 doi:10.20944/preprints202212.0027.v1

Article

A novel data-driven fuzzy for accurate coagulant dosage in drinking water treatment

A. Bressane [1,2]*, A.P.G. Goulart [2], I.G. Gomes [2], A.I.S. Loureiro [1], R.G. Negri [2,3], R. Moruzzi [1,2], A.G. dos Reis [1,2], J.K.S. Formiga[2], C.P. Melo [1] and G.H.R. da Silva [1]

[1] Civil and Environmental Engineering Graduate Program, College of Engineering, São Paulo State University, 14-01 Eng. Luiz E.C. Coube Avenue, Bauru, Brazil
[2] Environmental Engineering Department, São Paulo State University, Institute of Science and Technology, 500 Altino Bondensan Road, São José dos Campos, Brazil
[3] Natural Disasters Graduate Program, Brazilian Center for Early Warning and Monitoring for Natural Disasters, 500 Altino Bondensan Road, São José dos Campos, Brazil
* Corresponding author: A. Bressane, adriano.bressane@unesp.br

Abstract: Coagulation is the most sensitive step in drinking water treatment. Underdosing may not yield the required water quality, whereas overdosing may result in higher costs and excess sludge. Traditionally, the coagulant dosage is set based on bath experiments performed manually. Therefore, this test does not allow real-time dosing control, and its accuracy is subject to operator experience. Alternatively, solutions based on machine-learning (ML) have been evaluated as a computer-aided alternative. Despite these advances, there is open debate on the most suitable ML method applied to the coagulation process, capable of the most highly accurate prediction. This study addresses this gap, where a comparative analysis between ML methods was performed. As a research hypothesis, a novel data-driven fuzzy inference system (FIS) should provide the best performance due to its ability to deal with uncertainties inherent to complex processes. Although ML methods have been widely investigated, only a few studies report hybrid neuro-fuzzy systems applied to coagulation. Thus, to the best of our knowledge, this is the first study thus far to address the accuracy of this novel data-driven FIS for such application. The novel FIS provided the smallest error (0.86), indicating a promising alternative tool for real-time and highly accurate coagulant dosing in drinking water treatment.

Keywords: coagulant dosage; fuzzy; machine-learning; water treatment

Introduction

To remove contaminants such as suspended solids, colloidal material, and microorganisms, coagulation is among the primary processes for physical-chemical treatment of drinking water (Zhang and Luo, 2020; Wang et al., 2021). Jar tests are commonly used to determine the best dose of coagulant in drinking water treatment plants (WTPs) (Menezes et al., 2017; Jayaweera and Aziz, 2018). Considering the quality of raw water, the test simulates the coagulation step under laboratory conditions. Although this test has been used for many years, improving both its accuracy and response speed with respect to water quality changes remains very challenging (Narges et al., 2021).

Jar test experiments are manually performed and, hence, were not conceived for real-time decision-making. Additionally, coagulant dosing can become complex when raw water quality changes rapidly and substantially (Pandilov and Stojkov, 2019), particularly due to the critical influence of pH, turbidity, and color, among other properties of contaminants and hydraulic conditions, on coagulation performance (Oliveira et al., 2018; Zhang and Luo, 2020; Zhu et al., 2021). Therefore, the jar test is not feasible for real-time adjustment (Zangooei et al., 2016; Jayaweera and Aziz, 2018). On the other hand, reducing operating costs and improving efficacy in water treatment are some of the main challenges

 doi:10.20944/preprints202212.0027.v1

in the water sector, which also faces natural water degradation and strict standards and regulations. Therefore, the study and application of data-driven and real-time technologies, such as machine learning (ML), are essential to reduce costs and enhance water safety for the water industry (Kim and Parnichkun, 2017; Pandilov and Stojkov, 2019). However, the use of alternatives based upon mechanistic models for the coagulation process is a difficult task, as it is a complex system in which there are uncertainties since interactions between the mechanisms of transfer and kinetics are not yet deeply understood (Zhang et al., 2019; Zhu et al., 2021).

In several areas of knowledge, empirical models using ML methods have been evaluated with a good ability to model complex nonlinear problems (Kennedy et al., 2015). Among the advantages of this computer-aided alternative, the prevention of errors associated with the human operator and the reduction of response time can be highlighted (Zangooei et al., 2016). Another favorable factor is that the development of solutions based on ML only requires the availability of historical databases, which, in the case of drinking WTP, are usually stored in sufficient quantities for this alternative (Newhart et al., 2019). Thus, applications based on methods such as artificial neural networks (ANNs) have become increasingly popular (Zhang et al., 2019). However, even with continual progress in research, highlighted among the most recent studies by Pandilov and Stojkov (2019), Najafzadeh and Zeinolabedini (2019), Ju et al. (2019), Zhang et al. (2019), Zhang and Luo (2020), Wang et al. (2021), Narges et al. (2021) and Zhu et al. (2021), the results achieved on computer-aided coagulant dosing have not yet led to the replacement of the jar test, which is still widely performed in drinking WTP (Zhang and Luo, 2020). Therefore, additional studies are still needed to strengthen the evidence that makes it possible to reduce the dependence on bath experiments, enabling more accurate prediction in real time (Wang et al., 2021).

Several ML methods have been evaluated for coagulant dosing, with emphasis on different ANN architectures, such as the Levenberg–Marquardt neural network (Wu and Lo, 2008); inverse neural network (Robenson, 2009); generalized regression neural network (Heddam et al., 2011), adaptive neuro-fuzzy inference system (Pandilov and Stojkov, 2019; Narges et al., 2021), dynamic evolving neural-fuzzy system (Heddam and Dechemi, 2015), radial basis function (Zangooei et al., 2016; Kim and Parnichkun, 2017; Wang et al., 2021), multilayer perceptron (Zangooei et al., 2016; Menezes et al., 2018; Jayaweera and Aziz, 2018), genetic algorithm enhanced artificial neural network (Zhang et al., 2019), variable-structure neural network (Zhang and Luo, 2020), and backpropagation neural network (Zhu et al., 2021). Other tested ML methods include the linear regression model (Hernandez and Le Lann, 2006), k-nearest neighbors (Zhang et al., 2013), fuzzy linear and nonlinear regression models (Zangooei et al., 2016), *k-means* clustering (Kim and Parnichkun, 2017), and random forest (Wang et al., 2021).

