Chaos, Solitons and Fractals xxx (xxxx) xxx

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Indirect prediction system for variables that have gaps in their time series

Junio S. Bulhoes[a,*], Cristiane L. Martins[a], Marcia D. Oliveira[b], Debora F. Calheiros[b], Wesley P. Calixto[c,d]

[a] IFMT, Brazil
[b] Embrapa Pantanal, Brazil
[c] IFG, Brazil
[d] UFG, Brazil

ARTICLE INFO

Article history:
Received 25 April 2019
Revised 3 September 2019
Accepted 1 November 2019
Available online xxx

Keywords:
Filling gaps
System identification
Spectral analysis
Time series
Indirect predicting system
Correlated variables

ABSTRACT

Gaps in time series as well as the absence of such series make the implementation of prediction system difficult. This paper proposes a new methodology to fill gaps in time series that do not present fixed sampling rate. This paper also proposes the development of two forecast models for time series. The first model is based on autoregressive multilayer neural network that uses only the desired time series, while the second one is developed with multilayer neural network that uses pattern recognition in order to perform indirect predictions of a certain variable. Therefore, the second model does not need the variable time series to make predictions, but any time series that has correlation with the desired variable. The methodology is tested in limnological variables collected in the Paraguay River since 1987, and the results observed in each process are presented in order to validate the methodology of gap filling and forecast used.

1. INTRODUCTION

Studies, observations and conclusions about phenomena are possible thanks to past information obtained through collections. These collections usually have a fixed sampling rate and are stored in time series where the time position of each sample is preserved. If the time series that govern the behavior of specific phenomenon are available, it is possible to generate computational models that can predict the phenomenon behavior. Time series can be interpreted as a vector where each position represents the observed value for a particular system variable at a specific time point. Furthermore, the collections are ordered in the vector in a chronological way, which makes it easier to observe the behavior of the variable. [1] say that some temporal series derived from natural, physical or economic phenomena have characteristics that are repeated from time to time (seasonality), and these characteristics can or not be governed by linear patterns

In order to process and extract useful information from a time series, it is essential that all samples are equally spaced, the interval collected is sufficiently large and without gaps. Nevertheless, the presence of gaps in time series is unavoidable and affects the extraction of information from them [2]. [3,4] cite some reasons for the existence of gaps in time series such as: human faults, extreme weather, failure in measuring equipment and equipment failures that store these data. Several methods have been developed to fill gaps in time series, such as singular spectrum analysis [5], Kohonen self-organizing maps [6], multiple imputation method [7,8], multiple regression analysis [9,10] and parametric regression [11].

The singular spectrum analysis performed by [5] combines signal processing, system dynamics and multivariable statistics. This method presents good results for gap filling, but it requires that part of the existing signal present similar dynamics to what one wishes to complete and time series from periodic collections. [6] affirm that Kohonen self-organizing maps can fill gaps in data with high accuracy rate, depending on the amount of training data available. However, it is composed of very complex mathematical methods which end up requiring a high processing power. The multiple imputation method demonstrates efficiency for small data gaps, and it is not possible to reproduce with the same performance with the increase in the gap length, generating a limitation in the application of the method [8].

The multiple regression method used for filling gaps is a robust version of linear least squares prediction. According to [10], this method is extremely resistant to the presence of outliers, but

* Corresponding author.
E-mail address: junio.bulhoes@pdl.ifmt.edu.br (J.S. Bulhoes).

https://doi.org/10.1016/j.chaos.2019.109509
0960-0779/

it does not always demonstrate the best performance and does not present good results for few samples. Finally, parametric regression proposed by [11] is used when the gaps are generated by difficulties in collecting data in known systems, because it requires computational models to fill the gaps. Therefore, this method does not work in systems where the computational model does not exist.

Prediction systems are widely used in the most diverse areas in order to obtain future projections about a certain phenomenon, for instance: economic sciences to forecast stocks [12], engineering to forecast energy consumption [13], ecology to predict deforestation [14], and in medicine to predict the onset of diseases [15]. Time series are normally used in the implementation and execution of predicting systems. There are several prediction techniques that use time series such as: models based on artificial neural networks [12], models based on interconnected blocks [16] and NARMAX models [17]. Although the existence of time series facilitates the understanding of the phenomena which enable the development of prediction systems, not all variables have such time series, making it difficult to obtain prediction systems that perform good future projections.

[12] explain that models based on neural networks try to map the tendency and seasonality of the previous samples to generate model capable of predicting the next sample of the series. Besides, its performance is related to the quantity and quality of the samples submitted during the parameter estimation. According to [17], models that uses interconnected blocks are able to predict future values, but this kind of prediction system requires times series of other variables that have strong correlation with the analyzed variable, as well as time series of the analyzed variable itself. The NARMAX models used by [17], are predictive system options that use basically the same structure of the interconnected blocks models, and therefore can be substitutes to them.

The purpose of this article is to present a new methodology to fill gaps in time series through systems identification techniques and spectral analysis. In addition, this article proposes the use of these time series for the development of a prediction system for all variables. Some time series will be predicted through its own time series, and the other ones will be predicted indirectly through their correlation with the already predicted time series, proving the non necessity of the existence of time series of each variable in order to create prediction system. Finally, time series of physical and chemical variables of the Paraguay River, which have gaps and do not have fixed sampling rate, were made available by Embrapa Pantanal and the Brazilian Navy in order to validate the proposed methodology.

This paper is organized as follows: Section 2 describes the proposed methodology for filling gaps in time series and the prediction system. Section 3 presents the results obtained with the methodology proposed both for filling gaps and for the prediction system. Finally, Section 4 closes this article pondering the findings raised in this paper.

2. Methodology

This section presents the proposed methodology for filling gaps using spectral analysis and systems identification. It also presents an indirect prediction system and how to validate the proposed methodology. Fig. 1 presents the flowchart of the proposed methodology that contemplates all the steps to fill the gaps and development of the prediction system based on ANN.

2.1. Gap filling

There are several problems in time series that impair its use as a source of information for pattern recognition. There are two recurring problems in time series from collections. The first problem is the non-existence of a fixed sampling period and the second problem is the existence of collection failures in certain periods of time, which generates gaps in the time series. The proposed methodology to fill gaps in the time series and standardization of the sampling frequency is composed of the following steps: i) preprocessing of the data, ii) interpolation of the collected data and spectral analysis, iii) identification of the system and iv) validation of the model.

To perform analyzes of several time series with no fixed sampling rate and with gaps inside it using system identification, it is necessary to have at least one time series with fixed sampling rate, without gaps and with some relation with the other time series studied. when the time series with fixed collection frequency is found, this is called the input time series S_i, which is the basis for defining the fixed sampling frequency for the other time series, in order to standardize the existing data. The regularization of the spaces between the samples assists in locating the regions where the gaps are.

By analyzing the time series it is possible to find the fundamental frequency F_f of each one. Given the fundamental frequencies of the time series analyzed, it is possible to identify its gaps. The gaps are divided into two groups: Group I consisting of gaps with size $< 25\%$ of the fundamental period T_f of the time series; and ii) Group II consisting of gaps with gaps of $\geq 25\%$ of the T_f of the time series. In Group I, the gaps can be filled via the interpolation process. In Group II, the gaps can not be filled via interpolation process, since in this case, the interpolation method interferes in the dynamics of the analyzed variable. Fig. 2 illustrates the division of Group I and Group II.

In Fig. 2, the curve represents the values assumed by a given variable over time and the points represent the collected samples over the same time interval. It can be seen from Fig. 2 that the gaps from Group I (time interval represented by yellow color) can be filled by interpolation because of the proximity between the collections, while Gaps from Group II can not be filled by interpolation because it will not represent the dynamics of the variable.

After identifying each group of gaps in the time series, the linear interpolation process is performed. In this case, the interpolation does not change the dynamics of the series, since it only closes the gaps of Group I and standardizes the frequency in S_i. To close the gaps from Group II, it is proposed to carry out a spectral analysis of the interpolated time series. Since linear interpolation is used, there is no loss of time series dynamics and the interpolated time series from Group I can be considered as the original time series without the gaps.

Therefore, the fast Fourier transform (FFT) will be used to represent this time series interpolated in the frequency domain. In the frequency domain, the fundamental frequency F_f and all phases θ and frequencies ψ (multiples of F_f) with relevant amplitudes can be observed and extracted. With the values of F_f, θ and ψ, it is possible, using the inverse Fourier transform, to generate a new time series using the relevant parameters of the original time series. The purpose is to create the new time series based on the original series and replaces the regions where the gaps from Group II are located. Thus, with the Group II gaps closed through spectral analysis, all time series are with the same sampling frequency as the time series S_i and without a gaps.

In order to refine the results obtained with the spectral analysis, constructing the time series without gaps and with fixed sampling rate, a system identification method is going to be used. This method perform a fine adjustment and identifies possible errors obtained in the spectral analysis. In this way, the system identification is used to: i) construct the mathematical model that represents the time series of each analyzed variable and ii) verify the response of the system to the regions where they locate the

Please cite this article as: J.S. Bulhoes, C.L. Martins and M.D. Oliveira et al., Indirect prediction system for variables that have gaps in their time series, Chaos, Solitons and Fractals, https://doi.org/10.1016/j.chaos.2019.109509

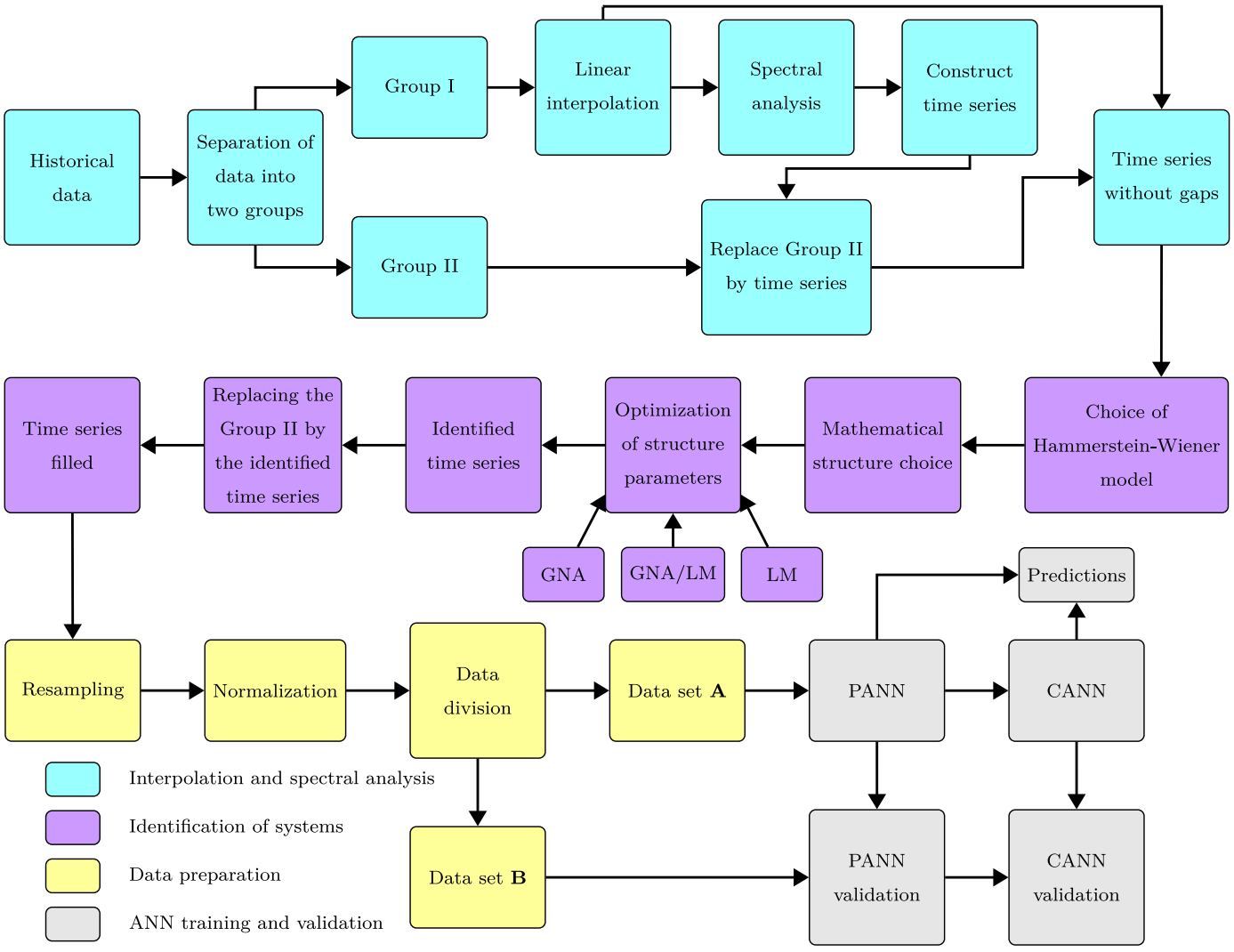

Fig. 1. Flowchart of the proposed methodology for gap filling in time series and prediction system for flooded areas.

