Online Journal of Biological Sciences

Original Research Paper

Prediction of Physicochemical Parameters in a Rainwater Harvesting and Treatment System Modeled Using the Yahtun Algorithm in Native Communities of Northern Peru

[1]Eli Morales-Rojas, [2]Edwin Adolfo Díaz Ortiz, [3]José Emmanuel Cruz de La Cruz, [4]Wilson Antony Mamani Machaca, [1]Ítalo Maldonado Ramírez, [5]Victor Yana-Mamani and [1]Gustavo Taboada

[1]Instituto de Investigación en Tecnología de Información y Comunicación (IITIC), Facultad de Ingeniería de Sistemas y Mecánica Eléctrica, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas, Jr. Libertad, Bagua, Amazonas, Peru
[2]Escuela Profesional de Ingeniería Civil, Facultad de Ingeniería Civil y Ambiental, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas, Triunfo, Chachapoyas, Amazonas, Peru
[3]Facultad de Mecánica Eléctrica, Electrónica y Sistemas, Universidad Nacional del Altiplano, Puno, Peru
[4]Universidad de Alicante, 03690, Alicante, Spain
[5]Escuela de Ingeniería de Sistemas e Informática, Facultad de Ingeniería, Universidad Nacional de Moquegua, Moquegua 18611, Peru

Article history
Received: 01-09-2024
Revised: 10-12-2024
Accepted: 10-01-2025

Corresponding Author:
Eli Morales-Rojas
Facultad de Ingeniería de Sistemas y Mecánica Eléctrica, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas; Jr. Libertad, Bagua, Amazonas. Perú
Email: eli.morales@untrm.edu.pe

Abstract: The inhabitants of the native communities of northern Peru do not have drinking water systems, generating serious health impacts. In this sense, rainwater harvesting systems have become an alternative for these areas. Therefore, the objective of this study was to predict the physicochemical parameters in a rainwater collection and treatment system, modeled by means of a novel "Yahtun" algorithm. The physicochemical quality data for the modeling were obtained from the PROLLUVIA project, executed during 2019 and 2020. With respect to the goodness of fit of the proposed model, the Mean Absolute Error (MAE), Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) metrics were used. For the increase of the data obtained, as well as for the generation of a regression model, data imputation techniques and an ensemble model were used for data prediction. The high-level programming language Python was used for the analysis and modeling. The results indicate that the Yahtun model is efficient for the prediction of water quality in the context of native communities, according to the MAE metrics was 84.16%, MSE was 86.21% and ASM was 83.63%. In this sense, it is important to apply data imputation techniques in order to estimate future predictions in the context of native communities.

Keywords: Data Imputation, Synthetic Data, Water Quality, Ensemble Method

Introduction

Water quality plays an important role in human health, however, in developing countries, the quality of drinking water is not adequate (Li and Wu, 2019) and can be reflected in the degree of contamination which is a worrying factor for humanity (Abdel-satar *et al.*, 2017). Water quality can be altered by different sources, such as factories, mining activities (heavy metals), food processing waste, agricultural runoff, animal waste, disposal of personal care products, and household chemicals (Rojas *et al.*, 2020).

In rural areas of Peru, 28.1% of people in rural areas do not have access to water by a public network, of which 16.9% access rivers, irrigation ditches, or spring water (INEI, 2018), this allows evaluating the search for new sources of supply (Gastañaga, 2018) such as rainwater that could be used for domestic purposes with minimal treatment (Adeyeye *et al.*, 2020; Roblero *et al.*, 2019). Rainwater can be used for human consumption, as long as it meets basic conditions in terms of physicochemical and microbiological parameters (Díaz Ortiz and Medina Tafur, 2021: MINSA, 2010).

In this sense, it is important to evaluate physicochemical parameters such as pH and turbidity, since inorganic particles are eliminated at approximately neutral pH values (LeChevallier *et al.*, 1981; Naceradska *et al.*, 2019), also pH plays a regulatory role in the chlorination and chlorination kinetics for ionizable toxic organic compounds in water

Science Publications

(Ge *et al*., 2006). On the other hand, aluminum is important to estimate because it may be associated with diseases such as Alzheimer's disease (Gauthier *et al*., 2000), other studies mention that exposure to high levels of aluminum can cause problems to the musculoskeletal system and kidneys (Rahimzadeh *et al*., 2022).