Despite advances in recent years, there are still gaps in terms of the best method of ML applied to coagulation control. We hypothesize that a novel data-driven fuzzy inference system (FIS), introduced in 2022, should provide the highest accuracy due to its ability to deal with intrinsic coagulation uncertainties that are not fully controlled during the WTP operation. To the best of our knowledge, this is the first study to date to assess the performance of this novel data-driven FIS in predicting coagulant dosage.

The theory of fuzzy sets was introduced by Lotfi Zadeh to address the uncertainties that arise in complex systems (Zadeh, 2012). To this end, inference systems based on fuzzy artificial intelligence with nonlinear functions and soft boundaries allow a gradual transition between intervals and degrees of truth, admitting partial membership in more than one set of linguistic values (Barros et al., 2017). Development of FISs that use data-oriented methods for regression tasks occurred relatively recently, but they have already become one of the most popular approaches in several areas (Zhang et al., 2018). Among the environmental applications reported in the literature, FIS has been developed to support participatory planning (Mehryar et al., 2017; Bressane et al., 2017), impact assessment (Caniani et al., 2016; Bressane et al., 2020), pattern recognition (Bressane et al., 2018, Bressane, 2017), and land reclamation (Zhang et al., 2016).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2022 doi:10.20944/preprints202212.0027.v1

Methods

The dataset used in this study was derived from the drinking water treatment plant Dr. Armando Pannunzio (WTP Cerrado) at Sorocaba, a city with a territorial area of 449.872 km² and 695,000 inhabitants (1,304.18 inhab/km²), one of the most important economic and technological hubs of São Paulo State (IBGE, 2022), in southwest Brazil (Figure 1).

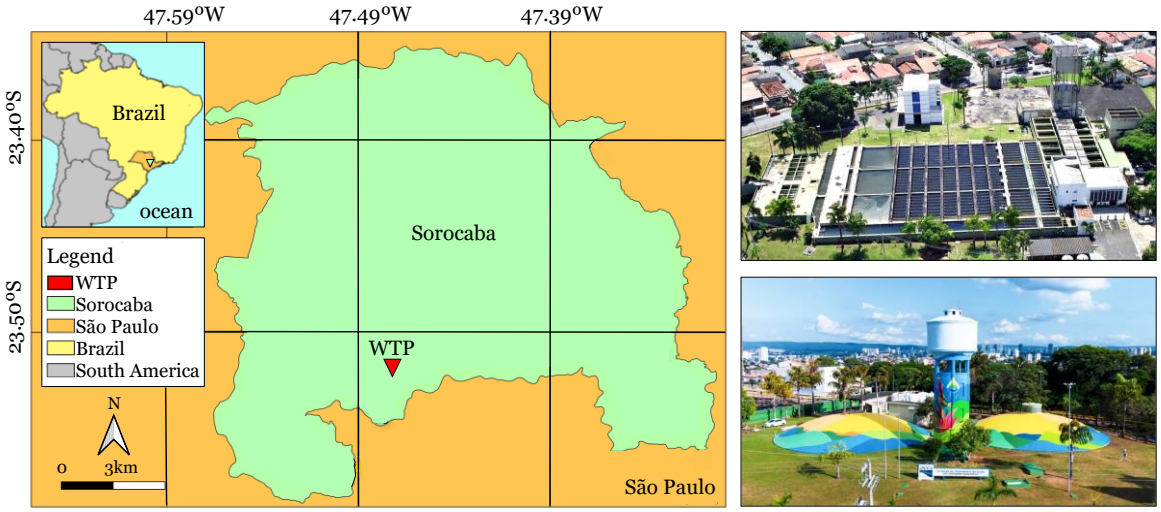

Figure 1. Drinking Water Treatment Plant - WTP Cerrado at Sorocaba city, São Paulo State, southwest, Brazil. Source: Modified from Santinon (2022).

The WTP Cerrado treats 2.2 m³/s of water via conventional treatment (coagulation - flocculation - sedimentation - filtration) using coagulant polyaluminum chloride (PAC) within the dose range of 30 to 40 mg/L (Figure 2).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2022 doi:10.20944/preprints202212.0027.v1

