39th IAHR WORLD CONGRESS GRANADA, SPAIN 2022 From Snow to Sea 19-24 June 2022

Proceedings of the 39th IAHR World Congress
19–24 June 2022, Granada, Spain
doi://10.3850/IAHR-39WC2521716X20221253

Application of Clustering Approach for Risk Assessment of Drinking Water Facilities Worldwide

Prisha Puri [(1)], Sudarshan Kurwadkar[(2)], and Doina Bein[(3)]

[(1)] California State University, Fullerton, United States of America
e-mail prisha@csu.fullerton.edu

[(2)] California State University, Fullerton, United States of America
e-mail skurwadkar@fullerton.edu

[(3)] California State University, Fullerton, United States of America
e-mail dbein@fullerton.edu

Abstract

United Nations recognized access to safe drinking water as a human right, yet many countries in the developing world lack access to potable water. Recurrent incidences of water-borne illnesses have a devastating effect on the morale and personal well-being of many people living in developing countries, contrasting the achievement of the UN's sustainable development goals. Qualitative and semi-quantitative approaches used for risk assessment are often ineffective, time-consuming, and do not discern the risk due to ingestion of unsafe drinking water at the global scale. This research utilizes a global dataset of drinking water facilities to evaluate the risks using a clustering approach. Data analysis involving predetermined criteria for three risk levels was performed using density-based spatial clustering of applications with noise (DBSCAN). The criteria for the three risk levels consist of population percentages associated with the categories of the Joint Monitoring Program service ladder for drinking water. The three risk levels are low risk, medium risk, and high risk. Of the dataset analyzed, 90 areas were designated as a low-risk category while 42 were medium-risk. Overall, the clustering approach is an excellent tool to analyze a large dataset for risk assessment which helps the potential stakeholder, including the water utility manager, to assess the potential risk due to declining water quality quickly. Additionally, the clustering approach can be further harnessed for better data visualization, long-term performance evaluation of water utility, and real-time drinking water quality monitoring.

Keywords: Clustering; Data Visualization; Risk Assessment; Drinking Water; Python Programming

1. INTRODUCTION

Global availability of drinking water is impacted by many factors, including the availability of clean source water, climatic conditions, infrastructure, and technological know-how. Recently gender disparity and access to safe drinking water were also highlighted as a factor determining accessibility to clean drinking water in rural and urban areas, with gender equality associated with greater access to drinking water (Wijesiri & Hettiarachchi, 2021). According to the recently published progress report on household drinking water, sanitation, and hygiene, between 2015-2020, only marginal improvement is made over the availability of drinking water through safely managed services. During these five-year periods, the global availability of safely managed services increased from 70 % to 74%, with rural and urban areas witnessing a 7% and 1% increase, respectively (WHO/ UNICEF, 2021). It is unclear if the report accounts for the rural-urban migration, potentially impacting the net availability of safely managed services in urban areas. While accessibility to safe drinking water remains a challenge, the risk associated with the ingestion of unsafe drinking water further exacerbates the situation.

Numerous approaches exist for risk assessment or drinking water quality at various publicly owned treatment works. The data clustering approach is a relatively new concept that has found its application in numerous engineering disciplines.

2. DATA COLLECTION AND METHODOLOGY

This section describes the data collection methodology, sources, and data quality used to conduct the clustering analysis.

2.1 Global Dataset Sources

The data was collected from the Joint Monitoring Program (JMP) established by the World Health Organization (WHO) and United Nations Children's Fund (UNICEF). The JMP database provides national, regional, and global drinking water quality through its Multiple Indicator Cluster Survey (MICS) program. Many countries are active participants in the MICS program, which includes a water quality-related questionnaire, a standardized water testing module, and the water quality testing manual. The JMP also collects data from the administrative, regulatory, research institutes, and international or regional initiatives (such as Eurostat) sources; nonetheless, household surveys are the primary sources of data collection. Current data on the quality and availability of drinking water available on JMP is obtained from the household surveys and regulatory agencies responsible for providing potable drinking water (JMP, 2021). While many challenges still exist in obtaining robust data due to the lack of information on sampling frame, obtaining a representative sample, and the effective implementation of MICS, data collected by JMP adheres strictly to its quality assurance and quality control protocol.

2.2 Data Quality Control and Quality Assurance

The JMP collects data worldwide that has been gone through the quality assurance protocol of the respective countries' national statistical offices. The JMP holds consultations with the countries to ensure that the submitted data has been processed by the official to be deemed reliable. It also establishes the acceptance criteria for the data to be representative, comparable, and of sufficient quality. Some of the datasets available through the JMP may not be representative or may represent only a subset of the population. In such circumstances, for datasets that represent less than 80% of the relevant population or are inconsistent with the dataset representing a similar population, JMP deemed such data unreliable and did not use it in its estimates. Although the JMP generates the dataset for several countries, for our study, we have accessed the dataset posted on the website – Our World in Data, a non-profit organization. This website hosts data on various critical issues in the world to assist other researchers in solving the world's most pressing problems (Ritchie & Roser, 2019).

