Water Supply

IWA PUBLISHING

Water Supply Vol 24 No 4, 1080 doi: 10.2166/ws.2024.057

Prediction of microbiological non-compliances using a Boosted Regression Trees model: application on the drinking water distribution system of a whole country

Mariana Barcia[a], Alexandra Sixto[a,b] and Maria Pia Cerdeiras[a,c,*]

[a] Unidad de Análisis de Agua, Facultad de Química, Universidad de la República, Montevideo, Uruguay
[b] Química Analítica, Facultad de Química, Universidad de la República, Montevideo, Uruguay
[c] Area Microbiología, Facultad de Química, Universidad de la República, Montevideo, Uruguay
*Corresponding author. E-mail: mcerdeir@fq.edu.uy

MPC, 0009-0001-5412-6771

ABSTRACT

Universal access to safe drinking water is a fundamental human right and a requirement for a healthy life. Therefore, monitoring the quality of the supplied water is of utmost importance. To achieve this goal, there is a need to develop tools that support monitoring activities and improve efficiency. Forecasting models enable the prediction of pollution levels and facilitate the implementation of action plans. In this study, the Boosted Regression Trees method was employed to investigate the variables influencing water quality failures (WQFs) due to microbial contamination at the delivery point. The dataset used was obtained from localities across the country's distribution systems. The variables under consideration included physicochemical parameters such as pH, turbidity (NTU), and free chlorine (mg L^{-1}), along with contextual parameters like the year, season, geographic location, and locality population. Indicators of microbial contamination assessed were the presence of total coliforms, *Escherichia coli*, and *Pseudomonas aeruginosa*. The most significant variables were geographic location, free chlorine content, and the population of the locality. The model achieved an AUC value of 0.77 and provided adequate predictions in the conducted tests. It enables the exploration of key factors affecting microbiological water quality, allowing for informed action to reduce associated risks.

Key words: Boosted Regression Trees, drinking water, machine learning, microbiological non-compliance

HIGHLIGHTS

- Boosted Regression Trees were employed to study the variables that influence water quality failures due to microbial contamination at the delivery point. Both posed greatest risk to the public.
- Drinking water suppliers can use this tool to improve their monitoring plans and public authorities can use this input to implement actions for preventing water contamination and to improve water safety plans.

1. INTRODUCTION

Universal access to safe drinking water is a fundamental need and human right (UN 2010). It is one of the main requirements for a healthy life. Monitoring the quality of the water supplied either by conventional water distribution systems or decentralized community systems has several challenges. Time, analysis capacity, human resources, and costs, to mention a few. Thus, developing different tools that support monitoring activities and improving their efficiency is needed.

There are International Guidelines for Drinking water quality (WHO 2022) as well as local regulations that establish maximum levels allowed for multiple parameters related to health and taste. A water quality failure (WQF) event is often defined as an exceedance value of one or more of these regulated parameters from specific legislations (Sadiq *et al.* 2008). One of the most frequent WQF is due to microbial contamination (Mian *et al.* 2020) and drinking this microbial-contaminated water is the greatest risk to public health due to water use. It is associated with acute human health effects causing gastrointestinal (GI) illnesses such as diarrhoea and nausea (Interior Health Authority. Office of the Medical Health Officer, 2017; WHO 2022).

Microorganisms can come from the water source and can also enter drinking water supply systems through contamination of storage facilities and distribution networks. Even though water can be contaminated at any stage in the water piping system, the quality failure in the distribution stage is considered the most serious, since it is the point of delivery to the public (WHO 2022). Besides this, microbial contamination along with changes in the water colour or turbidity has been reported as the main concern events by consumers (Benameur *et al.* 2022). For this reason, the monitoring of microorganisms in drinking water systems is generally established by law (European Commission 1998; EPA 2009; Uruguay Presidencia 2011; EPA 2018). Optimal management of the Drinking Water Distribution Systems (DWDSs) is a complex task and most of the surveillance programs are based on sampling bulk water tests which require a considerable time to detect the WQF. Thus, to develop technologies for Early Warning Systems has been an area of increasing interest from an environmental point of view and also from a socio-economic one (EPA 2006). Therefore, the prediction of the possibility of a WQF due to microbial contamination is of utmost importance for these programs.

Approaches aimed at early detection of microbial contamination in DWDS have been proposed in the literature. For example, Ikonen *et al.* (2017) developed an online monitoring system measuring pH and temperature as an alternative to traditional water quality monitoring frameworks to reveal bacterial intrusion. Meanwhile, Carpitella *et al.* use a two-fold multi-criteria decision-making approach, a tool to identify cause-effect elements of a complex decision-making problem, applied to the field of microbial management of DWDS (Carpitella *et al.* 2020). The aim of this study was to easily identify the presence of dominant members of microbial communities according to the pipe material used in the studied DWDS.