Consequently, traditional monitoring techniques exist and have demonstrated delayed results, labor-intensive processes and lack of real-time data (Satyanarayana Murthy and Ahamed, 2023). In addition, the small amount of data is a constraint for decision making, especially if native communities are involved and costly logistics are required to conduct water quality monitoring.

Therefore, in the scientific literature, the use of synthetic data has been suggested as an alternative to data transformation, as these are artificially created to preserve the statistical characteristics of the original data so that they exhibit a distribution and correlational structure similar to that of the original data.

These synthetic data have significant potential where real data sets may not be available. However, data generation will depend on the area of study, such as synthetic medical data faces unique challenges due to the inherent complexity and longitudinal nature of the data (Murtaza *et al*., 2023). Researchers are exploring various methods to generate realistic synthetic data and have proposed multiple quality assessment metrics to evaluate their suitability as a surrogate for real data.

There are several ways to generate synthetic data. A common method to generate new samples is the Synthetic Minority Oversampling Technique (SMOTE) algorithm and its variations, which generates new data by interpolating the available originals. Other common generative models are Generative Adversarial Networks (GAN) and Variational Automatic Encoders (VAE) (Espinosa and Figueira, 2023).

However, to ensure that these synthetic data are useful and do not just add noise to our real data set, it is important to verify and assess whether they are representative of the real sample. Therefore, we need objective tools to compare the synthetic data with the real data and then evaluate the differences (Arjovsky *et al*., 2017).

To determine the efficiency of the proposed Yathun framework, as well as when more than one forecasting technique is to be compared, forecast accuracy measures can also be used to discriminate between competing models (Chicco *et al*., 2021; Tanwar and Kakkar, 2017). Therefore, the following metrics were used: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE). In this sense, this study will serve to identify trends and predict physicochemical parameters (pH, turbidity, and aluminum) in the quality of rainwater and this research is one of the pioneers in Amazonas to address the problem of water pollution and provide rapid attention in order to improve the health problems of the Awajún population.

Based on the above, the objective of this project was to forecast the physicochemical parameters in a rainwater collection and treatment system installed in two native communities, modeled using the Yahtun algorithm.

Materials and Methods

Location

The study is located in two native communities, inhabited by the Awajún peoples (Tunants and Yahuahua), district of Nieva, province of Condorcanqui in the northern jungle of Peru, at an altitude of 196 masl, average temperature of 26°C and average annual rainfall of 3,121 mm (García, 2010). The data used to forecast the physicochemical parameters was obtained from the PROLLUVIA project executed during 2019 and 2020, which consisted of the construction and installation of 4 rainwater harvesting systems according to the coordinates shown in Table (1).

The installation was carried out ensuring that it complied with the minimum conditions of area (place and area of the systems) and number of users. The construction consisted of iron, cement and pipes (PVC) Fig. (1). The supporting structure of the tank was built with a mixture of concrete and cement, reinforced with corrugated steel (Morales Rojas *et al*., 2021).

Table 1: Coordinates of rainwater harvesting systems in native communities

Catchment systems	E coordinate	Coordinate N
System 1	830543	9481810
System 2	830880	9481944
System 3	832070	9482801
System 4	167331	9482999

Source: Prolluvia Project (contract no. 185-2018-FONDECYT-MB-IADT-SE)

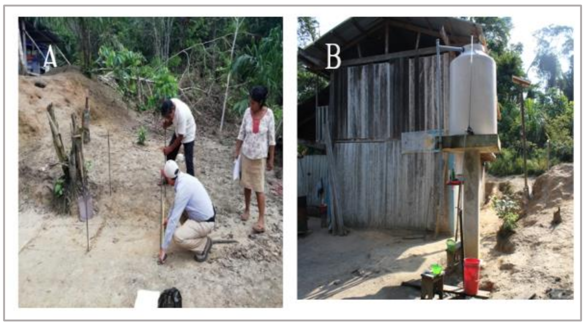


Fig. 1: Rainwater harvesting systems; (a) site selection; (b) water harvesting system installed; Source: Prolluvia Project (contract no. 185-2018-FONDECYT-MB-IADT-SE)