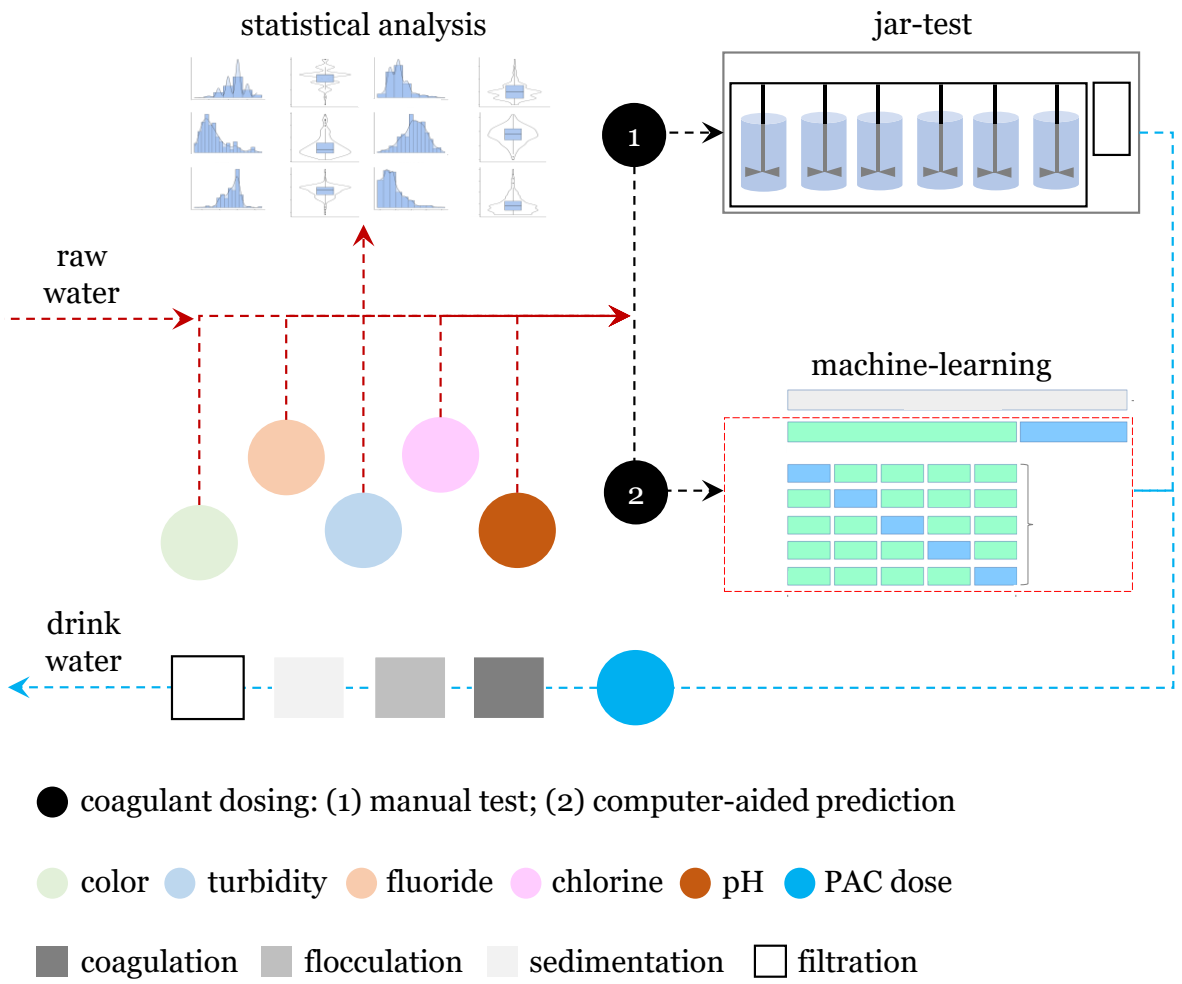

Figure 2. WTP via conventional treatment (coagulation - flocculation - sedimentation - filtration) with manual or computer-aided coagulant dosing.

A database equivalent to one year (January to December) of quasi-daily tests (n = 291) was used, with the dosage of coagulant polyaluminum chloride (PAC) and measurements of quality indicator parameters of raw water (pH, color, turbidity, fluoride, and chlorine) (Table 1 and Figure 3).

Table 1. Database with quality indicator parameters of raw water and PAC.

	pH	color (HU)	turbidity (NTU)	fluoride (mg/L)	chlorine (mg/L)	PAC (mg/L)
Average	6.77	2.01	0.248	0.688	1.840	32.0
Median	6.80	2.00	0.200	0.069	1.900	32.0
St. Deviation	0.109	1.18	0.155	0.029	0.237	1.84
Minimum	6.40	0.00	0.030	0.600	0.900	30.0
Maximum	7.00	7.00	0.780	0.760	2.700	40.0
Asymmetry	-0.247	1.14	1.29	-0.314	-0.628	1.40

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2022 doi:10.20944/preprints202212.0027.v1

Kurtosis	0.264	2.09	1.32	-0.026	1.680	2.39
Normality (p)*	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

* Shapiro-Wilk test.