The significant progress achieved with the aid of smart sensors for pollutant monitoring allowed the development of forecasting models for predicting pollution levels which enable action plans in advance (Henriques & Louis 2011; Mohammed *et al.* 2017; Imen *et al.* 2018; Mohammed *et al.* 2018; Mian *et al.* 2020; Podgorski & Berg 2020; Ahmed *et al.* 2021; Alsulaili & Alshawish 2021; Bong *et al.* 2021; Chen *et al.* 2021; Dawood *et al.* 2021; Li *et al.* 2021; Lobo *et al.* 2022; Schmidt *et al.* 2022; Xu *et al.* 2022). Many of these applications are for predicting different parameters in the water source and also chemical contaminants in the distribution system.

Related to the presence of microbiological non-compliance in distributed drinking water, Sadiq *et al.* (2008) use fault tree analysis to determine the causes of the distribution system failure identifying the particular sub-events that have a high impact on the failure. Fault tree analysis requires the assignment of crisp probabilities between events and the assumption of 'independence' between risk events. Alsulaili & Alshawish (2021) employed different multivariate statistical techniques to study spatial and temporal variations in the water quality distributed in hospitals, identifying the main parameters that explain these variations.

Notwithstanding, a lack of application in the field of microbial control and management strategies in drinking water systems exist.

Considering that these tools are helpful in the establishment of an effective management framework to achieve a reliable supply of safe drinking water, they are of high importance for public health. Also, international organizations have recommended the implementation of water safety plans based on the operational monitoring of the control measures in the drinking water supply (WHO 2005). Furthermore, the assessment of the whole system to determine whether the drinking water supply (from source through treatment, to the point of consumption) delivers quality water that meets the health-based criteria is of major relevance (IWA 2016). Regarding all that were stated, the objective of our study was to develop a model obtained through the monitoring data collected over 15 years to predict microbial contamination in the distribution system at the delivery point to the public. This enables to identify the most important variables which contribute to the WQF and the implementation actions to effectively overcome them.

Drinking water quality prediction uses an extremely imbalanced data set. The imbalance of the raw data set is one of the key reasons that severely restricts the thorough application of the learning models in many fields (Khalilia *et al.* 2011; Krawczyk *et al.* 2014) particularly in environmental quality monitoring and prediction (Cabaneros *et al.* 2019; Xu *et al.* 2020). As Boosted Regression Trees (BRT) is an algorithm that handles this type of data we decided to study its applicability in the DWDS to predict microbial WQF. BRT provide the possibility of handling different types of predictor variables and missing data. Other useful advantages are that they have no need for prior data transformation or elimination of outliers, and they can fit complex nonlinear relationships, and automatically handle interaction effects between predictors (Elith *et al.* 2008).

BRT is a combined method for fitting statistical models. It links two algorithms' strengths: regression trees (which relate a response to their predictors by recursive binary splits) and boosting (an adaptive method to combine many simple models).

Through this, the predictive performance of the resulting model is improved. It differs from conventional techniques in the fact that its aim is to fit a single parsimonious model (Elith *et al.* 2008). Previous works have shown it to be an extremely accurate tool for predicting the quality of groundwater (Nolan *et al.* 2015; Knierim *et al.* 2020; Stackelberg *et al.* 2020; Knierim *et al.* 2022) but it has not been used in the DWDS' WQFs.

The present work uses data obtained from the whole water distribution system of the country, including surface and groundwater sources. BRT was employed to study the variables that influence WQFs due to microbial contamination at the delivery point and to predict microbial WQF. This is the first time that this predictive tool for assessing water drinking quality is applied at the consumption point in order to help prevent WQF at this stage. This study constitutes an input for public health authorities who implement actions for water contamination prevention and water safety plans, and a useful tool for drinking water suppliers.

2. METHODS

2.1. Data collection

The drinking water quality dataset was obtained from the Water Analysis Unit of the Facultad de Química (UdelaR, Uruguay) between 2004 and 2020 in the framework of an agreement with the Regulatory Unit for Energy and Water Services (URSEA for its Spanish wording). Data from the analysis of 7,971 samples taken from localities ranging from less than 100 to over 200,000 inhabitants were employed. The whole water distribution system of the country was sampled and as it was stated before, it included surface and groundwater sources. Samples were taken every week according to SMEWW9060 A and preserved and stored pursuant to SMEWW9060 B.