Characterization was during the rainy season (December 2019 and January and February 2020) and two months of the dry season (September and October 2020). Sample collection, storage, and transfer, as well as laboratory analysis, were performed in accordance with APHA, AWWA, and WEF (Gilcreas, 1967). The pH analysis was in situ, with a Hanna multiparametric water meter model HI 98194, while turbidity and aluminum were analyzed at the Water and Soil Laboratory of the Research Institute for Sustainable Development of Ceja de Selva (INDES-CES) of the National University Toribio Rodríguez de Mendoza (UNTRM).

Comparison of Values with National Standards

The results of the rainwater characterization were compared with Supreme Decree No. 031-2010-SA (MINSA, 2010) Table (2).

Yahtun Framework Forecasting of Physical-Chemical Parameters

The Mean Absolute Error (MAE), Mean Squared Error, (MSE) and Mean Absolute Percentage Error (MAPE) metrics (Chatterjee and Byun, 2023; Tatachar, 2021)were used to determine the goodness of the proposed model, as well as the comparison with the reference model (Table 3).

The Yahtun framework was used to increase the data obtained, as well as to generate a regression model for the analysis of rainwater in native communities in northern Peru. The Yathun framework basically consists of data imputation techniques, synthetic data generation from random noise, and an ensemble model to perform data prediction (Fig. 2). This was analyzed from the original data (pH, Turbidity, and Aluminum) in order to discover if there are missing data. If this is the case, the Knn imputation technique is applied to complete the data. Subsequently, the data is separated into two groups: 50% for training and 50% for test. From the 50% training data, we proceeded to generate synthetic data using noise generation, with a low threshold to maintain the variability of the original data. A quantity of data equal to the training data was synthesized. To determine the quality of the generated data, an ensemble method was used: XGBoost for data regression.

The final results obtained with the synthetic data were compared with the original 50% of data separated for the Test in the first stage, to determine the efficiency of the system. We took as metrics the MAE, MSE and MAPE, which are metrics related to the error of the system.

Table 2: Water quality parameters according to Supreme Decree No. 031-2010-SA

Parameters	Unit	Maximum allowable limit
pH	-	6,5 - 8,5
Turbidity	UNT	5
Aluminum	mg Al L-1	0,2

Table 3: Metrics used in Yahtun framework forecasting

Metrics	Equation
Mean Absolute Error (MAE)	$MAE = \frac{1}{m}\sum_{i=1}^{m}\lvert X_i - Y_i\rvert$ Con Xi is the predicted ith value and the Yi element is the actual ith value.
Mean Squared Error (MSE)	$MSE = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2$
Mean Absolute Percentage Error (MAPE)	$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left\lvert\frac{Y_i - X_i}{Y_i}\right\rvert$