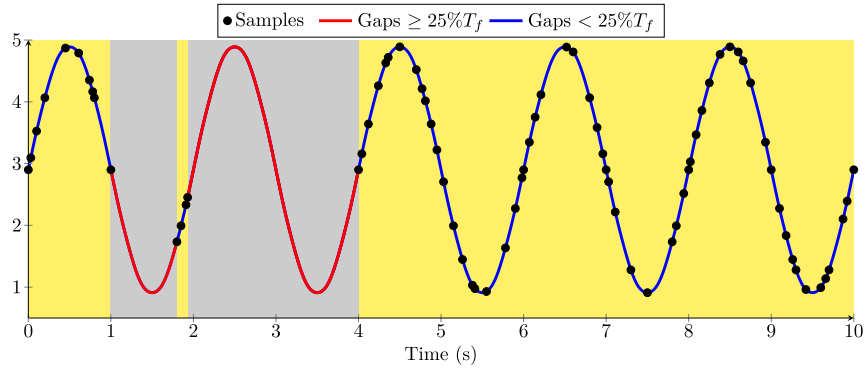

Fig. 2. Illustration of the method used for data division into Group I and Group II based on the fundamental period of the signal.

gaps from Group II. The system identification model used is the Hammerstein–Wiener block-oriented model because it can construct a mathematical structure capable of mapping both the nonlinearities in the inputs and the nonlinearities of the outputs.

In order to obtain the appropriate structures, the parameters from Hammerstein-Wiener block-oriented model should be optimized. Different methods of nonlinear optimization are implemented, aiming to find the mathematical structure that best represents the time series. The optimization methods used to estimate the structure parameters are: Gauss–Newton adaptive search method (GNA), Levenberg–Marquardt least squares method (LM), and the combination of these two ones GNA/LM. These methods were chosen because of their performance in identifying nonlinear systems.

The amount of input regressors R_i, as well as the quantity of output regressors R_o are variables that define the best structure for

Please cite this article as: J.S. Bulhoes, C.L. Martins and M.D. Oliveira et al., Indirect prediction system for variables that have gaps in their time series, Chaos, Solitons and Fractals, https://doi.org/10.1016/j.chaos.2019.109509

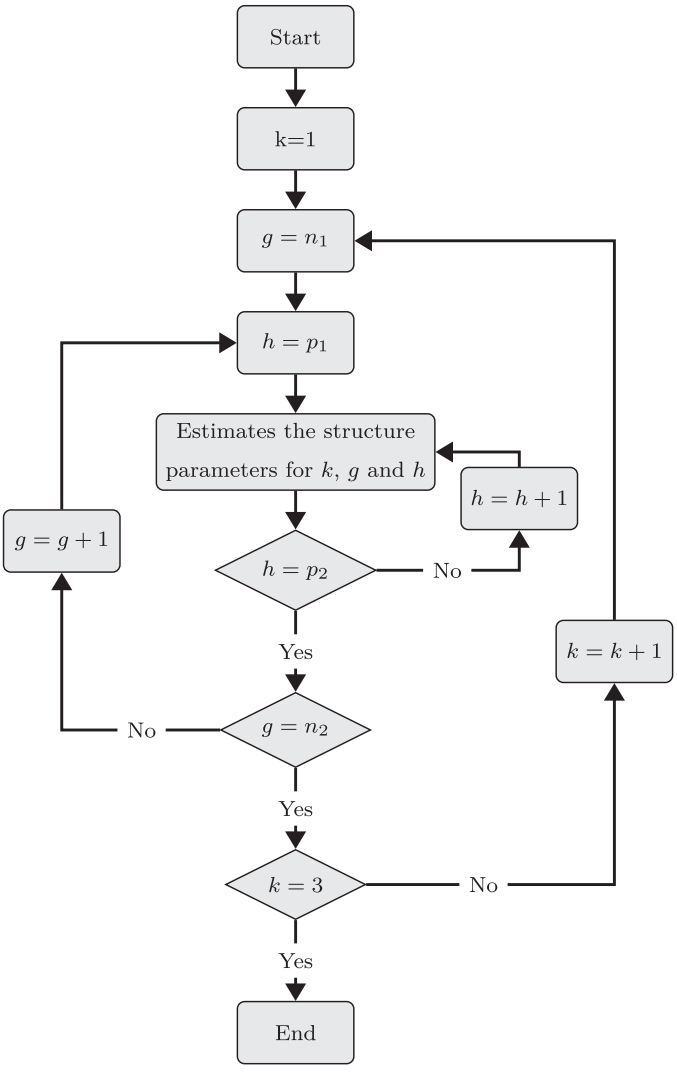

Fig. 3. Flowchart for estimating Hammerstein-Wiener model structures.

the Hammerstein-Wiener model. The amount of R_i used is limited in the closed range of n_1 to n_2, where n_1 and n_2 are the minimum and maximum numbers of R_i, respectively. The amount of R_o is limited in the closed range of p_1 to p_2, where p_1 and p_2 are the minimum and maximum numbers of R_o, respectively.

Several structures are generated for each time series (each series represents a variable), and the structure that best fits the data is chosen to represent the time series and to fine tune between the original time series and the time series constructed by the spectral analysis. The flowchart of Fig. 3 illustrates the proposal to estimate the parameters of the structure that represents the original time series. The k variable represents the optimization methods used, g is the amount of input regressors, and h is the amount of output regressors.

To estimate the structures according to Fig. 3, start with k assuming value 1, g assuming value n_1 and h assuming value p_1. After setting the values of k, g and h, the Hammerstein-Wiener model is generated with g input regressors, h output regressors and using the optimization method chosen in the parameter k. The three loop guarantee the creation and estimation of all combinations of possible structures with the parameters k, g and h within their respective intervals.

2.2. Prediction system using artificial neural networks

Having the time series with the gaps filled through the spectral analysis and treated through the system identification process, they can be used in the prediction system. However, if there is an excessive number of data samples in the time series, it is necessary to resample them in order to reduce the number of samples. This procedure is necessary to reduce the computational effort in the development of the forecasting system. Fig. 4 illustrates the process of resampling the time series, where the dots in blue represent the original time series. The resampled time series contains only the red dots, equally spaced and with a sampling period higher than the original series and obeying Nyquist-Shannon sampling theorem.

After the time series are resampled, the normalization process is required. This process standardizes the values of all the variables (time series) in a certain predefined interval. This amplifies the ability of the forecasting system to generalize. The expression (1) is used in order to perform the linear normalization process, where d_1 and d_2 are adjusted values in order to prevent ANN from obtaining negative values. The ANN used is multilayer Perceptron (MLP), composed of two configurations in cascade: prediction artificial neural network (PANN) based on information from the time series and classification artificial neural network (CANN) based on based on other time series patterns that influence the desired variable.

$$\tilde{x}_{(j)} = \frac{[x_{(j)} - x_{(j_{min})}](d_2 - d_1)}{x_{(j_{max})} - x_{(j_{min})}} \tag{1}$$

When analyzing variables (time series) that are correlated, there is a need to identify which variables influence the other variables. The PANN performs forecast only in the σ time series, where σ are the time series that represent the variables that exert some influence on the others. The CANN performs the forecast using the standards presented in the input layer, and thus, the σ time series forecast by PANN are the input presented to CANN. In this way, the other time series that receive influence of the *sigma* time series are predicted using the CANN.

For the training and validation of the PANN and CANN, the time series will be divided into two distinct sets: data set **A** and data set **B**, where the data set **A** is composed of the values of all time series from t_0 to t_1, and data set **B** is composed of the values of all time series from $t_1 + 1$ to t_2. The t_0 value represents the fist position of the vector t, while t_1 is an arbitrary value between t_0 and t_2. If t_2 is not the last position of the vector t, the remaining value of this vector are allocated in data set **A**.

2.2.1. Prediction artificial neural network

In the PANN is used autoregressive model (feedback). The PANN has a feedback structure whose main objective is to provide the inputs, past values of the analyzed variable (regressors). The PANN uses past information of the variable $x_i(t)$ in order to estimate the next value reached ($x_i(t + 1)$), where t represents the position of a sample in the time series, and $i = 1, 2, \cdots, \sigma$ are the variables to be predicted. Fig. 5 illustrates the PANN architecture.

In Fig. 5, the amount of PANN input is represented by r_i and each input is composed of a regressor. These regressors may or may not contain the delay d, for example: $\tilde{x}_i(t - d)$, $\tilde{x}_i(t - d - 1)$, $\tilde{x}_i(t - d - 2)$, $\ldots$, $\tilde{x}_i(t - d - r_i + 1)$, where the presence of the delay influences the PANN response and assists in the creation of long-term forecasts.

To train PANN it is necessary to first find the topology (PANN geometry) that best suits to solve the problem. This topology is found through several empirical tests varying the number of neurons per layer, number of hidden layers, and number of input regressors r_i required. By varying these parameters the performance

Please cite this article as: J.S. Bulhoes, C.L. Martins and M.D. Oliveira et al., Indirect prediction system for variables that have gaps in their time series, Chaos, Solitons and Fractals, https://doi.org/10.1016/j.chaos.2019.109509

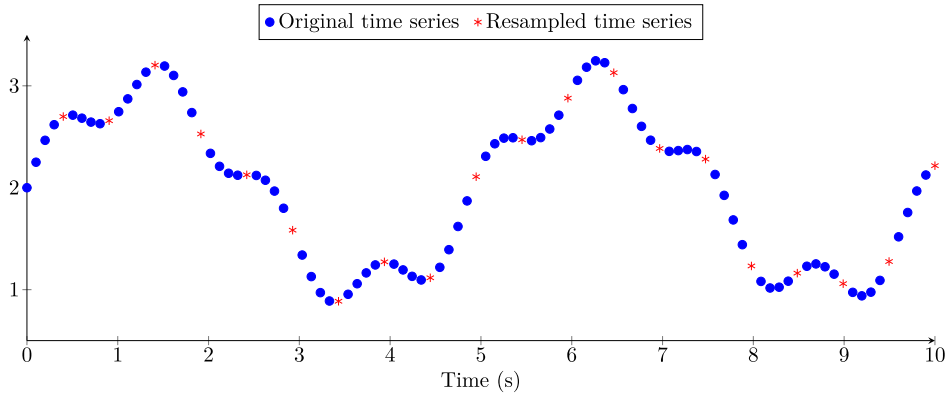

Fig. 4. Illustration of the method used to resample the time series of each variable, where the blue dots represent the existing samples and the red dots represent the resampled signal. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