The data collection was performed using a deliberate sampling approach in search of possible WQF as this is URSEA's decision in order to optimize its resources and fulfil its inspection objective. Therefore, it is not representative of the real water quality in the country.

2.2. Methodology

The variables studied were physicochemical parameters as pH, turbidity (NTU) and free chlorine (mg L^{-1}), and contextual parameters as year, season, geographic location, and locality population. These physicochemical variables were selected because they provide useful information about possible contamination or the potential of the water to support bacterial growth, regarding the contextual parameters they are linked with variables such as temperature, climate, purification process. Thus, it was expected that they would correlate with WQF due to microbial contamination. Besides this, these data were available for the majority of the analyzed samples, which represents an adequate number of data to create a predictive model.

The indicators of microbial contamination used were total coliforms and the presence of *Escherichia coli* and *Pseudomonas aeruginosa*. The response was a binary variable obtained by integrating the results of these microbiological analyses (compliant, non-compliant). Microbial quality was assessed according to SMEWW 9222B, 9213F, and 9223 methods (American Public Health Association, American Water Works Association and Water Environment Federation, 2017).

The predictor variables used in the case of the contextual parameters were year from 2004 to 2020 considering the alternate sea currents the Niño and the Niña; season (winter, spring, summer, and fall); department (19 different territorial units, indicated by letters from A to R); locality population (11 levels were considered: <100, between 100 and 500, 500 and 1,000, 1,000 and 2,000, 2,000 and 5,000, 5,000 and 10,000, 10,000 and 20,000, 20,000 and 50,000, 50,000 and 100,000, 100,000 and 200,000 and $>200,000$).

The physicochemical parameters employed were determined according to Standard Methods for the Examination of Water and Wastewater (SMEWW) methods: pH (4500H); turbidity (2130); and free chlorine (4500Cl-G) (American Public Health Association, American Water Works Association and Water Environment Federation, 2017).

To study the influence of these variables in quality failures by microbial contamination the BRT technique was used. In this supervised learning technique, a set of successive shallow and weak trees is built, with each tree learning and improving with respect to the previous one. When combined, these successive weak trees produce a powerful predictive tool.

The BRT model was built with open source R software version 3.6.1 (2019-07-05) (Ridgeway 2020) through the packages gbm (R-project.org n.a.) and dismo (Hijmans 2017).

The data set was divided into two subsets, one for training purposes (70% of the data) and the other for validation (30% of the data). In this division, we kept the proportion of defective results (about 8%) as in the original data set.

In this model, there are some hyper-parameters that require optimization.

The BRT model was run, using the gbm.step function that evaluates the optimal number of trees reinforcement by 10 times predetermined cross-validation, which is considered enough number of executions for an adequate optimization of the assigned hyper-parameters. The optimized hyper-parameters were bag.fraction, number of trees, tree complexity (tc) and learning rate. The model was fitted with different values for these parameters seeking the combination with the minimum predictive error.

For model optimization, regularization methods are used to constrain the fitting procedure so that it balances model fit and predictive performance (Hastie *et al.* 2009).

Bag.fraction is the hyper-parameter that controls stochasticity and specifies the proportion of data to be selected at each step. For example, if the bag.fraction is 0.5, this means that 50% of the data is taken randomly without replacement from the full training set at each iteration. Optimal bag.fractions can be established by comparing predictive performance and model-to-model variability under different bag.fractions. As discussed in the literature, stochasticity improved model performance and bag.fractions in the range 0·5–0·75 have given best results for presence–absence responses (Elith *et al.* 2008). A value of 0.8 for the bag.fraction was taken in accordance with other previous work (Vidal *et al.* 2018).

The learning rate (lr) determines the contribution of each tree when it is added to the model and the tree complexity controls if the interactions are fitted. These hyper-parameters determine the number of trees needed to optimize the model. Decreasing lr increases the number of trees (nt) required, and in general, a smaller lr (and larger nt) is preferable, depending on the number of data available and the time needed for computational analysis. These two hyper-parameters were estimated with an independent test set by cross-validation using the reduction of the deviance as the optimization goal. Cross-validation allows for testing the model on retained portions of data, while still using all data at some stage to fit the model.

The results of the cross-validation were used, systematically altering tc and lr while comparing the results. The learning rate took values of 0.001 and 0.005 and tc of 3 and 5. As was mentioned before, the number of trees was determined by 10 times predetermined cross-validation using the gbm.step function (R-project.org, n.d.) from the dismo package (Hijmans 2017).

