, Linear Regression, Decision Trees; and (ii) selecting the most relevant features to be used by that model, e.g. Univariate feature selection, Principal Component Analysis. The key benefits of this second stage are two-fold: (a) it helps capturing key features that can improve accuracy; (b) it avoids interference from irrelevant or redundant data (Guyon and Elisseeff 2003).

In light of the above, this article presents a feature selection strategy that aims to improve the performance of a predictor for flood forecasting. This feature selector extracts the key attributes among a set of 18 time series to train the predictor model. The model used is a simple Linear Regression as the focus of this research is how feature selection can impact the accuracy of the predictor, in that sense the use of a simple model facilitates such analysis.

During our study we employ a Genetic Algorithm (GA) to find a combination of attributes and parameters that would allow the algorithm to build a predictive model with higher accuracy than the predictive models generated from empirical configurations. We argue that evolutionary strategies are very suitable for this study because they have

Communicated by: H. Babaie

✉ Alen Costa Vieira
alen.vieira@sipam.gov.br

Extended author information available on the last page of the article.

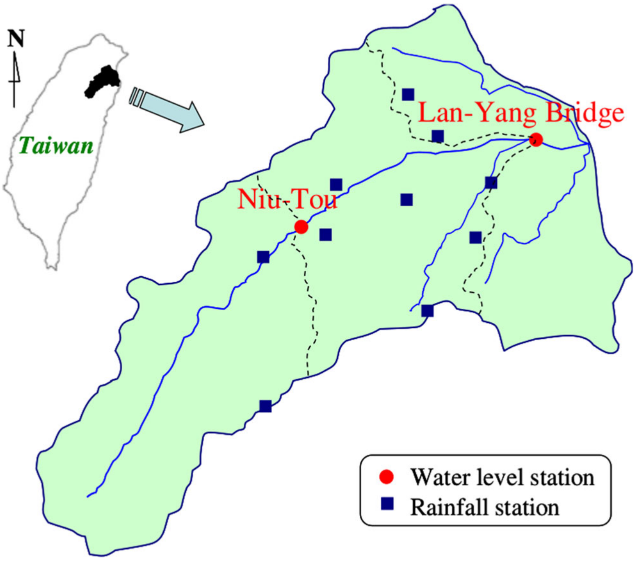


Fig. 1 Location of the rainfall and fluviometric stations used in the study carried out by Chen et al. (2007)

two key features: (i) the ability to cover a large search space (i.e. periodic river flow data that we read from sensors) and (ii) the ability to make fine adjustments that lead to optimal results; this is also shared by other work carried out by Eiben and Schippers (1998).

It should also be mentioned that when setting out the model in this article, we carried out experiments with real data from the Xingu River that surrounds the town of Altamira in Brazil (in the middle of Amazon rainforest and a tributary of Amazon river). The Belo Monte Hydroelectric Plant is also located in this region and this plant is considered to be the third largest in the world. In light of those features, this region has a high energy potential and attention is required to protect its natural resources; at the same time, it often suffers from floods that cause serious problems to the local population and severe damage to the environment.

The remainder of this article is structured as follows: first, Section "Related work" includes a discussion of related work. Then Section "Methodology" outlines our approach to improving flood forecasting. Following this, Section "Analysis of results" gives a performance evaluation and includes a discussion of the obtained results. Finally, Section "Conclusions" summarizes our conclusions of this research.

Related work

In this section, we outline strategies present in the literature for the task of hydrological forecasting using prediction models (Section "Prediction models for hydrological forecasting"); and feature selection aiming to improve those models (Section "Feature selection for hydrological forecasting").

Prediction models for hydrological forecasting

Several methods can be used for hydrological prediction. Dornelles et al. (2013) carried out a broad study on how different configurations of artificial neural networks (ANN) can be used to forecast river levels. Other studies also apply ANNs to forecast hydrological phenomena (Rodrigues et al. 2015; Furquim et al. 2016). Chen et al. (2007), Haddad and Rahman (2020), and Wu et al. (2019) use support vector machines (SVM) to perform flood forecasting. The use of simpler models as Linear Regression are also applied to river level forecasting by Rocha et al. (2007) and to flood forecasting by Franco (2014).

The data used by some studies (Chen and Yu 2007; Galelli and Castelletti 2013; Furquim et al. 2016) is the information of nearby stations. To illustrate this, Fig. 1 shows the rainfall and fluviometric stations used as support to carry out the prediction of the Lan-Yang River in the work undertaken by Chen and Yu (2007). In the absence of stations, ocean data can be used, as well as estimated

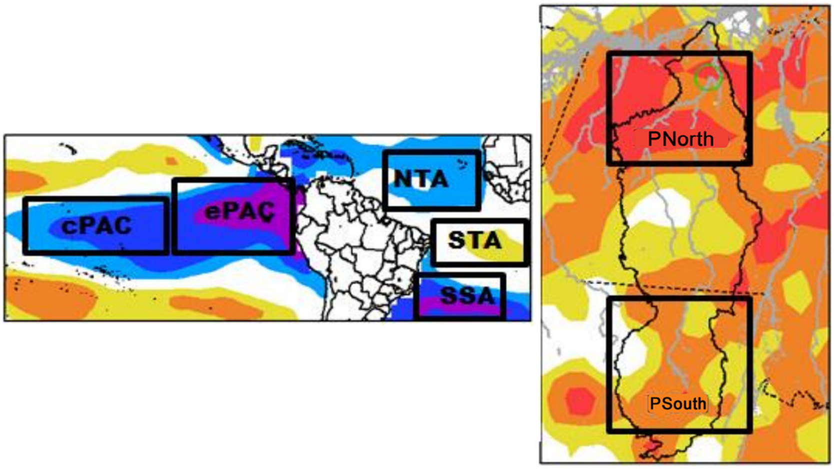


Fig. 2 Oceanic and precipitation data estimated by satellites used in the work by Franco (2014)

Table 1 Related work on variable selection methods

Related work	Method of selecting variables	Applications
Dornelles et al. (2013)	Empirical	Hydrological forecasting
Rodrigues et al. (2015)	Empirical	Hydrological Forecasting
Furquim et al. (2016)	Empirical	Hydrological Forecasting
Chen et al. (2007)	Cross correlation	Hydrological forecasting
Wu et al. (2019)	Cross correlation	Hydrological forecasting
Haddad and Rahman (2020)	Cross correlation	Hydrological forecasting
Rocha et al. (2007)	Linear correlation	Hydrological forecasting
Franco (2014)	Linear correlation	Hydrological forecasting
Galelli et al. (2013)	Partial information and iterative tree-based algorithm	Hydrological forecasting
de Lucena et al. (2012)	Multi-objective Genetic Algorithm	Prediction of protein concentration in wheat grains
Sumbana et al. (2012)	Genetic Algorithm	Detection of vandalism in Wikipedia
Tran et al. (2015)	Partial linear correlation, partial mutual information and genetic programming	Predicting multiple datasets

