PREDICTIVE MODELING IN THE TIETE RIVER (2014–2024), SÃO PAULO STATE, BRAZIL

[1]ISABELA GUIMARAES DE ARAUJO, [2]DANIEL MARCOS BONOTTO

[1],[2]Sao Paulo State University
E-mail: [1]isabela.guimaraes@unesp.br, [2]daniel.bonotto@unesp.br

Abstract - This study presents a predictive modeling analysis of potentially toxic elements (PTEs) in the Tietê River between 2014 and 2024. The Support Vector Machine (SVM) algorithm was employed to estimate future trends in contaminant concentrations using historical monitoring data. Although the limited dataset prevented the development of a fully conclusive predictive model, the results highlight the occurrence of concentrations exceeding the threshold values established by environmental standards, such as CONAMA Resolution No. 357 and the Brazilian Ministry of Health (2021). These findings emphasize the importance of continuous monitoring and management of water quality to prevent the persistence or intensification of contamination levels in the river system.

Keywords - Machine learning; Predictive modeling; Support Vector Machine; Tietê River.

I. INTRODUCTION

The importance of water for maintenance and continuity of life is widely recognized. Surface waters stand out not only for their relevance in supplying human populations but also for their accessibility, since their use does not require drilling processes for extraction [1]. However, this easy access also stimulates discussions on conservation, exploitation, and evaluation of water quality, as accessibility implies a higher susceptibility to environmental degradation. This reinforces the need for actions aimed at conservation, sustainable use, and continuous monitoring of water resources [2].

Beyond their vital importance, watercourses reflect the dynamics of their respective watersheds, encompassing natural attributes and the interactions among their various components, such as soil, water, geology, and vegetation cover, as well as the anthropogenic alterations to which they are subjected [3, 4]. In densely populated regions, particularly metropolitan areas, water bodies are frequently exposed to multiple sources of contamination and vulnerability resulting from domestic consumption, urban expansion, and industrial activities. These pressures often exceed the carrying capacity of ecosystems, contributing to the degradation of water sources and rendering them unsuitable for multiple uses [5, 6].

The Tietê River, one of the main watercourses in the state ofSão Paulo, extending for more than one thousand kilometers from the southwest to the northwest, is one of the most emblematic cases of how unregulated occupation of riverbanks contributes to the degradation of water sources. Human occupation of the Tietê River margins began as early as the seventeenth century, with the Paulista expeditions. During the nineteenth century, with industrialization, the river began receiving large volumes of untreated domestic and industrial wastewater. Despite Article 208 of the Constitution of the State of São Paulo [7], promulgated in 1989, which prohibits the discharge of untreated urban and industrial effluents into water bodies, less than 80% of the sewage collected in the Metropolitan Region of São Paulo is effectively treated, according to data from [8, 9].

In this context, the present study is based on the Support Vector Regression (SVR) methodology as a predictive modeling tool to better understand the behavior of both naturally occurring parameters and contaminant indicators present in the waters of the Tietê River. The application of this type of statistical modeling technique, combined with machine learning, in environmental studies has proven effective in identifying trends and anticipating impacts, as well as supporting mitigation actions [10, 11].

II. METHODOLOGY

The study area for this research is in the state of São Paulo along the Tietê River (Figure 1), covering the period from 2013 to 2024. All data used in this study were obtained from InfoÁguas, a publicly accessible initiative by CETESB.