The predictive capacity of the model was evaluated with the validation data. From these data, the rate of true positive results (TPR) and false positive results (FPR) is evaluated. A ROC (Receiver Operating Characteristic) curve plots TPR versus FPR at different classification thresholds. The area under the ROC Curve (AUC) measures the entire two-dimensional area below the ROC curve from (0.0) to (1.0).

As the AUC ranges in value from 0 to 1, a model whose predictions are 100% incorrect has an AUC of 0.0; another whose predictions are 100% correct has an AUC of 1.0, while a model with a performance equal to random guessing has an AUC of 0.5.

Once the model was optimized the relative importance of the variables (Figure 1) and the partial dependence graphics (Figure 2) were obtained in order to interpret the model. The AUC obtained for the model is reported in the Results and Discussion.

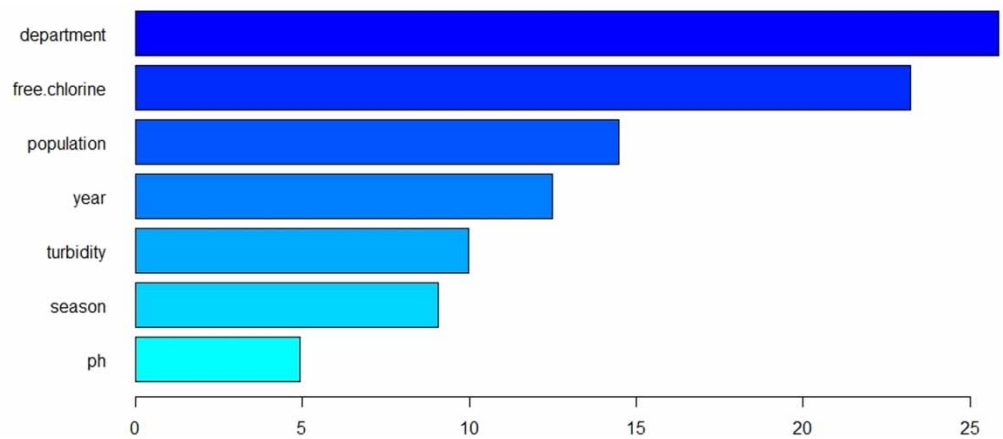

Figure 1 | Relative importance of each predictor variable in the predicted response.

