tecnia revista de educação, ciência e tecnologia do ifg

Volume 9, Número 2, 2024

Submetido 25/11/2023. Aprovado 20/05/2024
Avaliação: revisão duplo-anônimo

Bayesian Probabilistic Modeling applied to some environmental data using PyMC

MODELAGEM PROBABILÍSTICA BAYESIANA APLICADA A ALGUNS DADOS AMBIENTAIS USANDO PYMC

MODELADO PROBABILÍSTICO BAYESIANO APLICADO A ALGUNOS DATOS AMBIENTALES UTILIZANDO PYMC

Fernando Schimidt
Instituto Federal de Goiás (IFG)/Câmpus Goiânia
fernando.quimica@ifg.edu.br

Cleveland Lemos Freire
Instituto Federal de Goiás (IFG)
clevelandlemos@gmail.com

Luiza Cintra Campos
University College London
l.campos@ucl.ac.uk

Mariângela Fontes Santiago
Universidade Federal de Goiás (UFG)
mariangelafs@gmail.com

Abstract
Machine learning models can be employed to antecipate environmental problems and even oversee systems and processes. The Bayesian multivariate probabilistic model (BMP) is able to predict the concentration of Dissolved Oxygen (DO) in water. With the use of free software which, based on historical data from water analysis, we are able to achieve results similar to those produced by artificial neural networks. In this study, we applied a BMP model developed in Python using the PyMC3 library to two different nonlinear environmental data sets. In the first one, a multivariate calibration of DO was performed in raw water from the Piracicaba River (São Paulo - Brazil), and the BMP model obtained a Root Mean Square Error of prediction (RMSEp) of 0.835 mg L^{-1} for 20 samples tested. In the second set, the same modeling process was used to obtain an RMSEp of 0.839 mg L^{-1} for 20 samples tested. Both results comparable to those obtained with a back propagation neural network. Another non-linear environmental data set was tested with excellent results. A BMP model for the concentration of benzene in air, as used for monitoring air pollution, obtained a RMSEp of 0.584 $\mu g\ m^{-3}$.

Keywords: air pollution modeling. Bayesian modeling. dissolved oxygen. PyMC. Python. water quality modeling

Resumo
Modelos de aprendizado de máquina podem ser usados para prever problemas ambientais e até mesmo gerenciar sistemas e processos. O modelo probabilístico multivariado bayesiano (BMP) permite a construção de um modelo de previsão da concentração de Oxigênio Dissolvido (OD) na água. Com a utilização de softwares gratuitos que, a partir de dados históricos de análises de água, fornecem de

forma rápida e robusta, resultados semelhantes aos produzidos por redes neurais artificiais. Neste trabalho, um modelo BMP desenvolvido em Python utilizando a biblioteca PyMC3 foi aplicado a dois conjuntos de dados ambientais não lineares diferentes. No primeiro, a calibração multivariada do OD foi realizada em água bruta do Rio Piracicaba (São Paulo - Brasil), onde o modelo BMP obteve um Erro Quadrático Medio de previsão (RMSEp) de 0,835 mg L^{-1} para 20 amostras testadas. No segundo conjunto, o mesmo processo de modelagem foi realizado para o Rio Paraíba do Sul (São Paulo - Brasil), obtendo um RMSEp de 0,839 mg L^{-1} para 20 amostras testadas. Ambos os resultados foram equivalentes aos obtidos com uma rede neural de retropropagação. Outros dados ambientais não lineares foram testados com resultados muito bons. Um modelo BMP para concentração de benzeno no ar urbano, como monitoramento da poluição atmosférica, obteve um RMSEp de 0,584 $\mu g\ m^{-3}$.

Palavras-chave: Modelagem de poluição atmosférica. Modelagem bayesiana. Oxigênio dissolvido. PyMC. Python. Modelagem de qualidade da água

Resumen
Los modelos de aprendizaje automático se pueden utilizar para predecir problemas ambientales e incluso gestionar sistemas y procesos. El modelo probabilístico multivariado bayesiano (BMP) permite la construcción de un modelo de predicción de la concentración de Oxígeno Disuelto (OD) en agua. Utilizando software libre que, basándose en datos históricos de análisis de agua, proporciona de forma rápida y robusta resultados similares a los producidos por redes neuronales artificiales. En este trabajo, se aplicó un modelo BMP desarrollado en Python utilizando la biblioteca PyMC a dos conjuntos de datos ambientales no lineales diferentes. En el primero, se realizó la calibración multivariada de OD en agua cruda del río Piracicaba (São Paulo - Brasil), donde el modelo BMP obtuvo un error cuadrático medio de predicción (RMSEp) de 0,835 mg L-1 para 20 muestras analizadas. En el segundo conjunto, se realizó el mismo proceso de modelado para el río Paraíba do Sul (São Paulo - Brasil), obteniendo un RMSEp de 0,839 mg L-1 para 20 muestras probadas. Ambos resultados fueron equivalentes a los obtenidos con una red neuronal de retro propagación. Se probaron otros datos ambientales no lineales con muy buenos resultados. Un modelo BMP para la concentración de benceno en el aire urbano, como monitoreo de la contaminación del aire, obtuvo un RMSEp de 0,584 $\mu g\ m^{-3}$.