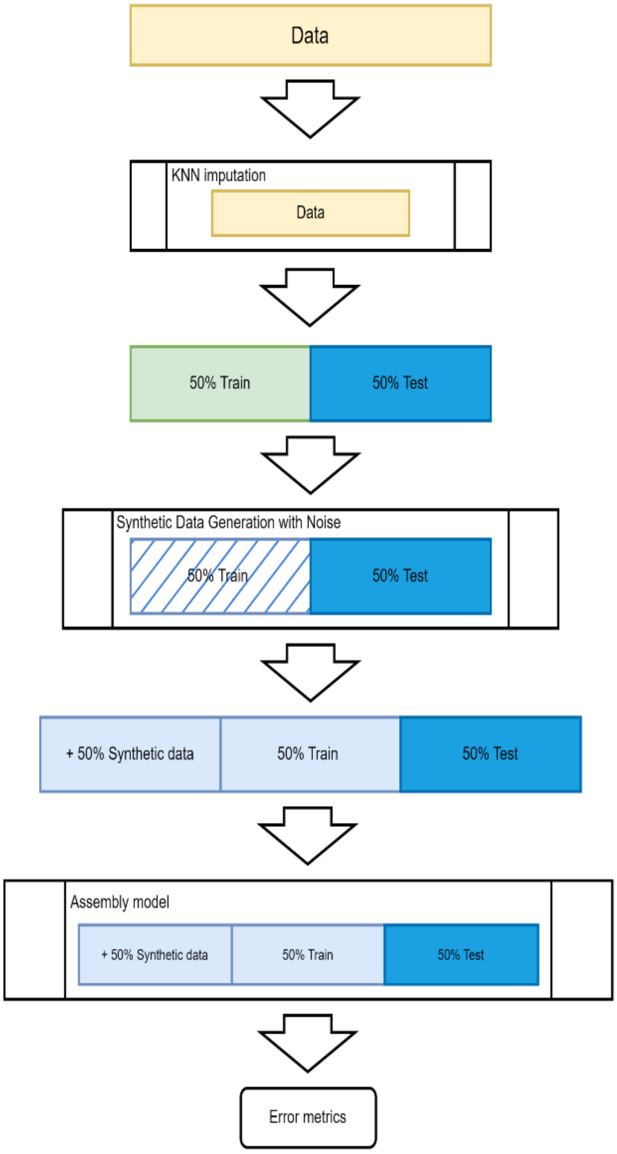


Fig. 2: Flow diagram of the Yahtun model

Data Analysis and *Modeling*

A computer with an Intel (R) Core (TM) i5-7200U processor, 24 GB RAM, and Windows 10 operating system was used for the data analysis. For the analysis and modeling, the high-level programming language Python was used with the following libraries: Pandas, seaborn, sklearn, numpy, and XGBoost.

Results and Discussion

Figure (3) presents the amount of missing data in the original data set. Where the variable "pH" was considered as the dependent variable, while the independent variables were: "Turbidity (UNT)" and "Aluminum". It was observed that there is 20% of missing data for the independent variables, therefore, Knn was applied as an imputation technique to complete the missing data. In water quality studies, data imputation is important because it will help to complete missing data that are not sampled due to economic issues and accessibility are not sampled, this will help decision-makers to formulate appropriate management strategies (Ratolojanahary *et al*., 2019; Rodríguez *et al*., 2021), in water quality, meteorology and water quantity (Pastorini *et al*., 2024). In addition, these results will be useful for adequate management, framed in the transfer of emerging technologies that will provide updated data for decision-making in native communities (Leutenegger *et al*., 2024).

Figure (4) shows the distribution of incomplete data over time. This representation is useful to identify the time periods where data are missing. In addition, it is positively observed that missing data occur at different time points for both variables that have incomplete data, as this suggests a lack of systematic bias in the absence of data.

Figure (5a) shows a boxplot graph, which allows visualizing the distribution of the data. Mainly, this type of graph helps to identify possible outliers through the analysis of quartiles. In that sense, the presence of an outlier in the variable "pH" is evident, which is located near the value 1. This outlier affects the normal distribution of the data. Therefore, it is necessary to eliminate these outliers so that the model can work correctly and make more accurate predictions with a lower error. This process was also performed for the other two variables (Turbidity and Aluminum). Figure (5b), shows the results generated after removing the outliers, where it exhibits a normal distribution, suggesting a more accurate capture of the patterns for prediction (Roy, 2003).