The dataset for various countries thus available is downloaded as a CSV file. Although a large CSV data file comprising 5,070 rows and eight columns was downloaded from the server, all data cells were not populated. For this reason, only data with no missing values were considered for clustering analysis. The data primarily comprises the entity’s name, year, demographic information, and the categorization of drinking water systems (safely managed, basic, limited, unimproved, and surface water). For example, drinking water systems that provide water from the improved water source that is readily available and accessible and free from fecal or priority chemical contamination as a safely managed system, while the unimproved systems are those that provide water from unprotected dug wells or springs (WHO/UNICEF, 2017).

2.3 Data Analysis – Python Programming Code

To conduct the clustering analysis, Jupyter Notebook – a web application tool for creating a computational analysis was downloaded from the jupyter.org website. The notebook offers a seamless experience to create live code, including equations, narratives, and visualization (Kluyver et al., 2016). Within the notebook, the IPython kernel was used for running the Python programming code. The notebook accessed the CSV file consisting of the drinking water dataset. The detailed information about the Python code is included in the supplementary information.
The Python program was downloaded from the freely available database (Python Software Foundation, 2016). The programming code comprises a series of subsets of smaller utility programs such as NumPy, Scikit-learn, and Plotly, which are freely available and used in developing the overall Python code. Of these, NumPy is a fundamental package routinely used by scientific communities for writing scripts in Python. At the same time, Scikit-learn is an open-source machine learning database with various built-in utilities such as model fitting, selection, and evaluation. Finally,, Plotly is an open-source data-visualization tool. It contains a series of app-building libraries for data science, engineering, and general sciences. After the clustering algorithm is applied, each data point within one row is assigned a value. If the data point is a negative one, then that means that the data point is an outlier. If the data point is a 0, then the point is not an outlier, and it is a core point. A neon blue color was assigned to outliers and blue color to core points based on those values.

2.4 Establishing Risk Assessment Criteria

The drinking water quality characterization used in conducting the clustering analysis is based on the JMP Ladder program. This study used a hierarchical risk categorization system that considers the service levels (safely managed, basic, limited, unimproved, and surface water) that provide drinking water and the

population being served. For example, a safely managed system that provides drinking water to 70% or more of its population is categorized as low-risk, a medium risk category entails service to less than 70% of the population or 70% or more of the population being served rely on drinking water from the basic, limited, or unimproved system. The high-risk category indicated the system that does 70% or more of its population with surface water (untreated water).

3. RESULTS AND DISCUSSION

3.1 Application of Clustering to Risk Analysis

A clustering algorithm and data visualization using Python and Jupyter Notebook was implemented in this research. Overall in this study, 132 sites/places were analyzed for risk assessment and categorization. The output obtained through the clustering analysis clearly shows that some countries still cannot provide an adequate supply of potable water as demonstrated by many countries being categorized as medium-risk countries (Table 1 and Table S1). On the brighter side, none of the countries whose data was analyzed was classified as high-risk. The global analysis also shows that 70% or more of the world population is within the low-risk category. The results also demonstrate apparent disparities between the developing and developed countries with regard to the number of safely managed systems. Figure 1. Shows the graphical representation of the data for selected countries

The DBSCAN algorithm analyzed data for 132 countries/places. Visualization tool Plotly was used to generate the plots for showing risk assessment for each country. Selected plots are shown in Figure 1. The comparison between the developed and developing countries with regard to access to safely managed drinking water sources shows apparent disparities. These results are consistent with the recent findings that show that 9% population of the world lacks access to the improved drinking water source while 29% do not have access to safe drinking water (Ritchie & Roser, 2019).

Clustering analysis shows the application of data science to evaluate one of the most pressing problems such as availability and accessibility of drinking water. The Python code used for conducting the clustering analysis can be further modified to achieve real-time monitoring of drinking water facilities when assuming real-time data is available to use from another dataset. Of all the available data in the CSV file of the online dataset, the data from only the year 2020 was used in this research. The criteria for the three risk levels were determined based on a surface-level analysis and other factors. However, the code can be changed to meet the user's needs. So, if the user has different criteria for the three risk levels, then the code can be changed to reflect those different criteria. As a result, this can allow the code to be used based on the circumstances. The data analysis and its visualization obtained through clustering analysis are easy to understand. The analysis is as good as the quality of the data.

3.2 Limitations of the Clustering approach Tables and figures

One of the significant limitations of the clustering approach is that it heavily relies on data quality. Missing data points makes analysis difficult, and as such, any inferences drawn from the incomplete dataset do not serve any worthwhile purpose. Some datasets have missing data, which restrict the calculations of risk levels for those specific areas; as such, it does not provide an accurate estimate of risks associated with drinking water facilities around the world. A clustering algorithm and data visualization using Python and Jupyter Notebook implemented in this research uses risk criteria that may change depending on how the risk is perceived.