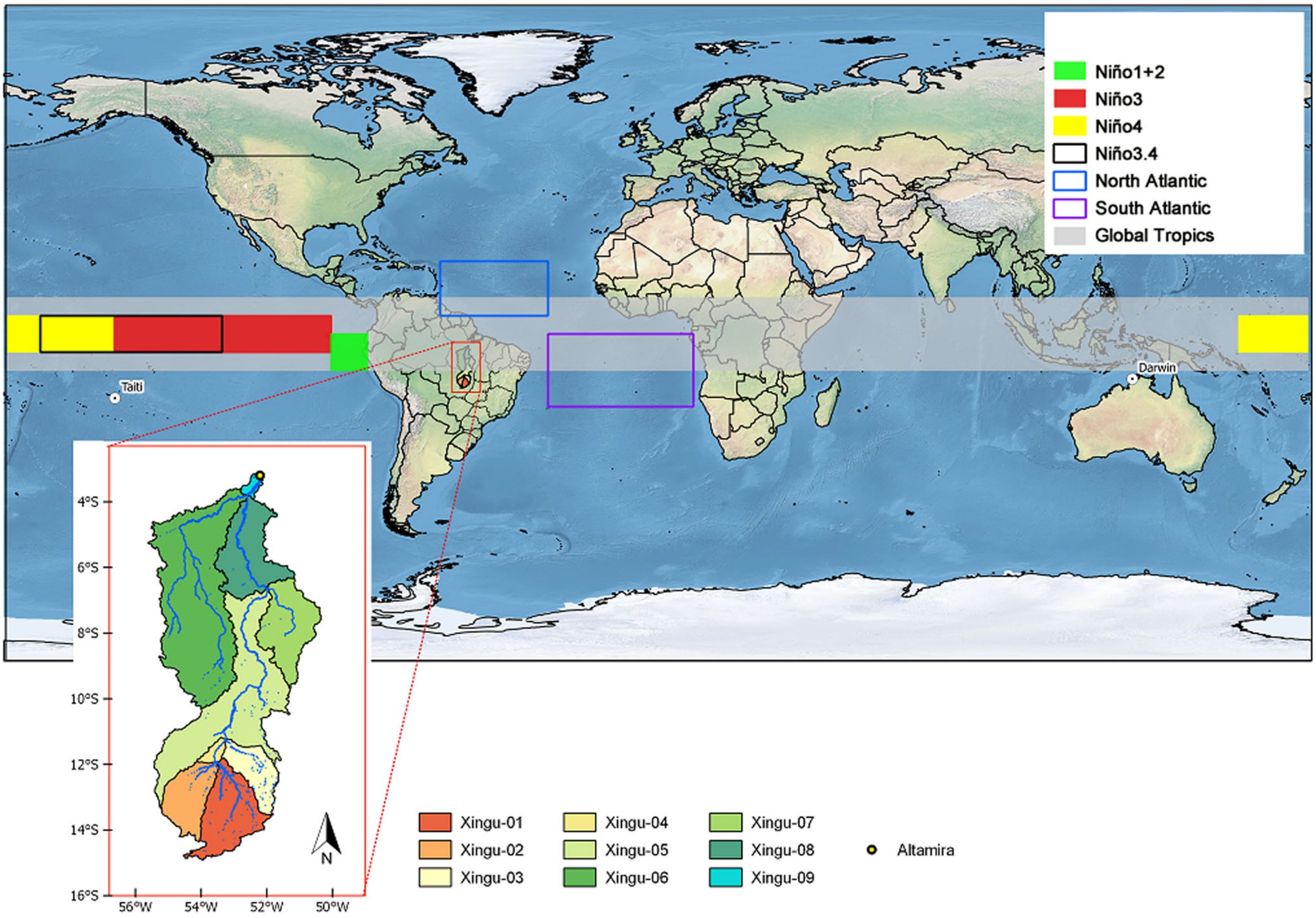


Fig. 3 Map showing the areas of 18 variables of interest zooming in on the sub-basins of the River Xingu and the 18 series of climate variables that can be used in the forecasting (with attributes and window sizes chosen by GA). These are as follows: (Niño1+2, Niño3, Niño4, Niño3.4, the North Atlantic, the South Atlantic, Global Tropics, the Atmospheric Pressure of Darwin and Tahiti, and Precipitation in the Xingu sub-basins (1· · · 9))

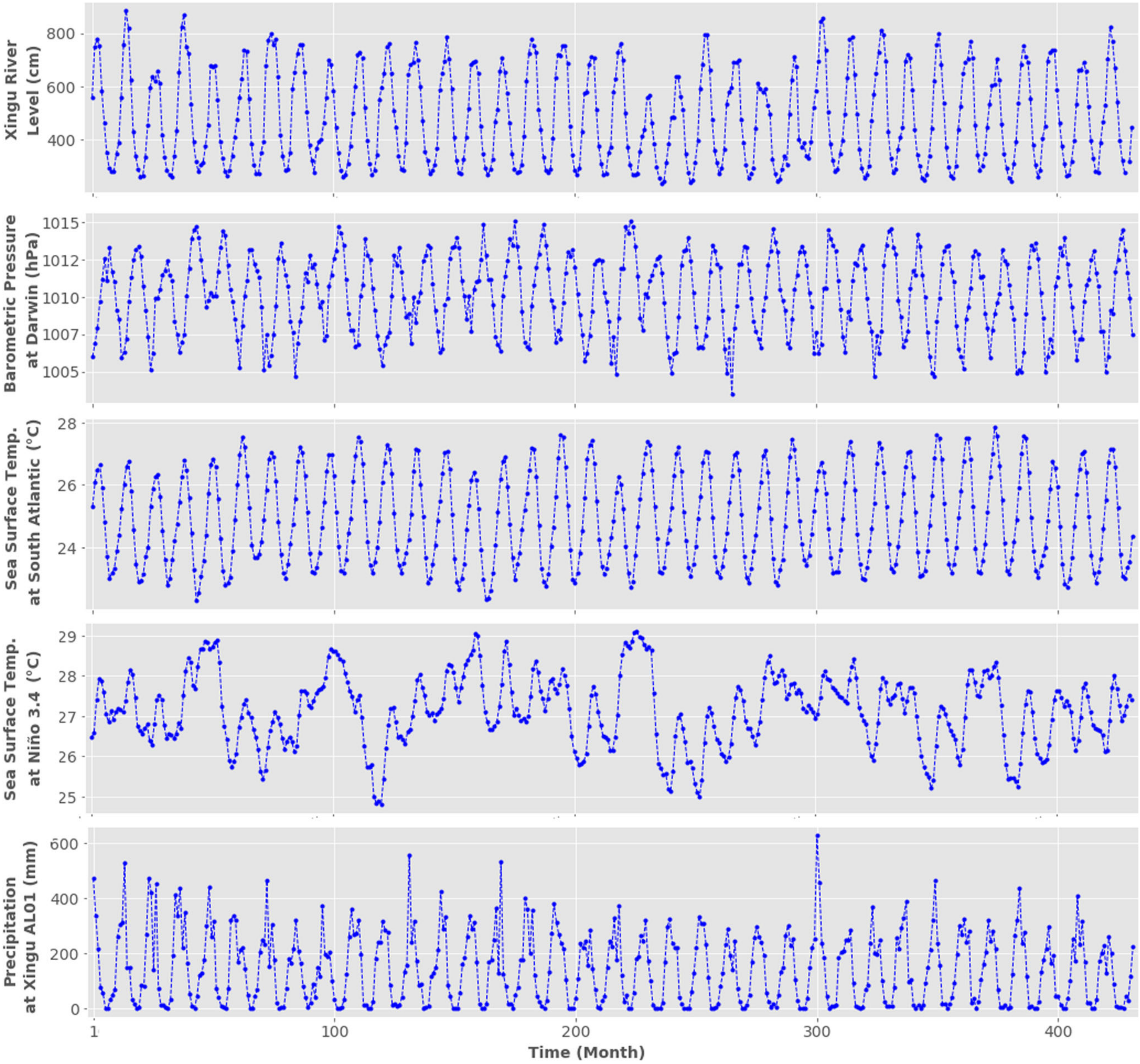


Fig. 4 Five examples of time series with monthly values (from 1979 to 2014) that were used in this study. The first time series (top-down) in this image is the level (cm) of the River Xingu (to be forecast). The other series are the barometric pressure (hPa) at Darwin, the sea surface temperature (°C) at South Atlantic, the sea surface temperature (°C) at Niño 3.4, and the and Precipitation (mm) in the Xingu Sub-basin 01

satellite precipitation data (Rocha et al. 2007; Dornelles et al. 2013; Rodrigues et al. 2015). Another example can be seen in Fig. 2, which shows which areas are used in the study conducted by Franco (2014) for the estimation of predicted levels in the Xingu River in Altamira.