Palabras clave: Modelado de la contaminación del aire. Modelado bayesiano. Oxígeno disuelto. PyMC. Python. Modelado de la calidad del agua

Introduction

It is essential that water resources have adequate physical, chemical and microbiological conditions for use. Water must contain essential substances essential to life and be free from other substances that may have detrimental effects on organisms in the food chains. Therefore, it must be available in sufficient quantity and quality to meet the needs of the biota (Baird e Cann, 2011). Aiming to improve water quality conditions, the monitoring data require and allow ways to monitor the variation of water quality indicators. The modeling of these data can help establish hypotheses about the structure or behavior of a physical system and can explain the properties of the system and predict reactions to stimulus (Emamgholizadeh *et al.*, 2014; Wooley; Lin, 2005). It is also important to build models to predict DO as a function of changes in other physical and chemical parameters. The modeling outcomes can facilitate a deeper understanding of the mechanisms that underlie diverse behaviors, thereby providing a foundation for decision-making in water quality control and hydrological system (Von Sperling, 2014; Qian *et al.*, 2005; Graf, 2018).

Bayesian statistics are conceptually very simple: there are some data that are fixed, in the sense that what is measured cannot be changed, and there are parameters whose values are of interest and, therefore, their possible values can be explored. All the uncertainties are modeled using probabilities. In other statistical paradigms, there are different types of unknown quantities; in the Bayesian structure, everything that is unknown is treated in the same way. If a quantity is not known, a probability distribution can be assigned to it. Then, Bayes' theorem is used to transform the previous probability distribution $p(\theta)$ (what is known about a given problem before looking at the data), into a posterior distribution $p(\theta|D)$ (what is known after observing D data). In other words, Bayesian statistics is a form of learning (Martin, 2016). Transcending the logical reasoning that ponders the prior knowledge about the problem, this knowledge base draws its conclusions, where the full scope of the problem is not previously known. Thus, probabilistic reasoning is necessary, and it can act in the face of uncertainty, assigning levels of reliability. A powerful tool in this reasoning is Bayesian inferences. The lack of information in the probability is the same as dealing with uncertainties, not just boolean values, type of primitive data that has true (1) and false (0), in Bayesian thinking, this view is updated after analyzing the evidence, even if contrary to what is believed a priori - prior or previous probability, and given the updated evidence, the posterior probability. The Bayesian worldview defines probability as the measure of credibility in an event - that is, how confident we are that a given event will occur (Davidson-Pilon, 2016).

Consequently, when casting a fair coin, the probability of obtaining the head is 50%, but assuming in a play, one of the opponents has spied the coin at the moment after the launch, so the certainty of the result is likely 100% for the head face. But knowledge of the outcome does not change the results of the coin. So, different probabilities to the outcome are assigned. Furthermore, if after 100 throws the result was 20 times for the head and 80 times for the crown, the probability, in the opponents' belief, will remain the same, in other words, 50% chance for each one. This only happens in the frequentist view of probability, which is given by the limit of the relative frequency of the occurrence of an event, within one of a sample space with "n" independent repetitions of the experiment, with this "n" tending to infinity. So, the so-called frequentists, attribute to the most classic version of statistics, an unconditional probability, assuming that the probability is the frequency of long-term events. Bayesians, on the other hand, define probability as the measure of belief, or confidence that an event occurs, that is, a conditional probability. So, a frequentist inference of a function returns a number, a value that represents an estimate, summarized as if it was the sample mean, while the Bayesian function returns probabilities (Martin, 2016; Davidson-Pilon, 2016). Then, in a previous belief, in the face of new evidence, a new belief, known as Bayes Theorem, defined by Thomas Bayes, moves on to the following equation:

$$P(\theta|\gamma) = \frac{(P(\gamma|\theta)P(\theta)}{P(\gamma)} \quad (1)$$

Where:
$P(\theta|\gamma)$: Posterior distribution of θ given the event γ;
$P(\theta)$: The priori (previous) distribution of the event θ;
$P(\gamma|\theta)$: Probability of event γ given the event θ (likelihood function);
$P(\gamma)$: Probability of event γ.

Interested in the proposition θ and knowing the event γ, the posterior $P(\theta|\gamma)$ is calculated. Inference is used to obtain it through computational processes (Davidson-Pilon, 2016).

Material and methods

Input and output data

The water quality data from the Piracicaba River (sampling point PCAB 02800) and the Paraíba do Sul River (PARB 02300), were provided by the Companhia Ambiental do Estado de São Paulo (Cetesb, 2023). The parameters: water temperature, turbidity, conductivity, total phosphorus concentration, ammoniacal nitrogen concentration, Kjeldahl nitrogen concentration, nitrate concentration, nitrite concentration, total dissolved solids concentration, biochemical oxygen demand (BOD), concentration of dissolved oxygen (DO), pH, and river flow were analyzed. The latter is available from Agência Nacional de Águas (Ana, 2023). All parameters from the samples analyzed were collected between 1990 and were used as historical data for building the models developed in this work. For the Piracicaba River, 148 data sets were provided, without any type of sample selection or removal of outliers for use in the models. For the calculations model, the physico-chemical parameters and river flow were arranged in a matrix X[148, 10], in which each row corresponds to the analyses made on a given day. The columns refer to the studied parameters, except for the DO concentration values, which were used exclusively as an output parameter, associated with a vector y[148, 1]. For the Paraíba do Sul River, 155 data sets were provided and the data was arranged in X[155, 11] and the output vector y[155, 1].