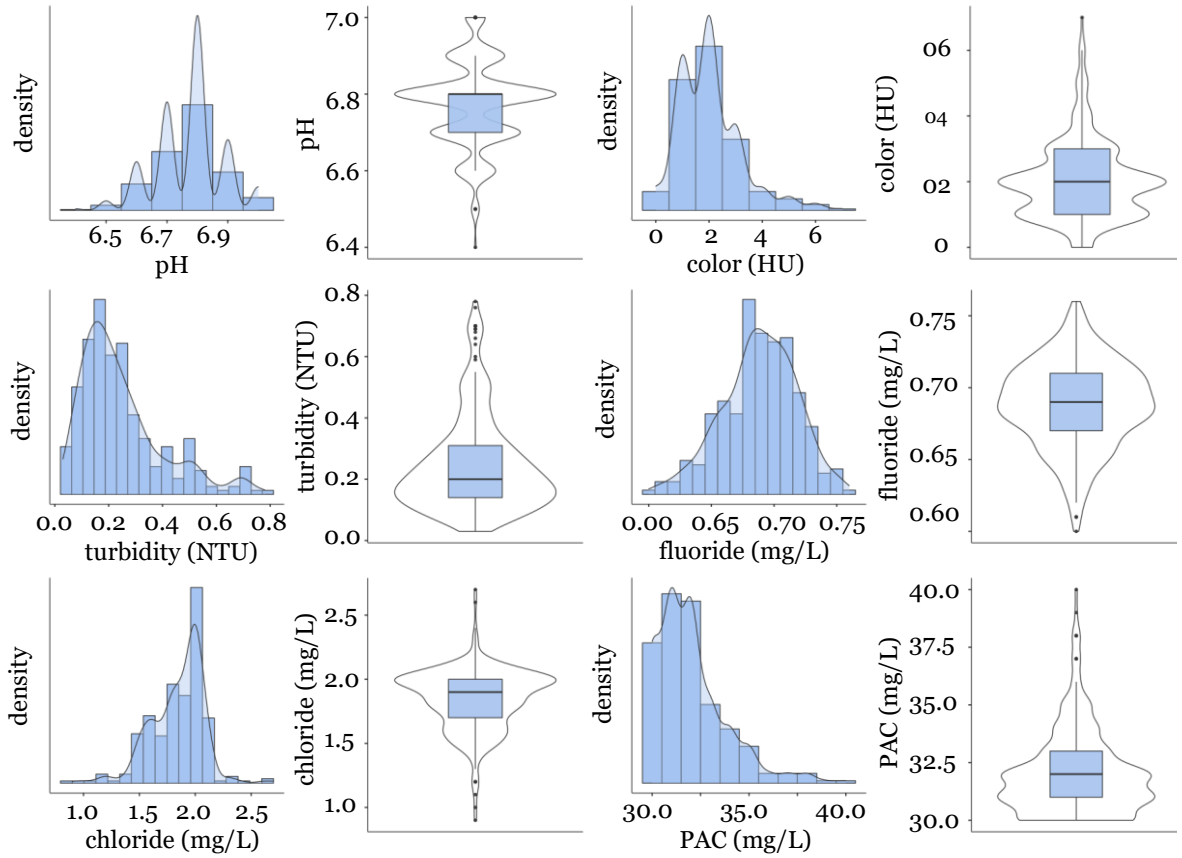

Figure 3. Exploratory analysis of quality indicator parameters of raw water and polyaluminum chloride (PAC).

As an artificial intelligence method specifically developed for data-driven fuzzy inference systems (FISs), the Wang & Mendel algorithm ('wm') was adopted in the present study. A novel ML method based on this algorithm was made available by Guillaume et al. (2022) in the package 'FisPro' in the R programming language, which was used in our research.

To test the research hypothesis, the accuracy of this novel FIS was compared to that obtained by some of the primary methods applicable to prediction tasks: cascade-correlation network (CCN), gene expression programming (GEP), polynomial neural network (GMDH), multilayer perceptron network (MLP), probabilistic neural network (PNN), radial basis function network (RBFN), stochastic gradient boosting (TreeBoost), and support vector machine (SVM).

As a standard way to measure the performance of a model in predicting quantitative data, the root mean square error ($RMSE$) was calculated to analyze the coagulant dosing accuracy:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i - O_i)^2}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2022 doi:10.20944/preprints202212.0027.v1

where O_i are the observations, s_i are the predicted values of the dose of the coagulant, and n is the number of observations. Considered the most common accuracy metric, $RMSE$ is widely considered a good measure for comparing different models or model settings (Hodson, 2022).

To avoid overfitting the model to the training data, the parameterization of the algorithms of each artificial intelligence method followed a grid search procedure (Yu and Zhu, 2022). Considering different combinations of parameters, the setting that minimized the RMSE was determined based on 5-*fold cross-validation*, using 70% of the dataset for the learning process and 30% for validation testing, as shown in Figure 4.

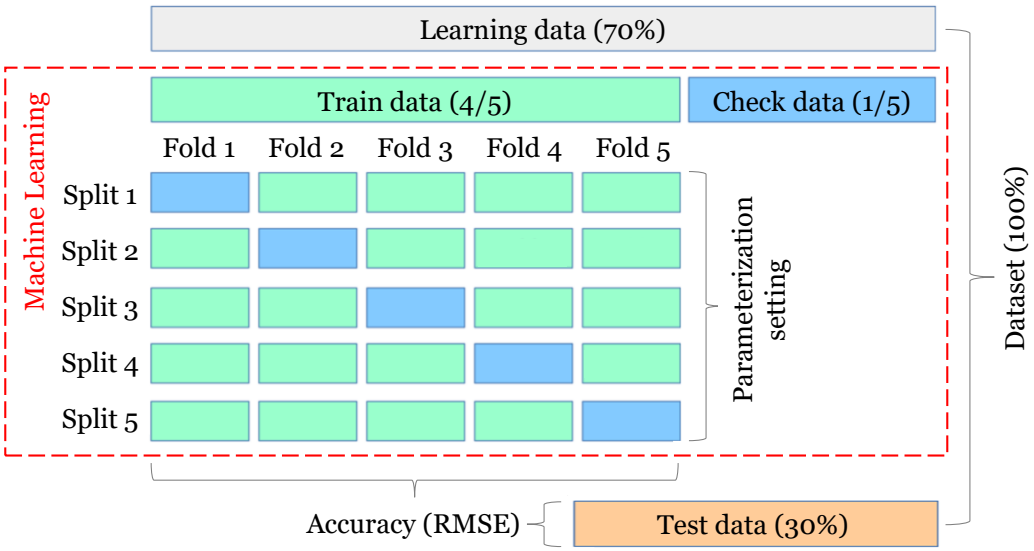

Figure 4. Determining parameterization settings based on 5-*fold cross-validation*. Source: Modified from Scikit-learn developers (2022).

Results and discussion

The performance of the ML methods is presented in Table 2, where the accuracy (RMSE) based on the testing data varies significantly between 1.28 (RBFN) and 0.86 (FIS). In general, although some ML algorithms stand out for their high performance in specific applications, it is essential to note that task accuracy is also highly associated with data behavior (Negri, 2021). Therefore, comparing several ML methods is important to verify the best alternative applicable to each case (Bressane et al., 2022). Analyzing Table 2, the results can be organized into three groups based on the performance of the ML methods during the tests. In the first group, with low performance (RMSE equal to or greater than 1.20), are the GMDH, TreeBoost and RBFN methods.

Wang et al. (2021) proposed a method for optimizing the coagulation process during drinking water treatment using distinct ML approaches, including the RBFN method. Although it delivers better performance compared to multiple linear regression models, the RBFN was outperformed by the random forest algorithm. In the present study, TreeBoost achieved the second-worst accuracy, with 1.25 RMSE. This algorithm develops a sequential training through which the decision trees grow in series. In this way, a tree is built to correct the errors of the previous one (boosting), which generally provides superior performance unless there is influence from noisy data (Wei et al., 2021; Bressane et al., 2018).

Table 2. Overall accuracy of each ML method based on the RMSE.