Feature selection for hydrological forecasting

The methods for feature selection can be categorized as (i) model-based, and (ii) model-independent (Guyon and Elisseeff 2003; Galelli et al. 2014).

Model-based methods involve designing a complete predictive or classification model to estimate the performance of the selected support variables through a metric. The most widely used examples of this method are evolutionary algorithms (Galelli et al. 2014; de Paula 2015).

Independent modelling methods are based on statistical solutions between the subsets of the variable that has to be predicted with the aid of auxiliary attributes. The objective is to use a metric to compare the auxiliary attributes with the attribute that has to be predicted and then select the best-performing attributes without employing a predictive

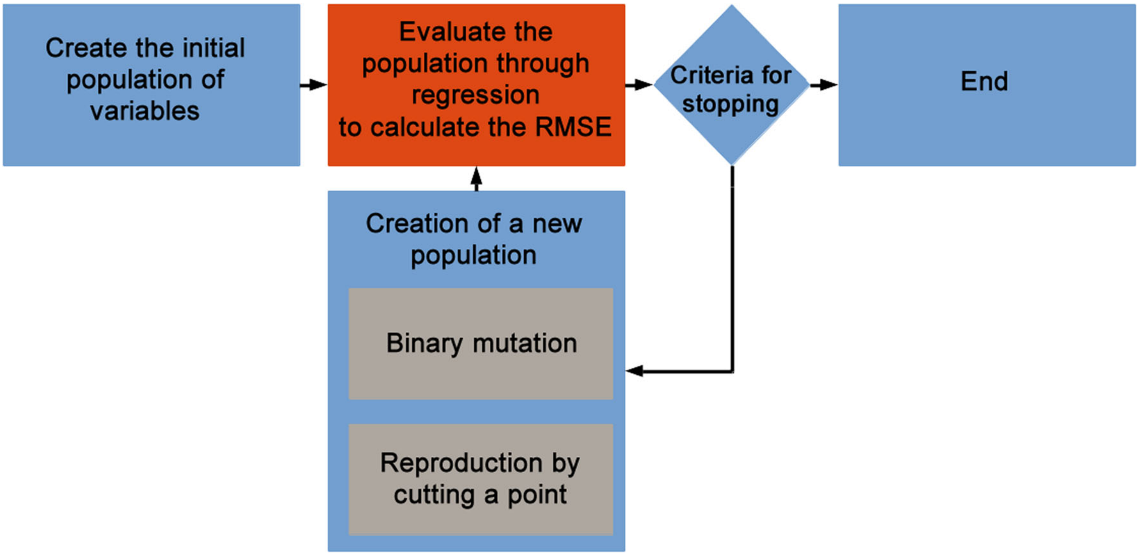

Create the initial population of variables
Evaluate the population through regression to calculate the RMSE
Criteria for stopping
End
Creation of a new population
Binary mutation
Reproduction by cutting a point

Fig. 5 Flowchart of the developed GA

model. The work by Souza and Araújo (2011) states that the main components of the analytical technique and correlation methods are the prime examples of these approaches.

In this study, we employ the model-based method by means of an evolutionary strategy that is widely used to tackle several types of optimization problems. Since it is a predictive problem, the purpose of the evolutionary strategy is to reduce errors in the forecasting.

Table 1 summarizes works on feature selection and their applications. The feature selection task to be performed is what time window of each variable should be used by the model, e.g. to use the data from X month of Y station.

Studies such as Franco (2014) and Rocha et al. (2007) define climatic features and their time windows to be used by the model using correlations for hydrological prediction in the Xingu River in Altamira. Wu et al. (2019) proposed a SVR to forecast flash floods at different lead times in small mountainous catchments from China. In the results, the method had a satisfactory predictive performance. Haddad and Rahman et al. developed models based on multidimensional scaling coupled to SVR to obtain improved flood quantile estimates at ungauged sites. The method was applied to 202 catchments from Australia and the SVR was able to capture efficiently the nonlinearity between the dependent and predictor variables. Chen et al. (2007) carried out a study that estimates the hourly levels in the Lan-Yang River in the North-East of Taiwan using support vector machine and choosing the variables and their time window through cross-correlation. This feature selection approach disregards interactions between variables that could improve the model accuracy. And, a good correlation or another statistical metric does not constitute by itself a cause-and-effect relationship. In the Genetic Algorithm outlined here, the selection evaluation method takes into account the interaction between the selected variables and their time windows.

The works by Dornelles et al. (2013), Rodrigues et al. (2015) and Furquim et al. (2016) investigate different predefined combinations of time windows in their predictions of river levels, and try to define the best features based on the analysis of each result. Due to the quadratic growing number of combinations between variables and their time windows it is clearly difficult to find the best combination using that method.

Sumbana et al. (2012) and de Lucena et al. (2012) state that they have achieved a good performance and solved their problems through a reduction of attributes. The selection of a smaller number of variables allows to achieve the

Table 2 Parameters for the population size of the Genetic Algorithm

Parameter	Value
Number of Generations	1.000 generations
Mutation rate	20%
Population size	25, 50, 100, 200 individuals
Elitism	1 individual
Initialization of the population	Random

best prediction through a multi-objective strategy. It can be concluded from this that, it would be worth checking the genetic mechanisms/operators that reduce the number of selected variables to determine if they meet the needs of the specific objectives of this study.

In this study, our aim is to devise strategies and parameters that can be used in evolutionary algorithms that help in the selection of variables, so that they can lead to the best possible performance and assist in the prevention of any type of disaster incident that may occur.

Methodology

Machine Learning (ML) is an effective approach to regression and the classification (whether supervised or unsupervised) of problems involving non-linear systems (Thomas 2017; Gonçalves et al. 2016). These systems can be intensely multivariate and include several or even thousands of descriptive attributes of one or more target attributes. In ML, a set of examples for training the employed technique is formed with a sufficiently comprehensive sample to generate predictive models of the target attributes. ML has proved to be useful in a large number of applications in natural environments (the earth, oceans and atmosphere), recovery algorithms, crop disease detection, new product creation, polarization correction and code acceleration (Silva et al. 2018; Gavriilidis et al. 2018; Jing et al. 2018).

