Revista Brasileira de Cartografia (2017), Nº 69/1, Edição Especial Geotecnologias e Desastres Naturais: 33-42
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto
ISSN: 1808-0936

A MACHINE LEARNING-BASED MODEL TO IMPROVE SHORT-TERM FORECASTS OF FLOODING IN NOVA FRIBURGO-RJ

Um Modelo Baseado em Aprendizado de Máquina para Melhorar as Previsões de Curto Prazo de Inundações em Nova Friburgo-RJ

Glauston Roberto Teixeira de Lima[1] & Graziela Balda Scofield[2]

[1]National Centre for Monitoring and Early Warnings of Natural Disasters - CEMADEN
Coordenação Geral de Pesquisa e Desenvolvimento
Estrada Doutor Altino Bondesan, 500, Distrito de Eugênio de Melo, CEP:12.247-016, São José dos Campos/SP, Brasil
glauston.lima@cemaden.gov.br

[2]National Centre for Monitoring and Early Warnings of Natural Disasters - CEMADEN
Coordenação Geral de Operação e Modelagem
Estrada Doutor Altino Bondesan, 500, Distrito de Eugênio de Melo, CEP:12.247-016, São José dos Campos/SP
graziela.scofield@cemaden.gov.br

Received on November 20, 2015/ Accepted on June 10, 2016
Recebido em 20 de Novembro, 2015/ Aceito em 10 de Junho, 2016

ABSTRACT

Machine learning use in hydrological modeling has intensified in recent decades given the potential of these techniques to produce in short time satisfactory solutions to support tasks such as early flooding warnings. In this context, this work reports the development and the results of a forecasting model built from a hydrometeorological database and using a regression tree. This regression tree-based model is intended to forecast, with hours in advance, the level of a river in Nova Friburgo-RJ, which was chosen as study area due to its recent history of major natural disasters. Rainfall and river level data used in the modeling were collected during the years of 2013 and 2014 in four stations located in the study area. The regression tree allowed more flexibility in the model design. The first regression tree results yielded Nash indexes above 0.75 indicating the feasibility of the approach. However, in order to be used as an operational decision-making support tool working in real time, the model should be improved with new studies and tests carried out with enlarged hydrometeorological databases.

Keywords: Natural Disasters, Hydrological Modeling, Early Flooding Warning, Machine Learning, Regression Tree, Nova Friburgo-RJ.

RESUMO

A utilização de aprendizado de máquina em modelagem hidrológica intensificou-se nas últimas décadas dado o potencial dessas técnicas de produzir em curto tempo soluções satisfatórias para suporte em tarefas como a emissão antecipada de alertas de inundações. Neste contexto, o presente trabalho relata o desenvolvimento e os resultados de um modelo preditivo construído com base em dados hidrometeorológicos e utilizando uma árvore de regressão. Este modelo baseado em árvore de regressão destina-se a predizer, com horas de antecedência, o nível de um rio em Nova Friburgo-RJ que foi escolhida como área de estudo pelo seu histórico recente de desastres naturais de grandes proporções. Os dados de chuva e nível de rio utilizados na modelagem foram coletados durante os anos de 2013 e 2014 em quatro estações de medição situadas na área estudada e a árvore de regressão permitiu maior flexibilidade na concepção do modelo. Os

primeiros resultados obtidos com a árvore de regressão mostraram índices Nash acima de 0,75 indicando a viabilidade da abordagem. Contudo, para ser utilizado como uma ferramenta operacional de suporte à decisão funcionando em tempo real, o modelo deverá ser aprimorado com novos estudos e testes realizados com bases de dados hidrometeorológicos mais abrangentes.

Palavras chaves: Desastres Naturais, Modelagem Hidrológica, Alerta Antecipado de Inundações, Aprendizado de Máquina, Árvore de Regressão, Nova Friburgo-RJ.

1. INTRODUCTION

Nova Friburgo is a municipality located in the mountainous region of Rio de Janeiro State, occupying a total area of 933.41 km^2. The estimated population in 2015, based in the Census of 2010, is 184.786 people (IBGE, 2015). The relief is very rough with altitude varying from 2310 (Três Picos) to 200 meters, in the border with Casimiro de Abreu municipality. The city and its urban área is completely surrounded by mountains of Serra dos Orgãos as Pico do Caledônia (in the south), Pedra do Imperador (in the southeast), among other, and the central area lies at an altitude of 846 meters (CORREIA, 2011).

Initially, the occupation of Nova Friburgo occurred along the floodplains of Bengalas River and of its tributaries Santo Antônio and Cônego Rivers. Due to population growth, there was an inordinate increase of housing construction on the slopes and on the banks of rivers and streams, especially by the low-income population. Furthermore, nowadays Bengalas River is channeled and/or rectified in all the urban area of Nova Friburgo and its margins are occupied with urban infrastructure such as roads, bridges, sidewalks and buildings. The presence of impermeable areas contributes to the rainwater drainage more quickly into the river, due to increased runoff, and consequently the peak of flooding is anticipated. Thus, occurrences of natural disaster caused by flooding increased in the city, especially during the summer that corresponds to the rainy season. According to Corrêa (2008), the first large flooding records occurred in the late 19th century, in an episode in which the rain lasted for three consecutive months.