ML method	Parameterization setting based on 5-folds cross-validation	RMSE	
		train	test

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2022 doi:10.20944/preprints202212.0027.v1

CCN	kernel: gaussian; minimum and maximum neurons range: [0, 10^3]; candidates: 10^2; epochs: 10^3; overfitting control: cross-validation	1.10	1.18
FIS	model: 'wm'; functions: 6; grid: hierarchy partition fuzzy of 150; conjunction: Lukasiewicz; out: crisp; defuzzification: maximum crisp; disjunction: sum	0.44	0.86
GEP	population: 50; maximum tries: 10^4; genes: 4; gene head length: 10; maximum: 2000; generations without improvement: 10^3	1.10	1.15
GMDH	maximum network layers: 20; maximum polynomial order: 16; neurons per layer: 20; function: linear; connections: to previous layer	1.17	1.20
MLP	number of layers: 03; hidden layer function: smooth; output layer function: linear; train: scaled conjugate gradient	1.11	1.17
PNN	type of kernel function: gaussian; steps: 20; sigma: each var. [10^{-4}, 10]; prior probability: frequency distribution	0.71	1.17
RBFN	max. neurons: 10^3; radius: [10^{-2}, 10^3]; population size: 200; maximum generations: 20; maximum generation flat: 5	0.96	1.28
TreeBoost	maximum trees: 300; minimum trees: 10; depth: 10; minimum size node: 5; shrink factor: auto; prune: minimum absolute error, smooth: 5	0.56	1.25
SVM	type: epsilon-SVR; kernel function: RBF; optimize: minimize total error; stopping criteria: 10^{-3}	1.12	1.17

In the second group, with intermediate performance (RMSE from 1.15 to 1.20), were GEP, MLP, PNN, SVM, and CCN. A CCN is a type of self-organizing neural network whose size and topology are determined by adding neurons to its architecture to guarantee improved learning over the training process (Mohamed et al., 2021). Consequently, this algorithm may overfit the training data and lose generalization ability. In this situation, we used an overfitting control pruning strategy to minimize the cross-validation error. Despite this, CCN's performance dropped from 1.10 in training to 1.18 RMSE during testing. Wadkar et al. (2021) also evaluated the CCN method to predict coagulant dose. The authors indicated that beyond large amounts of training data, as required by most ANN-based approaches, the CCN method showed a sensitive/fragile relationship between the network's architecture and the prediction error rates. Consequently, this method may demand great attention concerning its parametrization.

In turn, the MLP enables nonlinear mappings using activation functions based on the backward propagation of errors to adjust the ANN weight connections. Moreover, the network architecture of minimum training error during the ML process was considered to prevent model overfitting, delivering a 1.17 RMSE. Additionally, according to Jayaweera et al. (2018), although the MLP method has been useful for predicting the optimum coagulant dosage for water treatment, the high computational cost and requirement of sufficient training data are the primary drawbacks.

Almost all analyzed artificial neural networks (ANNs) achieved similar performance, approximately 1.17 RMSE. To minimize misclassification, PNN uses probability density functions to define complex decision boundaries, which generally improves its accuracy (Bressane et al., 2018a). Zhang et al. (2013) analyzed the performance of the SVM method applied to predict coagulant dosage in water treatment plants of distinct sizes and concluded that such a method performs better for large- and medium-sized water systems compared to small ones. Although it shares similarities with ANNs, SVM has better ability to deal with high dimensional data and is less prone to overfitting (Kalantar et al., 2018). Despite this, the SVM also achieved only 1.17 RMSE.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2022 doi:10.20944/preprints202212.0027.v1

Finally, with higher accuracy, the FIS reached 0.86 RMSE over the test data. Of note, this error is relatively low given that the PAC variation ranges from 30 to 40 mg/L, that is, less than ±0.9 in a variation range more than 10 times greater [30 40]. The occurrence of outliers makes the ability of FIS to handle data behaviors that are critical to other ML methods even more evident (Figure 5).

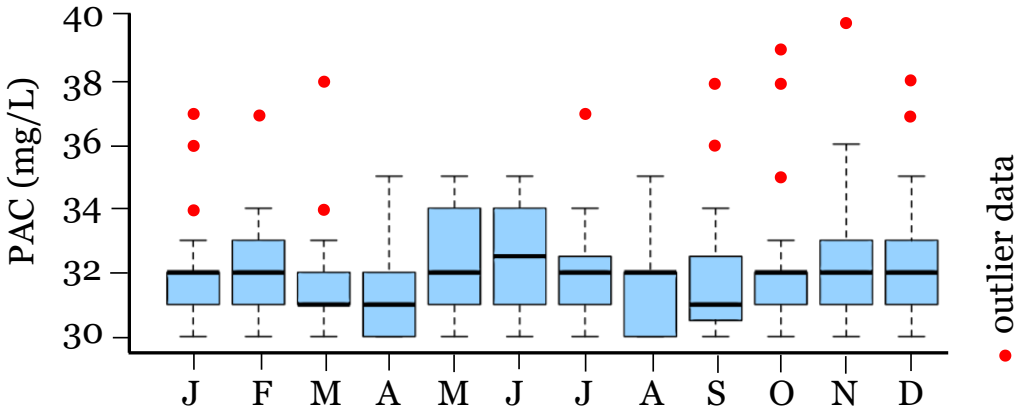

Figure 5. Data behaviors and occurrence of outliers related to coagulant dose throughout the year (January to December).