Genetic Algorithms (GAs) (Holland 1975; Eiben and Schippers 1998; Linden 2012) are global optimization techniques that employ parallel and structured search strategies, that allow multi-objective searches to be conducted in multidimensional space. GAs use iterative procedures that simulate the evolution of a population consisting of candidate solutions of a given problem. The evolutionary process is guided by a selection mechanism based on the suitability of individual structures. A new set of structures is created for each iteration of the algorithm (single generation), by exchanging information (in the form of bits or blocks) between the structures selected from the previous generation. Due to this crossing of information, the fitness of some structures happens to improve in relation to the previous generation. For this reason, GAs can be successfully employed to find a satisfactory solution to problems that have an exponential complexity. This approach is driven by the need to tackle exponential problems that have a non-polynomial resolution time, such as the optimization problem (selection of attributes and parameters) addressed here.

One factor that should be stressed is that the GAs are a practical evolving strategy extensively used by the scientific community, due to its good results in different applications and considerable flexibility (Mokadem et al. 2018; Francescomarino et al. 2018). Genetic Algorithms are techniques based on an evolving pattern guided by a metaphor of natural selection. The individuals in the GA are representations of how to solve the problem. Operators of reproduction and mutation are employed during the evolutionary process for the generation of descendants and the individuals are assessed by means of "aptitude functions" (such as fitness) (de Oliveira et al. 2018).

Data

The aim of this work is to improve flood prediction through feature selection using GA. To predict a flood is necessary

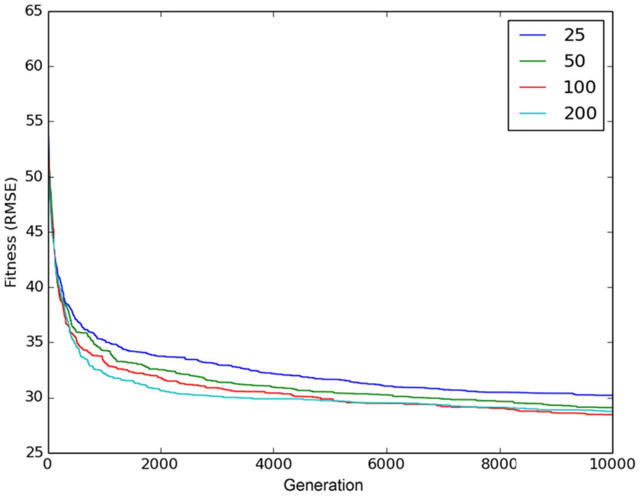


Fig. 6 Progression of fitness based on the number of generations and the different population sizes. Each line represents a population size and shows the average of 20 runs

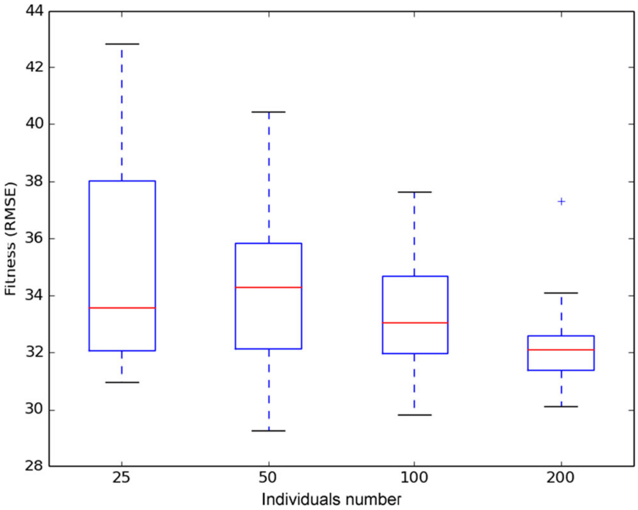


Fig. 7 Fitness of best individual from each population, in light of the twenty replications that were carried out

Table 3 Parameters for the mutation rate in the scenario and the elitism of the Genetic Algorithm

Parameter	Value
Number of generations	10.000 generations
Mutation rate	5%, 10%, 20%
Population rate	50 individuals
Elitism	1 individual, 10% and 20% of the population
Initialization of population	Random

to estimate what will be the maximum level of such river in a desired point based on information related to environment that surrounds the river. Therefore, this research focus on forecasting the maximum monthly values for Xingu River levels in the specific area of Altamira. The data used in this project is composed by 18 different time series. These time series are as follows: the sea-surface temperatures of the areas of Niño1+2, Niño3, Niño4, Niño3.4, and the North Atlantic and South Atlantic Ocean, together with the Global Tropics, the atmospheric pressure sea level of Darwin and Tahiti and precipitation in the Xingu River sub-basins $(1 \cdots 9)$.The sub-basins were defined in accordance to the coding system for hydrographic basins set out in the (Pfafstetter 1989). Figure 3 provides a map with the areas of 18 variables of interest zooming into the sub-basins of the River Xingu. Fig. 4 presents five examples of the time series (from 1979 to 2014).

The data to be predicted is the maximum monthly values of the River Xingu level at the Altamira Station, and these were made available by the hydro-meteorological database of the National Water Board (ANA). The data of the average monthly temperature and atmospheric pressure are derived from satellite observations and are interspersed with data obtained from the National Oceanic and Atmospheric Administration (NOAA) stations. All the variables relate to the period 1979 - 2014. The five last years of forecasting have been separated for this particular study.

Genetic algorithm for feature selection

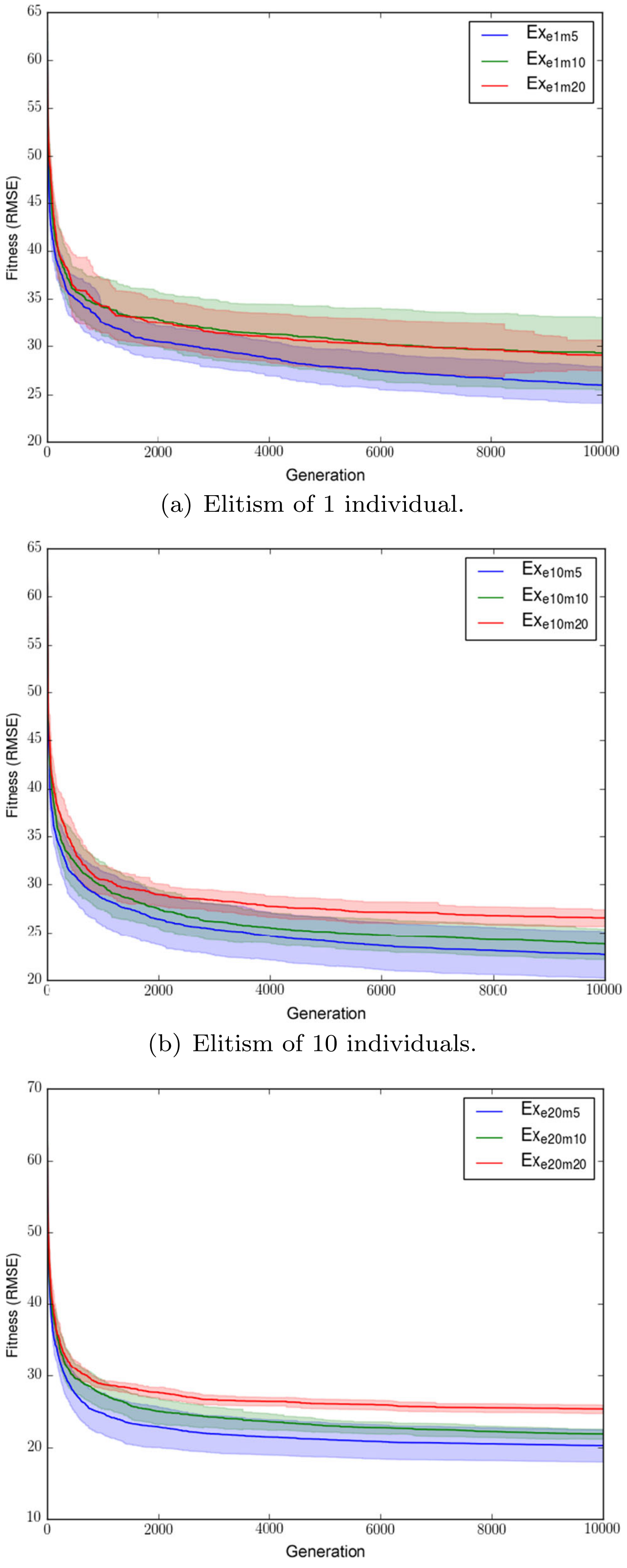
In this research, a Genetic Algorithm is created to perform a feature selection for a forecasting model. The GA makes a choice of the climate variables and their time windows of the 18 time series in order to improve the forecasting performed by a Linear Regression model. The Linear Regression uses the data selected by the GA to estimate the maximum monthly value of the river level.