From the above and the recent history of natural disasters in Nova Friburgo, it is clear the need to monitor the Bengalas River to predict in advance level rises exceeding its bankfull stage in order to issuing early flooding warnings. These warnings are important to Nova Friburgo civil defense that can take appropriate and immediate actions to evacuate the risk areas and prevent traffic in areas susceptible to flooding.

In this context, this work reports the development of a data-driven hydrological model based on a machine learning technique. In the case of this study, a regression tree (RT) was used. The proposed model is intended to make short-term forecasts of river level to improve the advance-accuracy tradeoff of flooding warnings. The model was built using rainfall and river level data collected at four monitoring stations of Instituto Estadual Ambiente (INEA), three of which are located in the urban area of Nova Friburgo and the remaining outside this area. Observational database cover the years 2013 and 2014. Figure 1 shows the location of the studied watershed in the Brazilian territory (Figure 1A), two maps with information about the altimetry (Figure 1B) and the population distribution in the watershed (Figure 1C). Both maps show Bengalas River, its two main tributaries (Santo Antônio and Cônego Rivers) and the four INEA monitoring stations.

Besides this introduction, the article is organized as follow: previous works using RT in hydrological modeling is presented in the section 2; Section 3 described the database used to develop the model; in Section 4, the modeling methodology is explained; Section 5 show the results and discussions and in the Section 6 the concluding remarks are addressed.

2. RELATED WORKS

The application of machine learning methods in hydrological modeling dates back to the ‘90s and since then has increased because it has proven to be fast and efficient in building solutions (such as forecasting models) in this field of study where, generally, the relationship among the different variables (climatic, topographical, vegetal cover and others) is highly nonlinear. This potential to satisfactorily model complex

nonlinear functional input-output relationships remains even if only a reduced database is available, as in the case of this work. Some previous works using RT (and other machine learning techniques) in hydrological modeling are listed below.

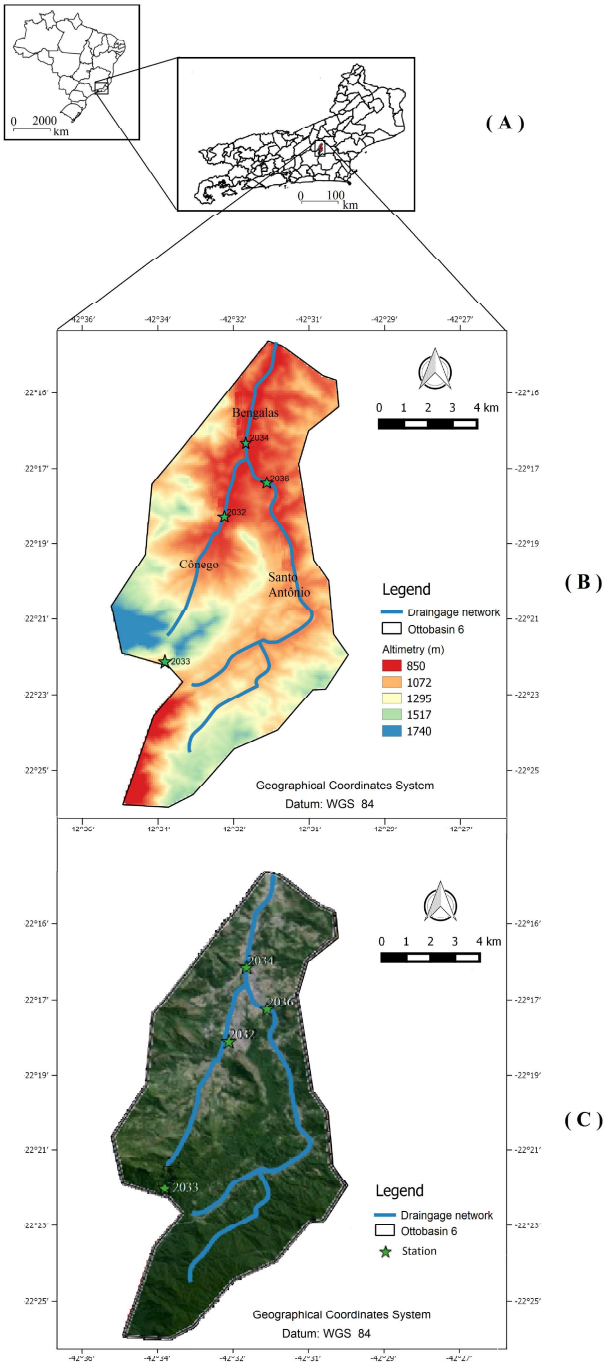

Fig. 1 - Maps of the studied area in the Bengalas River watershed in Nova Friburgo-RJ. (A) the location of the studied watershed in the Brazilian territory, (B) the altimetry of the watershed and (C) the population distribution in the watershed. The four INEA monitoring stations are identified by their numbers.

To model the rainfall-runoff relationship for two sub catchments in the United States, in order to deal with the underlying dynamic systems as a static ones, Iorgulescu and Beven (2004) took as predictors variables only linear combinations of input attributes observed at previous time steps in a ‘‘memory’’ window with predefined width and so it was possible to use RT as runoff forecasting approach in the modeling task.