Using the 'wm' rule induction technique, fuzzification of the linguistic values of each variable was performed using triangular functions (Figure 6), which is one of the most widely accepted and used fuzzy membership functions (Barros et al., 2017). The input space of the predictor variables shown in Figure 6 was partitioned into linguistic values (SS, S, M, L, LL, and XL) based on fuzzy soft boundaries. Khameneh et al. (2014) define a fuzzy soft boundary as a parameterization extension of the concept of a boundary in the classical sense. The properties associated with this extension allow a fuzzy model to make inferences based on partial degrees of certainty, which cannot be properly handled using traditional tools (Hussain, 2020). Considering these linguistic values and ranges, some examples of rules (R_i) generated by FIS during machine-learning are as follows:

R_1: if pH is S and $color$ is LL and $turbidity$ is XL and $fluoride$ is M and chlorine is SS , then the dosage of coagulant (PAC) = 37 mg/L;

R_7: if pH is L and $color$ is M and $turbidity$ is XL and $fluoride$ is SS and chlorine is LL , then the dosage of coagulant (PAC) = 30 mg/L.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2022 doi:10.20944/preprints202212.0027.v1

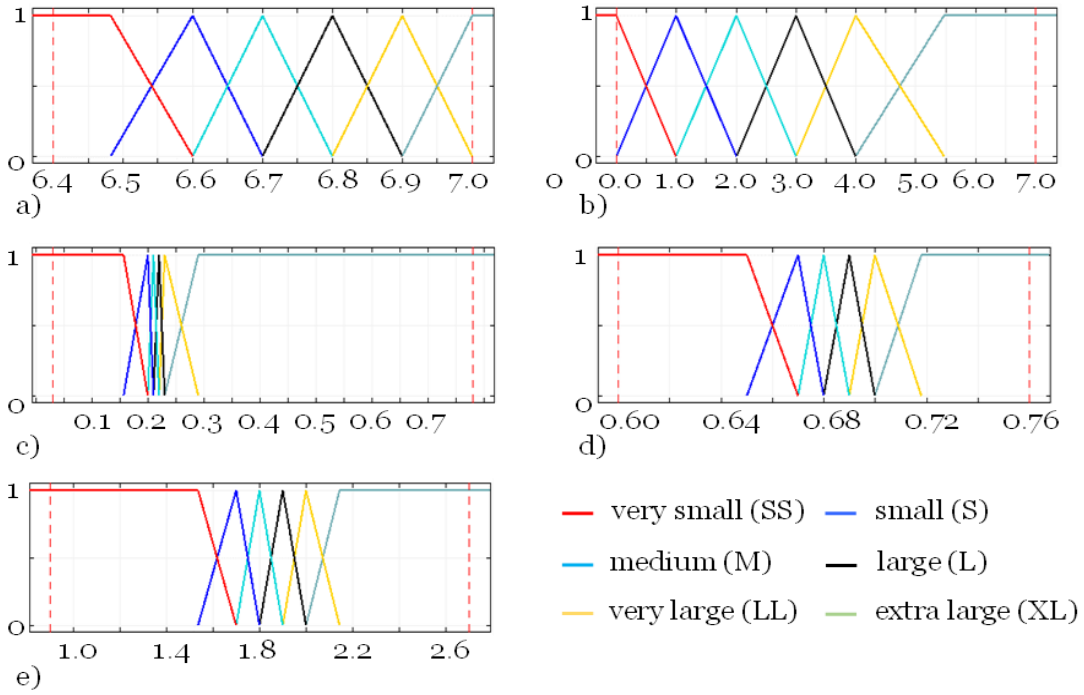

Figure 6. Fuzzification of the predictor variables in the input space using triangular membership functions and 6 linguistic values (SS, S, M, L, LL, and XL): a) pH, b) color (HU), c) turbidity (NTU), d) fluoride (mg/L), and e) chlorine (mg/L).

During parameterization in the ML process, the 'hfp' partitioning procedure provided the best fit of the data to the model. For the Lukasiewicz conjunction operator, 6 antecedent terms (linguistic values) were sufficient for the FIS to decrease the RMSE close to 0.44 during training. While some operators consider only the lowest membership in the disjunction step, the sum t-norm considers all membership values, which provides improved performance in the regression task (Ghodousian et al., 2018; Bressane et al., 2018).

From these results, computer-aided coagulant dosage can be highly accurately determined using the FIS approach proposed in this study. As a practical implication, this alternative avoids errors associated with the WTP operator's experience; it can predict dosages accurately and in real time, saving operational resources, the acquisition and maintenance of equipment, and the consumption of raw material required by jar tests.

Conclusions

In this study, experiments were conducted to test and compare the accuracy of several different ML algorithms, namely, a data-driven fuzzy inference system, cascade-correlation network, gene expression programming, polynomial neural network, multilayer perceptron network, probabilistic neural network, radial basis function network, stochastic gradient boosting, and support vector machine, for coagulant dosing of a drinking water treatment plant. As the main contributions from this comparative analysis, it is worth highlighting (*i*) filling the gap with the more suitable ML method applied to the coagulation process; (*ii*) identifying a promising alternative for computer-aided coagulant dosing; and (*iii*) stimulating further studies to assess the potential of data-driven FIS for the control and optimization of other unit operations in drinking water treatment.

From these findings, it was possible to confirm the research hypothesis that the fuzzy inference system (FIS) presented the highest accuracy due to its ability to deal with uncertainties inherent to complex processes. By constituting a solution based on nonlinear functions with soft boundaries, which allows the measurement of partial memberships (uncertainties), the FIS affords the best generalization ability and provides a highly accurate prediction. In conclusion, the accuracy of the FIS-based alternative (0.86 error) outperformed the other assessed ML algorithms, including ensemble models (1.25), ANNs (1.20),

and kernel-based methods (1.17), widely used in regression tasks. Therefore, FIS can be considered a promising alternative tool for real-time and highly accurate coagulant dosing in drinking water treatment.