Another non-linear environmental data was tested (supplementary material): air analysis with electronic nose of benzene (De Vito *et al.*, 2008; De Vito *et al.*, 2009), with a total of 9358 data sets, with 15 attributes containing missing values. All data containing missing values has been removed as well as the date and time attributes. The total samples were arranged in a matrix X [876, 12] with the variables from the gas sensors responses: true CO concentration, and PT08.S1 sensor response, true Non Metanic Hydrocarbons concentration, PT08.S2 sensor response, true NOx concentration, PT08.S3 sensor response, True NO_2 concentration, PT08.S4 sensor response, PT08.S5 sensor response, Temperature, Relative Humidity and Absolute Humidity. The output vector y [876, 1] is the true benzene concentration.

Mathematical modelling

This work was developed using the PyMC library (Salvatier; Wiecki; Fonnesbeck, 2016) which is programmed in the Python language through the Anaconda distribution (Anaconda, 2023). It allows the use of Bayesian inferences tools; for statistical treatment of water quality data. It could also use cloud resources. The Bayesian probabilistic model (BPM) is based on the construction of a mathematical relationship between independent variables $\boldsymbol{\gamma}_i$ and the other dependent variables $\boldsymbol{\theta}_i$, using specific Gaussian probability distribution functions (Normal) to characterize the priori distribution of each variable $\boldsymbol{\theta}_i$ in the model. Bayesian data analysis is based on the inference of unknown parameters for observed data models, returning to ***Equation 1***, where now:

$\boldsymbol{\theta}$: Unknown variable of the model to be estimated (DO);
$\boldsymbol{\gamma}$: Matrix of observed data X (physical-chemical parameters and river flow);
$P(\boldsymbol{\theta} \mid \boldsymbol{\gamma})$: Posterior distribution of DO ($\boldsymbol{\theta}$) given the data X (event $\boldsymbol{\gamma}$);
$P(\boldsymbol{\gamma} \mid \boldsymbol{\theta})$: Likelihood function, which models the probability of observing the data $\boldsymbol{\gamma}$ given the parameters in the probabilistic model for the DO concentration.

P(θ): Prior distribution of model parameters.
P(y): Probability distribution of data X, built from historical (temporal) data.

The adjustment parameters for BPM model are: the prior distributions parameters α, β and σ are gaussian using *Normal* (μ, 2), with μ from DO mean historical data and standard deviation σ as *Uniform* (0, 1); the sampling used are NUTS default method, with tune = 2000 and random seed. The comparison between the results of the BPM model and the artificial neural network was made using a back propagation neural network (*traingda* algorithm) present in Matlab, version R2016a. The training of the neural network considered all the adjustment parameters on the ***default*** condition. The validation/prediction sets for all models (for rivers) were built with 20 samples chosen using the original Kennard & Stone (Kennard-Stone […], 2022) algorithm. The models generated and optimized by the PyCaret library (Documentação, 2023) were also used in comparison with our BPM model using ***default*** conditions. One of the key features of PyCaret is its capability to automatically select and tune machine learning models for all kinds of data. All full models developed with their adjustment parameters and the original data can be accessed at https://github.com/Schimidt99/Bayesian-probabilistic-model.

Results and discussion

A relationship between the dependent variable **y** (i.e., DO), in relation to the other independent input variables **X** can be established by using the Bayesian multivariate linear model through the *probability density functions* (PDF):

$$y_i \sim Normal(\mu_i, \sigma) \quad \text{(likelihood function)} \tag{2}$$

The linear relationship between y_i and $X_{i,j}$ is:

$$\mu_i = \alpha + \beta_j X_{i,j} \tag{3}$$

Where:
$\alpha \sim Normal(4, 10)$ *:* the prior function α;
$\beta_j \sim Normal(0, 10)$ *:* the prior functions β;
$\sigma \sim Uniform(0, 1)$ *:* the prior function σ;
i : index for the number of samples;
j : index relative to the number of variables (**X** columns).

The fit values of the PDF functions are estimated and based on experimental data. The fit values are in parentheses. The variable $\mathbf{y}_i$ uses a gaussian probability distribution, based on experimental information, that defines the plausibility of the individual observations of river DO over the sample period studied. Other probabilistic distributions were also tested in this work, such as log-normal, but the OD prediction results were worse. The parameters α, β_j and σ are defined by *priori distributions* with Gaussian form as $Normal(\mu,\sigma)$, using mean μ and standard deviation σ. Gaussian distribution appears in many processes in nature. Measurement errors, variations in growth and the velocities of molecules tend to Gaussian distributions. These processes do this because, in the end, these processes add finite fluctuations to a distribution

of sums that aggregate information about the underlying process (Salvatier; Wiecki; Fonnesbeck, 2016; Mc Elreath, 2015; Lyon, 2014).

The values of μ and σ in the distributions are optimized for each parameter α, β_j and σ. This model can predict $\mathbf{y}_i$ results through the *posterior distribution*, as observations normally distributed through an expected value μ_i, which is a linear function of predictive variables $\mathbf{X}_j$, plus an observation error σ with uniform distribution (experimentally optimized).

At the PyMC library, the Bayesian model (Salvatier; Wiecki; Fonnesbeck, 2016) loads all the user-defined $\mathbf{X}_{ij}$ predictor (input) variables by associating them with a parameter β_j. It establishes a product between each vector (set) of $\mathbf{X}_j$ values and their respective coefficient β_j, being added to the coefficient α (or adjustment parameter) and establishing a mathematical relationship with the output $\mathbf{y}_i$ (or expected result) through the variable μ_i:

$$\mu_i = \alpha + \beta_0 X_{i1} + \beta_1 X_{i2} + \beta_2 X_{i3} + ... + \beta_i X_{ij} + \varepsilon \qquad (4)$$

The model considers μ to be *deterministic* random variable, which implies that its value is completely determined by the values of its 'parents'. In other words, there is no uncertainty beyond what is inherent in the parents values. Here (eq. 4), μ is just the sum of the intercept α and the products of the coefficients β_j representing the physico-chemical parameters from the analysis of the river water (Table 1) by the predictive variables $\mathbf{X}_j$, whatever their values are, plus the random error ε presents behavior compatible with $Normal(0,s^2)$.

