Dataset for Assessing Water Quality for Drinking and Irrigation Purposes using Machine Learning Models.

Olasupo Ajayi, Antoine Bagula, Hloniphani Maluleke

Department of Computer Science, University of the Western Cape,

Cape Town, South Africa.

Abstract

Access to potable water is a critical requirement for human survival. Beyond drinking, water is also necessary for animal consumption, irrigation, as well as domestic and commercial uses. Laboratory assessments of water samples to determine their fitness for use is a vital step in water quality assurance processes. However, laboratory assessments require adherence to stringent measures, which might be difficult to comply with. Machine learning (ML) has emerged in recent years as viable and cheaper solutions to complement (or replace) lab-based assessments, with a caveat of availability of sufficient data to train the ML models. Unfortunately, such data are not always (or sparsely) available, especially in less developed countries. To this end, the work attempts to fill this gap by creating ample sized datasets that can be used to train (and test) ML models. Two datasets are curated in this work, one for drinking water and the other for irrigation water. The datasets were curated by aggregating data from smaller datasets on related concepts, then processed and labelled to make them useful for supervised ML models. To prove the applicability of the curated datasets, they were used to train ML models in a related work and yielded good results.

Keywords

Drinking Water; Irrigation Water; Machine Learning; Physico-chemical parameters; Water Quality Index

1. Data Description

Water parameters are often categorised into three groups, which are physical characteristics, such as temperature and clarity; chemical constituent, such as the percentage of sodium present, amount of dissolved oxygen and solids, pH level etc.; and microbial characteristics, such as presence of microorganisms. The datasets provided in this work are for the physical and chemical components only, i.e., the physico-chemical parameters of drinking and irrigation water, and this is like work done in [1, 2, 3]. Two datasets are presented in work, one for drinking water and the second for irrigation water. Both datasets were created by aggregating smaller datasets scrapped from the Internet.

For the drinking water dataset, parameters considered were pH, Na, Mg, Ca, Cl, K, CO_3, HCO_3, SO_4, Turbidity, TDS, EC, and TH, while for irrigation water RSC, PI. KR. MH, Na%, SAR, and SSP were considered. For both datasets, an additional "label" column was introduced. This last column had values of "1" or "0", which respectively indicated if the sample was "fit for use" or not. For the drinking water dataset this additional column was labelled "Potability", while for the irrigation dataset it was termed "Usable".

2. Summary of Dataset

Subject	Computer Science, Hydrology, IoT
Specific subject area	Water Quality Prediction, Machine Learning, Water Sensor Network
How data was acquired	Data was curated through the aggregation of several smaller datasets. These smaller datasets were compared and combined based on common features (parameters) after converting all parameters to the same units. Python script was used to calculate the "fitness of use" of each water sample using regional or global accepted range. Finally, two datasets were obtained, the first for drinking water with about 720 samples and the second for irrigation water with 360 data samples.
Description of data	Physico-chemical parameters of water samples.
Data parameters	Potential of Hydrogen (pH), Sodium (Na), Magnesium (Mg), Calcium (Ca), Chloride (Cl), Potassium (K), Carbonate (CO_3), Bicarbonate (HCO_3), Turbidity, Total Dissolved Solids (TDS), Electrical conductivity (EC), Total Hardness (TH), Residual Sodium Carbonate (RSC), Permeability Index (PI), Kelly's Ratio (KR), Magnesium Hazard (MH), Sodium Percentage (Na%), Sodium Adsorption Ratio (SAR), Soluble Sodium Percentage (SSP)
Data source	Data on drinking and irrigation water scrapped from repositories on the Internet.
Related research article	Ajayi O, Bagula A, Maluleke C, Gaffoor Z, Jovanovic N, Pietersen K. WaterNet: A Network for Monitoring and Assessing Water Quality for Drinking and Irrigation Purposes. IEEE Access, vol. 10, pp. 48318-48337, 2022, doi: 10.1109/ACCESS.2022.3172274. [4]

3. Materials and Methodology

a. Data Aggregation

The emphasis of this work is on creating datasets that can be used to train and test machine learning models for water analysis. To curate these datasets, we carried out a systematic survey of relevant literature and aggregated data from them. The process involved searching through online repositories using keywords, sieving out unrelated articles, then comparing the selected articles for common feature sets (water parameters).