Each individual represents a set of attributes to be used by the forecasting model. The evolving algorithm makes the selection of variables to be used and then applies a Linear Regression method to carry out the prediction. Linear Regression is used in this research because it is a simple method with low computational costs and allows a better analysis of the feature selection. The objective function is the Mean Squared Error (RMSE) resulting from the Linear Regression compared to the real river level. In this way, the GA seeks to reduce the error (as the fitness) through multiple iterations; this continues until the stopping criterion is satisfied. We use the number of iterations as the stopping criterion. Figure 5 shows a flowchart of the developed GA.

The chromosome uses a binary code (0 or 1) with the aim of deciding whether variable and its time window will be used. The mutation method chosen is the simple binary mutation where the value of the gene is inverted through a certain degree of probability. The chosen selection

Table 4 Settings for the experiments in the 2nd evaluation stage

Experiments	Elitism	Mutation
Ex_{e1m5}	1 individual	5%
Ex_{e1m10}	1 individual	10%
Ex_{e1m20}	1 individual	20%
Ex_{e10m5}	10 individuals	5%
Ex_{e10m10}	10 individuals	10%
Ex_{e10m20}	10 individuals	20%
Ex_{e20m5}	20 individuals	5%
Ex_{e20m10}	20 individuals	10%
Ex_{e20m20}	20 individuals	20%



(a) Elitism of 1 individual.

(b) Elitism of 10 individuals.

(c) Elitism of 20 individuals.

Fig. 8 Fitness (RMSE) based on the number of generations, for different rates of mutation and elitism. Each row shows the average of 20 runs along with its standard deviation

method is the “roulette addict” or Restriction Endonuclease Fingerprinting (Holland 1975), in which the likelihood of an individual being selected is proportional to its aptitudes. The reproduction method is the crossing-point. A form of elitism is involved (Bhandari and Murthy 1996), where a number of individuals remain from one generation to the next based on their fitness.

Parameter selection for the GA

According to Linden (2012), the ability to ensure good results by means of an evolutionary strategy is directly linked to the parameters that are used. de Paula (2015) points out that choosing the right parameters is still a challenging task and that several works differ on the best configuration for them. Therefore, we use a flow to define the parameters to be used in the GA. That flow seeks to define (i) the population size, (ii) the mutation and elitism scenario, and (iii) the population initialization technique, in that order. As the GA presents a random component, we use the average of 20 runs for each scenario for a better evaluation. The process and the results to define each of the aforementioned parameters are presented in the sequel.

Population size

Population size is a sensitive configuration – if it is too small there will not be sufficient genetic information to arrive at a good solution and if it is too large it will incur in very high computational costs. This scenario seeks to investigate the most appropriate population size that gives the best results in the lowest time.

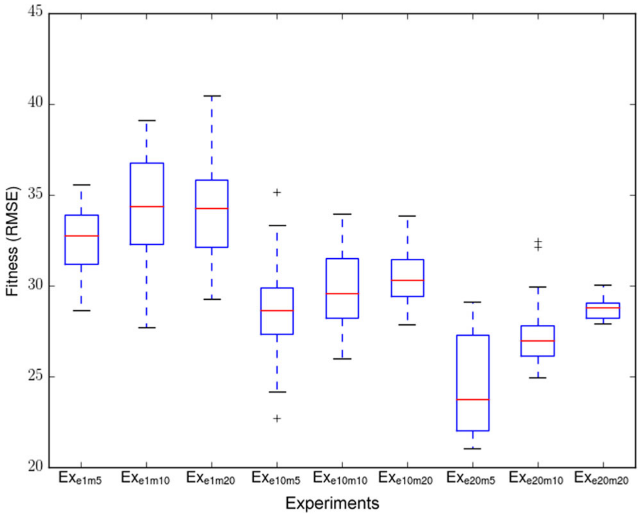


Fig. 9 The final results (fitness) for different mutation rates and elitism. Each boxplot displays the result of 20 executions

Table 5 Parameters for the initialization of the Genetic Algorithm population

Parameter	Value
Generation number	10.000 generation
Mutation rate	5%
Population size	50 individuals
Elitism	20% of the population
Initialization of population	Random 80% with one and 20% with zero 20% with one and 80% with zero

The parameters listed in Table 2 are used for this scenario, and the population size ranges between 25, 50, 100 and 200 individuals. Figure 6 shows the average fitness for the four population variations. It should be emphasized that the aim of our Genetic Algorithm is to reduce the Mean Squared Error (RMSE) of the prediction by means of selecting the most relevant attributes and their time windows represented by each individual.

It can be seen in Fig. 6 that the population with the largest size (200 individuals) has made a greater advance in its fitness in the first generations than the other configurations. However, with the advance of the generations, it is clear that the population with 100 individuals exceeds the level of fitness obtained by the largest population. In addition, the population with 50 individuals is approximately the same as the average fitness shown by the largest population and progressively maintains its fitness (such as the population with 100 individuals). This behavior is apparently opposite to the behavior of the configuration with a larger population where there is a trend towards stability in terms of average fitness.

Figure 7 shows the best individual fitness (20 replications), obtained from evaluating the population. It can be seen that there is a reduction in the variation of fitness achieved by the best individual as the size of the population increases. To ensure consistency and reliability in the analysis, we conducted several statistical tests. We use Shapiro-Wilk (1965) normality test to ascertain if the adequacy of results sets conformed to the normal distribution. The *p*-value obtained in all cases was greater than 0.05, which suggests that the hypothesis of adequacy to the normal distribution is acceptable. Thus, we proceed with Welch Two Sample *t*-test (parametric test) to compare the sets. The Welch Two Sample *t*-test (Montgomery 2013) shows , with a confidence interval of 99%, that there is no statistically significant difference between the results generated by populations with 50, 100 and 200 individuals (*p*-value greater than 0.01). However, there was a statistically significant difference when the population consisted of 25 individuals. These results provide evidence that increasing population to more than 50 does not necessarily improve the results. Since populations with 50, 100 and 200 individuals led to an equivalent behavior in the system, in