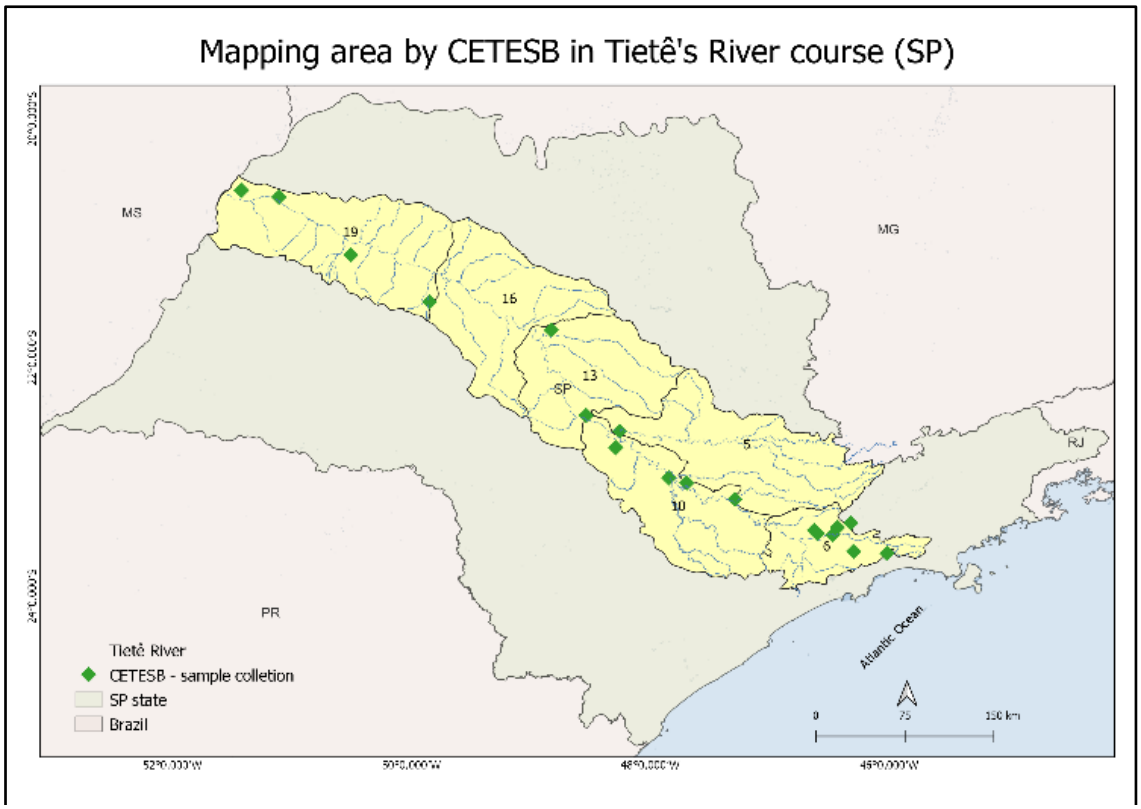

Figure 1- Mapping area in Tietê's River by CETESB. The yellow area is the Tietê Water Resources Management Units (UGRHI). The green points are the sample collected by CETESB.

Based on the obtained database, the data were tabulated and organized following the course of the river, from the point closest to its source (Biritiba Mirim) to its mouth at the Paraná River in Itapura. All sampling points include seasonal information, except for the year 2020 due to the COVID-19 pandemic.

The Tietê River comprises five Water Resources Management Units (UGRHIs), which extend from its headwaters to its downstream course. Table 1 presents the code of each unit, as well as the municipalities included and analyzed in this study.

Municipality	UGRHI Name	UGRHI Code
Biritiba Mirim	Alto Tietê	6
Mogi das Cruzes	Alto Tietê	6
Suzano	Alto Tietê	6
Itaquaquecetuba	Alto Tietê	6
Guarulhos	Alto Tietê	6
São Paulo[1]	Alto Tietê	6
São Paulo[2]	Alto Tietê	6
São Paulo[3]	Alto Tietê	6
Salto	Médio Tietê / Sorocaba	10
Tietê	Médio Tietê / Sorocaba	10
Laranjal Paulista	Médio Tietê / Sorocaba	10
Botucatu[1]	Médio Tietê / Sorocaba	10
Botucatu[2]	Médio Tietê / Sorocaba	10
Barra Bonita	Médio Tietê / Sorocaba	10
Ibitinga	Tietê-Jacaré	13
Promissão	Tietê-Jacaré	13
Araçatuba	Baixo Tietê	19
Pereira Barreto	Baixo Tietê	19
Itapura	Baixo Tietê	19

Table 1 - Tietê River basins analyzed in this study.