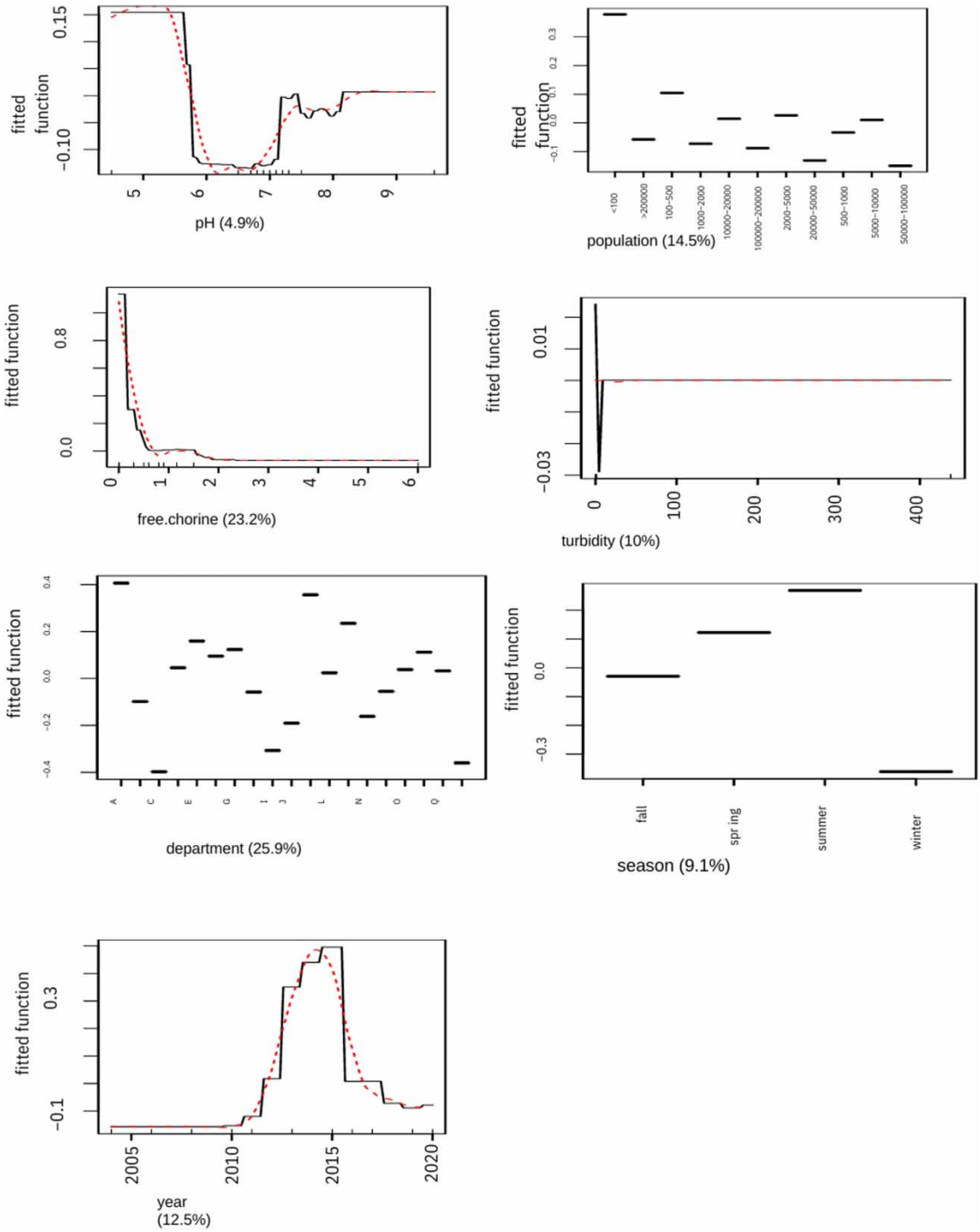

Figure 2 | Partial dependence graphs of the studied variables.

3. RESULTS AND DISCUSSION

The optimized values for the BRT model were tree.complexity = 5 and learning.rate = 0.005.

Regarding the studied variables, the importance of each one in the predicted response is shown in Figure 1.

pH, temperature, turbidity, and electrical conductivity were the most significant factors associated with the concentration of faecal indicators in a raw water source (Mohammed *et al.* 2018). Our conclusions are not exactly so. This can be explained

as we studied other variables that turned out to be more significant in the construction of the predictive model. Although pH and turbidity contribute significantly to the construction of the model, they are not the most influential among the variables studied. The type of water treatment plants of the different localities, mainly depends on the number of prospective users (that is inhabitants), therefore potentially influencing final drinking water quality.

The influence of the department includes the mentioned above about the purification process used, but also factors inherent to the available sources, capabilities of monitoring and doing the investment necessary to maintain the facilities and resulted to be the most important one.

Regarding the year, this predicts the influence of the alternate sea currents El Niño and La Niña that are the warm and cool phases of a recurring climate pattern across the tropical Pacific – the El Niño-Southern Oscillation, or 'ENSO' for short. The pattern shifts back and forth irregularly every 2–7 years, bringing predictable shifts in ocean surface temperature and disrupting the wind and rainfall patterns across the tropics. They have great incidence in rainfall throughout the year and the seasons and will, without doubt, set the median ambient temperature but the rainfall has an important effect on it.

Even though determining the predictive importance of the variables studied is of the utmost importance, once it is done, it is necessary to evaluate the relationship between them (or sub set thereof) and the response. This can be done through the construction of partial dependence plots.

Figure 2 shows the partial dependence graphs of the variables.

The figure shows that low chlorine values present greater microbiological non-compliance. As is expected, the increase in microbiological water failures is also seen during the seasons of higher temperatures. Between 2012 and 2015 there was an increase in the number of WQF which coincides with a change in the methodology for the determination of total and faecal coliforms. Until 2012 the determination was carried out by the membrane filtration method, since then the determination is carried out using the Colitag® (chromogenic substrate) kit. As it is known, coliform definition is a methodologic definition so the results depend on the method employed. For example, the genera *Pantoea*, *Leclercia* and *Lelliotia* belong to the new generation of coliforms included in the definition proposed by the new methods that detect the production of the enzyme *β*-D-galactosidase but do not produce acid or gas from lactose or characteristic colonies in culture media for coliforms such as lauryl tryptose broth or M-Endo medium. This could explain the increase in the total coliform detection. Notwithstanding this, the same methodology has been in use until now and the WQF have decreased. We could not find any reason that explained this fact.

Turbidity does not strongly correlate with microbiological non-compliances. The appearance of high turbidity could relate more to the presence of iron, manganese and aluminium oxides present in the source or in the distribution tube components.