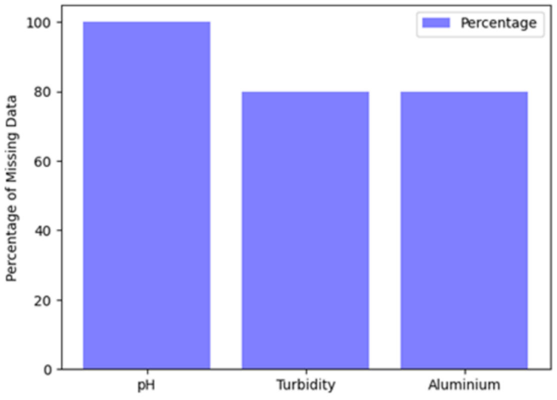


Fig. 3: Missing data for the variables under study

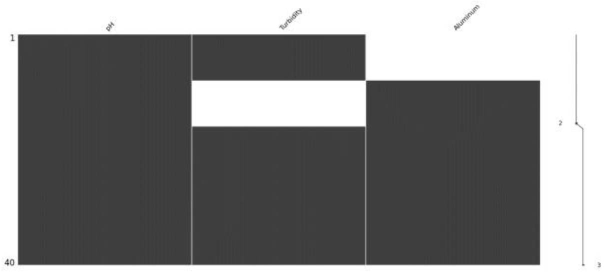
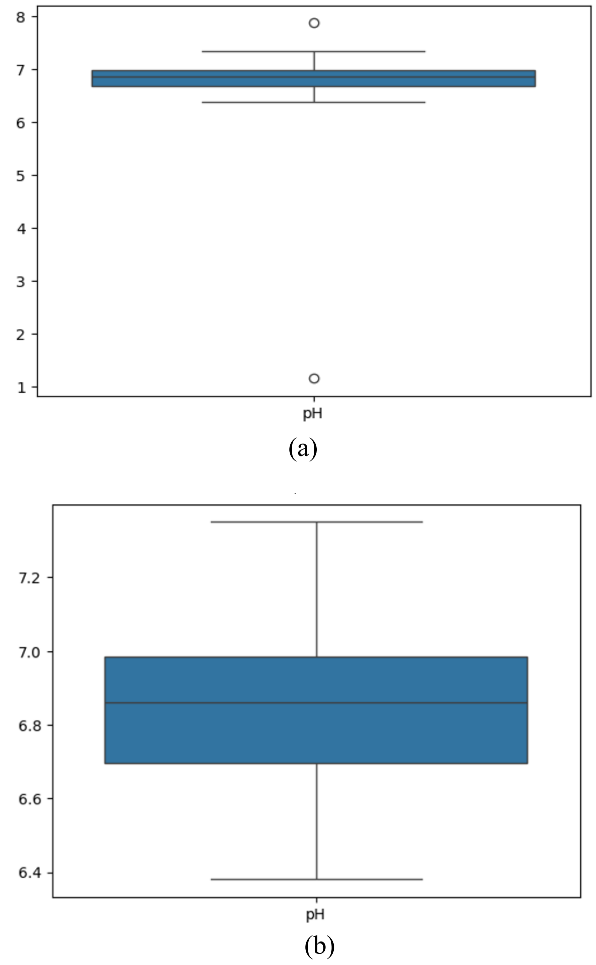


Fig. 4: Distribution of data over time



(a)

(b)

Fig. 5: Boxplot of original data (a); Boxplot of processed data (b)

Table 3: Descriptive statistics of the variables under study

Parameters	Count	MPL*	Mean	Std	Min	25%	50%	75%	Max
pH	38	6.5 an 8.5	6.878684	0.220934	6.380000	6.695000	6.860000	6.985000	7.350000
Turbidity	30	5	4.267667	15.637354	0.500000	1.025000	1.400000	1.800000	87.000000
Aluminum	30	0.2	0.295067	0.437833	0.000000	0.135000	0.205000	0.266750	2.280000

*LMP (D.S. N° 031-2010-SA).

Table (3), shows the behavior of rainwater parameters, where pH and turbidity comply with Peruvian regulations, however, aluminum is above the standard. In that sense, aluminum in rainwater should be addressed as it is a heavy metal directly exposed to human health and can cause lung and bladder cancer (Fathi *et al*., 2022). Therefore, accurate prediction of water quality will help in proper water management with the aim of keeping pollution within permissible limits (Najah *et al*., 2013).

For the simulations of the two models, i.e., the base model XGBoost and the proposed Framework Yathun, the same conditions were established for the training and test processes. In this sense, it is evident that the Yathun framework model is the most optimal, unlike the base model “XGboost” (Table 4).