Solomatine and Xue (2004) compared a M5 model tree (QUINLAN, 1992) with a multilayer perceptron (MLP) neural network in a flooding forecasting problem for the upper reach of the Huai River in China. The authors concluded that M5 is advantageous because the model is more interpretable, training is fast and always converges, and, in addition, M5 has predictive accuracy similar to MLP.

Siek and Solomatine (2007) used an approach for hydrological forecasting based on modular modeling in which each module (or sub-model) is an improved version of standard M5 algorithm and corresponds to a particular hydrological condition. This configuration allowed a better absorption of the knowledge of experts and also optimized the search in the solution space of each module. The results were more accurate than those obtained with overall global models approach.

In Londhe and Charhate (2010) some machine learning techniques for river flow forecasting are compared: artificial neural network (ANN), RT and evolutionary programming. It was found that the three techniques performed almost equally well.

In Tsai *et al.*, (2012) a hybrid model combining ANN’s (MLP and radial basis function (RBF)) with a CART tree (BREIMAN *et al.*, 1984) was used to perform river level forecasting under the interaction of upstream flows and tidal effects during typhoon attacks. CART was used to make a preliminary classification of the river level as high, medium or low and then ANN provided an estimate of the river level value. The results of CART-RBF were better than CART-MLP results.

Sattari *et al.*, (2013) compared M5 with support vector machine (SVM) in predicting, up to 7-day ahead, daily stream flow in Sohu River, Ankara, Turkey. The performance of

both approaches was similar, but to use M5 was considered more attractive by the possibility of applying a simpler linear model with lower computational cost.

Galelli and Castelletti (2013) developed a model-free algorithm to select the most relevant input attributes for hydrological modeling tasks using a regression tree-based structure.

The present study is part of a research line focused on construct data-driven hydrological models based on machine learning techniques which only recently started to be explored in the National Centre for Monitoring and Early Warning of Natural Disasters (Cemaden). So, it is understandable first explore modeling approaches simpler to handle and RT meets this requirement since it offers advantages such as: develop the algorithm based on which the RT is induced, it is the step that requires more effort and time, but once the induction algorithm is designed, the computational cost (i.e., the running time) of a RT is generally low, even when dealing with large databases; RT, usually, dispenses data pre-processing and prior knowledge on these data as, for instance, the underlying probability distributions; RT works well in solving non-linear "problems" (as hydrological modeling) by breaking it in minor "problems" which may be linearly treated; RT-based forecasting models are easily interpretable, which makes its operational use in early warning systems more friendly even for decision makers not expert in this type of modeling. Moreover, it is worthy to note that the in related works above, RT performance in hydrological modeling is reported as compatible with other machine learning techniques (e.g. ANN, SVM)

Particularly in our study, RT low computational cost afforded flexibility in the development of the hydrological model in the sense that many assumptions on the relationship between observed rainfall data and observed river level data could be investigated. Thus, we were able to test various alternatives of composing the input tuples (e.g. varying the number of samples in each input tuple - remember that each sample stand for observed values of rainfall and river level from four measuring stations) in order to find one that would produce a better fit to the output variable. We also tested training sets with different sizes (i.e., training sets corresponding to different time windows) to induce the RT, thus obtaining many insights on the temporal aspect of the inputs-output relationship.

At this early stage, our research aimed not algorithmic innovations and a basic algorithm was developed to induce RT. Thus, specifically from this point of view, there are no innovations in relation to the referenced related work. Nevertheless, we can quote others contributions: there are few studies focused on flooding forecasting and issuing early warnings in the studied area using data-driven models and machine learning techniques; this research line can improve the traditional hydrological modeling approaches based on calibration of physical models by experts interacting with them in a cooperative and complementary way.

Finally, we emphasize that the proposed forecasting model need to be improved to meet the reliability requirements of an early flooding warnings system. This purpose will be achieved, with additional studies and tests which include new enlarged hydrometeorological databases and other study areas.

3. DATABASE

The database used to build the model to forecast the level of the Bengalas River in the central area of Nova Friburgo is composed by two variables: rainfall and river's level. Values of these variables are given in mm and meters, respectively, and were collected with a temporal resolution of 15 minutes between February 2013 and November 2014 in four INEA measuring stations located in the study area (see Figure 1). Table 1 presents information on these four stations, including name, number, type, geographic coordinates and the river where the stations are installed.

Table 1: INEA monitoring stations in the studied area

station (number)	type	coordinates	monitored river
Suspiro (2034)	hydro	22º16'S/42º32'W	Bengala
Ypu (2036)	hydro	22º17'S/42º31'W	Santo Antônio
Olaria (2032)	hydro	22º18'S/42º32'W	Cônego
Caledônia (2033)	pluvio	22º21'S/42º34'W	-

As seen in Table 2, the pluviometric station 2033 (study area headwaters) provides only rainfall data while the hydrological stations 2032, 2034 and 2036 provide rainfall and river's level data. It should be emphasized that this work uses river's level data instead of flow data because the rating curves are no available to the conversion.

To induce the RT (our model to forecast Bengalas River's level), data were arranged in tuples containing 4 attributes concern observed values of rainfall and other 3 attributes concern observed values of river's level. Thus, each tuple in the database represents with its 7 attributes an observed rainfall-river's level condition for a specific day and hour in the study area.