Regarding the locality population, those with less than 100 inhabitants are the ones that have the major number of WQF. Among the possible causes to be considered, is that rural communities have less access to trained staff and state-of-the-art technologies to regulate and monitor their water treatment and distribution systems (Sadiq *et al.* 2008).

WQF together with the complexities of its distribution system makes risk analysis a highly complex process. As was mentioned before, it must be taken into account that drinking water quality prediction uses an extremely imbalanced data set. Boosting ensemble learning has yielded good results when solving the class imbalance problem in different domains. Examples of this are applications in areas such as fraud detection, medical diagnosis and manufacturing quality control (Sun *et al.* 2009; Kim *et al.* 2018). In our work, it was successfully applied for the prediction of WQF in the distribution system of our country.

BRT models give important advantages over other methods such as generalized linear models (GLM) (McCullagh & Nelder 1989) and generalized additive models (GAM; (Hastie & Tibshirani 1986) because they are able to select relevant variables, fit accurate functions and automatically identify and model interactions.

Other tools have been used for the prediction of drinking water quality (Dawood *et al.* 2021), but the aim was different as they were working on the design of a water distribution system. At the same time, they had to normalize their data, something we did not need to do.

When doing the model evaluation, we found that the model presents an AUC value of 0.77, which is considered acceptable (Elith *et al.* 2008). Although we did not find similar works dealing with the prediction of drinking water quality, similar results were obtained for models applied to groundwater microbiological results comparing different machine learning models (Wu *et al.* 2021).

The model used can effectively predict WQF from different geographical regions, different cities and towns, different source water treatments and different types of source water (surface and groundwater).

4. CONCLUSIONS

The model used is acceptable for the modelling of microbiological non-compliance, obtaining adequate predictions in the tests performed. Therefore, it is a useful tool for the prediction of microbiological non-compliance in other data sets. At the same time, we can state that the region, free chlorine content and number of inhabitants of the locality are the factors that most influence the appearance of microbiological non-compliance.

BRT technique used in this work can be used in the study and management of risks to be applied in water safety plans. While this model may need further improvement, the results of this study indicate that BRT models have high prospects as WQF prediction tools. The model allows exploring the key factors that affect microbiological water quality permitting the different actors involved to act accordingly and diminish the risk associated with it.

Because of this, drinking water suppliers can use this tool to improve their monitoring plans.

ACKNOWLEDGEMENTS

The authors thank Agencia Nacional de Investigación e Innovación (ANII), Comisión Sectorial de Investigación Científica-UdelaR, Unidad Reguladora de Servicios de Energía y Agua (URSEA) and Leticia Vidal for helpful discussion.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

There is no conflict of interest to declare

REFERENCES

Ahmed, M. F., Lim, C. K., Mokhtar, M. B. & Khirotdin, R. P. 2021 Predicting Arsenic (As) exposure on human health for better management of drinking water sources. *Int. J. Environ. Res. Public Health* **18**, 1–14.

Alsulaili, A. & Alshawish, S. 2021 Spatial and temporal multivariate statistical analysis to assess drinking water quality in medical services. *J. Eng. Res.* **9**, 211–233.

APHA/AWWA/WEF 2017 *Standard Methods for the Examination of Water and Wastewater*, 23rd edn. American Public Health Association/American Water Works Association/Water Environment Federation, Washington, DC, USA.

Benameur, T., Benameur, N., Saidi, N., Tartag, S., Sayad, H. & Agouni, A. 2022 Predicting factors of public awareness and perception about the quality, safety of drinking water, and pollution incidents. *Environ. Monit. Assess.* **194** (22), 1–26.

Bong, T., Kang, J.-K., Yargeau, V., Nam, H.-L., Lee, S.-H., Choi, J.-W., Kim, S.-B. & Park, J.-A. 2021 Geosmin and 2-methylisoborneol adsorption using different carbon materials: Isotherm, kinetic, multiple linear regression, and deep neural network modeling using a real drinking water sourcenetwork modeling using a real drinking water source. *J. Clean. Prod.* **314**, 127967.

Cabaneros, S. M., Calautit, J. K. & Hughes, B. R. 2019 A review of artificial neural network models for ambient air pollution prediction. *Environ Model Softw.* **119**, 285–304.

Carpitella, S., Del Olmo, G., Izquierdo, J., Husband, S., Boxall, J. & Douterelo, I. 2020 Decision-Making tools to manage the microbiology of drinking water distribution systems. *Water* **12**, 1247.