São Paulo[1]: Aricanduva Avenue bridge
São Paulo[2]: Remédios Bridge, Marginal Avenue
São Paulo[3]: Bandeiras Bridge, Santos Dumont Avenue
Botucatu[1]: Bridge on SP-191 highway
Botucatu[2]: Downstream of the confluence between Tietê and Piracicaba Rivers

The water samples included chemical and physicochemical parameters, encompassing metals and metalloids, hydrocarbon derivatives, field parameters, bioanalytical, microbiological, ecotoxicological, and hydrobiological indicators. However, for the predictive modeling purposes of this study, only physicochemical and chemical parameters were considered, including metals, metalloids, hydrocarbon derivatives, and compounds used in agricultural pesticides.

For the modeling, the Support Vector Regression (SVR) methodology was adopted. SVR is an extension of the Support Vector Machine (SVM) and is a well-established technique originally developed in Russia by [12, 13]. The model has been refined over the decades with contributions from other scientists in the field, and the current implementation follows the concept developed by [14]. The choice of this algorithm is justified by the nature of the available dataset: although the database is rich in information, it is relatively small in terms of the number of observations, making SVR particularly suitable for the predictive modeling in this study.

The basic idea of SVR is to find a function f(x) that deviates by at most ε from the target values yi for all training data, while keeping the function as flat as possible, meaning that errors should not exceed the deviation ε [15]. Equation 1 describes the general form of the function:

$$f(x) = \langle w, x \rangle + b, w \in X, b \in R \qquad \text{(Eq. 1)}$$

where $\langle w, x \rangle$ denotes the dot product in X representing the linear model. The objective is to find w with minimal variation, expressed as $\| w \|^2 = \langle w, w \rangle$. This problem can be formulated as a convex optimization task, aiming to minimize the norm of w while ensuring that deviations between the predicted values and actual y_i do not exceed ε. Equation 2 formalizes this problem:
Minimize:

$$\frac{1}{2} \| w \|^2$$

Subject to:

$$y_i - \langle w, x_i \rangle - b \leq \varepsilon$$
$$\langle w, x_i \rangle - b \leq \varepsilon \langle w, x_i \rangle + b - y_i \leq \varepsilon \qquad \text{(Eq. 2)}$$

The formulation presented by [14], described in Equation 3, introduces slack variables ξi and ξi ∗ to allow some errors, making the problem feasible and flexible, as it is intended for prediction. These slack variables act similarly to a loss function described by [16].

Minimize:

$$\frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i *)$$

Subject to:

$$y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i y$$
$$\langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i *$$
$$\xi_i, \xi_i * \geq 0 \quad \text{(Eq. 3)}$$

Smola &Schölkopf [15] define the constant C as a trade-off between the flatness of f(x) and the tolerance for deviations larger than ε. This concept is related to the ε -insensitive loss function, shown in Equation 4, where ξ is the slack variable associated with the prediction error for any point (x_i, y_i) Errors smaller than or equal to ε are not penalized, while errors larger than ε are penalized linearly. The constant C controls this penalization; smaller values of C allow higher tolerance for errors.

$$|\xi|_\varepsilon = \{0, \text{if } |\xi| \leq \varepsilon;\ |\xi|, \text{if } |\xi| > \varepsilon\} \quad \text{(Eq. 4)}$$

Figure 2 visually illustrates the four equations described above.

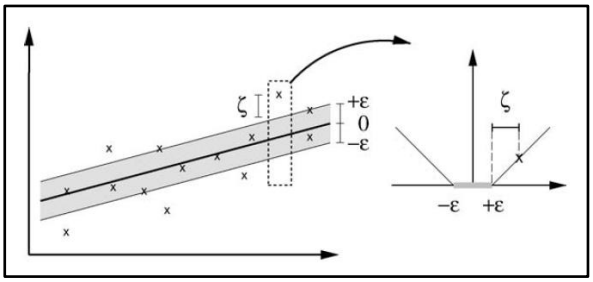

Figure 2-Soft margin model illustrating the flatness of a function $f(x)$. The image demonstrates how errors within the tolerated ε range are important for predicting values using machine learning[15].

III. RESULTS AND DISCUSSION

Initially, the data obtained from the InfoÁguas system were categorically organized by municipality, covering the period from 2013 to 2024. However, the first notable issue was that the municipality of Itapura, where the Tietê River discharges into the Paraná River, only has data available starting in 2014. This results in a gap between municipalities and hinders the agency's monitoring.