Figure (6) evidences a percentage improvement (decrease in error) for MAE of 84.16%, for MSE of 86.21%, and finally for MAPE of 83.63%, reflecting the remarkable performance of the developed model (Yathun). These metrics are widely used to measure the accuracy of models (Khullar and Singh, 2022). The models can be categorized as excellent prediction (MAPE ≤10%), good prediction (10%< MAPE ≤20%), acceptable prediction (20%< MAPE ≤50%), and inaccurate prediction (50%< MAPE) (Lewis, 1982). The limitations of the work focus mainly on the number of parameters trained, therefore, future work is needed that focuses on the prediction of heavy metals in the main surface sources of water for human consumption that are contaminated by anthropogenic activities, such as informal mining.

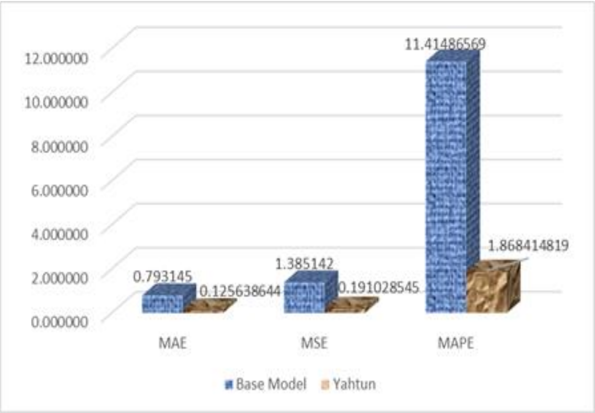


Fig. 6: Comparison of Yahtun model metrics with the base model

Table 4: Results of models: Base and Yahtun

Metrics	Base Model	Yahtun
MAE	0.793145	0.12563864
MSE	1.385142	0.19102854
MAPE	11.4148657	1.86841482

Conclusion

It is evident that the imputation technique to complete the missing data is important to estimate prediction models in rainwater quality, in that sense it was evidenced that the new proposed model "Yahtun" showed good efficiency reaching the MAE of 84.16% the MSE was 86.21% and MAPE of 83.63% being solid results and can be used to verify the water quality in native communities and avoid high logistical costs.

Our current results are promising and future work will address a greater number of physicochemical parameters such as lead, cadmium, and mercury, taking into account that native communities do not have access to basic sanitation. In these communities, water contamination by heavy metals constitutes a serious public health problem, due to the food trophic chain and its final arrival to humans, which can cause chronic intoxication.

Finally, these results are of great importance for local governments, researchers and policy makers to make the right decisions at the right time.

Acknowledgment

The authors acknowledge and thank the support of the Instituto de Investigación para el Desarrollo Sustentable de Ceja de Selva (INDES-CES) of the Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM).

Funding Information

The authors acknowledge and appreciate the support of the Instituto de Investigación para el Desarrollo Sustentable de Ceja de Selva (INDES-CES) of the Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM). This research was funded by Fondo Nacional de Desarrollo Científico, Tecnológico y de Innova-ción Tecnológica—FONDECYT, contract no. 185-2018-FONDECYT-MB-IADT-SE (PROLLUVIA project).

Declaration of Competing 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 study.

Author's Contributions

Eli Morales-Rojas: Conceptualization, drafting, and revision of the final version.

Edwin Adolfo Díaz Ortiz: Conceptualization, drafting, and review of final version.

José Emmanuel Cruz de La Cruz: Conceptualization, drafting, and review of the final version.

Wilson Antony Mamani Machaca: Manuscript preparation, statistical analysis, and final version editing.

Ítalo Maldonado Ramírez: Methodology, project management, resources, software, validation and data visualization.

Victor Yana-Mamani and Gustavo Taboada: Manuscript writing, data collection, analysis, and project management. Finally, all authors have read and accepted the final version of the manuscript.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all authors have read and approved the manuscript and that there are no ethical issues.

References

Abdel-Satar, A. M., Ali, M. H., & Goher, M. E. (2017). Indices of Water Quality and Metal Pollution of Nile River, Egypt. *Egyptian Journal of Aquatic Research*, *43*(1), 21–29. https://doi.org/10.1016/j.ejar.2016.12.006

Adeyeye, J. A., Akintan, O. B., & Adedokun, T. (2020). Physicochemical Characteristics of Harvested Rainwater Under Different Rooftops in Ikole Local Government Area, Ekiti State, Nigeria. *Journal of Applied Sciences* and *Environmental Management*, *23*(11), 2003–2008. https://doi.org/10.4314/jasem.v23i11.15

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein Generative Adversarial Networks. *34th International Conference on Machine Learning*, *ICML*, 214–223.