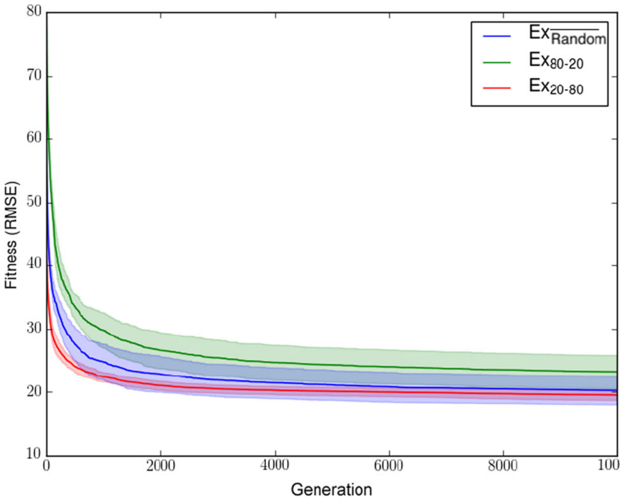


Fig. 10 Fitness (RMSE) by number of generations, when there are different types of population initialization. Each row shows an average of 20 executions along with their standard deviation

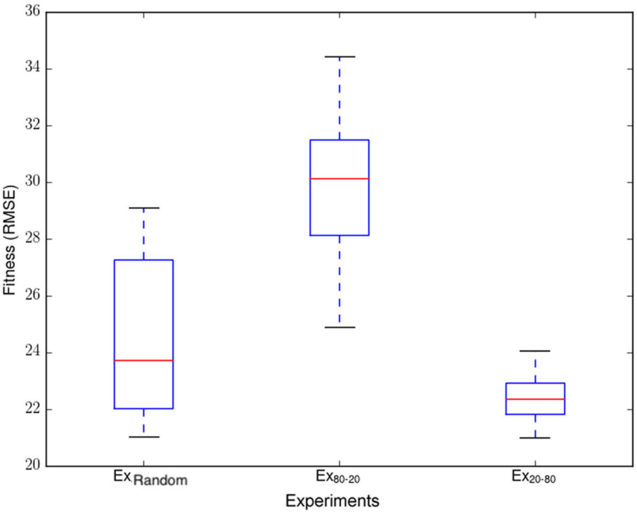


Fig. 11 Results (fitness) for different population initializations. Each boxplot displays the result of 20 executions

Table 6 Final parameters for the Genetic Algorithm

Parameter	Value
Generation number	10.000 generation
Mutation rate	5%
Population size	50 individuals
Elitism	20% of the population
Initialization of population	20% with one and 80% with zero

terms of predictive accuracy. Therefore, we select a population with 50 individuals as it incurs in lower computational costs.

The mutation scenario and elitism

Mutation is a genetic operator that ensures the genetic diversity of the population (Linden 2012). If the rate is too low, the population might stagnate quickly, and if it is too high the algorithm can behave as a random search. Elitism is one way to ensure that the best genes will be preserved for the next generation. However, it can cause premature convergence in a minimum local. The parameters defined in Table 3 are used for the mutation scenarios.

In this scenario we seek to evaluate the influence of mutation and elitism on fitness optimization. Table 4 shows the combinations of elitism and mutation rates evaluated at this stage. In view of the results achieved previously, we fixed the population size at 50 individuals.

Figure 8 displays three graphs referring to the advance of fitness based on the number of generations. Each subfigure fixes the rate of elitism and diversifies the rate of mutation, as can be seen in the respective captions. The three graphs behave similarly and the configurations with the lowest mutation rate present a better performance.

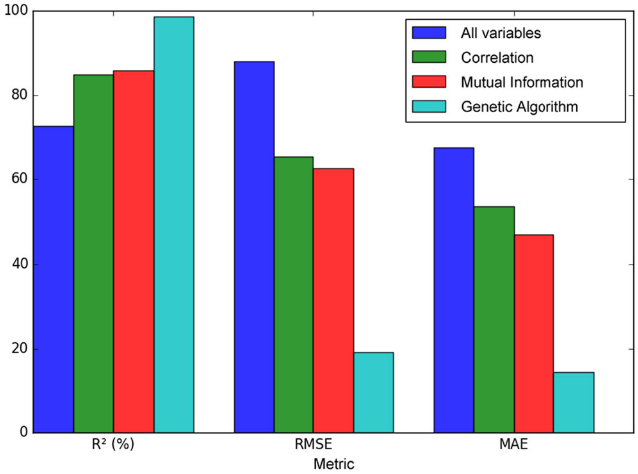


Fig. 12 Statistical analysis of the prediction performed by three different methods for feature selection and the use of all variables

Figure 9 shows the result of 20 executions, obtained from the experiments. It can be noted that higher elitism (20%) obtains a superior performance (lower fitness) for all rates of mutation. Thus, the configuration with a better performance is a lower mutation rate with a higher rate of elitism. We use Shapiro-Wilk normality test to ascertain adequacy and normality of results, and only Ex_{e20m5} and Ex_{e20m10} experiments do not present normal distribution. Then, we apply *t*-test through Welch Two Sample *t*-test. After the analysis, we conclude that Ex_{e20m5} experiment is better because it is more effective than others.

Population initialization techniques

In general, the population in GA is initialized at random. However, a knowledge of the problem can be obtained through different initializations (Rahnamayan et al. 2007; Khaji and Mohammadi 2014), so long as there is no premature convergence of the model. For this reason, we use the parameters shown in Table 5, where three population initialization policies are explored.

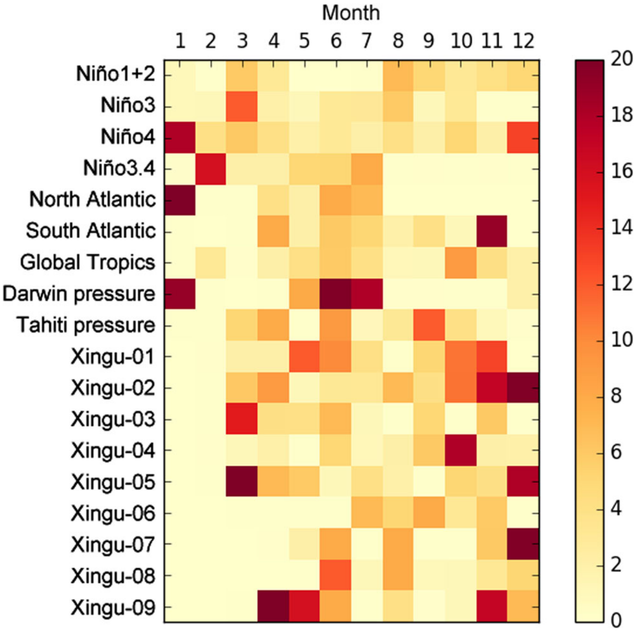


Fig. 13 The intensity of each pixel represents the number of times each variable was selected in the 20 replicate feature selection experiments

Table 7 Prediction accuracy indices using Linear Regression and feature selection through GA on the basis of 20 executions

	R^2	MAE	RMSE
Average	0.987	14.558	19.278
Maximum	0.988	16.612	20.136
Minimum	0.986	13.435	18.572
Standard deviation	0.000	0.698	0.453

Figure 10 shows the fall in fitness in terms of the number of generations, and includes the following: *i*) the different initialization of the initial population, *ii*) a Ex_{Random} experiment using random initiation, *iii*) an $Ex_{80\text{-}20}$ experiment using the initialization of the chromosome with 80% of values 1 and 20% of values 0 and *iv*) an $Ex_{20\text{-}80}$ experiment using chromosome initialization with 20% of 1 values and 80% of 0 values. It can be seen in Fig. 10 that the fall is greater when fewer genes start at 0. The difference in the $Ex_{80\text{-}20}$ experiment is greater than in the other experiments.