Basic physicochemical and chemical parameters, as expected in natural water composition, were analyzed across all sixteen municipalities included in the InfoÁguas database. However, certain chemical parameters are indicative of anthropogenic contamination. Nevertheless, despite the lack of analyses in certain regions, predictive modeling of contamination was still performed, along with the modeling of naturally occurring chemical parameters

in water. These results were compared against regulatory limits established by public agencies, including the Brazilian Ministry of Health and CONAMA Resolution 357 [2, 17] as illustrated in Tables 2 and 3.

Municipality	Chromium (mg/L)	Nickel (mg/L)	Potassium (mg/L)
Biritiba Mirim	0.020	0.020	2.010
Mogi das Cruzes	0.022	0.020	3.030
Suzano	0.023	0.020	8.510
Itaquaquecetuba	0.022	0.020	8.556
Guarulhos	0.030	0.060	13.114
São Paulo[1]	0.032	0.030	15.147
São Paulo[2]	0.034	0.030	12.725
São Paulo[3]	0.034	0.040	14.668
Salto	0.050	0.020	11.323
Tietê	0.020	0.020	11.523
Laranjal Paulista	0.020	0.020	10.325
Botucatu[1]	0.040	0.020	8.365
Botucatu[2]	0.020	0.020	7.425
Barra Bonita	0.020	0.020	7.740
Ibitinga	0.020	0.010	6.651
Promissão	0.020	0.010	6.060
Araçatuba	0.020	0.010	5.989
Pereira Barreto	0.020	0.010	5.345
Itapura	0.020	0.020	5.445
CONAMA 357 Limit (mg/L)	0.050	0.025	N.E.
GM/MS Nº888 Limit (mg/L)	0.050	0.070	N.E.

Table 2- Analysis of Parameters Indicative of Anthropogenic Contamination Along the Tietê River Compared with Limit Values Established by CONAMA and the Brazilian Ministry of Health [2, 17].

Table 2 presents the average concentrations of Chromium, Nickel, and Potassium in comparison with the limit values established by CONAMA and GM/MS No. 888. While both Chromium and Nickel, as transition metals, generally show low concentrations below the public health thresholds, Nickel levels at the sampling sites in Guarulhos and São Paulo exceed the CONAMA limits, indicating potential contamination sources in these areas. Additionally, Chromium in Salto reaches the established limit, which may also reflect a contamination source and highlights a concern for public health.Potassium, although not subject to any established regulatory limits, exhibits elevated concentrations that suggest anthropogenic activity, potentially related to agricultural practices.

Municipality	Aluminum (mg/L)	Lead (mg/L)	Phosphorus (mg/L)
Biritiba Mirim	0.18	0.01	0.06
Mogi das Cruzes	0.18	0.01	0.1
Suzano	0.2	0.01	0.71
Itaquaquecetuba	0.15	0.01	0.69
Guarulhos	0.15	0.01	1.7
São Paulo[1]	0.14	0.01	1.8
São Paulo[2]	0.15	0.01	1.74

São Paulo[3]	0.13	0.01	1.64
Salto	0.16	0.01	1.53
Tietê	0.12	0.01	1.33
Laranjal Paulista	0.15	0.01	1.18
Botucatu[1]	0.14	0.01	0.35
Botucatu[2]	0.13	0.01	0.43
Barra Bonita	0.12	0.01	0.21
Ibitinga	0.08	0.01	0.08
Promissão	0.08	0.01	0.04
Araçatuba	0.08	0.01	0.03
Pereira Barreto	0.08	0.01	0.43
Itapura	0.08	0.01	0.43
CONAMA 357 Limit (mg/L)	0.1	0.01	0.025
GM/MS Nº888 Limit (mg/L)	0.2	0.01	N.E.

Table 3- Analysis of Parameters Indicative of Anthropogenic Contamination Along the Tietê River Compared with Limit Values Established by CONAMA and the Brazilian Ministry of Health.