Chatterjee, S., & Byun, Y.-C. (2023). A Synthetic Data Generation Technique for Enhancement of Prediction Accuracy of Electric Vehicles Demand. *Sensors*, *23*(2), 594–1048. https://doi.org/10.3390/s23020594

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The Coefficient of Determination R-Squared is More Informative Than SMAPE, MAE, MAPE, MSE, and RMSE in Regression Analysis Evaluation. *PeerJ Computer Science*, *7*, 623. https://doi.org/10.7717/peerj-cs.623

Díaz Ortiz, E. A., & Medina Tafur, C. A. (2021). Demand, Collection, and Quality of Rainwater in the Native Community Yahuahua, Nieva, Amazonas (Perú). *REBIOL*, *40*(2), 188–205. https://doi.org/10.17268/rebiol.2020.40.02.07

Espinosa, E., & Figueira, A. (2023). On the Quality of Synthetic Generated Tabular Data. *Mathematics*, *11*(15), 3278–3436. https://doi.org/10.3390/math11153278

Fathi, A., Mahmoud, N., Issam, A. A.-K., & Yung-Tse, H. (2022). Heavy Metals in Harvested Rainwater Used for Domestic Purposes in Rural Areas: Yatta Area, Palestine as a Case Study. International Journal of Environmental Research and Public Health, 19(5), 2683–3146. https://doi.org/10.3390/ijerph19052683

García, J. M. (2010). Hidrografía - Zonificación Ecológica Económica del departamento de Amazonas. *Journal of Chemical Information* and *Modeling*, *53*(9), 1689–1699.

Gastañaga, M. del C. (2018). Agua, Saneamiento Y Salud. *Revista Peruana de Medicina Experimental y Salud Pública*, *35*(2), 181–182. https://doi.org/10.17843/rpmesp.2018.352.3732

Gauthier, E., Fortier, I., Courchesne, F., Pepin, P., Mortimer, J., & Gauvreau, D. (2000). Aluminum Forms in Drinking Water and Risk of Alzheimer's Disease. *Environmental Research*, *84*(3), 234–246. https://doi.org/10.1006/enrs.2000.4101

Ge, F., Zhu, L., & Chen, H. (2006). Effects of pH on the Chlorination Process of Phenols in Drinking Water. *Journal of Hazardous Materials*, *133*(1–3), 99–105. https://doi.org/10.1016/j.jhazmat.2005.09.062

Gilcreas, F. W. (1967). Future of Standard Methods for the Examination of Water and Wastewater. *Health Laboratory Science*, *4*(3), 137–141.

INEI. (2018). Perú: formas de acceso a agua y saneamiento básico. *Instituto Nacional Estadística E Informática. Instituto Nacional Estadística E Informática*, 1–69.

Khullar, S., & Singh, N. (2022). Water Quality Assessment of a River Using Deep Learning Bi-LSTM Methodology: Forecasting and Validation. *Environmental Science* and *Pollution Research*, *29*(9), 12875–12889. https://doi.org/10.1007/s11356-021-13875-w

LeChevallier, M. W., Evans, T. M., & Seidler, R. J. (1981). Effect of Turbidity on Chlorination Efficiency and Bacterial Persistence in Drinking Water. *Applied* and *Environmental Microbiology*, *42*(1), 159–167. https://doi.org/10.1128/aem.42.1.159-167.1981

Lewis, C. D. (1982). Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting. *Butterworth Scientific*, 111–153.