In addition, once the regression tree is intended to forecast the level of Bengalas River at Suspiro station (study area outlet), each input tuple was associated to an output attribute (the target attribute to be forecast) corresponding to the level of Bengalas River measured at this station t-minutes later (which means that "t minutes" is the forecast horizon of the model).

To illustrate, Figure 2 shows an example of input tuple and its output value recorded in Nova Friburgo at midnight and at 2 a.m., respectively, of July 15, 2013 with high rainfall values.

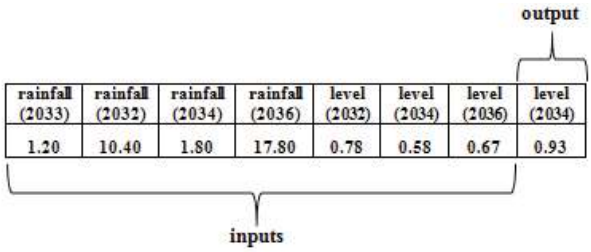

Fig. 2 – Tuple of rainfall and river's level values recorded in Nova Friburgo; the inputs were measured at midnight while the output was measured at 2 a.m., of July 15, 2013.

Approximately 62000 tuples with rainfall and river's level data (as exemplified in Figure 2) were used to inducing the RT, according to the methodology described in the Section 4.

4. METHODOLOGY

Standard nonlinear regression technique models the behavior of a dependent variable Y (output attribute) as a function of multiple independent variables **X** (X_1 ... X_i, ... X_n), (input variables) using a single relation, as given by Eq.

$$Y = a_0 + a_i^T \mathbf{X} + b\mathbf{X}^T + \text{å} \quad (1)$$

where

a_i are the coefficients (or weights) assigned to individual influence of each variable X_i on the Y behavior,

b are coefficients (or weights) attributed to the influence of the correlations among independent variables **X** on Y, and

ε is the error of the model.

RT's (BREIMAN *et al.*, 1984) are also a kind of nonlinear prediction approach, but instead use an unique general model (like Eq.1) for linking Y with the entire input attribute space defined by **X**, RT apply the called recursive partitioning to the attribute space and, for each resulting terminal subdivision, use simpler models to relate Y with **X**. The underlying idea is that within these subdivisions the relationship between Y and **X** is more manageable and it can be treated as linear. Therefore, the induction of a regression tree is made in two stages: (1) the recursive partitioning and (2) attaching a model to the terminal subdivisions.

In this work a binary RT is induced, that is, the original attribute space is successively divided into 2, 4, 2^i, ..., 2^n smaller regions and this process ends when a previously established stopping criteria is met. A tree of nodes can represent this recursive partitioning. The original attribute space is associated with the so-called root node of the tree and, similarly, the internal nodes of the tree are associated with the successive subdivisions. Figure 3 shows a simplified representation of the binary tree structure with its nodes as well as the resulting partitioning of the original attribute space.

In order to clarify the understanding of the tree representation in Figure 3A, we provide the meaning of the literal symbols in the figure: the X_i's represent any of the input variables, with i = 1, ... , n, (for example, in our database i = 1, ... , 7); the v_j's (v_1, … , v_6) are the so-called cutoff values (split values in the RT terminology); and the m_k's (m_1, … , m_8) are values that the RT can estimates for the output variable Y. Also, remember that each point in the feature space in Figure 3B is a tuple with its input and output values.