Piracicaba river	Paraiba do Sul river
β_0: BOD	β_0: BOD
β_1: Water temperature	β_1: Water temperature
β_2: River flow	β_2: River flow
β_3: Water pH	β_3: Water pH
β_4 : Total phosphorus concentration	β_4 : Ammonia nitrogen concentration
β_5: Nitrate NO_3^- concentration	β_5: Total phosphorus concentration
β_6: Nitrite NO_2^- concentration	β_6: Condutivity
β_7: Ammonia nitrogen concentration	β_7: Nitrate NO_3^- concentration
β_8: Kjeldahl nitrogen concentration	β_8: Total dissolved solids conc.
β_9: Turbidity	β_9: Nitrite NO_2^- concentration
	β_{10}: Turbidity

Table 1 - The β_j coefficients of the Bayesian model related to predictor variables for rivers models
Fonte: Authors own elaboration.

The coefficients values generated by the Markov Chain Monte Carlo (MCMC) algorithm, based on calibration data, are depicted graphically in Figure 1, which was generated by the PyMC library.On the left side of this figure, histograms with the randomly generated values obeying the characteristic reference values of mean of the chosen distributions, are presented. On the right side, the values generated (in sequential order) by Monte Carlo simulation are shown.

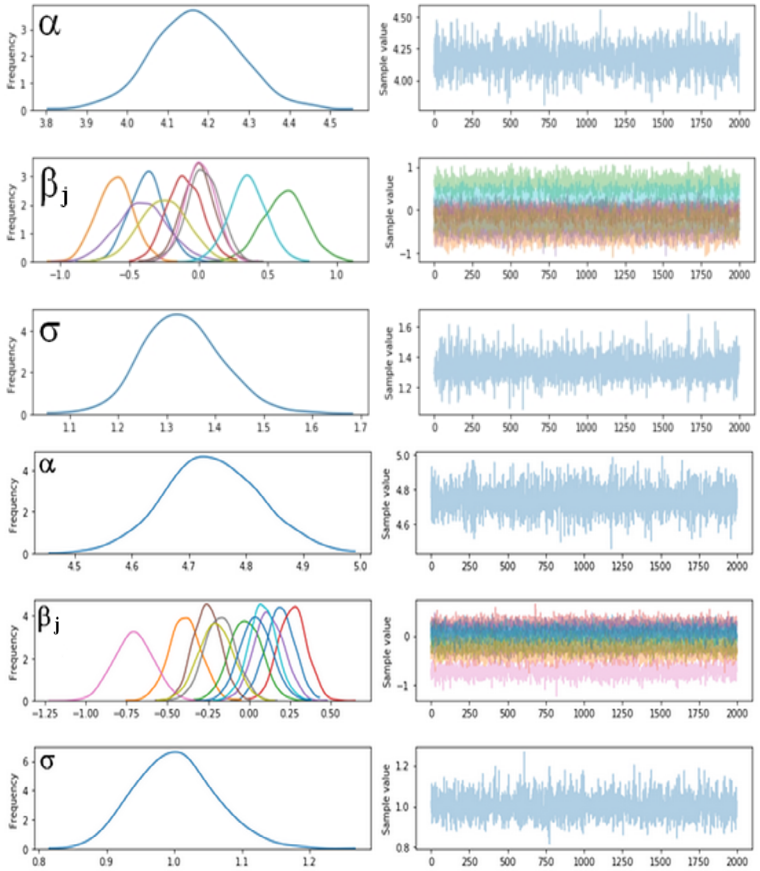

Figure 1. Gaussian FDP functions used in linear models parameters (left) randomly generated by Monte Carlo simulations, using values on the right (Piracicaba river – top, Paraiba do Sul river - bottom)
Fonte: Authors own elaboration.

Graphically, the model also determines the meaning of the coefficients according to the Gaussian distributions as shown in Figure 2 (graphics generated by PyMC library, for better data visualization and understanding of the model). Tables 2 and 3 show the values generated in Figure 2, for Piracicaba and Paraiba do Sul rivers respectively, including the means, standard deviations, standard error of the Monte Carlo simulation, the lower and upper limits of 95% confidence interval, all referring to Gaussian functions by *forestplot* PyMC function (Figure 2).

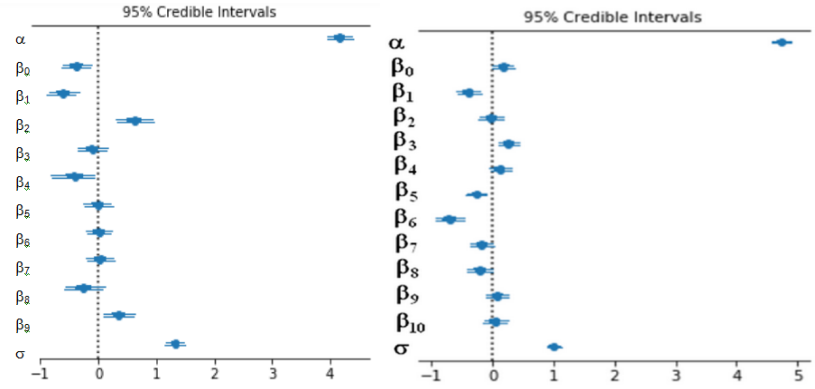

Figure 2. Mean values with error bars for to the coefficients used in the linear DO prediction model generated in Python by PyMC library (Piracicaba River – left, Paraiba do Sul River - right)
Fonte: Authors own elaboration.