Li, P., & Wu, J. (2019). Drinking Water Quality and Public Health. *Exposure* and *Health*, *11*(2), 73–79. https://doi.org/10.1007/s12403-019-00299-8

MINSA. (2010). Reglamento de la Calidad del Agua para Consumo Humano. DS N° 031-2010-SA, 20–25. https://doi.org/10.1130/micro18-p20

Morales Rojas, E., Díaz Ortiz, E. A., Medina Tafur, C. A., García, L., Oliva, M., & Rojas Briceño, N. B. (2021). A Rainwater Harvesting and Treatment System for Domestic Use and Human Consumption in Native Communities in Amazonas (NW Peru): Technical and Economic Validation. *Scientifica*, *2021*, 1–17. https://doi.org/10.1155/2021/4136379

Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., & Bano, A. (2023). Synthetic Data Generation: State of the Art in Health Care Domain. *Computer Science Review*, *48*, 100546. https://doi.org/10.1016/j.cosrev.2023.100546

Naceradska, J., Pivokonska, L., & Pivokonsky, M. (2019). On the Importance of pH Value in Coagulation. *Journal of Water Supply: Research* and *Technology-Aqua*, *68*(3), 222–230. https://doi.org/10.2166/aqua.2019.155

Najah, A., El-Shafie, A., Karim, O. A., & El-Shafie, A. H. (2013). Application of Artificial Neural Networks for Water Quality Prediction. *Neural Computing* and *Applications*, *22*(1), 187–201. https://doi.org/10.1007/s00521-012-0940-3

Pastorini, M., Rodríguez, R., Etcheverry, L., Castro, A., & Gorgoglione, A. (2024). Enhancing Environmental Data Imputation: A Physically-Constrained Machine Learning Framework. *Science of The Total Environment*, *926*, 171773–171951. https://doi.org/10.1016/j.scitotenv.2024.171773

Rahimzadeh, M. R., Rahimzadeh, M. R., Kazemi, S., Amiri, R. J., Pirzadeh, M., & Moghadamnia, A. A. (2022). Aluminum Poisoning with Emphasis on Its Mechanism and Treatment of Intoxication. *Emergency Medicine International*, *2022*(1), 1–13. https://doi.org/10.1155/2022/1480553

Roblero, J. U. A., Bravo, J. R. S., Acevedo, A. D., Cruz, C. L. de la, & Villa, O. R. M. (2019). Validation of a Prototype of Rainwater Harvesting System for Domestic use and Human Consumption. *Idesia*, *37*(1), 53–59. https://doi.org/10.4067/S0718-34292019005000302

Ratolojanahary, R., Houé Ngouna, R., Medjaher, K., Junca-Bourié, J., Dauriac, F., & Sebilo, M. (2019). Model Selection to Improve Multiple Imputation for Handling High Rate Missingness in a Water Quality Dataset. *Expert Systems with Applications*, *131*, 299–307. https://doi.org/10.1016/j.eswa.2019.04.049

Rodríguez, R., Pastorini, M., Etcheverry, L., Chreties, C., Fossati, M., Castro, A., & Gorgoglione, A. (2021). Water-Quality Data Imputation with a High Percentage of Missing Values: A Machine Learning Approach. *Sustainability*, *13*(11), 6318–6507. https://doi.org/10.3390/su13116318

Rojas, E. M., Rascón, J., Huatangari, L. Q., Quintana, S. C., Oliva, M., & Pino, M. E. M. (2020). Mixed Greywater Treatment for Irrigation Uses. *Ambiente e Agua - An Interdisciplinary Journal of Applied Science*, *15*(6), 1–11. https://doi.org/10.4136/ambi-agua.2599

Roy, D. (2003). The Discrete Normal Distribution. *Communications in Statistics - Theory* and *Methods*, *32*(10), 1871–1883. https://doi.org/10.1081/sta-120023256

Satyanarayana Murthy, N., & Ahamed, S. (2023). Water Quality Monitoring and Measuring Physicochemical Parameters Using Wireless Sensor Networks. *African Journal of Aquatic Science*, *48*(4), 366–373. https://doi.org/10.2989/16085914.2023.2277955

Tanwar, H., & Kakkar, M. (2017). Performance comparison and future estimation of time series data using predictive data mining techniques. *2017 International Conference on Data Management, Analytics* and *Innovation (ICDMAI)*, 9–12. https://doi.org/10.1109/icdmai.2017.8073477

Tatachar, A. V. (2021). Comparative Assessment of Regression Models Based On Model Evaluation Metrics. *International Research Journal of Engineering* and *Technology*, *8*(9), 853–860.