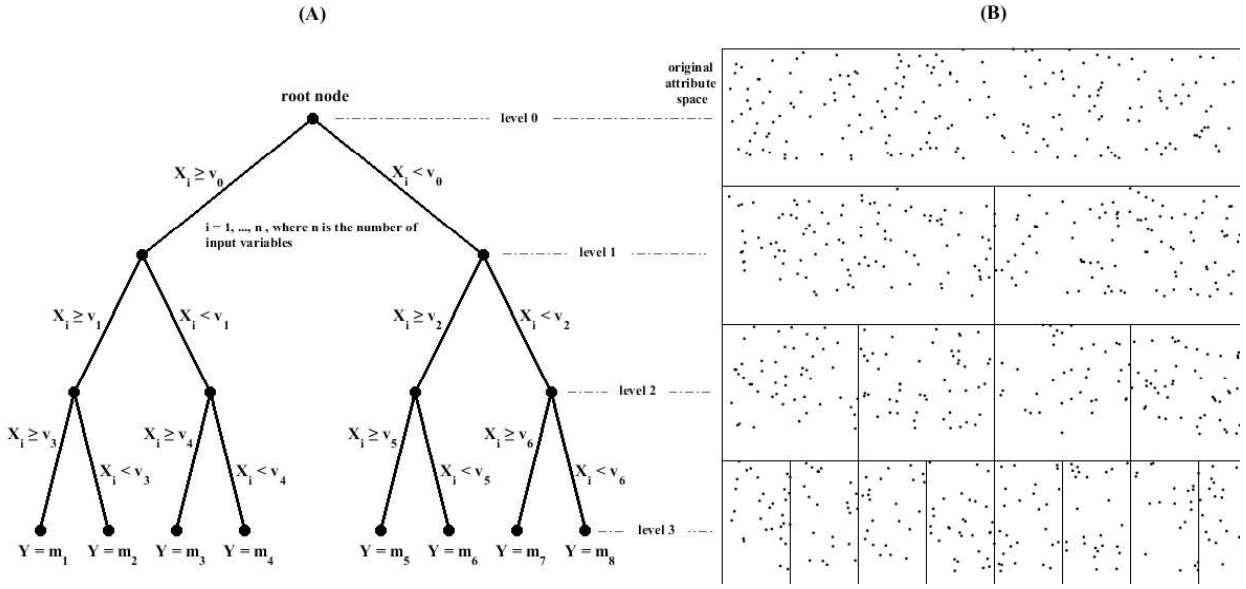

Fig. 3 – Representations of (A) a binary regression tree and (B) successive partitioning of the attribute space.

In the tree induction process, the divisions applied to the attribute space are chosen to reduce the prediction error of the two sets of output attributes resulting from this division compared to the prediction error of the previous set of output attributes. This variation in the prediction error is calculated as a function of the sample variance as given by Eq. 2:

$$\Delta E = \sum_{i=1}^{N}(y_i - \bar{y})^2 - \sum_{j=1}^{L}(y_{Lj} - \bar{y}_L)^2 - \sum_{k=1}^{R}(y_{Rk} - \bar{y}_R)^2 \quad (2)$$

where ΔE is the variation of the prediction error, the first summation represents the variance of the primitive set of output attributes (with N values) before division and the other two summations represents the variances of the two sets of output attributes derived by the division (with L and R elements, respectively).

For a set of data (tuples of input attributes and its respective output attributes) that reach a node of the tree, Eq. 2 is applied in order to assess which division of the output attribute (in two new sets) yields greater ΔE. The best division is determined by a specific value of a specific input variable. So, the binary test (if $X_i \geq v_j$ or if $X_i < v_j$, as illustrated in Figure 3A) made in this node is associated with this variable (X_i) and the value (v_j).

To reinforce the idea of this partitioning scheme using Figure 3, consider the original attribute space in Figure 3B and the root node of the tree in Figure 3A. In this node, after all the possible divisions being tested, it was found that the division of the database which produced the two sets of output attributes with less sample variance (what implies the greater ΔE of Eq. 2) was one set composed of all tuples whose value X_i is greater than or equal to v_0 and other set composed of all tuples whose value of X_i is less than v_0. In the same way, in the sequence each of these sets will be analyzed separately and will be also split. Still based on Figure 3A, the first will be split in two news sets considering the value v_1 of X_i as cutoff while the second will be split in two news sets considering the value v_2 of X_i as cutoff (just remember that X_i can be any of the input variables).

When a set of data reaches a node, this data is not split into two sets in two cases: (1) if ΔE provided by the best division represent a percentage decrease below a pre-defined threshold, and/or (2) if the number of data instances reaching the node is smaller than or equal to a pre-defined threshold. In these cases, the node is called a leaf (a terminal node of the tree). In the tests conducted with RT (the results of which are shown in section 4), the two thresholds above mentioned have been set, respectively, as 5% and 2.

If all nodes at a certain level of RT are leaves (or terminal nodes), the recursive partitioning stage ends and begins the phase of assigning a model to every leaf of the tree. In this work, we

simplified and the sample mean of the output attributes that reached each leaf was adopted as the "forecasting model" for that leaf. In Figure 3A this corresponds to the values m_k assigned to the output variable Y ($Y=m_1$, $Y=m_2$, $Y=m_3$, $Y=m_4$).

Thus, to obtain a forecast for a test tuple (i.e, a tuple not used in the tree induction process), this tuple makes a path that starts at the root node, goes through some internal nodes (these are determined by the attributes of the test tuple and the binary tests on the nodes of the induced tree) and ending on a leaf. The value (m_k) associated with the leaf, is the forecast for the test tuple.

5. RESULTS

A RT induced according to the algorithm described in section 3 was tested to forecast the level of the Bengalas River with 2 hours in advance. Our RT has been also tested for shorter forecast horizons. But in this section emphasis was placed on the 2 hours forecast results considering that, in a real-world situation, the greater the warning in advance, the better is the mitigation of the damages for Nova Friburgo Civil Defense. The experiment carried out is detailed below.

Let Ti to be a given tuple of the database used as test tuple. A RT was then induced using as "training inputs" all the tuples measured 4 hours immediately before Ti. Whereas the tuples in the database were collected every 15 minutes, this 4-hours time window corresponded to a training set with 16 tuples.