Figure 11 shows the result of the best individual (20 executions), obtained with different types of initialization of the population. On the basis of the data displayed in Fig. 11, we carried out a series of statistical evaluations to determine the difference between the types of initialization. Initially, we use Shapiro-Wilk normality test to calculate the accuracy of the results. The results for all tests obtain *p*-values higher than 0.05, which can be interpreted as a normal distribution. In view of this, a comparison of the results are made with Welch Two Sample *t*-test. The test shows that there is no statistical difference in the sets (i.e. the p-value was less than 0.05).

After analysis, we conclude that the $Ex_{20\text{-}80}$ experiment is better because its performance is more satisfactory than others.

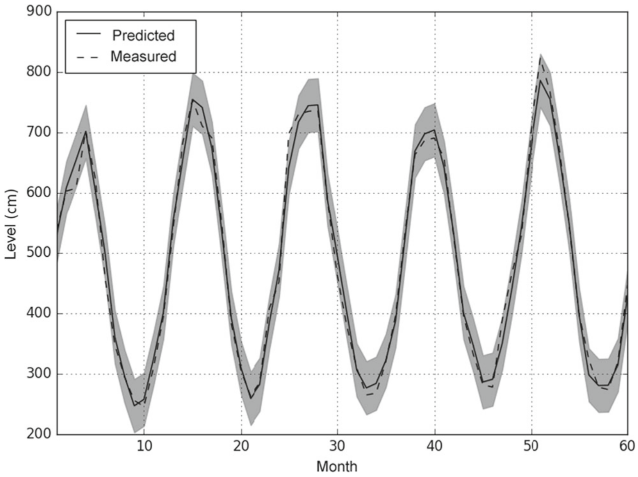


Fig. 14 Observed and predicted values with a 95% confidence interval using Linear Regression with the input values selected by the Genetic Algorithm

Analysis of results

The experiments previously discussed had the objective of finding optimal parameters among the ones evaluated to the application of feature selection using a Genetic Algorithm. The final parameters are a combination of the best parameters found and are exposed in Table 6. Using those parameters, the next step is to perform an extensive analysis of the prediction and feature selection results. In view of the stochastic behavior of the evolutionary algorithm, the experiments were replicated twenty times so that the results could be more reliable. Subsequently, the results were submitted to statistical tests for a more detailed analysis as discussed in the sequel.

The Genetic Algorithm performs the feature selection and a Linear Regression model is used to predict the river levels. The following comparative metrics are used to evaluate the performance of the GA: *i*) Coefficient of Determination (R^2); *ii*) Root Mean Square Error (RMSE); and *iii*) Mean Absolute Error (MAE). In our case, the Coefficient of Determination[1] provides a measure of how dependent the variable to be predicted is from the independent variables used. It is a measure used to evaluate how relevant are the features used to the prediction model. The RMSE[2] is the standard deviation of the errors of prediction, i.e. the square root of the average of the squared differences between the river level predicted and the measured. And the MAE[3] is the average of the errors.

Analysis 1: attribute selection

In order to evaluate the performance of the Genetic Algorithm against other feature selection methods, the GA

[1] Equation for the Coefficient of Determination:

1. $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{true}-y_{pred})^2}{\sum_{i=1}^{n}(y_{true}-\bar{y})^2}$, where y_{true} is the data set, y_{pred} is the prediction, $\bar{y}$ is the average of y, and n is number of the observations.

[2] Equation for the Root Mean Square Error:

2. $RMSE = \sqrt{\frac{1}{n}\sum_{i=i}^{n}(y_{true} - y_{pred})^2}$

[3] Equation for the Mean Absolute Error:

3. $MAE = \frac{1}{n}\sum_{i=i}^{n}|y_{true} - y_{pred}|$

is assessed against two common feature selections strategies and the use of all variables. One approach is selecting the most relevant features through correlation, as proposed by Hall (1999). And the other one is using mutual information to select the features to be used, as in Peng et al. (2005). All strategies are evaluated based on the results of twenty different executions.

Figure 12 presents the results for the experiments performed. The coefficient of determination for the GA is 98.7%. What means that the features used are 98.7% relevant for the problem tackled. The other strategies achieved between 72% and 86% of coefficient of determination. What demonstrates that the Genetic Algorithm is capable of selecting the best features in order to predict the river level. Also, the mean square errors and the mean absolute errors of the Genetic Algorithm are significantly smaller than those made by the other approaches.

Figure 13 shows the decoding of the final chromosome results (20 executions). We noticed that some variables were selected during the 20 executions such as Tropical North Atlantic, Darwin Pressure, Xingu-05, Xingu-07 and Xingu-09 in different time windows. Whereas, some features are rarely used, as Xingu-04 and Xingu-06. Another point to be observed is that the features used change according to the month, i.e. a feature can be relevant in a month and irrelevant in another month. What enhances the contribution of the feature selection stage.

Analysis 2: prediction

The second step aims to evaluate the quality of the Linear Regression predictions. Analyzing the RMSE and MAE metrics in Fig. 12 one can note that the Genetic Algorithm presents a considerably smaller error. Table 7 provides the statistical analysis for 20 executions predicting the river level with Linear Regression after selecting the features using GA. The results of the forecasting clearly demonstrated that with the aid of Linear Regression, feature selection by GA, and environmental data it is possible to estimate the maximum levels of the Xingu River in Altamira in a satisfactory way.

Figure 14 shows the values predicted by the model that presented the best performance of the 20 executions based on RMSE, together with the observed values. It can be seen that the lines are similar, and that the expected results generally fall within a 95% confidence interval.

Conclusions

Hydrological forecasting of river water levels is crucial to prevent or mitigate the effects of possible disasters in towns and agricultural areas such as those encountered close to the Amazon river. The use of artificial intelligent systems can help to predict the floods and reduce the amount of losses caused.

This article outlines a strategy for flood prediction and applies it to forecasting the maximum levels of the Xingu River in Altamira. We propose a predictor based on a Linear Regression model that uses information of different climate variables. Our approach applies a Genetic Algorithm with the purpose of selecting the most relevant features and their time windows between a set of 18 time series of climate variables. The set of features selected are used by the Linear Regression model to predict the river level.

The Genetic Algorithm was tuned based on a flow aiming to define the population size, mutation and elitism rates, and initialization scenario. The parameters are selected based on a complete statistical analysis seeking for fast execution and high accuracy. With the best parameters selected, the feature selection proposed method is compared to other common approaches. The Genetic Algorithm presents the best results and proved to be a feasible and robust method for feature selection for flood forecasting.

The final Linear Regression model using the most relevant features selected by the Genetic Algorithm is assessed by a statistical analysis. Based on that analysis we verify that the proposed model is able to successfully predict the monthly maximum level in the Xingu River within a good confidence interval.

Acknowledgments The authors would like to thank the following colleagues due to help revising the manuscript and providing ideas to its best organization: Bruno S. Faiçal, Leandro Y. Mano, Vinícius Gonçalves and Pedro H. Gomes. The authors would like also to thank Márcio Nirlando Gomes Lopes due to his help in the development of Figure 3. Dr. J. Ueyama would like to acknowledge FAPESP, process 2018/17335-9.