Table 3 presents the average concentrations of Aluminum, Lead, and Phosphorus. It can be observed that, except for Phosphorus, the average values are generally low. However, the mean Aluminum concentrations exceed the limits established by CONAMA at nearly all sampling points. Lead exhibits mean values equal to the established limits, and due to their toxicity, it should be continuously monitored to prevent alarming levels. Phosphorus concentrations surpass the limits set by CONAMA; the Ministry of Health (2021) has not established a safe limit for water. Given its low mobility, the elevated concentrations are considered anomalies and may result from human activities aimed at soil enrichment for agricultural purposes.

Upon examining Tables 2 and 3, it is noteworthy that although the average values generally exhibit homogeneity, certain results stand out due to elevated concentrations and, in some cases, surpass at least one of the threshold values established by regulatory agencies. Figures 3 to 7 present the main variations together with their respective predictive modeling.

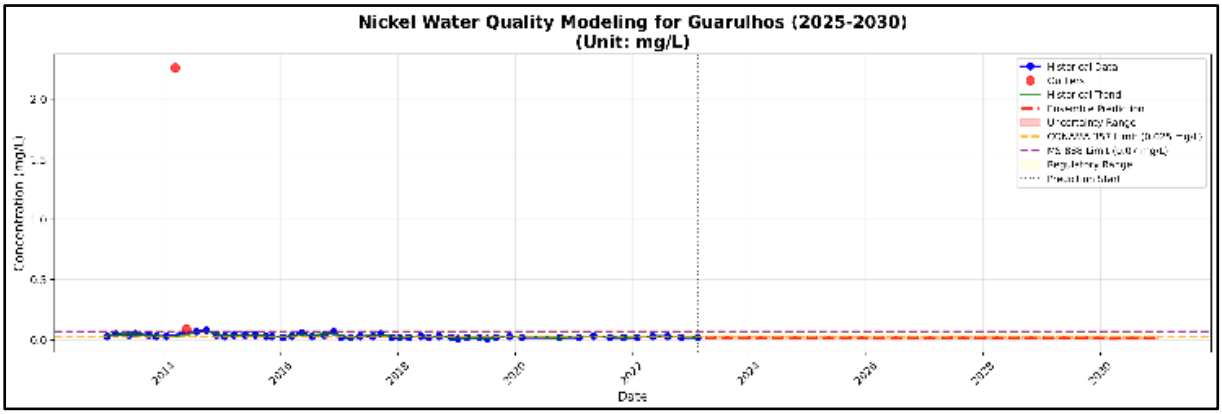

Figure 3 - Variation and predictive modeling of nickel in Guarulhos based on the CETESB database.

Figure 3 illustrates the variation of Nickel concentrations in Guarulhos (2013–2024) along with the predictive modeling. Given the municipality's historical record showing generally homogeneous concentrations of this parameter, with only a single outlier, the model tends to remain relatively flat, assuming the discharge patterns of the river remain consistent. Continuous monitoring of Nickel is crucial, as it is a toxic metal that poses risks to both human health and the environment.

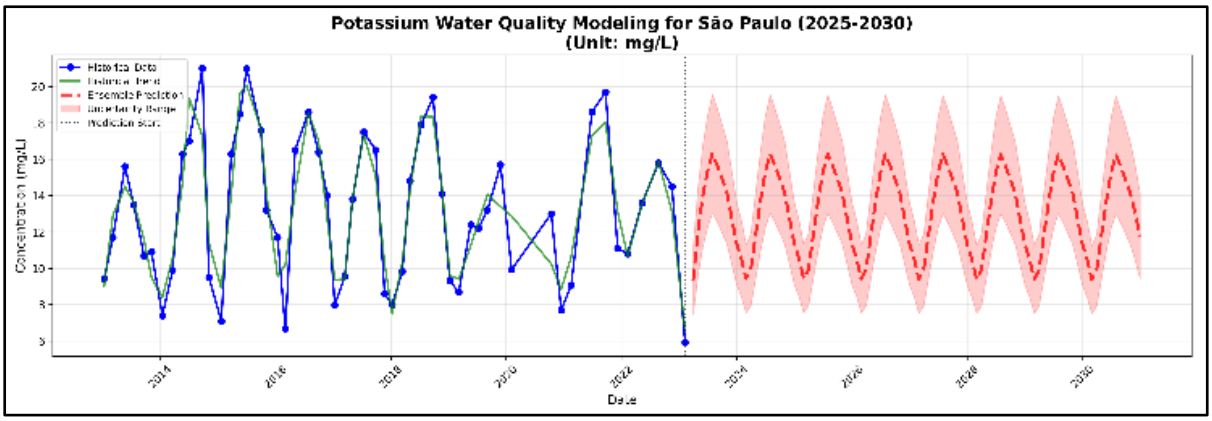

Figure 4 - Variation and predictive modeling of potassium in São Paulo based on the CETESB database.