To the first 9 training tuples (measured in the first two hours of the 4-hours window), are associated, as output attributes, the level of the Bengalas River measured in the station 2034 two hours later. However, for the remaining 7 training input tuples (measured in the final hour and forty-five minutes of the 4-hours window), there are no output values yet, since the experiment must simulate a real situation. To clarify this point, suppose, for example, that Ti (our test tuple) has just been measured at 12:00. Therefore, the 4-hours time window goes from 08:00 to 11:45. The last value of Bengalas River's level, measured at 12:00, is se t as output attribute for the training tuple measured at 10:00. For the training tuples measured after 10:00, between 10:15 and 11:45, the output attributes are not measured yet. To circumvent this problem, the solution was to use the own input attributes of these 7 tuples related to the Bengalas River's level (input attribute: level/2034) as proxy for its output attributes. For the above example, the training tuples set is as illustrated in Figure 4.

time	inputs		outputs	time
	rainfall	level	level	
08:00	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	10:00
08:15	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	10:15
08:30	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	10:30
08:45	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	10:45
09:00	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	11:00
09:15	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	11:15
09:30	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	11:30
09:45	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	11:45
10:00	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	12:00
10:15	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	10:15
10:30	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	10:30
10:45	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	10:45
11:00	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	11:00
11:15	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	11:15
11:30	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	11:30
11:45	2033, 2032, 2034, 2036	2032, 2034, 2036	2034	11:45

Fig. 4 – Set of training tuples used to induce RT to make a forecast for a test tuple (Ti) measured at 12:00. Note that for the first 9 tuples the measurement times of inputs and outputs are delayed by 2 hours, while for the last 7 tuples the value of the input attribute (level/2034) was repeated as output attribute.

This experiment was repeated considering each tuple T_i in the database as a test tuple (obviously our first test tuple T_i was the seventeenth since this is the first tuple for which it is possible to form a 4-hours time window). The results obtained with our RT in these experiments are shown in figures 5 and 6. Besides, Table 2 presents two descriptive metrics, the mean error (ME) and the Nash index (NASH), calculated for the same experiment performed considering several forecast horizons. ME measures the systematic error of the forecasting model while NASH indicate the accuracy. NASH index (NASH & SUTCLIFFE, 1970) is widely used in hydrological modeling. Still regarding metrics to evaluate the results of the RT 120-minutes, the mean absolute error (for all tested tuple) was 0.1043 m and the absolute error was less than 0.25 m in more than 98% of the cases.

Table 2: Descriptive metrics used to evaluate the model

Forecast horizon (min)	Descriptive metric	
	ME (m)	NASH
15	-0.0005	0.9919
30	-0.0002	0.9362
60	0.0002	0.8697
120	0.0004	0.7630

5.1 Discussions

The first issue to be addressed is the choice of the short time window of training data used to induce the RT. It was based on a heuristic assumption that recently observed rainfall and river level data is the most relevant and reliable to perform short-term forecasts of the river level dynamic behavior. Considering this assumption, time windows of various sizes (as mentioned in section 2), ranging from 2 to 8 hours, were tested and the 4-hour window has provided the best results.

RT results were very good in some cases, such as the level peaks shown in panels A, B and C of Fig. 6 but also results only reasonable such as in the other level peaks shown in Figure 6 (panels D, E and F). In the latter case, the curve of the RT forecasts seems to have a delay relative to the observed curve (approximately equal to the forecast horizon of 2 two hours).

Once the RT was used to make forecasts with two hours in advance, the two types of results (good and not so good) can be understood considering two scenarios: (1º) it occurred no rainfall or only low intensity rainfall in the watershed or (2º) it occurred rainfall of medium or high intensity in the watershed within two hours.

In the first scenario, the river level dynamic behavior remained the same during the next two hours due to the lack of a disturbing factor (rainfall) in the watershed. Thus, the training set (even using surrogates for the output attribute of the last 7 tuples, as we did in this study) was adequate to provide the induced RT with sufficient information to make a good forecast with two hours in advance.