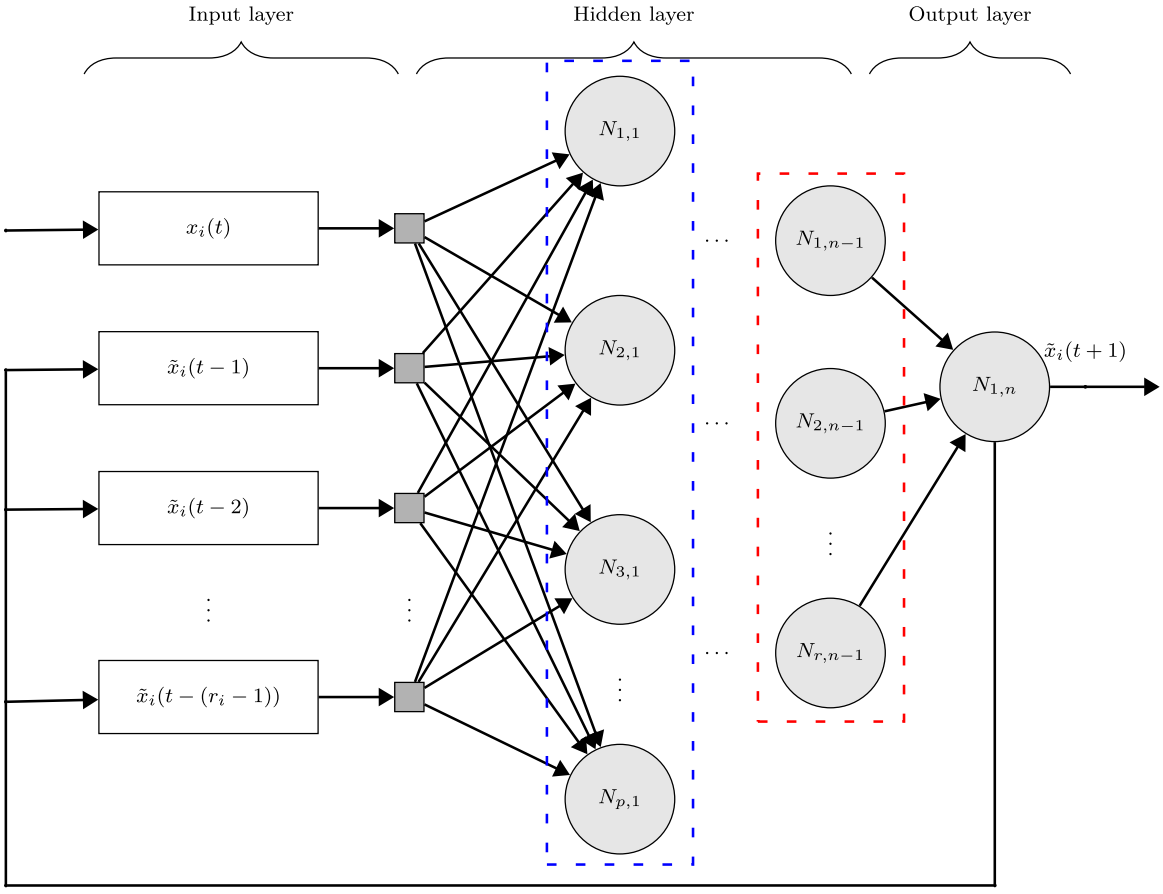

Fig. 5. Prediction artificial neural network (PANN) based on autoregressive model.

of PANN is observed in order to find the best configuration. For σ time series (variables) it is necessary σ PANN, where the topology may or may not be the same, however the weights and bias are different since each time series has its own characteristics. In this way, *sigma* PANN creation, training and testing are something that demands time and computational effort.

The PANN training uses the supervised backpropagation algorithm in order to estimate the weights and bias. The data set **A** is randomly subdivided into two subsets: training subset $\mathbf{A_1}$ and validation subset $\mathbf{A_2}$. The subset $\mathbf{A_1}$ is reserved for PANN training while the subset $\mathbf{A_2}$ is used as one of the PANN stopping criteria because it can indicate the ability to recognize the variable's patterns.

2.2.2. Classification artificial neural network

The goal of the CANN is to predict the ρ time series that are influenced by the σ time series already predicted. Thus, the ρ time series forecasts are receiving the standards of the σ time series. In conclusion, the inclusion of the CANN in cascade with the PANN, the prediction system reduces the computational effort due to the reduction of the number of PANN and it inserts the characteristics of the independent variables in the prediction CANN = f(PANN).

The CANN is also a multilayer Perceptron where its inputs are the outputs of PANN. The outputs of the CANN are the predictions of the time series. In addition, The best network topology for CANN is found by empirically varying the parameters: number of neurons

Please cite this article as: J.S. Bulhoes, C.L. Martins and M.D. Oliveira et al., Indirect prediction system for variables that have gaps in their time series, Chaos, Solitons and Fractals, https://doi.org/10.1016/j.chaos.2019.109509

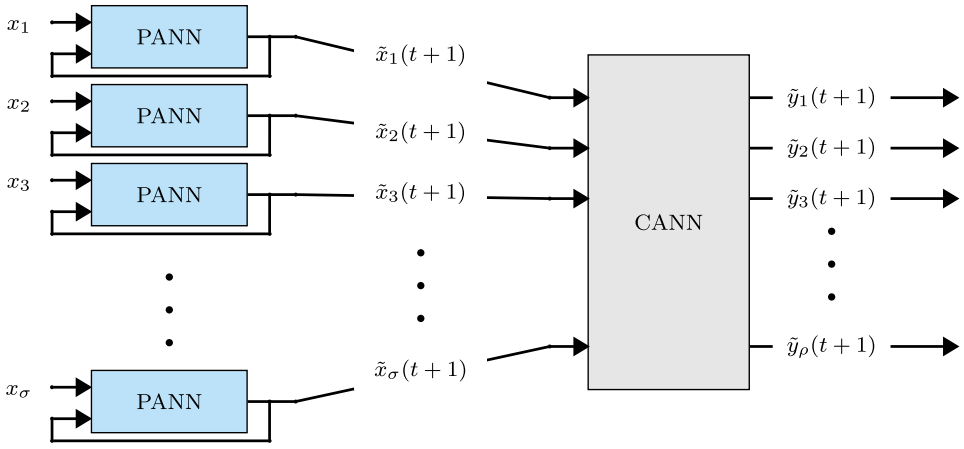

Fig. 6. Classification artificial neural network (CANN) used to perform indirect prediction form other variables.

per layer and number of hidden layers. Fig. 6 illustrates the CANN model, where $\tilde{y}_i(t+1)$ are the predicted values.

In Fig. 6, the input $x_1, x_2, \ldots, x_\sigma$ are the σ time series (variables that exert influence on the others), which are predicted by the PANN. The outputs $\tilde{y}_1(t+1), \tilde{y}_2(t+1), \ldots, \tilde{y}_\rho(t+1)$ are the ρ time series predicted indirectly by CANN (variables that are influenced by σ time series). The CANN training also uses the Levenberg–Marquardt supervised backpropagation algorithm to find the weights and biases.

2.3. Model validation

In order to validate the method of filling in the gaps, samples are taken for later comparison with those obtained by the identified model. The relation between the collected samples and the values found by the proposed method gives the percentage of the error committed. The validation of the forecasting system composed of PANN, Fig. 5 and CANN, Fig. 6 uses the data set **B** which consists of the samples between t_1+1 and t_2 from all time series. The expression (2) calculates the mean square error between the output of PANN/CANN and the data set **B**.

$$E_s = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_{r(i)}^2} \tag{2}$$

3. Results

This section presents the results obtained from the proposed methodology and is structured in: presentation of the database, application of the proposed method to fill gaps in time series and application of time series in the development of the prediction system.

3.1. Database

In this work two databases were used: i) Paraguay River level database provided by the VI Naval District of the Brazilian Navy and physical-chemical database of water quality of the Paraguay River provided by Embrapa Pantanal located in the city of Corumba-Brazil. The analysis period was from October 1st, 1987 until May 10th, 2018, totaling 30 years of collection of the physical and chemical variables of the water of the largest river in the Pantanal biome. Pantanal is one of the largest areas subject to annual flood of the world and its diversity attracts researchers from all over the world. Fig. 7 illustrates the point P_1 where the water collections are made. The Brazilian Navy located in the city of Ladario monitors the level of the Paraguay River daily since 1900. This variable has a fixed sampling period and no gaps.

All collections were carried out following protocols such as the same time and place (P_1) of sampling, same methods of collection and analysis, which guarantees high reliability of the data. The variables collected by Embrapa Pantanal and marine used in this work are: i) Paraguay River level (L_R), ii) water temperature (T_W), iii) dissolved oxygen (O_D), iv) potential of hydrogen (pH), v) electrical conductivity of water (C_E), vi) free carbon dioxide (D_C), vii) alkalinity (A_L), viii) total dissolved nitrogen (N_D), ix) inorganic suspended matter (M_I), x) transparency (T_R), xi) turbidity (T_U) and xii the flooded area (A_F). The A_F variable is used only in the prediction system. The time series formed by the chemical and physical variables provided by Embrapa Pantanal do not contain a fixed frequency, and it is necessary to use the process of filling in gaps so that the data can be used in the development of the prediction system to assist in the monitoring of water quality of the Paraguay River.

3.2. Gap filling

The L_R time series was the only one that did not present gaps, with sampling period T_s and one sample per day, totaling 10,304 samples equally spaced. As the L_R time series contains T_s fixed and has relation with the other time series, it could be used in the system identification process and for this reason, it is called the input time series S_i. Observing the existence of gaps and finding S_i, the next step was to find the fundamental frequency of the all time series. Since all the time series worked has its dynamics controlled by the flood pulse, the fundamental period T_f of the other time series has approximately the same value of S_i, being its value approximately 365 days. This is because the flood cycle is annual due to several factors such as rainy season, relief among others.

After finding the fundamental frequency of all time series analyzed, it is possible to locate the gaps. The sampling frequency F_s of all time series should be set to one sample per day, which totals 10,304 samples, equal to the S_i used. Table 1 provides the information collected for each of the time series analyzed, where Q_s is the quantity of existing samples of each variable.

The gaps were divided into two groups: Group I, composed of gaps $<25\%$ of T_f of the time series, and ii) Group II, consisting of gaps $\geq 25\%$ of T_f of the time series. Since the T_f of all time series is equal to 365 days, 25% of this value is equivalent to approximately 91 days. Therefore, Group I contains gaps up to 91 days and Group II has gaps longer than 91 days. After identifying each group of gaps in the time series, the linear interpolation in the Group I was

Please cite this article as: J.S. Bulhoes, C.L. Martins and M.D. Oliveira et al., Indirect prediction system for variables that have gaps in their time series, Chaos, Solitons and Fractals, https://doi.org/10.1016/j.chaos.2019.109509

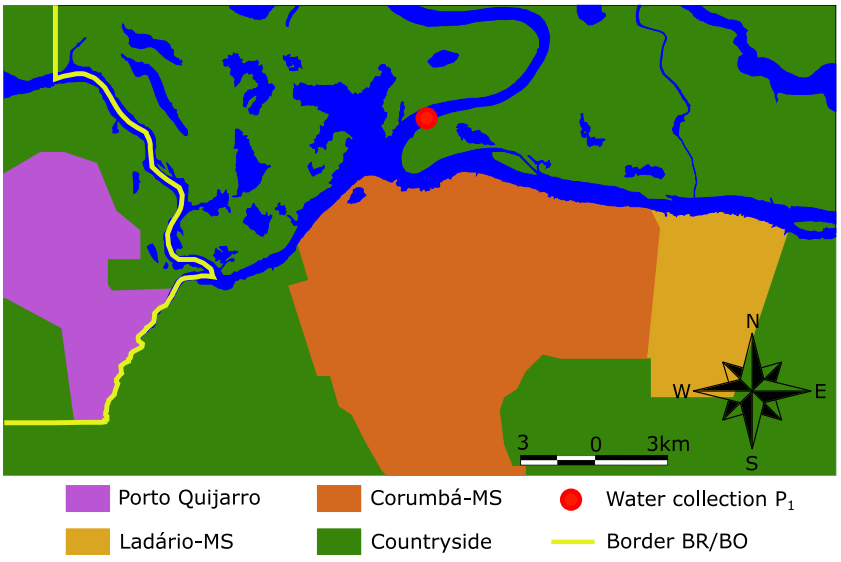

Fig. 7. Illustration of the region P_1, where the water samples were collected in the Paraguay River.

Table 1
Gap analysis for the used limnological variables.

Var.	Q_s	Largest gap [days]	Gap $\geq 25\%T_f$	Gap $< 25\%T_f$
L_R	10304	–	–	–
T_W	323	1199	2	320
O_D	331	1045	3	327
pH	344	685	4	339
C_E	349	303	6	342
D_C	326	773	4	321
A_L	339	685	5	333
N_D	261	2011	6	354
M_I	241	482	1	339
T_R	340	685	5	334
T_U	228	369	2	325

Table 2
Spectral analysis of the used time series.