Figure 4 depicts the temporal variation of Potassium in the city of São Paulo (2014–2024). The concentrations of this parameter show considerable variability, and due to the relatively short time span, the predictive algorithm yields a model of limited accuracy, as it interprets the contamination pattern as fluctuating between higher and lower values.

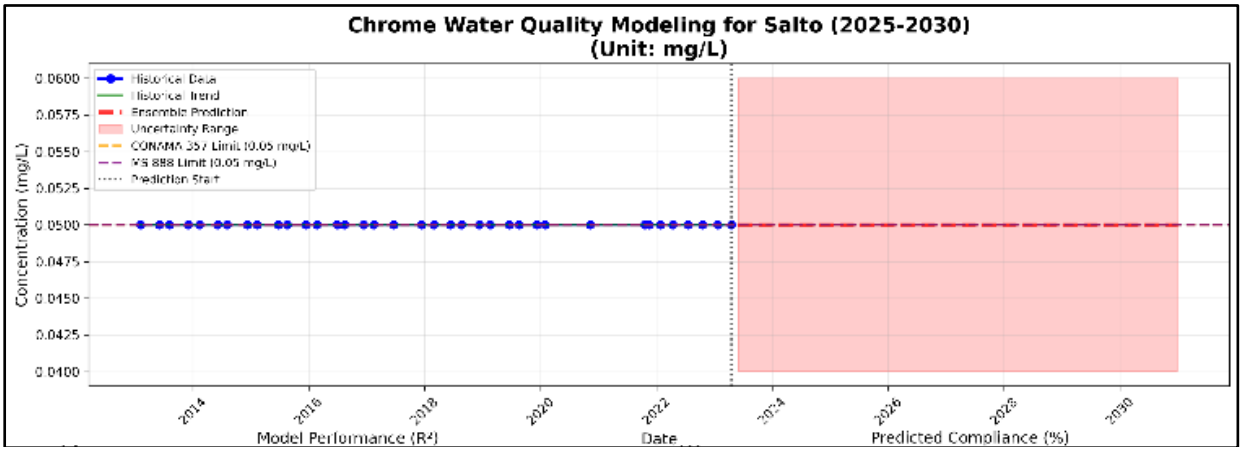

Figure 5 - Variation and predictive modeling of chromium in Salto based on the CETESB database.

Figure 5 shows the variation and predictive modeling of Chromium in the municipality of Salto. Like Figure 3, the annual mean values are relatively homogeneous, resulting in a nearly flat model. However, the acceptable range of variation introduces uncertainty, making the predictive trend somewhat inconclusive.