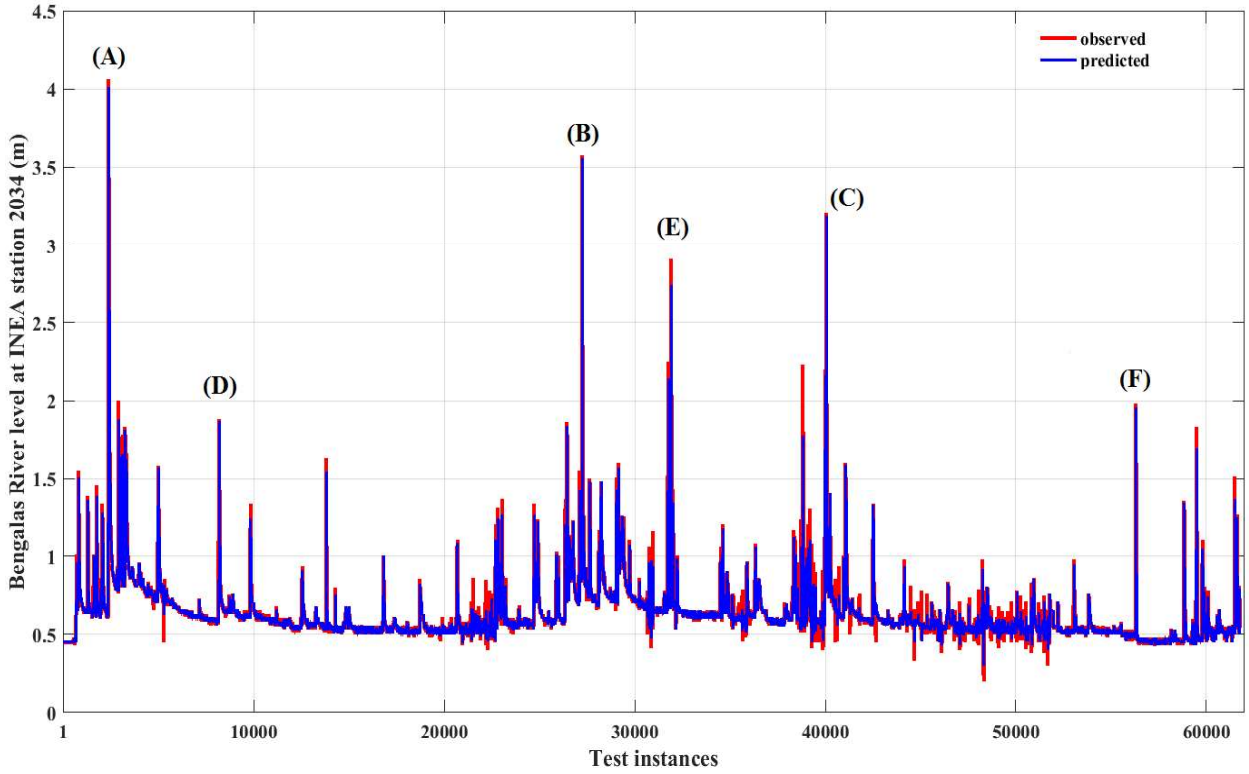

Fig. 5 – Comparison between observed level of Bengalas River and forecasts made by a RT two hours in advance. Peaks indicated as (A, B, C, D, E, F) are shown highlighted in Figure 6.

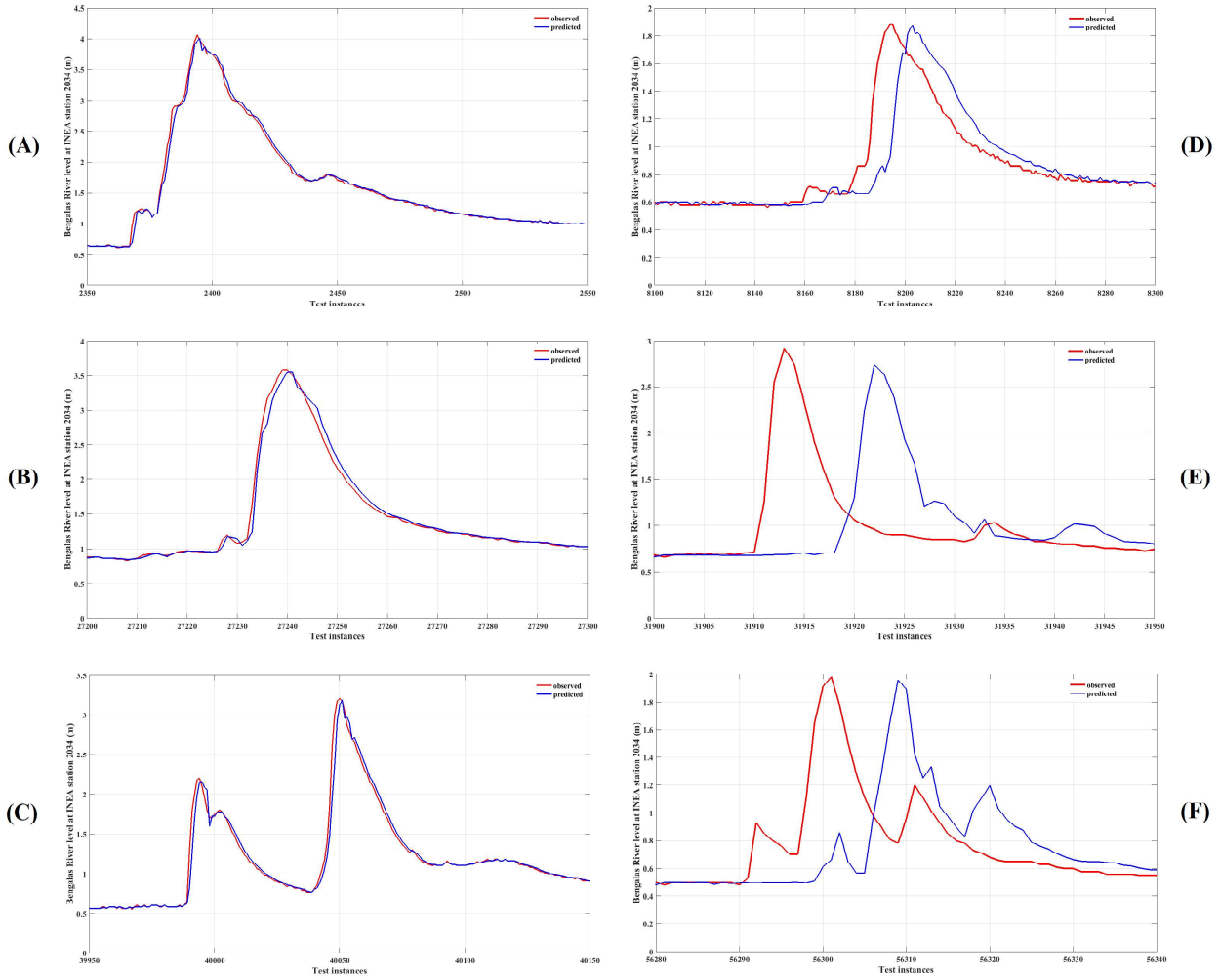

Fig. 6 – Zoom in on some parts of the curves in Figure 5 where occurred level peaks.