In selecting relevant articles, we used "irrigation water", "potable water", "groundwater", and "drinking water" in "developing countries" as search phrases. We then filtered the resulting articles by discarded articles that did not include dataset(s), either as table(s) within the context or as attached supplementary file(s). After these processes, we were left with 11 publications, 7 of which also had data on irrigation water. Of the 11 articles, 4 were related to water bodies in India [1, 2, 3, 5], 3 in Iran [6, 7, 8], 2 in South Africa [9, 10], 1 in Bangladesh [11] and Ethiopia [12] each. Finally, only articles having at least 3 of the parameters in common were considered. Tables 1 and 2 show the distribution of features across the articles considered for both drinking and irrigation water respectively. Data were then scrapped from these articles, combined, and saved into two csv files, for drinking and irrigation respectively.

Table 1: Distribution of features for drinking water

Ref	pH	Na	Mg	Ca	Cl	K	CO_3	HCO_3	SO_4	NTU	TDS	EC	TH	Label
[1]	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓
[2]	✓	✓	✓	✓	✓				✓		✓	✓	✓	
[3]	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓
[5]		✓	✓	✓	✓	✓		✓	✓					
[6]	✓	✓	✓	✓			✓			✓	✓			
[7]	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓	
[8]	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	
[9]	✓	✓	✓		✓	✓	✓		✓	✓		✓	✓	
[10]	✓									✓		✓		
[11]	✓		✓	✓	✓	✓			✓		✓	✓	✓	
[12]	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		

Table 2: Distribution of features for irrigation water

Ref.	RSC	PI	KR	MH	Na%	SAR	SSP	Label
[1]	✓	✓	✓	✓	✓	✓	✓*	
[3]	✓	✓	✓	✓	✓	✓	✓	
[5]	✓	✓	✓	✓	✓	✓	✓	
[6]		✓*	✓*	✓*	✓	✓	✓*	✓
[7]	✓	✓	✓	✓	✓	✓	✓	
[8]	✓	✓	✓	✓	✓	✓	✓	
[11]		✓	✓	✓	✓	✓	✓	

"✓*" are parameters not explicitly included in the original dataset but calculated using other parameters present.

b. Standardising Units

The unit of measurement of certain water parameters differed across some of the datasets, hence there was a need to standardise the units. For instance, EC was measured in microsiemens/centimetre (μS/cm) in [1, 2, 10, 12], in millisiemens/metre (mS/m) in [9], and in

μmhos/cm in [2, 3, 7, 8]. To address this, μmhos/cm values were directly changed to μS/cm (as 1 μmhos/cm = 1 μS/cm), while values in μS/cm were divided by 10 to convert them to mS/m.

4. Data Processing & Labelling

For machine learning, especially supervised learning models [13], labelled data is very important, as the "labels" are used to train the models. In a bid to create datasets suitable for ML we had to label the data. For this work, the "label" represents the status of the water sample, i.e., fit for use (1) or not (0). From tables 1 and 2, only [1, 3, 6] were pre-labelled data, hence we had to manually include labels for the other unlabelled data samples. To obtain these labels we first calculated the Water Quality Index (WQI), then mapped it to 1 or 0. Algorithm 1 (adapted from [14]) shows the steps followed to achieve this.