Var.	A_f	F_f [10^{-3}/day]	θ_f [°]	M	B
T_W	3.35	2.76	−95.42	1	27.77
O_D	1.64	2.76	−6.14	1	4.32
pH	0.22	2.51	108.02	3	6.37
C_E	4.57	2.51	−54.30	5	49.30
D_C	25.93	2.74	54.27	10	30.81
A_L	55.87	2.76	120.05	5	412.05
N_D	46.60	2.77	−73.75	1	83.00
M_I	13.29	2.67	−151.09	1	23.35
T_R	29.56	2.76	120.07	1	61.64
T_U	22.47	2.66	−154.08	1	31.13

performed to standardize the F_s of the time series in 1 sample per day. This linear interpolation is given by (3), where the points $[t_i, x(t_i)]$ and $[t_f, x(t_f)]$ represent the samples before and after the interpolated gap, respectively.

$$\check{x}_i(t) = \frac{[x(t_f) - x(t_i)]t}{t_f - t_i} + x(t_i) - \frac{[x(t_f) - x(t_i)]t_i}{t_f - t_i} \tag{3}$$

The largest collection period of Group I of each time series is used to perform the spectral analysis. This interval is interpolated and subsequently represented in the frequency domain, where F_f and all phases θ, frequencies ψ, and the displacement of the y-axis B could be analyzed in order to generate a new signal. The Table 2 provides the results for all variables, where A_f is the amplitude of fundamental frequency, F_f is the fundamental frequency, θ_f is the phase of fundamental frequency, M is the number of ψ analyzed frequencies, and B is the displacement of the signal in the axis of the ordinates. The variables pH, C_E, D_C and A_L used more ψ, since it was observed that some relevant characteristics of these variables were not observed with low values of ψ.

The extracted frequencies were used to construct new time series that have the characteristics of the analyzed time series. Fig. 8 illustrates the separation of the groups, the linear interpolation of the Group I (blue curve), and the reconstructed signal via spectral analysis for the filling gaps on the Group II (green curve) for the T_W variable. The red curve is presented to show how the linear interpolation does not represent the real dynamics of the water temperature variable in the Group II.

With all the time series filled by linear interpolation and spectral analysis, the next step was to identify the systems using a Hammerstein-Wiener block-oriented model. This identification promotes the search for a signal closer to the real one (fine tuning) to replace the signal inserted in the gaps from Group II. Fig. 9 shows the time series of Paraguay River level measured in the Ladario ruler, which was chosen earlier as Si due to its relation to the other time series.

The structure of the Hammerstein–Wiener model chosen was the sigmoidal function due to its characteristic of being limited between two points, which provides a desired representation for the variables with periodic dynamics. The parameters estimation for each structure was performed with the implementation of three different optimization methods: Gauss-Newton adaptive search method (GNA), Levenberg-Marquardt least squares method (LM), and the combination of these two ones GNA/LM.

The search for the best structure for the Hammerstein–Wiener model was performed with different amounts of input and output regressors. The number of input regressors was limited in the closed range of n_1 to n_2, adopting $n_1 = 60$ and $n_2 = 90$, totaling 31 different possibilities. The amount of output regressors was limited to the closed range of p_1 a p_2, where $p_1 = 2$ and $p_2 = 11$, totalizing ten distinct possibilities. In the 31 variations of input regressors, ten variations of output regressors and the three methods of parameter optimization, 930 structures were estimated for each of the ten time series analyzed, using all possible combinations as shown in the flowchart of Fig. 3.

Please cite this article as: J.S. Bulhoes, C.L. Martins and M.D. Oliveira et al., Indirect prediction system for variables that have gaps in their time series, Chaos, Solitons and Fractals, https://doi.org/10.1016/j.chaos.2019.109509

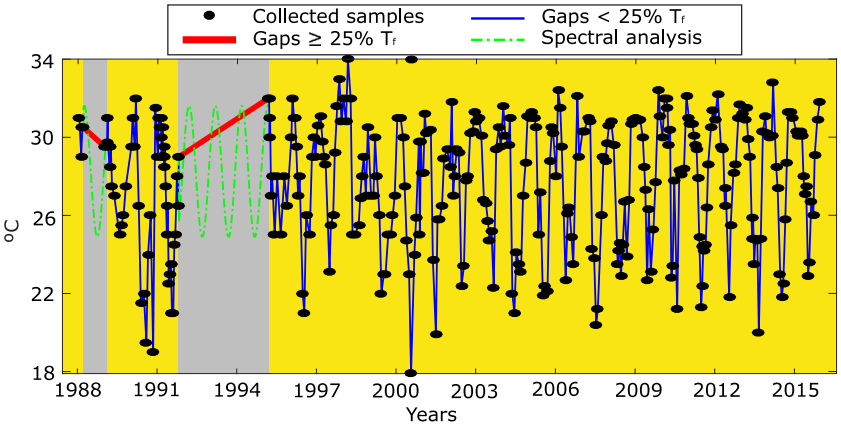

Fig. 8. Spectral analysis applied to the time series of the water temperature variable where the black dots are the collected samples, the red curve are the interpolated signal on Group II, the blue curve represents the interpolated signal on Group I, and the green curve is the result of the spectral analysis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

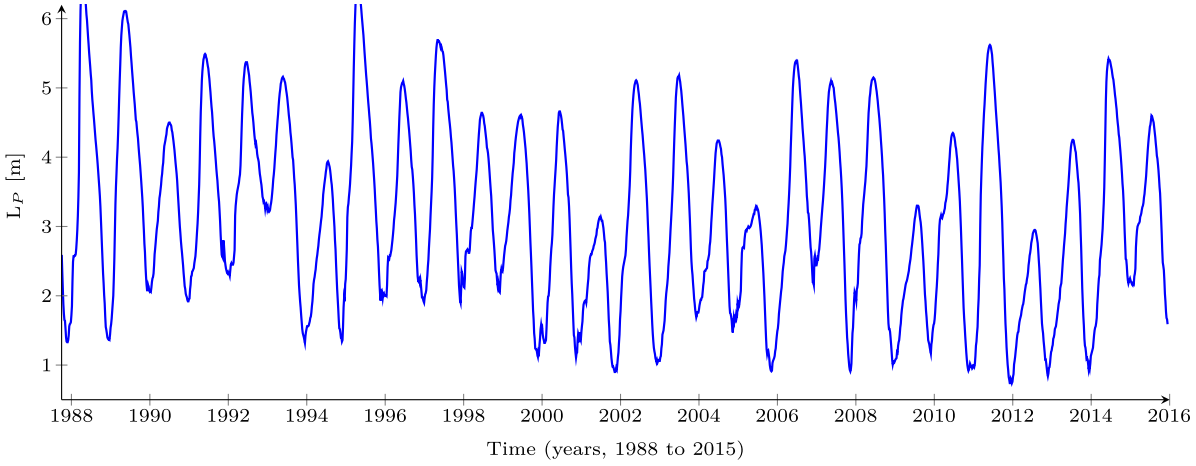

Fig. 9. Time series of Paraguay River Level (L_R) measured by Brazilian Navy in Ladario city.

3.3. Gap filling validation

The validation of the proposed method for filling the gaps uses two original samples of each series, and the mean squared error E_s of all collected samples given by (2). For each analyzed series, two structures (E_1 and E_2) of the 930 structures that obtained the best results were separated for analysis and are shown in Fig. 10 through Fig. 14, where the *Ref* curve represents the reference time series, the curves E_1 and E_2 represent the Hammerstein-Wiener time series for each structure. Finally, the validation and choice of the best structure was conditioned to its performance in approaching the values obtained to the two original values. The local error E_l is defined by (4), where m_l is the value returned by the mathematical model and s_l is the sample value taken from the initial time series in order to perform this validation. The parameters $s_{\max}$ and $s_{\min}$ are the maximum and minimum values present in the time series s, respectively. In addition, the samples taken for this validation were the last sample before the largest gap and the first sample after it. Only two samples were used because there were few samples for each variable.

$$E_l = \frac{|m_l - s_l|}{s_{max} - s_{min}} \tag{4}$$

It was observed that all variables presented the same optimization method for E_1 and E_2, which is probably an inherent characteristic of each variable to adapt better to a given method. In addition, the difference in the amount of input regressors presented by the E_1 and E_2 structures was small for most of the series analyzed. For the output regressors, six of the ten time series mapped, presented their two best structures with the same amount of regressors, which indicates precision in the system identification process, since both structures needed the same information to be able to map through S_i the desired variable.Tables 3–7 show the amount of input regressors R_i and the amount of output regressors R_o for the structures E_1 and E_2, as well as the optimization method M_o used in the parameter estimation for every analyzed variable.

Table 3
Analysis of the parameters obtained for the water temperature and dissolved oxygen.

Parameter	T_W		O_D	
	E_1	E_2	E_1	E_2
R_i	84	84	88	89
R_o	8	7	2	2
M_o	LM	LM	GNA	GNA
E_l	8.97	7.71	25.50	23.85
E_s	3.70	3.88	4.50	4.93
E_a	0.60 °C	0.62 °C	0.41 mg/L	0.45 mg/L

For T_W variable, E_2 was chosen as the best structure, since it had the lowest E_l as set out in Table 3. Fig. 10(a) shows the time series obtained by E_1 and E_2 for this variable. The O_D variable presented higher values of E_l among all variables, reaching 25.50% for E_1, and E_s was restricted to 4.50%. In view of this information E_2

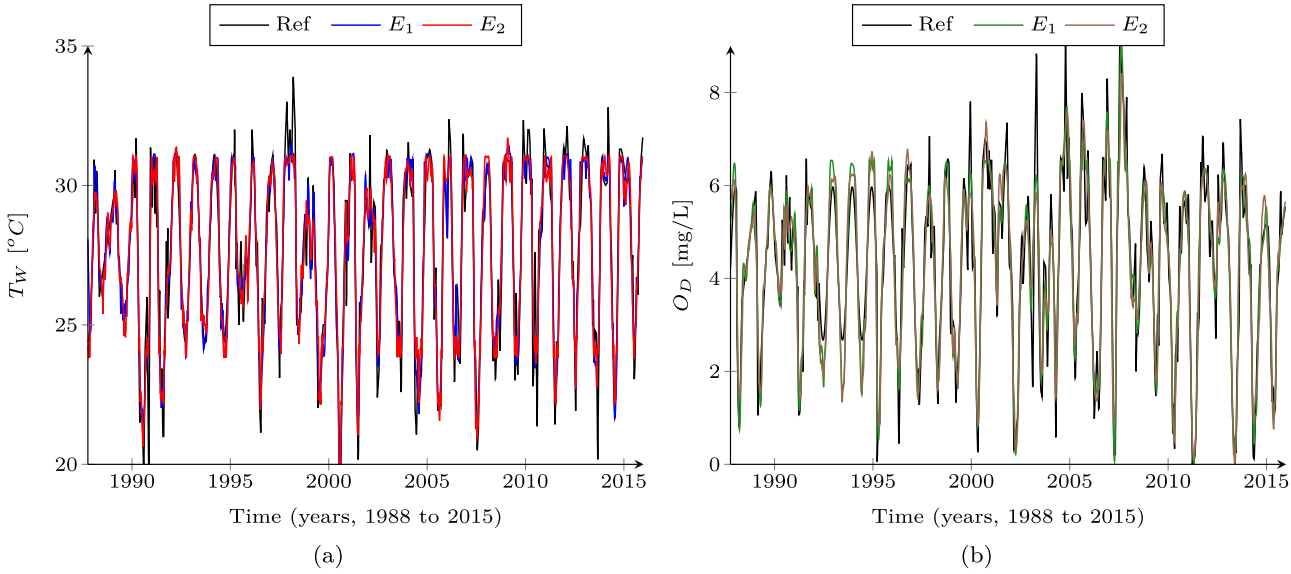

Fig. 10. Estimated Hammerstein-Wiener model for: (a) water temperature and (b) dissolved oxygen.

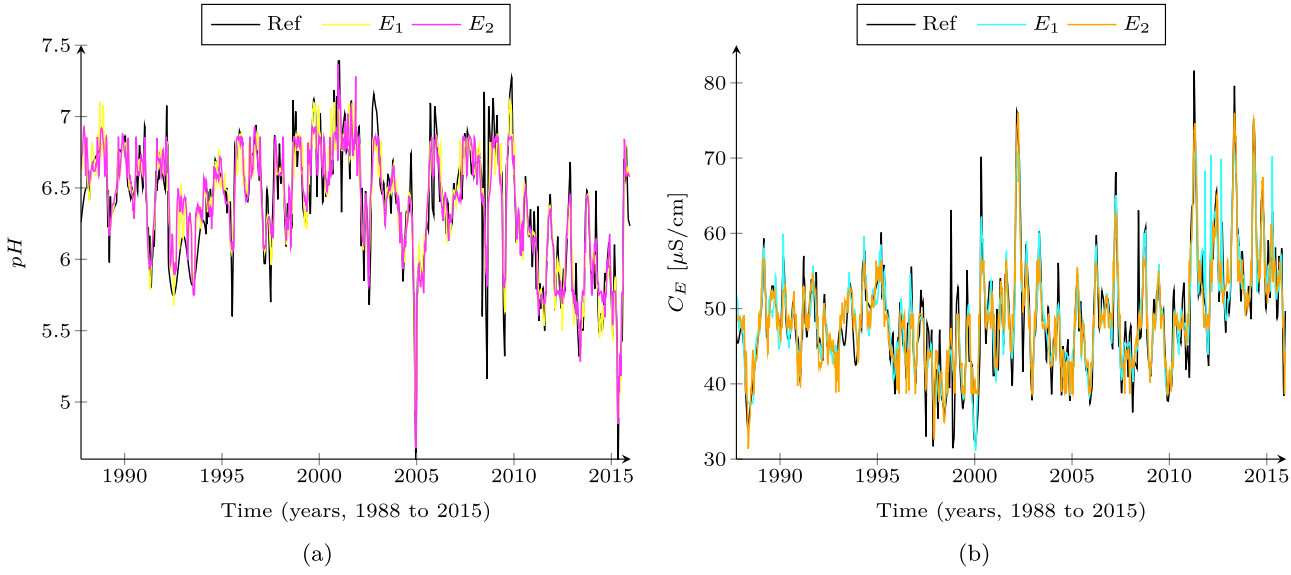

Fig. 11. Estimated Hammerstein–Wiener model for: (a) potential of hydrogen and (b) electrical conductivity of water.

was chosen, since its E_l was almost 2% below E_1, as presented in Table 3. Fig. 10(b) shows the time series obtained by E_1 and E_2 for the O_D variable.