Chen, X., Liu, H., Liu, F., Huang, T., Shen, R., Deng, Y. & Chen, D. 2021 Two novelty learning models developed based on deep cascade forest to address the environmental imbalanced issues: A case study of drinking water quality prediction. *Environ Pollut.* **291**, 118153.

Dawood, T., Elwakil, E., Novoa, H. M. & Gárate Delgado, J. F. 2021 Toward urban sustainability and clean potable water: Prediction of water quality via artificial neural networks. *J. Clean. Prod.* **291**, 125266.

Elith, J., Leathwick, J. & Hastie, T. 2008 A working guide to boosted regression trees. *J. Anim. Ecol.* **77** (4), 802–813.

EPA 2006 Technologies and Techniques for Early Warning Systems to. Available from: https://nepis.epa.gov/Exe/ZyPDF.cgi/P1005TJT.PDF?Dockey=P1005TJT.PDF (accessed 3 September 2023).

EPA 2009 *National Primary Drinking Water Regulations – EPA*. Available from: https://www.epa.gov/sites/default/files/2016-06/documents/npwdr_complete_table.pdf (accessed 3 September 2023).

EPA 2018 *2018 Edition of the Drinking Water Standards and Health*. Available from: https://www.epa.gov/system/files/documents/2022-01/dwtable2018.pdf (accessed 03 September 2023).

European Commission 1998 Council Directive 98/83/EC on the quality of water intended for human consumption. *Off. J. Eur. Commun.* **41**, 32–54.

Hastie, T. & Tibshirani, R. 1986 Generalized additive models. *Stat. Sci.* **1**, 297–318.

Hastie, R., Tibshirani, R. & Friedman, J. 2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. *Springer Series in Statistics*. Springer-Verlag, NY, USA, pp. 295–336.

Henriques, J. J. & Louis, G. E. 2011 A decision model for selecting sustainable drinking water supply and greywater reuse systems for developing communities with a case study in Cimahi, Indonesia. *J. Environ. Manage.* **92**, 214–222.

Hijmans, R. 2017 Package 'dismo'. Available from: https://mran.microsoft.com/snapshot/2017-02-04/web/packages/dismo/index.html (accessed 3 September 2023).

Ikonen, J., Pitkänen, T., Kosse, P., Ciszek, R., Kolehmainen, M. & Miettinen, I. T. 2017 On-line detection of Escherichia coli intrusion in a pilot-scale drinking water distribution system. *J. Environ. Manage.* **198**, 384–392.

Imen, S., Chang, N.-B. & Jeffrey Yang, Y. 2018 Developing a model-Based drinking water decision support system featuring remote sensing and fast learning techniques. *IEEE Syst J.* **12** (2), 1358–1368.

Interior Health Authority. Office of the Medical Health Officer. 2017 *Interior Health Clean Drinking Water*. Available from: https://drinkingwaterforeveryone.ca//files/IH_Drinking_Water_Report.pdf (accessed 3 September 2023).

IWA 2016 *The Bonn Charter for Safe Drinking Water*. Available from: https://iwa-network.org/wp-content/uploads/2016/06/Bonn-Charter-for-Safe-Drinking-Water.pdf (accessed 3 September 2023).

Khalilia, M., Chakraborty, S. & Popescu, M. 2011 Predicting disease risks from highly imbalanced data using random forest. *BMC Med Inform Decis. Mak.* **11**, 51–64.

Kim, A., Oh, K., Jung, J.-Y. & Kim, B. 2018 Imbalanced classification of manufacturing quality conditions using cost-sensitive decision tree ensembles. *Int. J. Comput. Integr. Manuf.* **31**, 701–717.

Knierim, K. J., Kingsbury, J. A., Haugh, C. J. & Ransom, K. M. 2020 Using boosted regression tree models to predict salinity in Mississippi embayment aquifers, Central United States. *J. Am. Water Resour. Assoc.* **56** (6), 1010–1029.

Knierim, K. J., Kingsbury, J. A., Belitz, K., Stackelberg, P. E., Minsley, B. J. & Rigby, J. R. 2022 Mapped predictions of manganese and arsenic in an alluvial aquifer using boosted regression trees. *Groundwater* **60** (3), 362–376.

Krawczyk, B., Woźniak, M. & Schaefer, G. 2014 Cost-sensitive decision tree ensembles for effective imbalanced classification. *Appl. Soft Comput.* **13**, 554–562.

Li, R. A., McDonald, J. A., Sathasivan, A. & Khan, S. J. 2021 A multivariate Bayesian network analysis of water quality factors influencing trihalomethanes formation in drinking water distribution systems. *Water Res.* **190**, 116712.