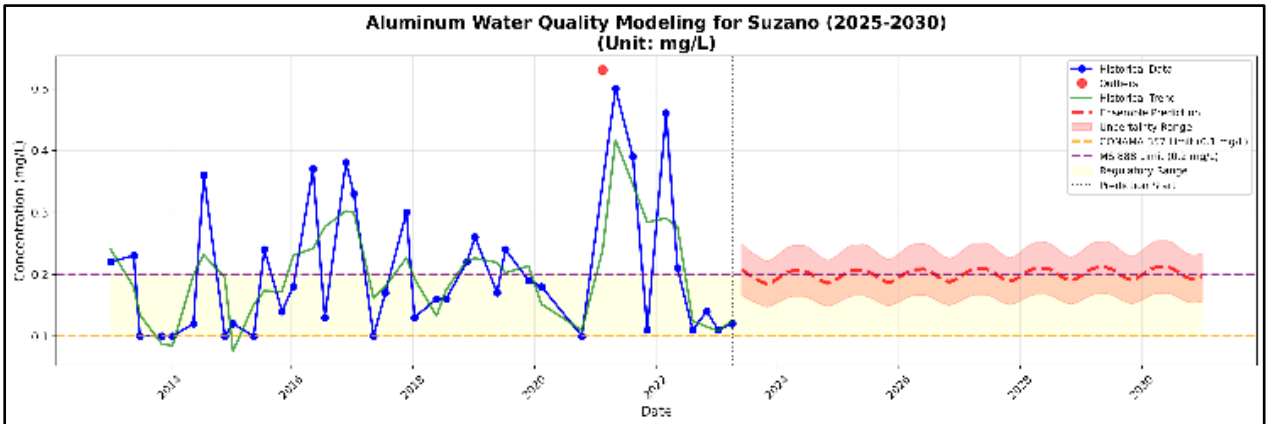

Figure 6 - Variation and predictive modeling of aluminum in Suzano based on the CETESB database.

Figure 6 presents the variation and predictive modeling of Aluminum in the municipality of Suzano. Initially, the concentrations appear highly heterogeneous, with values equal to or above the thresholds established by regulatory agencies. The model also projects concentrations above these limits but without pronounced peaks,

indicating that the predictive performance was limited, as it reflects mainly the average annual behavior rather than capturing the variability of the data.

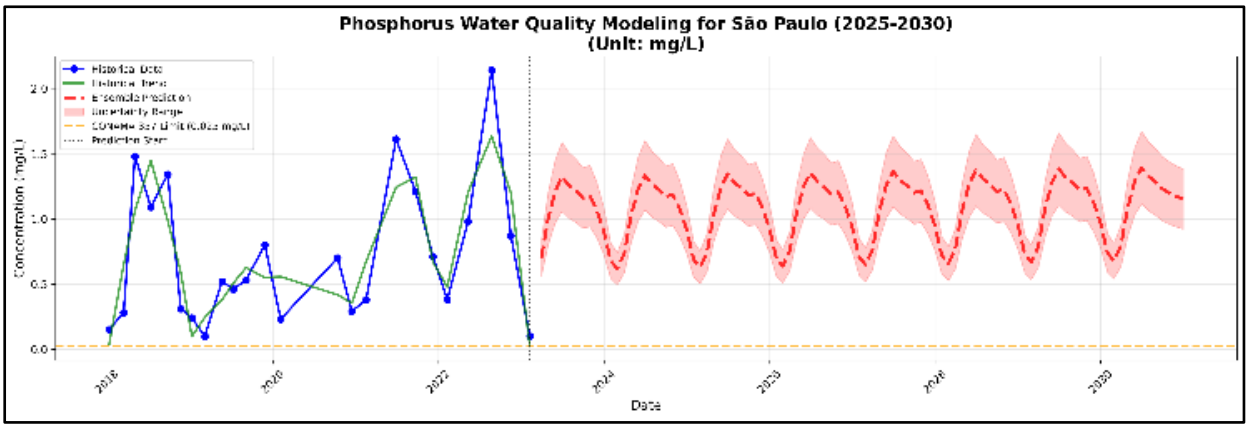

Figure 7- Variation and predictive modeling of phosphorus in São Paulo based on the CETESB database.

Figure 7 shows the temporal variation of Phosphorus in the city of São Paulo between 2014 and 2024, along with its predictive modeling based on historical data. The concentrations are not homogeneous and exceed the pre-established threshold values. However, similar to Figure 6, the model does not provide reliable results, as it reproduces a "zigzag" pattern of the data, making the prediction uncertain.

Overall, the predictive modeling was not particularly successful. This limitation is primarily due to the database, which, although covering seasonal values over a decade, was reduced to annual averages, thereby disregarding seasonality. In addition, external factors that may influence concentration levels were not considered. Events such as the COVID-19 pandemic and contamination episodes associated with both public entities, such as SABESP [18] and private sectors also play a significant role.

The authors would like to acknowledge funding from FAPESP (São Paulo State Research Support Foundation) and PRH 40 (Human Resources Improvement Program).