The structures E_1 and E_2 of the *pH* presented satisfactory results in the analyzes, however E_2 presented E_s value higher than E_1. In addition, the E_l of E_1 was lower than that of E_2, and thus, E_1 was chosen to represent the *pH* variable. The Fig. 11(a) presents the time series obtained by E_1 and E_2 for this variable. The E_2 structure of the C_E variable presented $E_l = 6.90\%$ as presented in Table 4 and thus, E_1 which presented E_l 1% less than E_2 was chosen in order to represent the C_E variable. Fig. 11(b) shows the time series obtained by E_1 and E_2 for the C_E variable.

The E_1 and E_2 structures of the D_C variable obtained E_s less than 2.8%. This suggests that the spectral analysis would be a reasonable approximation to fill in the existing gaps. The E_1 was chosen as it reached the lowest E_l, as shown in Table 5. Fig. 12(a) shows the time series obtained by E_1 and E_2 for the D_C variable. The E_2 presented E_l smaller than E_1 and therefore was chosen to represent the A_L variable. The Table 5 provides the values obtained for the A_L variable, and Fig. 12(b) presents the time series obtained by E_1 and E_2.

Table 4
Analysis of the parameters obtained for the potential of hydrogen and electrical conductivity of water.

Parameter	*pH*		C_E	
	E_1	E_2	E_1	E_2
R_i	85	76	88	75
R_o	8	8	10	8
M_o	GNA	GNA	GNA/LM	GNA/LM
E_l	2.18	3.88	5.25	6.90
E_s	4.11	4.59	3.94	4.55
E_a	0.12	0.13	2.03 µS/cm	2.35 µS/cm

The structure chosen for N_D was E_1, since the performance of both structures for E_l and E_s were similar and E_1 has fewer input regressors, being therefore model with lower computational cost as observed in Table 6. Fig. 13(a) shows the time series obtained

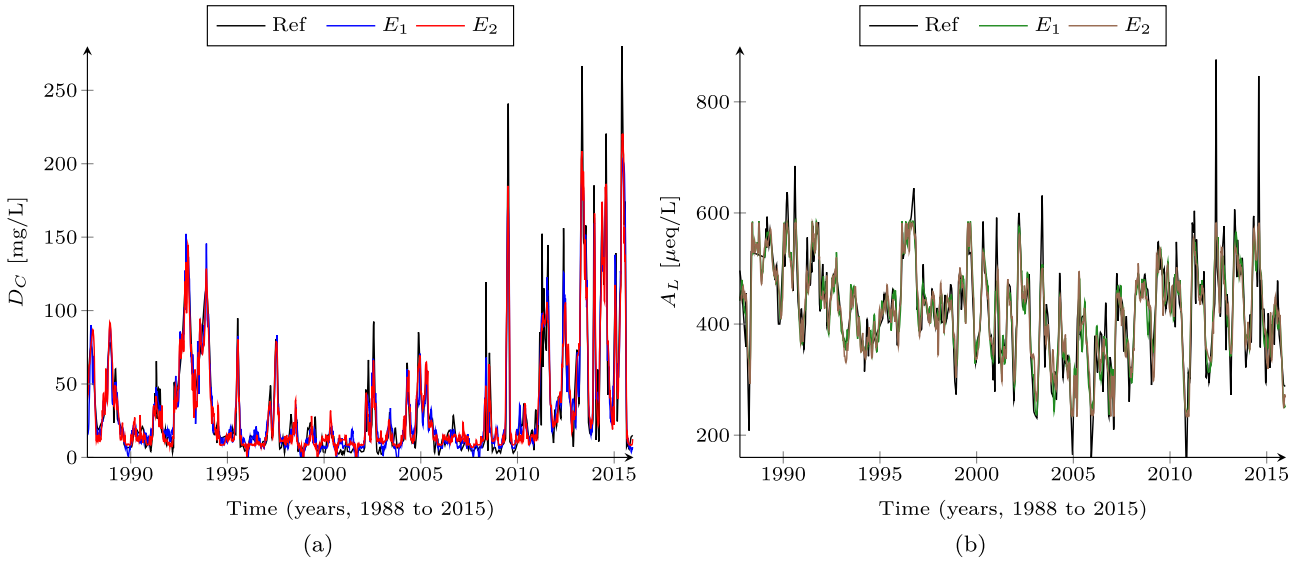

Fig. 12. Estimated Hammerstein–Wiener model for: (a) free carbon dioxide and (b) alkalinity.

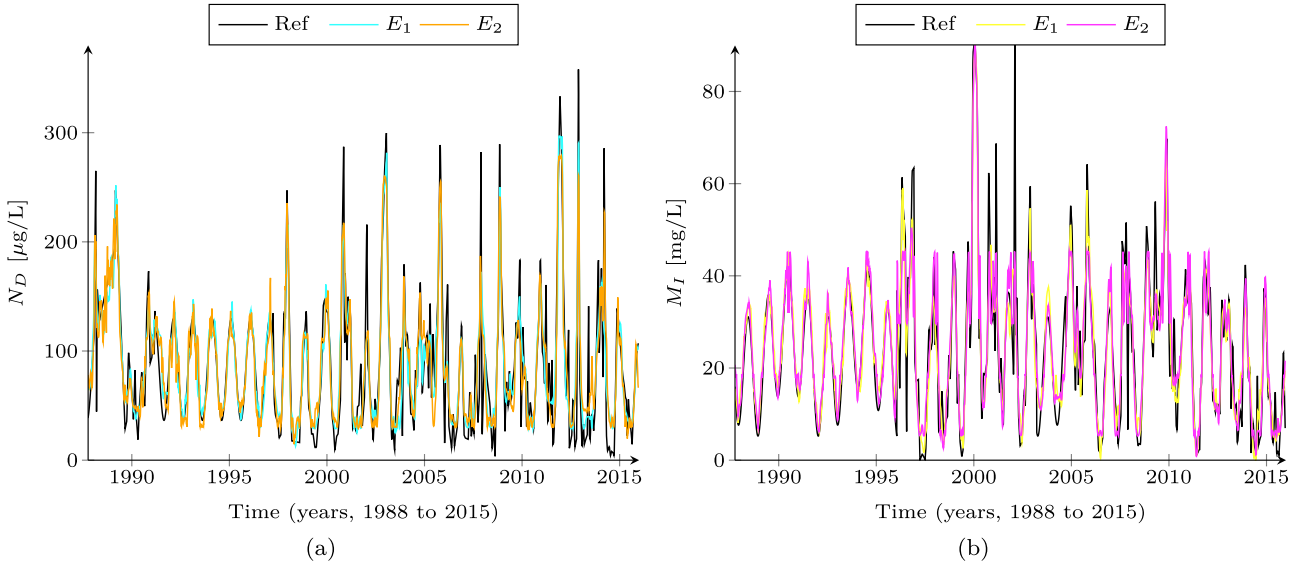

Fig. 13. Estimated Hammerstein–Wiener model for: (a) total dissolved nitrogen and (b) inorganic suspended matter.

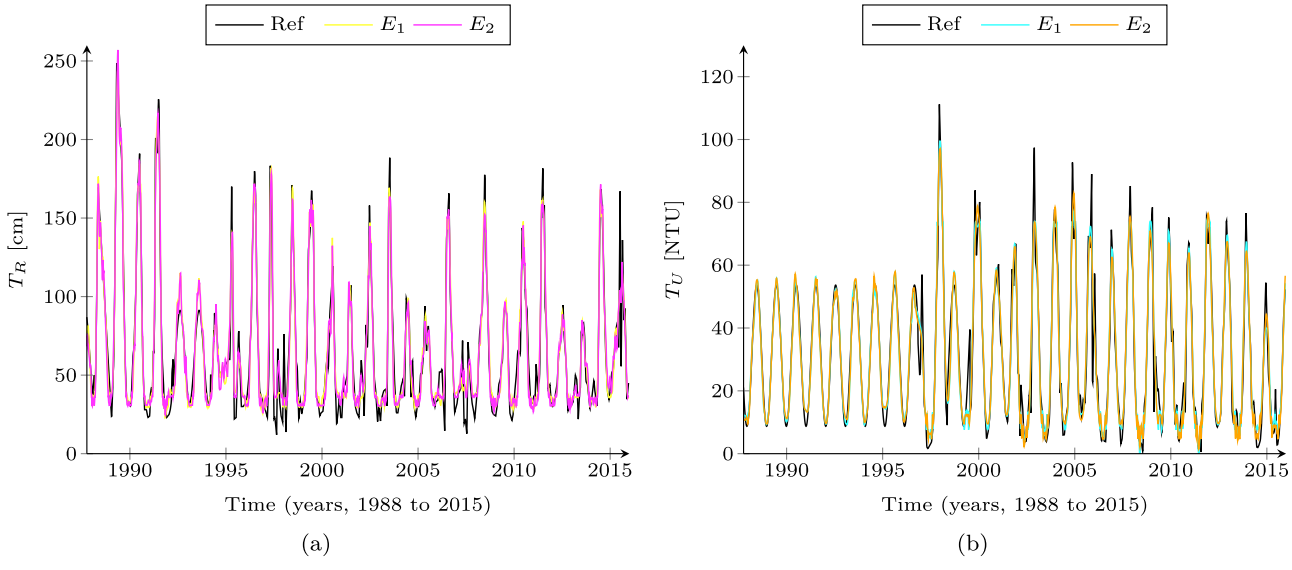

Fig. 14. Estimated Hammerstein–Wiener model for: (a) transparency and (b) turbidity.

Please cite this article as: J.S. Bulhoes, C.L. Martins and M.D. Oliveira et al., Indirect prediction system for variables that have gaps in their time series, Chaos, Solitons and Fractals, https://doi.org/10.1016/j.chaos.2019.109509

Table 5
Analysis of the parameters obtained for the free carbon dioxide and alkalinity.

Parameter	D_C		A_L	
	E_1	E_2	E_1	E_2
R_i	90	88	85	88
R_o	11	8	9	9
M_o	GNA	GNA	LM	LM
E_l	2.92	3.71	8.71	7.46
E_s	2.60	2.77	3.36	3.52
E_a	7.43 mg/L	7.92 mg/L	26.15 µeq/L	27.39 µeq/L

Table 6
Analysis of the parameters obtained for the total dissolved nitrogen and inorganic suspended matter.

Parameter	N_D		M_I	
	E_1	E_2	E_1	E_2
R_i	72	90	77	79
R_o	5	5	9	9
M_o	LM	LM	GNA	GNA
E_l	10.48	10.33	4.02	2.49
E_s	4.37	4.68	4.23	4.23
E_a	16.42 µg/L	17.60 µg/L	4.02 mg/L	4.02 mg/L

Table 7
Analysis of the parameters obtained for the transparency and turbidity.

Parameter	T_R		T_U	
	E_1	E_2	E_1	E_2
R_i	87	90	87	87
R_o	8	8	10	6
M_o	GNA	GNA	GNA/LM	GNA/LM
E_l	5.06	5.62	9.07	10.25
E_s	3.14	3.26	3.00	3.02
E_a	7.46 cm	7.75 cm	3.32NTU	3.34NTU

by E_1 and E_2 for the N_D variable. The E_l of the E_1 and E_2 structures for the M_I variable were below 5%, with the E_s also limited to this value. However, E_2 presented better results as analyzed in Table 6. Fig. 13(b) shows the time series obtained by E_1 and E_2 for the M_I variable.