On the other hand, in the second scenario the information about disturbing factors that will occur within two hours in the watershed cannot be incorporated into the induced RT using only past data. That is, the values assumed as observed values for the outputs attributes of the last 7 training tuples do not represent the information about what it really occurred in terms of rainfall intensity in the watershed and this explains that the regression tree fails to make a good forecast in these cases. In this regard, it is noteworthy that as the forecast horizon increases, the quality of the results yielded by RT decreases, as demonstrated by the NASH indexes in Table 2 for results obtained with RT considering different forecast horizons. Still with reference to Table 2, the ME indexes seemed indicate that the model has a systematic error that overestimates the level for smaller forecast horizons while underestimates for larger ones.

An alternative to improve the RT forecasts, even for larger forecast horizons, is to use, in addition to the historical series of hydrometeorological data, rainfall estimates provided by radar, small or medium scale numerical weather models and satellite. Thus, instead of compose the training sets of the RT as illustrated in the Figure 3, the rainfall estimates (from radar, numerical models and satellite) could be used as a proxy for the output attributes of the last 7 tuples of the training set. So, despite the occurrence of a few "not so good" results, the hydrological modeling approach based on RT looks promising and could be improved with new data and tests including other study areas.

6. CONCLUDING REMARKS

This work presented the development of a hydrological forecasting data-driven model based on a RT. The model is intended to improve the accuracy of the flooding warnings in the city Nova Friburgo-RJ for river level forecasts made two hours in advance.

Rainfall and river's level data, collected

between February-2013 and November 2014 in four INEA monitoring stations located in Nova Friburgo, were used to build the model. More than 98% of the model's forecasts agreed closely with the observed values of Bengalas River's level. However, in a few cases the model was unable to make good forecasts due, probably, to the occurrence of rainfall in the 2 hours forecast horizon.

To further improve the results, more historical rainfall and river's level data and also rainfall estimates from radar, numerical models and satellite can be included in the future versions of the model. In addition, the model may be tested also for other study areas.

REFERENCES

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R.A., STONE, C.I. **Classification and regression trees**. Belmont, Calif.: Wadsworth. 1984. 369p.

CORRÊA, M. J. B. **O Cotidiano de Nova Friburgo no Final do Século XIX, Práticas e Representação Social.** EDUCAM, 2008. 502 p.

CORREIA, E. F. G. **Modelagem hidrológica da bacia hidrográfica do rio Bengalas, Nova Friburgo, RJ, utilizando o potencial de geotecnologias na definição de áreas de risco de inundação**. Dissertação em Engenharia de Computação. Universidade do Estado do Rio de Janeiro. 2011.

GALELLI, S.; CASTELLETTI, A. Tree-based iterative input variable selection for hydrological Modeling. **Water Resources Research**, 49, 4295–4310, 2013. doi:10.1002/wrcr.20339.

Instituto Brasileiro de Geografia e Estatística-IBGE. Disponível em: http://cidades.ibge.gov.br/xtras /perfil.php?codmun=330340. Acesso em 25/10/2015.

IORGULESCU, I.; BEVEN, K. J. Nonparametric direct mapping of rainfall-runoff relationships: An alternative approach to data analysis and modeling?. **Water Resources Research**, 40, W08403, 2004. doi:10.1029/2004WR003094.

LONDHE, S.; CHARHATE, S. Comparison of data-driven modelling techniques for river flow forecasting. **Hydrological Sciences Journal**, 55:7, 1163-1174, 2010. doi: 10.1080/02626667.2010.512867.

NASH, J. E.; SUTCLIFFE J. V. River flow forecasting through conceptual models part I — A discussion of principles. **Journal of Hydrology**, 10 (3), 282–290, 1970.

QUINLAN, J. R. Learning with continuous classes. Proceedings **5th Australian Joint Conference on Artificial Intelligence**. World Scientific, Singapore, 343-348, 1992.

SATTARI, M. T.; PAL M.; , APAYDIN H,; OZTURK, F. M5 model tree application in daily river flow forecasting in Sohu Stream, Turkey. **Water Resources**, 40(3), 233-242, 2013.

SIEK, M.; SOLOMATINE, D. P. Tree-like machine learning models in hydrologic forecasting: optimality and expert knowledge. **Geophysical Research Abstracts**, v.9(p.09665), 2007.

SOLOMATINE, D. P.; XUE, Y. M5 Model Trees and Neural Networks: Application to Flood forecasting in the Upper Reach of the Huai River in China. **ASCE Journal of Hydrologic Engineering**, 9(6), pp. 491-501, 2004. doi 10.1061/(ASCE)1084-0699(2004)9:6(491)).

TSAI, C. C.; LU, M. C.; WEI, C. C. Decision Tree–Based Classifier Combined with Neural-Based Predictor for Water-Stage Forecasts in a River Basin During Typhoons: A Case Study in Taiwan. **Environmental Engineering Science**, 29(2), 2012. Doi: 10.1089/ees.2011.0210.