IV. CONCLUSIONS

This study assessed predictive modeling of selected parameters. The results showed that while Chromium and Nickel generally presented low average concentrations, but exceedances were observed in Guarulhos, São Paulo, and Salto, indicating potential contamination risks. Aluminum displayed heterogeneous values, frequently surpassing regulatory thresholds, whereas Phosphorus consistently exceeded established limits, and Potassium, although lacking defined guidelines, revealed high concentrations linked to agricultural practices.

The predictive modeling, however, yielded limited performance, as it relied on annual averages that disregarded seasonal variability and did not account for external influences such as the COVID-19 pandemic or contamination episodes involving public and private entities. Despite these limitations, the findings underscore the importance of continuous monitoring of toxic and nutrient-related parameters in the Tietê River to better support public health protection and water resources management.

REFERENCES

[1] C. E. M. Tucci, Urban Waters.Porto Alegre, Brazil: UFRGS Press, 2006.
[2] National Council for the Environment (CONAMA), Resolution No. 357, of March 17, 2005.Official Gazette of the Federative Republic of Brazil, no. 053, pp. 58–63, Mar. 18, 2005.
[3] A. Jenkins, N. E. Peters, and A. Rhode, "Hydrology," in Biogeochemistry of Small Catchments: A Tool for Environmental Research, B. Moldan and J. Cerny, Eds. Chichester, UK: John Wiley & Sons, 1994.
[4] CPRM – Geological Survey of Brazil, Hydrogeological Map of the State of São Paulo. Brasília, Brazil: CPRM, 2011.
[5] Environmental Company of the State of São Paulo (CETESB), Report on the Quality of Inland Waters of the State of São Paulo – 2022. São Paulo, Brazil, 2023.
[6] P. C. S. Mancuso and H. F. Santos, Water Reuse.São Paulo, Brazil: Manole, 2003.
[7] São Paulo State, State Constitution of São Paulo, October 5, 1989. Last update: Constitutional Amendment No. 55, Nov. 27, 2024. São Paulo, Brazil: Legislative Assembly. Available: https://www.al.sp.gov.br/repositorio/legislacao/constituicao/1989/compilacao-constituicao-0-05.10.1989.html
[8] SABESP, Sustainability Report 2023. São Paulo, Brazil: São Paulo State Basic Sanitation Company, 2023.
[9] National Water Agency (ANA), Sewage Atlas: Watershed Depollution. Brasília, Brazil: ANA, 2021.
[10] O. Chapelle et al., "Choosing multiple parameters for support vector machines," Machine Learning, vol. 46, no. 1–3, pp. 131–159, 2002.
[11] Y. Zhang, H. Shi, and M. Liu, "Prediction of water quality using support vector machine," Environmental Monitoring and Assessment, vol. 135, pp. 391–401, 2008.
[12] V. N. Vapnik and A. Y. Lerner, "Recognition of patterns with help of generalized portraits," Avtomat. I Telemekh, vol. 24, no. 6, pp. 774–780, 1963.
[13] V. N. Vapnik and A. Y. Chervonenkis, "A class of algorithms for pattern recognition learning," Avtomat. I Telemekh, vol. 25, no. 6, pp. 937–945, 1964.
[14] V. Vapnik, The Nature of Statistical Learning Theory. New York, USA: Springer, 1995.

[15] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," Statistics and Computing, vol. 14, pp. 199–222, 2004. Kluwer Academic Publishers. Available: https://doi.org/10.1023/B:STCO.0000035301.49549.88

[16] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," Optimization Methods and Software, vol. 1, pp. 23–34, 1992.

[17] Ministry of Health (Brazil), Ordinance No. 888, May 4, 2021: Procedures for Control and Surveillance of Drinking Water Quality and Potability Standards. Brasília, Brazil, 2021, 33 pp.

[18] G. Bergamim, "Sabesp admits it's discharging untreated sewage into the Tietê River to drain broken pipeline in Marginal," G1 São Paulo / TV Globo, Jun. 27, 2025. [Online]. Available: https://g1.globo.com/sp/sao-paulo/noticia/2025/06/27/sabesp-admite-que-esta-lancando-esgoto-sem-tratamento-no-rio-tiete-para-esvaziar-tubulacao-rompida-na-marginal.ghtml. [Accessed: Sep. 18, 2025].

★★★