References

Bhandari D, Murthy CA (1996) Genetic algorithm with elitist model and its convergence. IJPRAI 10(6):731–747

Chen ST, Yu PS (2007) Pruning of support vector networks on flood forecasting. J Hydrol 347(1):67–78

de Lucena DV, de Lima TW, Soares AS, Coelho CJ (2012) Multi-objective evolutionary algorithm nsga-ii for variables selection in multivariate calibration problems. Int J Natural Comput Res 3:43–58

de Oliveira LL, Freitas AA, Tinós R. (2018) Multi-objective genetic algorithms in the study of the genetic code's adaptability. Inf Sci 425:48–61

de Paula TI (2015) Avaliação da influência de parêmetros do algoritmo genético na otimização de um problema multiobjetivo utilizando-se arranjo de misturas. Master's thesis, PPGEP, Univesidade Federal de Itajubá

Dornelles F, Goldenfum JA, Pedrollo OC (2013) Artificial neural network methods applied to forecasting river levels. Revista Brasileira de Recursos Hídricos 18:45–54

Eiben AE, Schippers CA (1998) On evolutionary exploration and exploitation. Fundamenta Informaticae 35(1-4):35–50

EM-DAT (2016) The international disaster database. Emdat Advanced Search. Available at www.emdat.be/advanced_search/index.html

Francescomarino CD, Dumas M, Federici M, Ghidini C, Maggi FM, Rizzi W, Simonetto L (2018) Genetic algorithms for hyperparameter optimization in predictive business process monitoring. Inf Syst 74(Part):67–83

Franco VS (2014) Previsao hidrológica de cheia sazonal do rio xingu em altamira-pa. Master's thesis, PPGCA, Universidade Federal do Pará

Furquim G, Pessin G, Faiçal BS, Mendiondo EM, Ueyama J (2016) Improving the accuracy of a flood forecasting model by means of machine learning and chaos theory. Neural Comput & Applic 27(5):1129–1141

Galelli S, Castelletti A (2013) Tree-based iterative input variable selection for hydrological modeling. Water Resour Res 49(7): 4295–4310

Galelli S, Humphrey GB, Maier HR, Castelletti A, Dandy GC, Gibbs MS (2014) An evaluation framework for input variable selection algorithms for environmental data-driven models. Environ Model Softw 62:33–51

Gavriilidis A, Velten J, Tilgner S, Kummert A (2018) Machine learning for people detection in guidance functionality of enabling health applications by means of cascaded SVM classifiers. J Franklin Institute 355(4):2009–2021

Gonçalves VP, Giancristofaro GT, Geraldo Filho P, Johnson T, Carvalho V, Pessin G, de Almeida Neris VP, Ueyama J (2016) Assessing users emotion at interaction time: a multimodal approach with multiple sensors. Soft Comput 21(18): 5309–5323

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn 3:1157–1182

Haddad K, Rahman A (2020) Regional flood frequency analysis: evaluation of regions in cluster space using support vector regression. Nat Hazards 102:489–517

Hall MA (1999) Correlation-based feature selection for machine learning

Holland JH (1975) Adaptation in natural and artificial systems. The University of Michigan Press

IPCC (2013) Climate change 2013: the physical science basis. contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge

Jing M, Jie Y, Shou-yi L, Lu W (2018) Application of fuzzy analytic hierarchy process in the risk assessment of dangerous small-sized reservoirs. Int J Mach Learn Cybern 9(1):113–123

Khaji E, Mohammadi AS (2014) A heuristic method to generate better initial population for evolutionary methods. CoRR arXiv:1406.4518

Linden R (2012) Algoritmo genetico editora ciencia mordena

Mokadem D, Amine A, Elberrichi Z, Helbert D (2018) Detection of urban areas using genetic algorithms and kohonen maps on multispectral images. IJOCI 8(1):46–62

Montgomery DC (2013) Design and analysis of experiments, 8th edn. Wiley, New York

Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238

Pfafstetter O (1989) Classificação de bacias hidrográficas - Metodologia de classificação Departamento Nacional de Obras de Saneamento (RJ)

Rahnamayan S, Tizhoosh HR, Salama MMA (2007) A novel population initialization method for accelerating evolutionary algorithms. Comput Math Applic 53(10):1605–1614

Rocha EJP, Rolim PAM, Santos DM (2007) Modelo estatístico hidroclimático para previsão de níveis em Altamira-PA. In: XVII Simpósio brasileiro de recursos hídricos

Rodrigues MM, Costa MGF, Filho CFFC (2015) Proposta de um método para previsão de cheias sazonais utilizando redes neurais artificiais: Uma aplicação no rio amazonas. In: Workshop de computação aplicada a gestão do meio ambiente e recursos naturais (WCAMA)

Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52:591–611

Silva B, Netto MAS, Cunha RLF (2018) Jobpruner: a machine learning assistant for exploring parameter spaces in HPC applications. Future Gen Comp Sys 83:144–157

Souza F, Araújo R (2011) Variable and time-lag selection using empirical data. In: IEEE 16th conference on emerging technologies & factory automation, ETFA 2011, pp 1–8

Sumbana MIM, Silva AJC, Gonçalves MA, Almeida JM, Pappa GL (2012) Seleção de atributos utilizando algoritmos genéticos para detecção do vandalismo na wikipedia. In: XXVII Simpósio brasileiro de banco de dados - short papers, São Paulo, São Paulo, Brasil, October 15-18, 2012, pp 209–216

Thomas JM (2017) Complex network embedding in the hyperbolic space using non-linear unsupervised machine learning techniques. Ph.D. thesis, Dresden University of Technology, Germany

Tran H, Muttil N, Perera B (2015) Selection of significant input variables for time series forecasting. Environmental Modelling & Software 64(C):156–163

Ueyama J, Faiçal BS, Mano LY, Bayer G, Pessin G, Gomes PH (2017) Enhancing reliability in wireless sensor networks for adaptive river monitoring systems: reflections on their long-term deployment in Brazil. Computers, Environment and Urban Systems 65:41–52

UFSC (2013) Atlas Brasileiro de Desastres Naturais: 1991 a 2012. Centro Universitario de Estudos e Pesquisa sobre Desastres. Universidade Federal de Santa Catarina

Wu J, Liu H, Wei G, Song T, Zhang C, Zhou H (2019) Flash flood forecasting using support vector regression model in a small mountainous catchment. Water 11:1327

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Alen Costa Vieira[1] · Gabriel Garcia[2] · Rosa E. C. Pabón[3] · Luciano P. Cota[3] · Paulo de Souza[4] · Jó Ueyama[5] · Gustavo Pessin[3]

Gabriel Garcia
gabriel.garcia@vale.com

Rosa E. C. Pabón
rosa.correa@itv.org

Luciano P. Cota
luciano.p.cota@itv.org

Paulo de Souza
paulo.desouza@griffith.edu.au

Jó Ueyama
joueyama@icmc.usp.br

Gustavo Pessin
gustavo.pessin@itv.org

[1] Management and Operational Center of the Amazon Protection System (Censipam), Brasília, DF, Brazil

[2] Vale S.A, Carajás, PA, Brazil

[3] Robotics Lab, Instituto Tecnológico Vale, Ouro Preto, MG, Brazil

[4] School of Information and Communication Technology, Griffith University, Nathan, QLD, Australia

[5] Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, SP, Brazil