Algorithm 1: Labelling the Dataset

1. Select n relevant parameters ($P = [P_1, P_2, P_3 \ldots P_n]$)
2. Obtain the concentration level of each parameter C_i ($1 < i < n$)
3. Assign weights to each parameter w_i, ($1 < i < n$)
4. Calculate the relative weight $R_i = w_i / (Sum_{i=1}^{n} w_i)$
5. Calculate quality index $Q_i = C_i / (S_i * 100)$
6. Obtain the Water Quality Index $WQI = Sum_{i=1}^{n} R_i * Q_i$
7. Label the data sample

a. For drinking water:
   i. if WQI < 50, label = 1
   ii. else if WQI >= 50, label = 0
b. For irrigation water:
   i. if IWQI >= 50, label = 1
   ii. else if IWQI < 50, label = 0

Where S_i is the accepted value of parameter i as stipulated by the World Health Organisation (WHO) [15] or local adaptation of the value, as is the case with the South African National Standard (SANS) [17].

The WQI and Irrigation WQI (IWQI) are unitless metrics used for assessing the suitability of water for human use (and/or consumption) and for irrigation purposes respectively. Both values are well used in literature to assess water quality and are calculated using steps spelt out in [14] (for WQI) and [16] for IWQI. Though WQI and IWQI are widely used, they have some limitations, some of which are discussed in [4].

For this work and for consistency purposes, we used the same method to calculate both WQI and IWQI. We also set the threshold value of "fitness for use" to 50 for both drinking and irrigation water. The value of 50 was chosen because it was also used in most of the articles on Tables 1 and 2 to indicate water of good or acceptable quality [2, 3, 6]. All parameters were also assigned equal weights. This was done to remove bias and ensure that the final datasets are generic enough to be applied anywhere. Finally, in determining the acceptable value for each parameter, i.e., S_i in Algorithm 1, Table 3 was used. The table summarises the acceptable range for various water parameters as stipulated by the WHO [15] and the SANS [17].

Table 3: Acceptable Range for Values for Water Parameters

Parameter	Unit / Formula	Range
pH		5 - 9.7
Na	mg/L	< 200
Mg	mg/L	< 50
Ca	mg/L	< 75
Cl	mg/L	< 300
K	mg/L	< 12
SO_4	mg/L	< 500
CO_3	mg/L	1.2 - 2
HCO_3	mg/L	120 - 200
Turbidity	NTU	< 5
TDS	mg/L	< 1200
EC	mS/m	< 170
TH	mg/L	100 - 300
RSC	$CO_3 + HCO_3 - Ca - Mg$	< 1.25
PI	$\frac{Na + \sqrt{HCO3}}{Na + Ca + Mg} \times 100$	< 70
KR	$\frac{Na}{Ca + Mg}$	< 1.5
MH	$\frac{Mg}{Ca + Mg} \times 100$	< 50
Na%	$\frac{Na + K}{Na + Ca + Mg + K} \times 100$	< 40
SAR	$\frac{Na}{\sqrt{(Ca + Mg)/2}}$	< 10
SSP	$\frac{Na}{Na + Ca + Mg} \times 100$	< 50

5. Conclusion

The classic way of assessing quality of a water sample is through laboratory analyses. This can be laborious, as it involves adhering to strict quality control measures during sample collection, transportation, and testing. The data provided by these laboratory analyses can be used with cheaper and less strenuous alternatives such as machine learning (ML) models. The data provided in this work can be used to train (and test) ML models for classifying and/or predicting "fitness of use" of water samples for human consumption or irrigation. It also helps address the absence of open and sufficiently large datasets on water quality parameters, especially in developing countries of Africa, for computational analyses. Finally, by simply changing the "accepted range" of each parameter, models developed using this data can be easily localised and adapted to different regions globally.

6. Acknowledgements

The authors acknowledge reviewers from the department of Water resources and Earth Sciences at the University of the Western Cape, Cape Town, South Africa.

7. References

[1] Divahar R, Raj P, Sangeetha S, Mohanakavitha T, Meenambal T. Dataset on the assessment of water quality of ground water in Kalingarayan Canal, Erode district, Tamil Nadu, India. Data in brief. 2020 Oct 1;32:106112.