The E_1 and E_2 structures of the T_R variable obtained similar results among extracted samples. However, the E_1 presented slightly lower E_l as provided in Table 7. Fig. 14(a) shows the time series obtained by E_1 and E_2 for the T_R variable. The E_2 of the T_U variable presented $E_l = 10.25\%$ while the E_1 got $E_l = 9.07\%$, as shown in Table 7. Fig. 14(b) shows the time series obtained by E_1 and E_2 for the T_U variable.

The results on all ten variables identified had E_s less than 5%. Fig. 15 presents the distribution of E_a in all available collections, in other words, the probability of occurrence of the absolute error E_{a1} for each time series, where E_{a1} is the difference between the value obtained by the model and the value in the database.

3.4. Prediction system using artificial neural networks

Before using the time series filled through system identification and spectral analysis, they need to undergo a resampling process to reduce excess data and normalization to standardize all time series.

The resampling aims to reduce the size of the time series, since time series with excessive data can make the training of artificial neural networks unfeasible. The resampling of the series used approximately 11,000 elements with frequency of the samples fixed in one sample per day and reduced to a single sample for every five days, which reduces the size of the time series to 2200 elements, and consequently decreases in approximately five times the training time. It is important to note that there are still about 73 samples per T_f of the time series. Finally, the time series resampled were normalized using Expression (1) where x represents the time series, $d_1 = 100$, and $d_2 = 200$.

As suggested in the methodology, not all time series were predicted with prediction artificial neural network (PANN). Among the m time series existing, only σ time series were predicted and with the σ PANN implemented, the ρ time series remaining used the classification artificial neural network (CANN). The *sigma* chosen for this work was four, that is, four PANN were implemented to predict four time series among the twelve existing ones. The series chosen to be used in the PANN were the Paraguay River level, the dissolved oxygen, the water temperature and the turbidity, since they have a significant influence on the other time series proved by Table 8.

The time series of the level interferes in all the other eleven variables, the temperature of the water influences in the dissolved oxygen, in the free carbonic gas and in the *pH*. The turbidity has a strong relationship with the material in suspension and also in the replacement of oxygen in the aquatic environment. Finally, dissolved oxygen was also one of the predicted time series with PANN due to its high correlation with the other variables. The Table 8 shows the correlation of all the eleven variables obtained from the analysis of the time series filled previously.

To carry out the long-term prediction, the delay in the input signal d of 146 samples was implemented. The 146 samples are equivalent to two years of data, since the sample period was adjusted to a sample every 5 days in the data preparation stage. The PANN implemented took 73 input regressors because it is long-term forecast, which plans to predict events with at least two years beyond the trained data. Therefore, the prediction $\tilde{x}_i(t+1)$ depends on the previous values $\tilde{x}_i(t-146)$, $\tilde{x}_i(t-147)$, $\tilde{x}_i(t-148)$, ..., $\tilde{x}_i(t-218)$.

The PANN implemented for the long-term prediction system have two hidden layers, where the first contains 150 neurons and the second with 30 neurons. The output layer has only one neuron to provide the predicted value of the analyzed variable. The activation function used for the two hidden layers of PANN was the hyperbolic tangent and for the output layer the linear activation function was used. Finally the training algorithm used was the Levenberg–Marquardt.

The CANN were implemented with only one hidden layer of 150 neurons and one neuron in the output layer. The best tested configuration was also the hyperbolic tangent and linear activation function for the hidden and output layer respectively. Finally, PANN and CANN were trained using the training algorithm based on gradient descendent with momentum.

To validate the long-term forecasting system, the time series were divided into: Data set **A** and data set **B**. The data set **A** was responsible for training stage and contains data from 1987 to 2010 and from 2014 to 2017. Finally, the data set **B** was responsible for the validation of the prediction system and contain the data from 2011, 2012 and 2013, since these three years were atypical years in the pantanal which will help to verify if the system is able to predict scenarios not presented.

The variable L_R had E_s less than 8% in all years belonging to data set **B** as observed in Table 9. Fig. 16(a) presents the predicted time series of L_R variable. The prediction was carried out until the year 2022, and since the PANN has delay d of only two years, the other two years were estimated using previously predicted values, which can increase the prediction errors.

The T_W variable obtained the lowest E_s among all predicted variables as analyzed in Table 9. Fig. 16(b) shows the output of

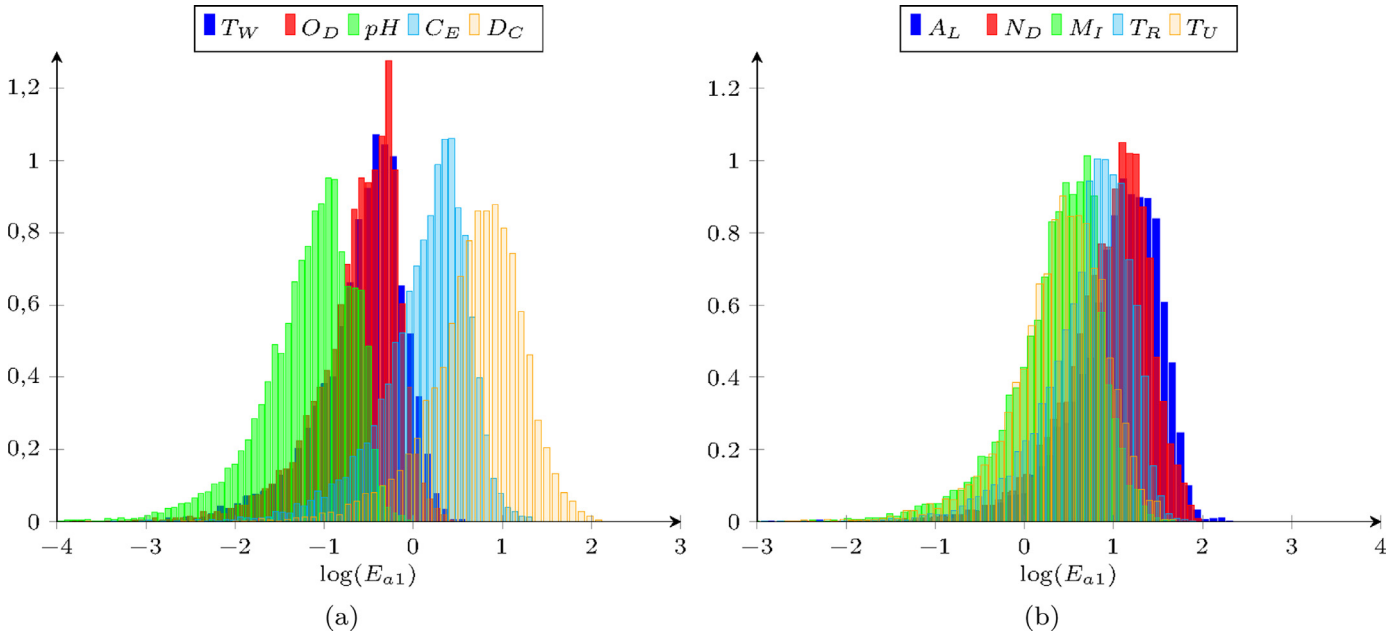

Fig. 15. Distribution of the mean absolute error in the structure chosen for the Hammerstein–Wiener model.

Table 8
Correlation between all analyzed variables.

Var.	L_R	T_W	O_D	pH	C_E	D_C	A_L	N_D	M_I	T_R	T_U
L_R	1.00	−0.56	−0.46	−0.26	0.16	0.18	0.41	−0.46	−0.37	0.72	−0.69
T_W	−0.56	1.00	−0.02	0.06	−0.01	−0.06	−0.29	0.40	0.17	−0.53	0.58
O_D	−0.46	−0.02	1.00	0.38	−0.37	−0.28	−0.23	0.23	0.28	−0.31	0.37
pH	−0.26	0.06	0.38	1.00	−0.42	−0.72	−0.03	0.23	0.30	−0.20	0.27
C_E	0.16	−0.01	−0.37	−0.42	1.00	0.39	0.33	−0.23	−0.29	0.03	−0.34
D_C	0.18	−0.06	−0.28	−0.72	0.39	1.00	0.25	−0.19	−0.30	0.13	−0.25
A_L	0.41	−0.29	−0.23	−0.03	0.33	0.25	1.00	−0.29	−0.15	0.36	−0.41
N_D	−0.46	0.40	0.23	0.23	−0.23	−0.19	−0.29	1.00	0.36	−0.31	0.62
M_I	−0.37	0.17	0.28	0.30	−0.29	−0.30	−0.15	0.36	1.00	−0.21	0.45
T_R	0.72	−0.53	−0.31	−0.20	0.03	0.13	0.36	−0.31	−0.21	1.00	−0.57
T_U	−0.69	0.58	0.37	0.27	−0.34	−0.25	−0.41	0.62	0.45	−0.57	1.00

Table 9
Mean square error for the prediction system.

Var.	2011	2012	2013	2014	2015	2016	2017	2018	Total
L_R	7.84	7.19	6.48	12.15	7.94	5.39	5.87	4.59	7.18
T_W	4.72	4.19	5.70	9.54	4.17	4.29	4.30	4.72	5.20
O_D	11.93	11.67	8.46	14.79	6.05	9.89	–	–	10.46
T_U	4.31	6.64	5.37	10.76	6.68	12.69	4.32	2.92	6.71
pH	14.41	12.19	7.84	4.76	7.83	5.65	3.57	3.60	7.48
C_E	8.91	8.87	8.37	9.38	9.79	5.43	6.59	5.28	7.83
D_C	18.97	8.46	5.21	4.15	5.74	3.14	3.32	1.65	6.33
A_L	11.35	9.84	4.91	6.86	6.61	6.71	6.57	2.98	6.98
N_D	13.10	14.26	10.30	7.56	4.07	6.84	5.73	4.47	8.29
M_I	8.18	8.74	8.32	8.34	7.84	7.52	6.81	–	7.96
T_R	6.30	2.61	3.32	3.59	7.07	5.75	5.12	9.27	5.37

the PANN for the T_W variable. The results of the O_D variable can be observed in Table 9, and it is important to know that there are no values for comparison in 2017 and 2018. This was the variable that presented the worst performance among those predicted with PANN. Fig. 17(a) displays the output of the PANN for the O_D variable.

The T_U variable had E_s less than 7% for the data set **B** as shown in Table 9. It is observed that among the four predicted variables with PANN, turbidity is the one that presented the best results in this validation set. The Fig. 17(b) displays the predicted values for the T_U variable. The long-term prediction for the other variables was performed using the CANN, which has as input the four series forecast by PANN. This methodology reduces the time of training and execution of the network. As the inputs are the PANN with predicted values for two years, the CANN also obtain prediction for two years of each manipulated variable.

The pH variable presented one of the highest E_s for the year 2011 as observed in Table 9. Fig. 18(a) presents the pre-

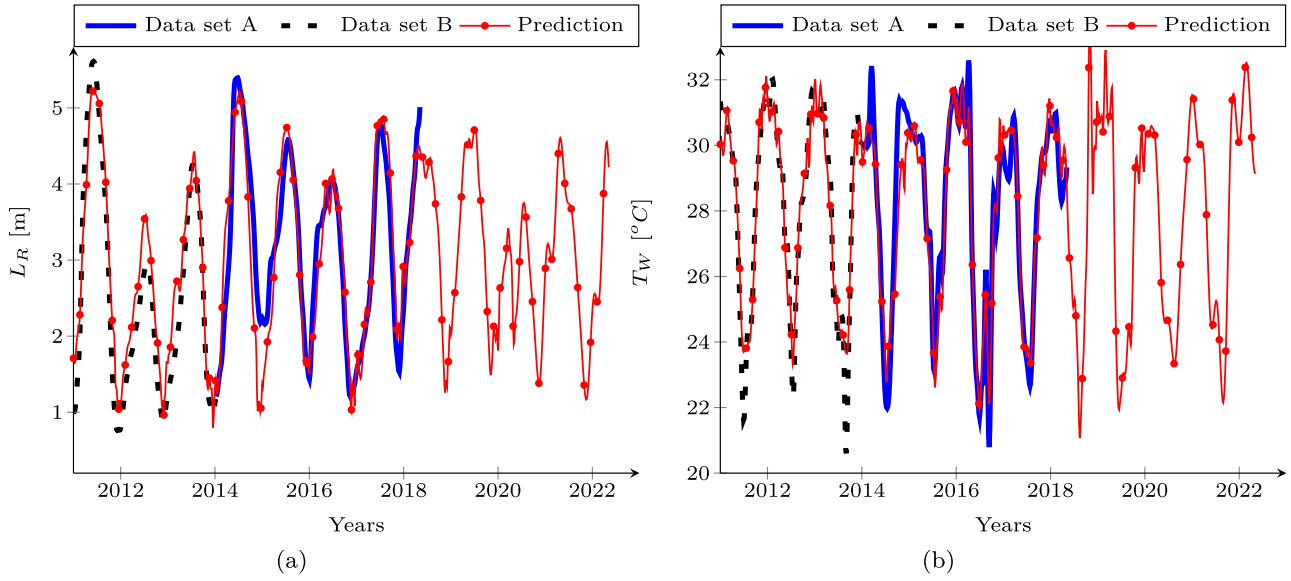

Fig. 16. Long-term forecast for the variables: (a) Paraguay River level and (b) water temperature.