Author Contributions: Conceptualization, Adriano Bressane, Jorge K. S. Formiga and Gustavo Silva; Data curation, Rogério Negri and Jorge K. S. Formiga; Formal analysis, Adriano Bressane, Ana Paula Goulart, Isadora Gomes, Anna Isabel Loureiro, Rogério Negri, Rodrigo Braga Moruzzi, Carrie Melo and Gustavo Silva; Funding acquisition, Ana Paula Goulart and Gustavo Silva; Investigation, Anna Isabel Loureiro, Rodrigo Braga Moruzzi and Adriano Reis; Methodology, Adriano Bressane, Ana Paula Goulart, Isadora Gomes, Anna Isabel Loureiro, Rogério Negri, Rodrigo Braga Moruzzi, Adriano Reis, Jorge K. S. Formiga and Carrie Melo; Resources, Adriano Bressane, Adriano Reis and Jorge K. S. Formiga; Software, Ana Paula Goulart, Isadora Gomes and Carrie Melo; Supervision, Adriano Bressane; Validation, Rodrigo Braga Moruzzi, Carrie Melo and Gustavo Silva; Writing – original draft, Adriano Bressane, Ana Paula Goulart, Isadora Gomes, Anna Isabel Loureiro, Rogério Negri, Adriano Reis, Jorge K. S. Formiga, Carrie Melo and Gustavo Silva; Writing – review & editing, Rodrigo Braga Moruzzi.

Funding : The São Paulo Research Foundation – FAPESP. Grant number #2022/03675-8.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Barros, L.C., Bassanezi, R.C. & Lodwick, W.A. 2017. *A First Course in Fuzzy Logic, Fuzzy Dynamical Systems, and Biomathematics*. Springer-Verlag Berlin, Heidelberg.
2. Bressane, A., Biagolini, C.H., Mochizuki, P.S., Roveda, J.A.F. and Lourenço, R.W. 2017. Fuzzy-based methodological proposal for participatory diagnosis in linear parks management. *Ecological Indicators*, **80**, 153-162.
3. Bressane, A., Fengler, F.H., Roveda, J.A.F., Roveda, S.R.M.M. and Martins, A.C.G. 2018. Arboreal identification supported by fuzzy modeling for trunk texture recognition. *Trends in Applied and Computational Mathematics*, **19**, 111-126.
4. Bressane, A., Silva, P.M., Fiore, F.A., Carra, T.A., Ewbank, H., De-carli, B.P. and Mota, M.T. 2020. Fuzzy-based computational intelligence to support screening decision in environmental impact assessment: A complementary tool for a case-by-case project appraisal. *Environmental Impact Assessment Review*, **85**, e106446.
5. Bressane, A., Spalding, M., Zwirn, D., Loureiro, A.I.S., Bankole, A.O., Negri, R.G., de Brito Junior, I., Formiga, J.K.S., Medeiros, L.C.dC, et al. 2022. Fuzzy Artificial Intelligence—Based Model Proposal to Forecast Student Performance and Retention Risk in Engineering Education: An Alternative for Handling with Small Data. *Sustainability* **14**, e14071.
6. Caniani, D., Labella, A., Lioi, D.S., Mancini, I.M. and Masi, S. 2016. Habitat ecological integrity and environmental impact assessment of anthropic activities: A GIS-based fuzzy logic model for sites of high biodiversity conservation interest. *Ecological Indicators*, **31**, 238-249.
7. Ghodousian, A., Naeeimi, M. and Babalhavaeji, A. 2018. Nonlinear optimization problem subjected to fuzzy relational equations defined by Dubois-Prade family of t-norms. *Computers & Industrial Engineering*, **119**, 167-180.
8. Guillaume, S., Charnomordic, B., Lablée, J., Jones, H. and Desperben, L. 2022. FisPro: Fuzzy Inference System Design and Optimization. R package version 1.1.1. https://CRAN.R-project.org/package=FisPro (accessed 11 september 2022).
9. Heddam, S., Bermad, A. and Dechemi, N. 2011. Applications of radial-basis function and generalized regression neural networks for modeling of coagulant dosage in a drinking water-treatment plant: Comparative study, *J. Environ. Eng*. **137**, 1209–1214.
10. Heddam, S. and Dechemi, N. 2015. A new approach based on the dynamic evolving neural-fuzzy inference system (DENFIS) for modelling coagulant dosage (Dos): case study of water treatment plant of Algeria, *Desalination, and Water Treatment*, **53**, 1045-1053.
11. Hernandez, H. and Lann, MV. 2006. Development of a neural sensor for on-line prediction of coagulant dosage in a potable water treatment plant in the way of its diagnosis. *Iberamia Sbia*, **4140**, 249-257.
12. Hodson, T.O. 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geosci. Model Dev.*, **15**, 5481–5487.
13. Hussain, S. 2020. On some properties of intuitionistic fuzzy soft boundary. *Commun. Fac. Sci. Univ. Ank. Ser. A1 Math. Stat*., **69**, 1033-1044.
14. IBGE - Brazilian Institute of Geography and Statistics (2022) Sorocaba city. https://www.ibge.gov.br/cidades-e-estados/sp/sorocaba.html (accessed 11 september 2022).
15. Jayaweera, C.D. and Aziz, N. 2018. Development and comparison of Extreme Learning machine and multi-layer perceptron neural network models for predicting optimum coagulant dosage for water treatment, *Journal of Physics*, **1123**, e012032.