[2] Balamurugan P, Kumar P, Shankar K. Dataset on the suitability of groundwater for drinking and irrigation purposes in the Sarabanga River region, Tamil Nadu, India. Data in brief. 2020 Apr 1;29:105255.

[3] Verma A, Yadav B, Singh N. Data on the assessment of Groundwater Quality in Gomti-Ganga alluvial plain of Northern India. Data in Brief. 2020 Jun 1;30:105660.

[4] Ajayi O, Bagula A, Maluleke C, Gaffoor Z, Jovanovic N, Pietersen K. WaterNet: A Network for Monitoring and Assessing Water Quality for Drinking and Irrigation Purposes. IEEE Access, vol. 10, pp. 48318-48337, 2022. doi: 10.1109/ACCESS.2022.3172274.

[5] Acharya S, Sharma SK, Khandegar V. Assessment of groundwater quality by water quality indices for irrigation and drinking in South West Delhi, India. Data in brief. 2018 Jun;18:2019.

[6] Abbasnia A, Radfard M, Mahvi A, Nabizadeh R, Yousefi M, Soleimani H, Alimohammadi M. Groundwater quality assessment for irrigation purposes based on irrigation water quality index and its zoning with GIS in the villages of Chabahar, Sistan and Baluchistan, Iran. Data in brief. 2018 Aug 1;19:623-31.

[7] Soleimani H, Abbasnia A, Yousefi M, Mohammadi A, Khorasgani F. Data on assessment of groundwater quality for drinking and irrigation in rural area Sarpol-e Zahab city, Kermanshah province, Iran. Data in brief. 2018 Apr 1;17:148-56.

[8] Jafari K, Asghari F, Hoseinzadeh E, Heidari Z, Radfard M, Saleh H, Faraji H. Groundwater quality assessment for drinking and agriculture purposes in Abhar city, Iran. Data in brief. 2018 Aug 1;19:1033-9.

[9] Masindi V. Dataset on physicochemical and microbial properties of raw water in four drinking water treatment plants based in South Africa. Data in brief. 2020 Aug 1;31:105822.

[10] Sithole M, Nwulu N, Dogo E. Dataset for a wireless sensor network based drinking-water quality monitoring and notification system. Data in brief. 2019 Dec 1;27:104813.

[11] Hasan M, Khan M, Islam M, Hasan M, Hossain M, Monir M, Samad M, Ahmed M. Dataset on the evaluation of hydrochemical properties and groundwater suitability for irrigation purposes: South-western part of Jashore, Bangladesh. Data in brief. 2020 Oct 1;32:106315.

[12] Fenta MC, Anteneh ZL, Szanyi J, Walker D. Hydrochemical data on groundwater quality for drinking and irrigation use around Dangila town, Northwest Ethiopia. Data in brief. 2020 Aug 1;31:105877.
[13] Liu Q, Wu Y. Supervised Learning. In: Seel N.M. (eds) Encyclopedia of the Sciences of Learning. Springer, Boston, MA. 2012. https://doi.org/10.1007/978-1-4419-1428-6_451. Accessed 03/05/2022
[14] Horton RK. An index number system for rating water quality. J. Water Pollution Control Fed. 1965 Mar;37(3):300-6.
[15] World Health Organisation. Guidelines for drinking-water quality. World Health Organisation; 1993. http://apps.who.int/iris/bitstream/handle/10665/44584/9789241548151-eng.pdf. Accessed 12/01/2022
[16] Meireles A, Andrade E, Chaves L, Frischkorn H, Crisostomo L. A new proposal of the classification of irrigation water. Revista Ciência Agronômica. 2010;41:349-57
[17] SABS Standards Division. South African National Standard (SANS) Drinking water - Part 2: Application of SANS 241-1 (2015). https://alabbott.co.za/wp-content/uploads/2020/02/abbott_sans_241_test_requirements.pdf Accessed 03/05/2022