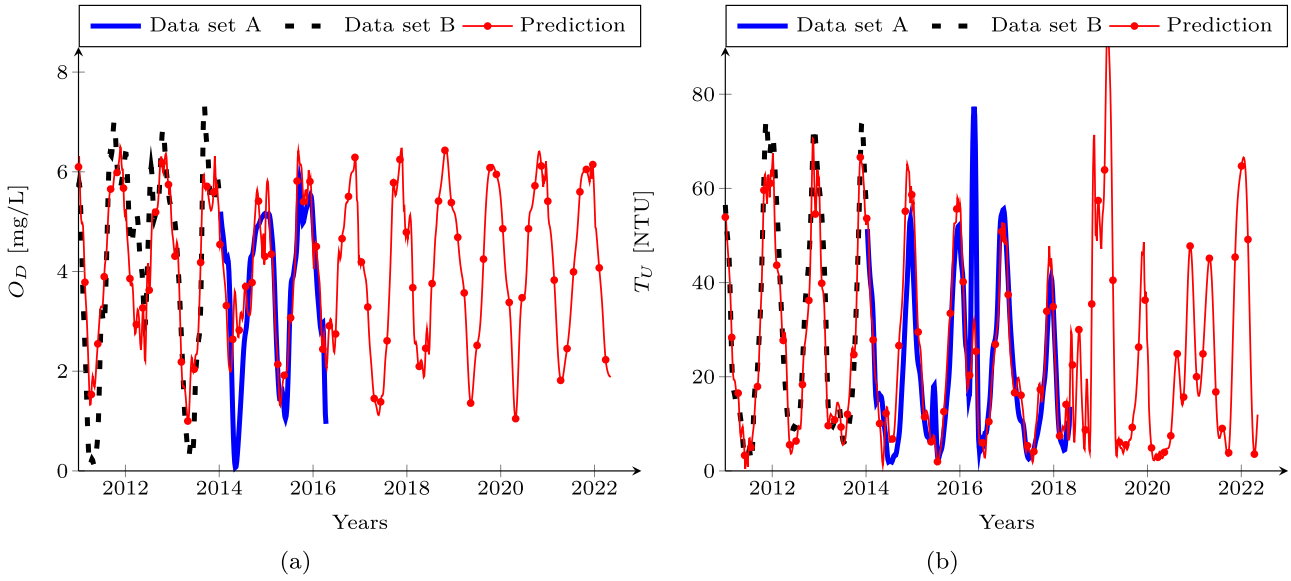

Fig. 17. Long-term forecast for the variables: (a) dissolved oxygen and (b) turbidity.

dicted time series for the *pH* variable. The C_E variable obtained E_s constant in the three years belonging to the data set **B** as demonstrate in Table 9. The Fig. 18(b) shows the output of the CANN for the C_E variable. The C_E variable has a strong correlation with the O_D (−0.37) and with the T_U (−0.34) which helped in the prediction performance of this variable.

The D_C variable presented the highest E_s for the year of 2011, which has diminished in subsequent years as observed in Table 9. Fig. 19 (a) shows the output of the CANN for the D_C variable. The D_C variable has a correlation with the O_D (−0.28) and with the T_U (−0.25) which does not explain its good results due to weak correlation with the input variables. Like the D_C variable, the A_L variable obtained E_s greater in 2011 and decreases in the following years. Fig. 19(b) shows the output of the CANN for the A_L variable. The A_L variable correlates with L_R (0.41) and T_U (−0.41), which corroborates the recognition of the patterns of this variable.

The N_D variable obtained the highest Total E_s among all the predicted variables with the CANN, however the graphic result of the forecast presented in Fig. 20(a) shows that the prediction system was able to observe all the dynamics of this variable. The N_D has a strong correlation with all input variables, mainly L_R (-0.46) and T_U (0.62). The M_I variable had E_s less than 9% for the whole data set**B**. Fig. 20(b) displays the predicted time series for the M_I variable.

The T_R variable obtained the best results among the other variables predicted by the CANN. Its performance is related to a strong correlation with the four input variables, mainly L_R (0.72) and T_U (−0.57). Fig. 21 shows the prediction of this variable.

Embrapa Pantanal also made available the flooded area (A_F) estimated through satellite images from February 2000 to December 2009 [18]. The important characteristic of the A_F variable is that it has a strong correlation with the variable NV, which facilitates the use of the proposed methodology for CANN. [18] related the flooded area to the level of the Paraguay River through

Please cite this article as: J.S. Bulhoes, C.L. Martins and M.D. Oliveira et al., Indirect prediction system for variables that have gaps in their time series, Chaos, Solitons and Fractals, https://doi.org/10.1016/j.chaos.2019.109509

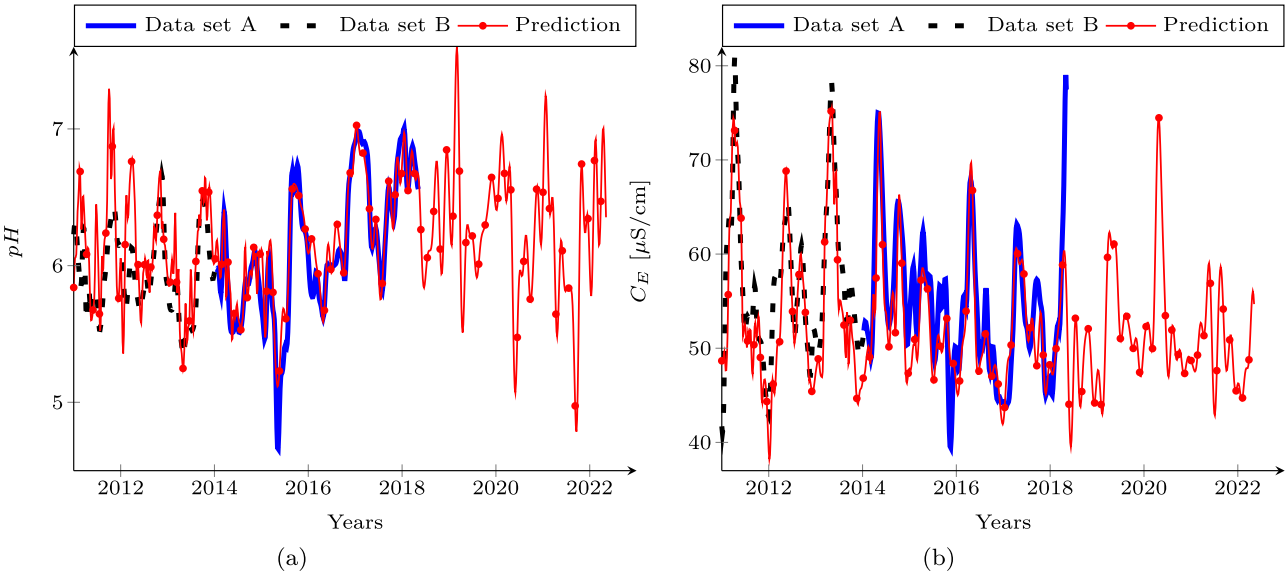

Fig. 18. Long-term forecast for the variables: (a) potential of hydrogen and (b) electrical conductivity of water.

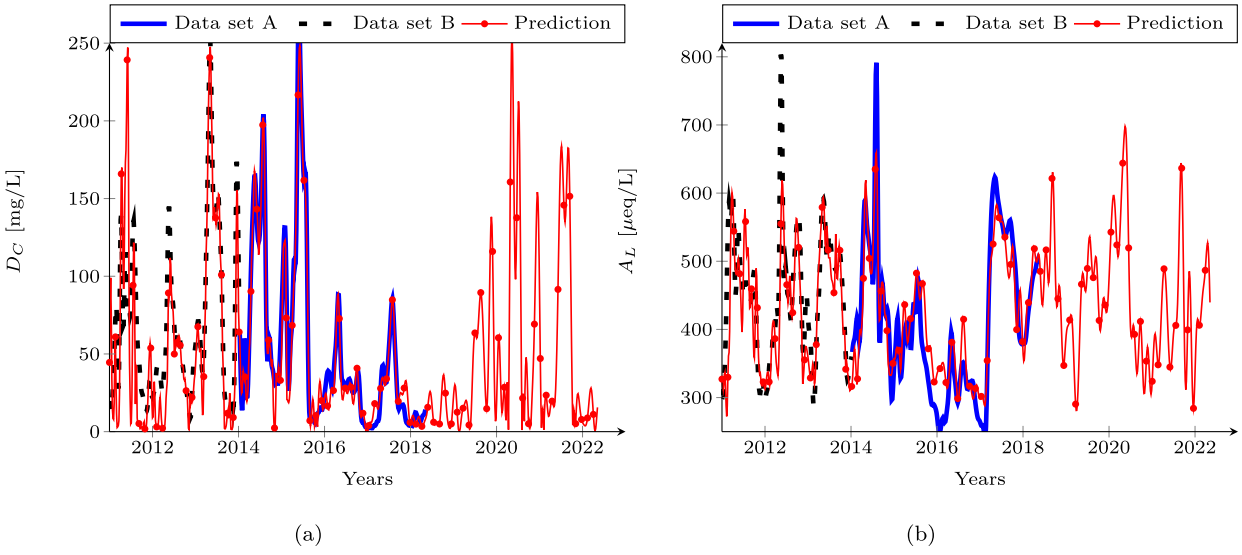

Fig. 19. Long-term forecast for the variables: (a) free carbon dioxide and (b) alkalinity.

Table 10
Comparison between the actual values of the flooded area variable and the values obtained in the [18] and CANN models.

Model	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	Total
Padovani	15.56	12.65	8.12	10.22	6.54	11.12	14.29	9.26	10.41	14.11	11.20
CANN	4.54	2.83	3.72	3.03	3.75	2.68	5.01	3.49	4.40	6.74	4.02

the expression (5), where $\tilde{A}_s(t)$ represents the flooded area measured through satellite images and $\tilde{N}_p(t+60)$ is the level of the Paraguay River in [cm] 60 days after the instant t. Therefore, the flooded area can be expressed by (6), where $\tilde{A}_p(t)$ is the estimated flooded area for the instant t and $N_l(t+60)$ is the level in [cm] observed in the Ladario rule 60 days after the instant t.

$$\tilde{N}_p(t+60) = \frac{537.1}{1+4.4443e^{-0.00022\tilde{A}_s(t)}} \tag{5}$$

$$\tilde{A}_p(t) = \frac{-ln\left(\frac{537.1 - N_l(t+60)}{4.4443N_l(t+60)}\right)}{0.00022} \tag{6}$$

Fig. 22 presents the results obtained for the A_F variable and the prediction for the next four years using the CANN and the expression (6). The flooded area given from (6) was presented by the brown curve, where the results of the CANN adjusted better to the actual values of the flooded area, curve in blue color (data set **A**) and dotted black (data set **B**). Although Fig. 22 presents the values