	Mean	Std. Dev.	MC Error	Lower Lim.	Upper Lim.
α	4.150	0.125	0.002	3.928	4.395
β_0	-0.355	0.133	0.002	-0.604	-0.112
β_1	-0.673	0.150	0.002	-0.945	-0.388
β_2	0.647	0,171	0,003	0,322	0,967
β_3	-0.116	0,135	0,003	-0,379	0,130
β_4	-0.383	0,196	0,003	-0,767	-0,026
β_5	0.010	0,129	0,002	-0,229	0,251
β_6	0.019	0,121	0.001	-0.193	0.261
β_7	0.018	0.134	0.002	-0.239	0.267
β_8	-0.268	0.187	0.003	-0.596	-0.101
β_9	0.343	0.141	0.002	0.071	0.600
σ	1.401	0.093	0.001	1.239	1.590

Table 2 - Numerical values of the Bayesian model coefficients for Piracicaba river (Fig. 2 left)
Fonte: Authors own elaboration.

	Mean	St. Dev.	MC Error	Lower Lim.	Upper Lim.
α	4.708	0.089	0.001	4.542	4.874
β_0	-0.013	0.103	0.001	-0.213	0.170
β_1	-0.382	0.106	0.001	-0.577	-0.175
β_2	0.242	0.095	0.001	0.058	0.420
β_3	0.188	0.096	0.001	0.016	0.373
β_4	-0.692	0.127	0.002	-0.931	0.453
β_5	0.143	0.109	0.002	-0.065	0.350
β_6	-0.214	0.104	0.001	-0.405	-0.202
β_7	0.090	0.091	0.001	-0.075	0.268
β_8	-0.239	0.093	0.001	-0.408	-0.060
β_9	0.026	0.105	0.001	-0.157	0.234
β_{10}	-0.169	0.113	0.002	-0.157	0.234
σ	1.028	0.066	0.001	0.910	1.156

Table 3 - Numerical values of the Bayesian model coefficients for Paraiba do Sul river (Fig. 2 right)
Fonte: Authors own elaboration.

Figure 2 allows assessing the importance of the contribution of each predictor variable $\mathbf{X}_j$ in the model as a function of its associated β_j coefficient. The further away from the vertical line of ***zero***, the greater the contribution with positive or negative values of the predictive variables $\mathbf{X}_j$. For example, to the Piracicaba River, the variables X_5 (NO_2^-), X_6 (NO_3^-) and X_7 (ammoniacal nitrogen) have a small contribution to the model output, i.e. DO, in relation to the variables X_0 (BOD), X_1 (water temperature) and X_2 (river flow), which have higher average values. These variables X_5, X_6 and X_7 could even be removed from the model without significantly affecting the results. There is a similar behavior for the β_j coefficients for the Paraiba do Sul River. Figure 2 allows an individual assessment of each variable $\mathbf{X}_j$ and its importance for the model. This is a unique feature of the PyMC library.

Finally, the Bayesian model calculates the sample distribution of the Y results, in the data set, using a likelihood normal distribution function and mean parameter μ and standard deviation σ (eq. 2). Then, it estimates the posterior probability for the unknown DO variables in the model (generically, it calculates the distribution of probabilities of the coefficients, which are used to predict via the regression model, the unknown Y values for new values of the regressors $\mathbf{X}_j$). Optimization methods based on samples taken from the posterior distribution using Markov Chain Monte Carlo (MCMC) sampling are used - NUTS default method (modification of the Hamiltonian Monte Carlo) at PyMC library (Salvatier; Wiecki; Fonnesbeck, 2016).

In this work, 128 samples were used to calibrate the Piracicaba River, and 20 samples to predict DO values and obtained a RMSEp of 0.835 mg L^{-1} and a coefficient R^2 = 0.6736, when comparing the real values against those calculated, as shown in Figure 3. For the Paraiba do Sul River, 135 samples were used to calibrate the model and 20 samples to predict DO values and it obtained a RMSEp of 0.839 mg L^{-1} and a coefficient R^2 = 0.7677, when comparing the real values against those calculated, as shown in Figure 4. Silva e Schimidt (2016) obtained results very similar to this one, including the same dispersion of Figure 4, but using an artificial neural network back propagation model, for the same data. The dispersion observed in Figures 3 and 4 shows that a large majority of the calculated values are very close to the line and, within the expectation, they should have a low prediction error. There are also several values far from it contributing to the increase of the RMS error that could have been discarded in the construction of the model, but all original samples were used.