Lobo, G. P., Laraway, J. & Gadgil, A. J. 2022 Identifying schools at high-risk for elevated lead in drinking water using only publicly available data. *Sci. Total Environ.* **803**, 150046.

McCullagh, P. & Nelder, J. 1989 *Generalized Linear Models*. Chapman & Hall, 2nd edn. CRC, Philadelphia, PA, USA.

Mian, H. R., Chhipi-Shrestha, G., Hewage, K., Rodriguez, M. J. & Sadiq, R. 2020 Predicting unregulated disinfection by-products in small water distribution networks: An empirical modelling. *Environ. Monit. Assess.* **192** (497), 1–20.

Mohammed, H., Hameed, I. A. & Seidu, R. 2017 Random forest tree for predicting fecal indicator organisms in drinking water supply. In: *International Conference on Behavioral, Economic, Socio-cultural Computing (BESC)*. Krakow, Poland, pp. 1–6.

Mohammed, H., Hameed, I. A. & Seidu, R. 2018 Comparative predictive modelling of the occurrence of faecal indicator bacteria in a drinking water source in Norway. *Sci. Total Environ.* **628–629**, 1178–1190.

Nolan, B. T., Fienen, M. N. & Lorenz, D. L. 2015 A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. *J. Hydrol.* **531** (3), 902–911.

Podgorski, J. & Berg, M. 2020 Global threat of arsenic in groundwater. *Science* **368** (6493), 845–850.

Ridgeway, G. 2020 Generalized Boosted Models: A *Guide to the gbm Package*. Available from: https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf (accessed 3 September 2023).

R-project.org. N.A. *R: gbm step*. Available from: https://search.r-project.org/CRAN/refmans/dismo/html/gbm.step.html (accessed 3 September 2023).

Sadiq, R., Saint-Martin, E. & Kleiner, Y. 2008 Predicting risk of water quality failures in distribution networks under uncertainties using fault-tree analysis. *Urban Water J.* **5**, 287–304.

Schmidt, A., Ellsworth, L. M., Tilt, J. H. & Gough, M. 2022 Predicting conditional maximum contaminant level exceedance probabilities for drinking water after wildfires with Bayesian regularized network. *Mach. Learn. Appl.* **7**, 100227.

Stackelberg, P. E., Belitz, K., Brown, C. J., Erickson, M. L., Elliott, S. M., Kauffman, L. J., Ransom, K. M. & Redd, J. E. 2020 Machine learning predictions of pH in the glacial aquifer system, Northern USA. *Groundwater* **59** (3), 352–368.

Sun, Y., Wong, A. & Kamel, M. 2009 Classification of imbalanced data: A review. *Intern. J. Pattern Recognit. Artif. Intell.* **23**, 687–719.

UN 2010 Resolution Adopted by the General Assembly on 28 July 2010. Available from: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N09/479/35/PDF/N0947935.pdf?OpenElement (accessed 3 September 2023).

Uruguay Presidencia 2011 Decreto 375/011. | Uruguay Presidencia – GUB.UY. Available from: http://archivo.presidencia.gub.uy/sci/decretos/2011/11/msp_291.pdf (accessed 3 September 2023).

Vidal, L., Antúnez, L., Rodríguez-Haralambides, A., Giménez, A., Medina, K., Boido, E. & Ares, G. 2018 Relationship between astringency and phenolic composition of commercial Uruguayan Tannat wines: Application of boosted regression trees. *Food Res. Int.* **112**, 25–37.

WHO 2005 Water Safety Plans: Managing drinking-water quality from catchment to consumer. Available from: https://iris.who.int/bitstream/handle/10665/42890/WHO_SDE_WSH_05.06_eng.pdf?sequence=1.

WHO 2022 *Guidelines for Drinking-water Quality*: Fourth edition incorporating the first and second addenda. Geneva: World Health Organization. Licence: CC BY-NC-SA 3.0 IGO.

Wu, J., Song C, C., Dubinsky, E. & Stewart, J. 2021 Tracking major sources of water contamination using machine learning. *Front. Microbiol.* **11**, 616692.

Xu, T., Coco, G. & Neale, M. 2020 A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning. *Water Res.* **177**, 115788.

Xu, Z., Shen, J., Qu, Y., Chen, H., Zhou, X., Hong, H., Sun, H., Lin, H., Deng, W. & Wu, F. 2022 Using simple and easy water quality parameters to predict trihalomethane occurrence in tap water. *Chemosphere* **286**, 131586.

First received 15 September 2023; accepted in revised form 27 February 2024. Available online 20 March 2024