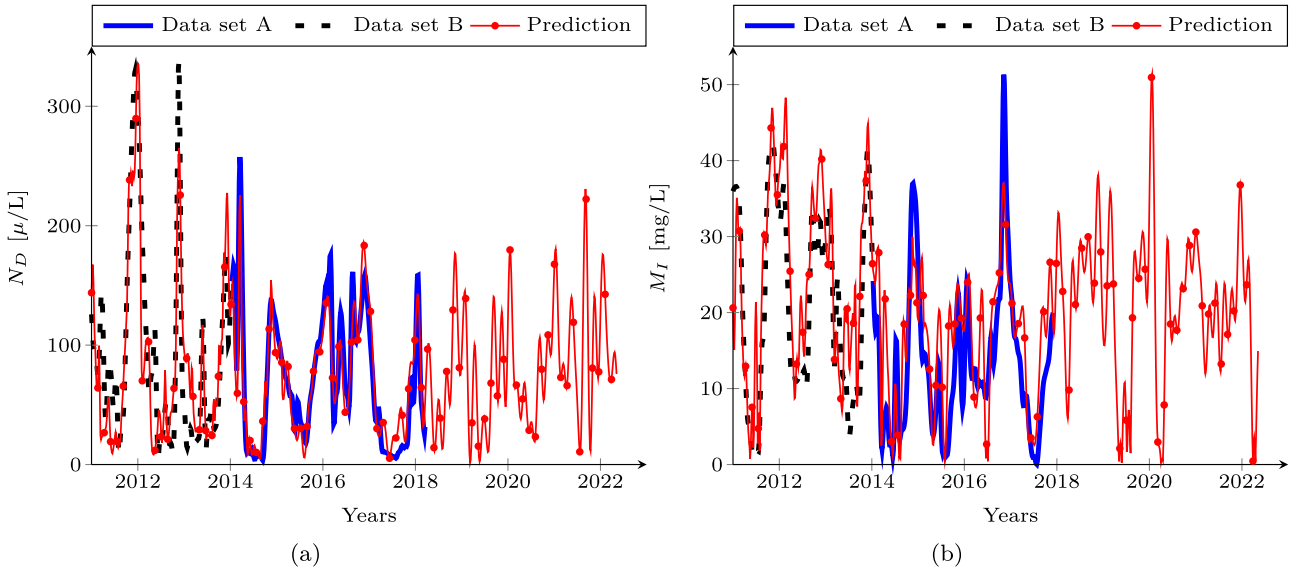

Fig. 20. Long-term forecast for the variables: (a) total dissolved nitrogen and (b) inorganic suspended matter.

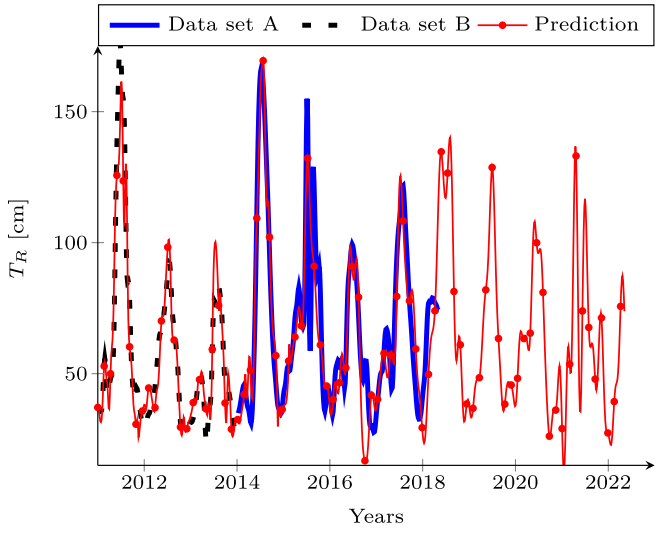

Fig. 21. Long-term forecast for the transparency.

of the flooded area from 2005, there are data since February 2000, and with these data it is possible to compare the model proposed by [18] and the model proposed in this article. The Table 10 provides the comparative on both models, where it is observed that the CANN obtained better performance.

4. Conclusion

This work proposed a method that uses spectral analysis and system identification to fill gaps in time series with no fixed sampling rate. Besides, it used the multilayer neural networks for the development of a prediction system applied to physico-chemical variables of the Paraguay River. The difference of this methodology of filling of gaps with those existing in the literature is the combination of two filling techniques, in order to improve the final performance of the process. In relation to the prediction system, the differential is the indirect prediction of variables through the classification neural network, which is capable of recognizing the dynamics of a variable from other variables with some correlation with each other.

The variables predicted indirectly with the classification neural network had satisfactory performance, producing results that validate the proposed method, even those that presented weak correlation with the input variables. The classification neural network used four time series, which are: i) Paraguay River level, ii) water temperature, iii) dissolved oxygen and iv) turbidity. These time series were already predicted by PANN. This is one of the strengths of this methodology, because even for series with low correlation with the four input variables of the CANN, the results obtained were satisfactory. Therefore, with this type of prediction system, it is not necessary to have the time series of all the variables, just having only the time series of variables that influence the others, as observed in the case study of the flood area variable.

This work contributes with: new methodology of gap filling in time series and the development of indirect prediction system. For new studies, it will be interesting to research about the impact of different optimization techniques for parameter estimation in Hammerstein-Wiener models. Also, it will be useful to test the gap filling methodology in other time series, as well as to perform new indirect predictions in other databases.

Please cite this article as: J.S. Bulhoes, C.L. Martins and M.D. Oliveira et al., Indirect prediction system for variables that have gaps in their time series, Chaos, Solitons and Fractals, https://doi.org/10.1016/j.chaos.2019.109509

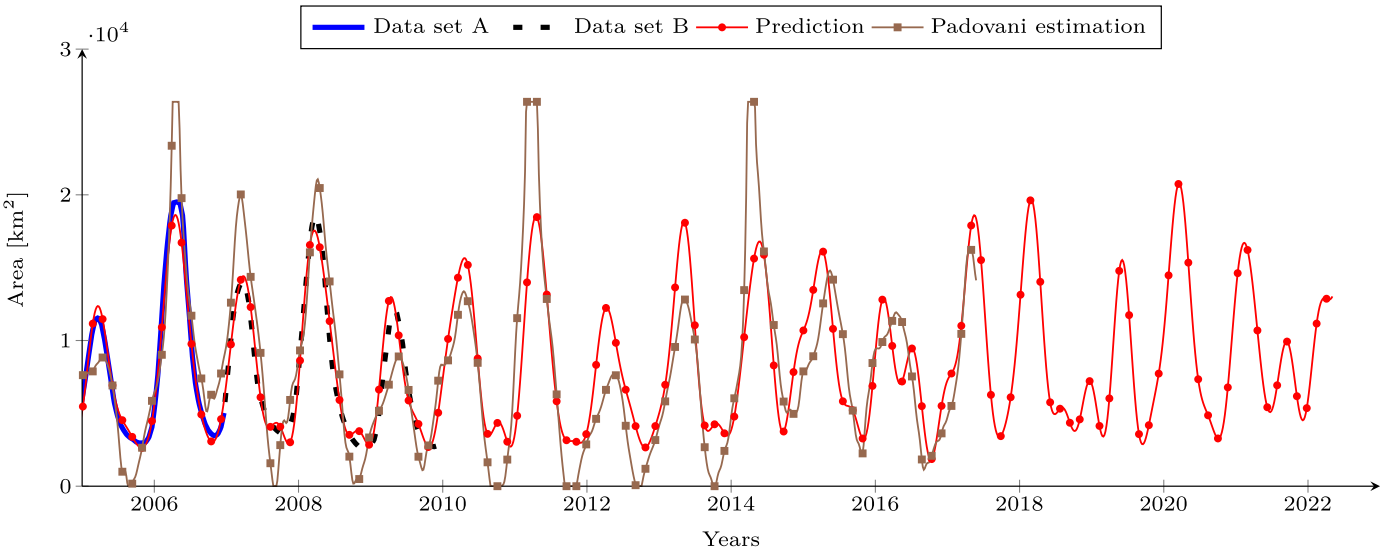

Fig. 22. Prediction of the flooded area through the CANN model and [18] estimation.

Declaration of Competing Interest

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

CRediT authorship contribution statement

Junio S. Bulhoes: Conceptualization, Data curation, Formal analysis, Methodology, Writing - original draft, Writing - review & editing. **Cristiane L. Martins:** Conceptualization, Data curation, Formal analysis, Methodology, Writing - original draft, Writing - review & editing. **Marcia D. Oliveira:** Conceptualization, Data curation, Formal analysis, Methodology, Writing - original draft, Writing - review & editing. **Debora F. Calheiros:** Conceptualization, Data curation, Formal analysis, Methodology, Writing - original draft, Writing - review & editing. **Wesley P. Calixto:** Conceptualization, Data curation, Formal analysis, Methodology, Writing - original draft, Writing - review & editing.

Acknowledgment

The authors thank the Coordination for the Improvement of Higher Education Personnel (CAPES) for the support and financial support for the development of this research.

References

[1] Box GE, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control. 5th. John Wiley & Sons; 2015. ISBN 978-1-118-67502-1.
[2] Tardivo G, Berti A. A dynamic method for gap filling in daily temperature datasets. J Appl Meteorol Climatol 2012;51(6):1079–86. doi:10.1175/JAMC-D-11-0117.1.
[3] Ustoorikar K, Deo M. Filling up gaps in wave data with genetic programming. Mar Struct 2008;21(2–3):177–95. doi:10.1016/j.marstruc.2007.12.001.
[4] Moffat AM, Papale D, Reichstein M, Hollinger DY, Richardson AD, Barr AG, et al. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. Agric Meteorol 2007;147(3):209–32. doi:10.1016/j.agrformet.2007.08.011.
[5] Kondrashov D, Shprits Y, Ghil M. Gap filling of solar wind data by singular spectrum analysis. Geophys Res Lett 2010;37(15). doi:10.1029/2010GL044138.
[6] Skakun SV, Basarab RM. Reconstruction of missing data in time-series of optical satellite images using self-organizing kohonen maps. J Autom Inf Sci 2014;46(12). doi:10.1615/JAutomatInfScien.v46.i12.30.
[7] Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. BMC Med Res Methodol 2017;17(1):162. doi:10.1186/s12874-017-0442-1.
[8] Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? some practical clarifications of multiple imputation theory. Prevent Sci 2007;8(3):206–13. doi:10.1007/s11121-007-0070-9.
[9] Pashova L, Koprinkova-Hristova P, Popova S. Gap filling of daily sea levels by artificial neural networks. TransNav Int J MarNavig Saf od Sea Transp 2013;7(2). doi:10.12716/1001.07.02.10.
[10] Eischeid JK, Bruce Baker C, Karl TR, Diaz HF. The quality control of long-term climatological data using objective data analysis. J Appl Meteorol 1995;34(12):2787–95. doi:10.1175/1520-0450(1995)034<2787:TQCOLT>2.0.CO;2.
[11] Magalhaes AS, Bulhões JS, Furriel GP, Reis MR, Alves AJ, Silva AH, et al. Parametric regression in synchronous and induction generators. In: 2017 18th International Scientific Conference on Electric Power Engineering (EPE). IEEE; 2017. p. 1–6. doi:10.1109/EPE.2017.7967310.
[12] Kimoto T, Asakawa K, Yoda M, Takeoka M. Stock market prediction system with modular neural networks. In: Neural Networks, 1990., 1990 IJCNN International Joint Conference on. IEEE; 1990. p. 1–6. doi:10.1109/IJCNN.1990.137535.
[13] Ekonomou L. Greek long-term energy consumption prediction using artificial neural networks. Energy 2010;35(2):512–17. doi:10.1016/j.energy.2009.10.018.
[14] Vieilledent G, Grinand C, Vaudry R. Forecasting deforestation and carbon emissions in tropical developing countries facing demographic expansion: a case study in madagascar. Ecol Evol 2013;3(6):1702–16. doi:10.1002/ece3.550.
[15] Ibrahim MM, Vanini R, Ibrahim FM, Martins WdP, Carvalho RTdC, Castro RSd, et al. Epidemiology and medical prediction of microbial keratitis in southeast brazil. Arq Bras Oftalmol 2011;74(1):7–12. doi:10.1590/S0004-27492011000100002.
[16] Palanthandalam-Madapusi HJ, Ridley AJ, Bernstein DS. Identification and prediction of ionospheric dynamics using a hammerstein-wiener model with radial basis functions. In: American Control Conference, 2005. Proceedings of the 2005. IEEE; 2005. p. 5052–7. doi:10.1109/ACC.2005.1470814.
[17] Fung EH, Wong Y, Ho H, Mignolet MP. Modelling and prediction of machining errors using armax and narmax structures. Appl Math Model 2003;27(8):611–27. doi:10.1016/S0307-904X(03)00071-4.
[18] Padovani CR. Dinâmica espaço-temporal das inundações do pantanal. Universidade de São Paulo; 2010. Ph.D. thesis. doi: https://doi.org/10.11606/T.91.2010.tde-14022011-170515