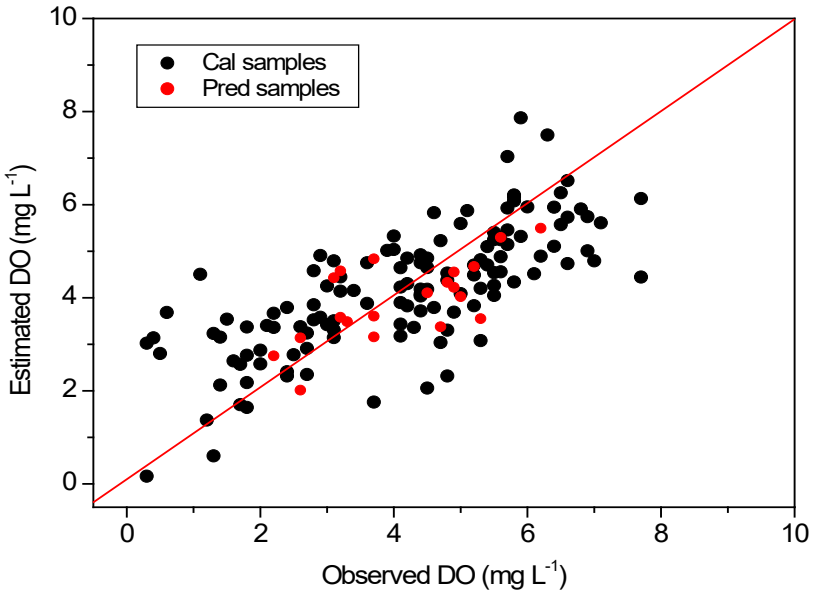

Figure 3. Dispersion of 148 samples in relation to the DO obtained from the Bayesian probabilistic model for the Piracicaba River, comparing the observed and the calculated values
Fonte: Authors own elaboration.

To compare the efficiency and robustness of Bayesian probabilistic models, some statistical parameters were compared (Table 4). The RMS calibration and prediction errors, in addition to the R^2 coefficient (Rcal and Rpred) between the predicted values by the models against the real values (Figures 3 and 4) and the application of a back propagation neural network to compare the same parameters. The architecture of the neural networks were built for the Piracicaba River was 10-5-1, and for the Paraiba do Sul River was 11-6-1. The number of neurons in the hidden layer is the meaning

of the neurons in the sum of input and output layers (Heaton, 2011). The box plot of residuals from the BPM models are described at https://github.com/Schimidt99/Bayesian-probabilistic-model.

	BPM model		ANN model		PyCaret library	
	Piracicaba River	Paraiba do Sul River	Piracicaba River	Paraiba do Sul River	Piracicaba River	Paraiba do Sul River
RMSEc (mg L-1)	1.330	0.972	1.403	1.222	1.327	0.844
RMSEp (mg L-1)	0.835	0.839	0.855	1.05	1.326	0.984
Rcal	0.7067	0.7686	0.6825	0.6538	0.4524	0.6194
Rpred	0.6736	0.7677	0.6305	0.7017	0.4521	0.5140

Table 4 - Comparative statistical results between the models
Fonte: Authors own elaboration.

Results show that the RMS errors and the R coefficients for the Bayesian model and for the neural network are similar (Table 4). Considering that every neural network has several initialization parameters, and all must be adjusted, mainly to avoid problems of underfitting and overfitting, the Bayesian model can be built and adjusted in a much simpler and faster way, in relation to an artificial neural network. The models chosen by the PyCaret library, using the same proportion of samples for training/prediction, presented bad results for the statistical parameters evaluated in both 'modern' optimized models, mainly the R coefficient. The best models optimized were Random Forest Regressor, for data from the Piracicaba River, and Extra Tree Regressor, for data from the Paraiba do Sul River.

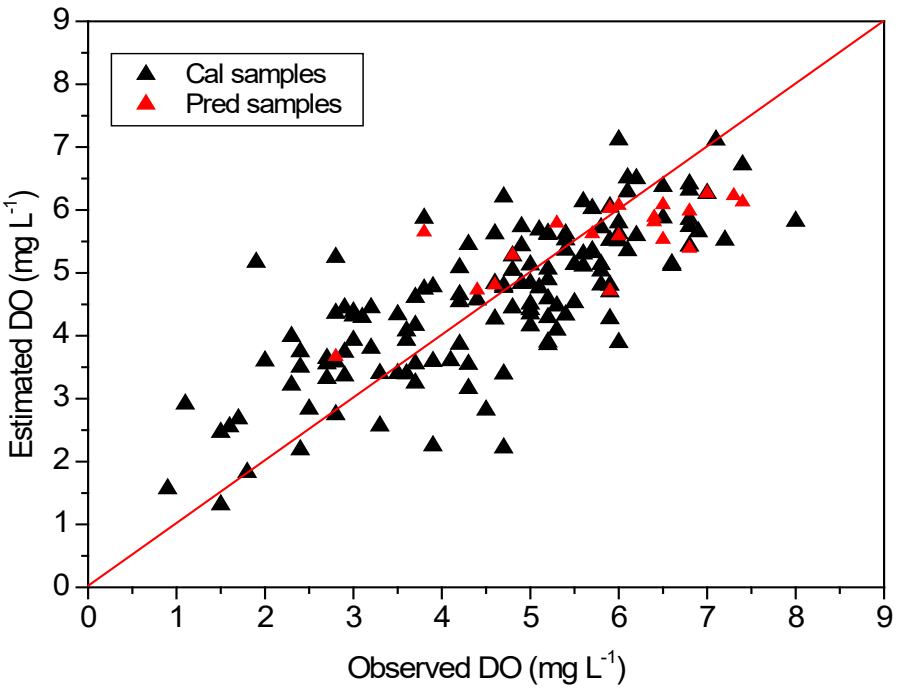

Figure 4. Dispersion of 155 samples in relation to the DO obtained from the Bayesian probabilistic model for the Paraiba do Sul River, comparing the observed and the calculated values
Fonte: Authors own elaboration.