16. Ju, J., Park, Y., Choi, Y. and Lee, S. 2019. Comparison of statistical methods to predict fouling propensity of microfiltration membranes for drinking water treatment. *Desalination and water treat*, **143**, e716.
17. Kalantar, B., Pradhan, B., Naghibi, S.A., Alireza, M. and Mansor, S. 2018. Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN), *Geomatics, Natural Hazards and Risk*, **9**, 49-69.
18. Kennedy, M.J., Gandomi, A.H. and Miller, C.M. 2015. Coagulation modeling using artificial neural networks to predict both turbidity and DOM-PARAFAC component removal. *Journal of Environmental Chemical Engineering*, **3**, 2829–2838.
19. Khameneh, A.Z., Kiliçman, A. and Salleh, A.R. 2014. Fuzzy soft boundary. *Annals of Fuzzy Mathematics and Informatics*, **8**, 687–703.
20. Kim, C.M. and Parnichkun, M. 2017. Prediction of settled water turbidity and optimal coagulant dosage in drinking water treatment plant using a hybrid model of k-means clustering and adaptive neuro-fuzzy inference system. *Applied Water Science*, **7**, e3902.
21. Mehryar, S., Sliuzas, R., Sharifi, A., Reckien, D. and van Maarseveen, M. 2017. A structured participatory method to support policy option analysis in a social-ecological system. *Journal of Environmental Management*, **15**, 360-372.
22. Menezes, F.C., Fontes, R.M., Oliveira-Esquerre, K.P. and Kalid, R. (2018) Application of uncertainty analysis of artificial neural networks for predicting coagulant and alkalized dosages in a water treatment process. *Brazilian Journal of Chemical Engineering*, **35**, 1369-1381.
23. Mohamed, S.M., Mohamed, M.H. and Farghally, M.F. 2021. A New Cascade-Correlation Growing Deep Learning Neural Network Algorithm. *Algorithms*, **14**, 1-18.
24. Moher, D., Liberati, A., Tetzlaff, J. and Altman, D.G. 2022. The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. http://www.prisma-statement.org (accessed 11 september 2022).
25. Najafzadeh, M. and Zeinolabedini, M. 2019. Prognostication of wastewater treatment plant performance using efficient soft computing models: an environmental evaluation. *Measurement*, **138**, 690-701.
26. Narges, S., Ghorban, A., Hassan, K. and Mohammad, K. 2021. Prediction of the optimal dosage of coagulants in water treatment plants through developing models based on artificial neural network fuzzy inference system (ANFIS), *Journal of Environmental Health Science and Engineering*, **19**, 1543–1553.
27. Negri, R.G. 2021. *Pattern recognition: a directed study*. Edgard Blucher, São Paulo.
28. Newhart, K.B., Holloway, R.W., Hering, A.S. and Cath, T.Y. 2019. Data-driven performance analyses of wastewater treatment plants: a review. *Water Research*, **157**, 498e513.
29. Oliveira, A.S., Lopes, V.S., Coutinho Filho, U., Moruzzi, R.B. and Oliveira, A.L. 2018. Neural network for fractal dimension evolution. *Water Science & Technology*, **78**, 795-802.
30. Pandilov, Z. and Stojkov, M. 2019. Application of intelligent optimization tools in determination and control of dosing of flocculant in water treatment. *International Journal of Engineering*, **3**, 109-116.
31. Robenson, A., Shukor, S.A. and Aziz, N. 2009. Development of process inverse neural network model to determine the required alum dosage at Segama water treatment plant Sabah, Malaysia. *Computer Aided Chemical Engineering*, **27**, 525-530.
32. Santinon, E. 2022. Drinking water treatment plant Dr. Armando Pannunzio at Sorocaba city, São Paulo State, Brazil [Photograph]. https://noticias.sorocaba.sp.gov.br/saae-sorocaba-realiza-manutencao-preventiva-na-eta-cerrado-neste-domingo-21/ (accessed 11 september 2022).
33. Scikit-learn developers. 2022. Cross-validation: evaluating performance. https://scikit-learn.org/stable/modules/cross_validation.html (accessed 11 september 2022).
34. Wadkar, D.V., Karale, R.S. and Wagh, M.P. 2022. Application of cascade feed forward neural network to predict coagulant dose, *Journal of Applied Water Engineering and Research*, **10**, 87-100.
35. Wang, D., Wu, J., Deng, L. and Li, Z. 2021. A real-time optimization control method for coagulation process during drinking water treatment, *Nonlinear Dyn*, **105**, 3271-3283.
36. Wei, Y., Ding, J., Yang, S., Yang, X. and Wang, F. 2021. Comparisons of random forest and stochastic gradient treeboost algorithms for mapping soil electrical conductivity with multiple subsets using Landsat OLI and DEM/GIS-based data at a type oasis in Xinjiang, China, *European Journal of Remote Sensing*, **54**, 158-181.
37. Wu, G.D. and Lo, S.L .2008. Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system. *Engineering Applications of Artificial Intelligence*, **21**, 1189-1195.
38. Yu, T. and Zhu, H. 2022. Hyper-parameter optimization: a review of algorithms and applications. *Computer Science*, IN PRESS.
39. Zadeh, L.A. 2012. *Computing with words*. Berlin: Springer Berlin Heidelberg.
40. Zangooei, Z., Delnavaz, M. and Asadollahfardi, G. 2016. Prediction of coagulation and flocculation processes using ANN models and fuzzy regression. *Water Science & Technology*, **74**, 1296-1311.
41. Zhang, K., Achari, G., Li, H., Zargar, A. and Sadiq, R. 2013. Machine learning approaches to predict coagulant dosage in water treatment plants. *Int. J. Assur. Eng. Manag.*, **4**, 205-214.
42. Zhang, H., Sun, T., Shao, D. and Yang, W. 2016. Fuzzy logic method for evaluating habitat suitability in an estuary affected by land reclamation. *Wetlands*, **36**, 19-30.
43. Zhang, J., Qiu, H., Li, X., Niu, J., Neyers, M.B., Hu, X. and Phanikumar, M.S. 2018. Realtime nowcasting of microbiological water quality at recreational beaches: a wavelet and artificial neural network-based hybrid modeling approach. *Environ. Sci. Technol.*, **52**, 8446-8455.

 doi:10.20944/preprints202212.0027.v1

44. Zhang, Y., Gao, X., Smith, K., Inial, G., Liu, S., Conil, L.B. and Pan, B. 2019. Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network. *Water Research*, **164**, 1-12.
45. Zhang, J. and Luo, Y. 2020. Multimodal Control by Variable-Structure Neural Network Modeling for Coagulant Dosing in Water Purification Process. *Complexity*, **20**, 1-12.
46. Zhu, G., Xiong, N., Wang, C., Zhongwu, L. and Hursthouse, A.S. 2021. Application of a new HMW framework derived ANN model for optimization of aquatic dissolved organic matter removal by coagulation. *Chemosphere*, **262**, 1-12.