The prediction errors obtained by the Bayesian probabilistic model are equivalent to other works done with artificial neural networks. Khani e Rajaee (2017) studied the DO concentration in a river, testing several mathematical models, including some neural networks with peculiar algorithms and obtaining RMS prediction errors of 0.744 mg L^{-1} and 0.803 mg L^{-1}. Chen e Liu (2014) studied the water quality of reservoirs in Taiwan through DO concentration. They obtained RMS prediction errors from 0.52 to 0.98 mg L^{-1} among the different neural networks' algorithms. Antanasijević et al. (2013) studied different algorithms of neural networks to predict DO concentration in the Danube River in Serbia. They obtained RMS prediction errors from 0.59 to 0.83 mg L^{-1} among the tested algorithms. Khan e Valeo (2017) did two regression models, a Bayesian linear and a fuzzy linear, which were constructed for different scenarios. They used an autoregressive model to uncertainty quantification in DO prediction at Bow River in Calgary, Canada, using only DO historic data. They obtained Mean Squared Errors from 0.55 to 36.5 mg L^{-1} among the two tested models.

Another non-linear environmental data was tested: a BPM model for concentration of benzene in air, measured by electronic nose sensors, as monitoring air pollution (De Vito, 2008), it was also constructed with very good results, RMSEp of 0.584 µg m^{-3} and R coefficient of 0.9961. The 826 samples were used to calibrate the model and 50 samples for prediction/testing, which are detailed in the Supplementary Material. All full Bayesian models developed in Python can be accessed at https://github.com/Schimidt99/Bayesian-probabilistic-model.

In both models for the two rivers studied, there is a great mathematical contribution of the input variables BOD, water temperature and river flow, in the relationship with DO. According to Ross and Stock (2019) and Emamgholizadeh et al. (2014), the variables BOD and water temperature are important environmental parameters that affect the direct measurement of DO in a river. The influence of water temperature on the dissolution of gases is also well known (Ross; Stock, 2019; Manaham, 2013). Other variables used in the models showed different importance for each of the rivers, probably due to the specific characteristics of each river, geographic location and the soil characteristics (chemical composition).

Conclusion

The Python programming language in conjunction with its mathematical packages provides powerful tools for data modeling. One of the significant advantages of Bayesian regression is its ability to provide uncertainty estimates along with predictions, for example, by previously choosing values for the standard deviation σ (eq. 2). Traditional regression models, including neural networks, often output a point estimate without any indication of the confidence or uncertainty associated with the prediction. Bayesian models, on the other hand, provide a probability distribution over possible parameter values, allowing for a more nuanced understanding of uncertainty. Bayesian models can be easily adapted to different problem settings by choosing appropriate prior distributions. This makes them flexible and suitable for a wide range of applications. Finally, Bayesian models naturally incorporate regularization through the choice of prior distributions. This regularization helps prevent problems with overfitting situations.

Mathematical statistical models associated with Bayesian inferences can be applied to the water quality analysis of a river, allowing to relate time series information from analyzed variables with the water quality. The PyMC library is easier and simpler

to use than other equivalent languages such as WinBUGS, JAGS and Stan. This library allows a simple and fast visualization of the results through its own graphs, tables and the export of the same in TXT or CSV files. Based on the results above, we can say that the Bayesian probabilistic model allows us to analyze how each input variable X_j relates to (affects) the output Y (Figure 2). This is not possible in other multivariate models, except for the Principal Component Analysis (PCA) where the detailed study of the *loadings* can allow these observations. The use of these Bayesian probabilistic models in water quality studies could allow the visualization of extreme changes due to severe climate processes or anthropogenic activities (simulating different probability distributions of variables) and consequently changes in prediction of the results. Although the models used in this paper all have variables with gaussian distribution, the probabilistic programming of the PyMC library allows other distributions to be used, not being limited only to gaussian processes. Knowledge of the behavior of the variables helps in this step. It is possible to study the temporal variation of the variable to be analyzed. A histogram of the distribution (range of values) of the variable X_j as a function of the frequency count of the values must be constructed. The distribution of the values will make it possible to identify the most appropriate PDF, although other processes to do this can be used. The physico-chemical parameters of the existing Brazilian environmental legislation by Conama, and any other control agencies in the world, could be followed in real time, as well as projections through predictions of anomalous situations. The models would allow predictions about the concentration of dissolved oxygen (based on other physico-chemical parameters analyzed in the water) and even projections of strategies to minimize the cost of water quality treatment and to mitigate situations of resource degradation, contamination and preservation.

References

ANA (Agência Nacional de Águas). *Hidroweb Mobile*, [*s. l.*], 2023. Disponível em: https://www.snirh.gov.br/hidroweb-mobile/mapa. Acesso em: nov. 2023.

ANACONDA. *Anaconda Distribution*, [*s. l.*], 2023. Disponível em: https://www.anaconda.com/download. Acesso em: nov. 2023.

ANTANASIJEVIĆ, D.; POCAJT, V.; POVRENOVIĆ, D.; PERIĆ-GRUJIĆ, A.; RISTIĆ, M. Modelling of dissolved oxygen content using artificial neural networks: Danube River, North Serbia, case study. *Environmental Science and Pollution Research*, [*s. l.*], v. 20, p. 9006-9013, 2013. DOI: https://doi.org/10.1007/s11356-013-1876-6.

BAIRD, C.; CANN, M. *Química Ambiental*. 4. ed. Porto Alegre: Bookman, 2011.

CETESB (Companhia Ambiental do Estado de São Paulo). *InfoÁguas*, São Paulo, 2023. Disponível em: https://cetesb.sp.gov.br/infoaguas/. Acesso em: nov. 2023.

CHEN, W. B.; LIU, W. C. Artificial neural network modeling of dissolved oxygen in reservoir. *Environmental Monitoring and Assessment*, [*s. l.*], v. 186, n. 2, p. 1203-1217, 2014. DOI: https://doi.org/10.1007/s10661-013-3450-6.

CONAMA (Conselho Nacional do Meio Ambiente). *Resolução CONAMA nº 357, de 17 de março de 2005*. Brasília, DF: Ministério do Meio Ambiente, 2012. Disponível em: http://conama.mma.gov.br/images/conteudo/LivroConama.pdf. Acesso em: 27 nov. 2023.

DAVIDSON-PILON, C. *Bayesian Methods for Hackers*: Probabilistic Programming and Bayesian Inference. Crawfordsville: Addison-Wesley, 2016.

DE VITO, S.; MASSERA, E.; PIGA, M.; MARTINOTTO, L.; DI FRANCIA, G. CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic Bayesian regularization. *Sensors and Actuators B: Chemical*, [*s. l.*], v. 143, n. 1, p. 182-191, 2009. DOI: https://doi.org/10.1016/j.snb.2009.08.041.

DE VITO, S.; MASSERA, E.; PIGA, M.; MARTINOTTO, L.; DI FRANCIA, G. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, [*s. l.*], v. 129, n. 2, p. 750-757, 2008. DOI: https://doi.org/10.1016/j.snb.2007.09.060.

EMAMGHOLIZADEH, S.; KASHI, H.; MAROFPOOR, I.; ZALAGHI, E. Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models. International Journal of Environmental Science and Technology, [*s. l.*], v. 11, p. 645-656, 2014. DOI: https://doi.org/10.1007/s13762-013-0378-x.

GRAF, R. Distribution Properties of a Measurement Series of River Water Temperature at Different Time Resolution Levels (Based on the Example of the Lowland River Notec, Poland). *Water*, [*s. l.*], v. 10, p. 203-210, 2018. DOI: https://doi.org/10.3390/w10020203.

HEATON, J. *Programming Neural Networks with Encog3 in C#*. 2. ed. Chesterfield: Heaton Research Inc, 2011.

KENNARD-STONE Algorithm. *Nirpy Research*, [*s. l.*], 2022. Disponível em: https://nirpyresearch.com/kennard-stone-algorithm/. Acesso em: nov. 2023.

KHAN, U. T.; VALEO, C. Comparing A Bayesian and Fuzzy Number Approach to Uncertainty Quantification in Short-Term Dissolved Oxygen Prediction. *Journal of Environmental Informatics*, [*s. l.*], v. 30, n. 1, p. 1-16, 2017. DOI: https://doi.org/10.3808/jei.201700371.

KHANI, S.; RAJAEE, T. Modeling of Dissolved Oxygen Concentration and Its Hysteresis Behavior in Rivers Using Wavelet Transform-Based Hybrid Models. *Clean Soil Air Water*, [*s. l.*], v. 45, n. 2, 2017. DOI: https://doi.org/10.1002/clen.201500395.

LYON, A. Why are Normal Distributions Normal?. *The British Journal for the Philosophy of Science*, [*s. l.*], v. 65, p. 621-649, 2014. DOI: https://doi.org/10.1093/bjps/axs046.

MANAHAM, S. E. *Química Ambiental*. 9. ed. Porto Alegre: Bookman, 2013.

MARTIN, O. *Bayesian Analysis with Python*. Birmingham: Packt Publishing, 2016.

MC ELREATH, R. *Statistical Rethinking*: A Bayesian Course with Examples in R and Stan. New York: Chapman and Hall/CRC, 2015. 483 p.

DOCUMENTAÇÃO. *PyCaret Library*, [*s. l.*], 2023. Disponível em: https://pycaret.gitbook.io/docs/get-started/tutorials. Acesso em: 27 nov. 2023.

QIAN, S. S.; RECKHOW, K. H.; ZHAI, J.; MCMAHON, G. Nonlinear regression modeling of nutrient loads in streams: A Bayesian approach. *Water Resources Research*, [*s. l.*], v. 41, 2005. DOI: https://doi.org/10.1029/2005WR003986.

ROSS, A. C.; STOCK, C. A. An assessment of the predictability of column minimum dissolved oxygen concentrations in Chesapeake Bay using a machine learning model. *Estuarine, Coastal and Shelf Science*, [*s. l.*], v. 221, p. 53-65, 2019. DOI: https://doi.org/10.1016/j.ecss.2019.03.007.

SALVATIER, J.; WIECKI, T. W.; FONNESBECK, C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, [*s. l.*], v. 2, e55, 2016. DOI: https://doi.org/10.7717/peerj-cs.55.

SILVA, S. R.; SCHIMIDT, F. Reduction of Artificial Neural Network Input Variables from Principal Component Analysis Data in Dissolved Oxygen Modeling. *Química Nova*, São Paulo, v. 39, p. 273-278, 2016. DOI: https://doi.org/10.5935/0100-4042.20160024.

VON SPERLING, M. *River Water Quality Studies and Modeling*. 2. ed. Belo Horizonte: Editora UFMG, 2014.

WOOLEY, J. C.; LIN, H. S. (eds.). *Catalyzing Inquiry at the Interface of Computing and Biology*. Washington (DC): National Academies Press (US), 2005.