Table 1. Risk characterization using cluster analysis approach for selected countries. For a complete list of countries, please refer to the supplementary information table S1.

Low Risk	Medium Risk	High Risk
Austria	Afganistan	
Belgium	Bangladesh	
Canada	Bhutan	
Chile	Cambodia	
Colombia	Central African Republic	
Costa Rica	Chad	
Czech Republic	Congo	
Denmark	Cote d'Ivoire	
Finland	Ecuador	
France	Ethiopia	
Germany	Gambia	
Greece	Georgia	
Hong kong	Ghana	
Hungary	Guatemala	
Iceland	Guinea-Bissau	None
Ireland	Iraq	
Israel	Kirbati	
Italy	Laos	
Japan	Lesotho	
Jordan	Madagascar	
lebanon	Mexico	
Netherlands	Mongolia	
New Zealand	Myanmar	
South Korea	Nepal	
Spain	Nicaragua	
Sweden	Nigeria	
Switzerland	Rwanda	
United Kingdom	Suriname	
United States	Uganda	
World	Zimbabwe	

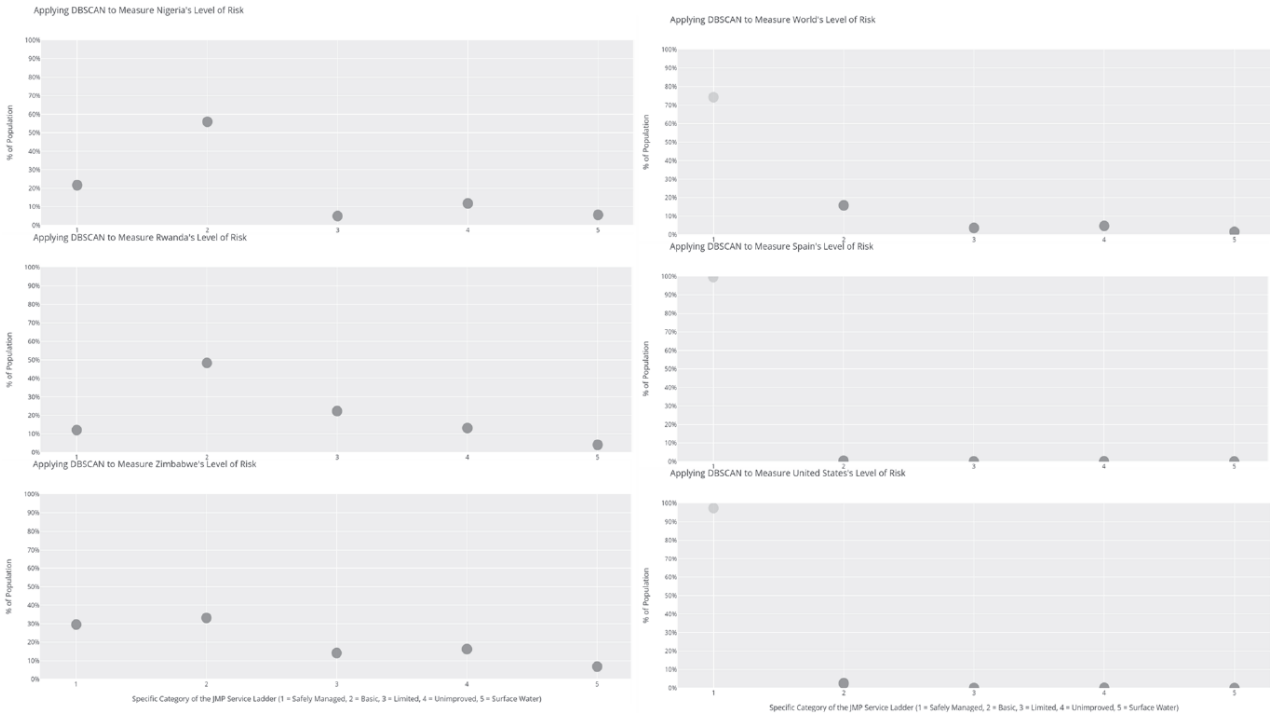

Figure 1. Risk categorization through clustering analysis of selected countries. Systems having the percentage of the low, medium, and high-risk countries across the world.

4. CONCLUSIONS

The data clustering analysis was conducted for several countries using the data collected in 2020. The study was conducted using Python programming shows a straightforward application of the computing ability of Python programming. Of the countries for which the complete dataset was available, Guam is a low-risk category, which means that the system provides drinking water to more than 70% of its population using the safely managed (potable) water system. At the same time, other countries such as Nepal, Ecuador, and Mexico have some systems categorized as low to medium risk. In general, some systems worldwide still cannot meet the sustainable development goals as enumerated in the WHO/UNICEF protocol. At least 15% of the system worldwide provides only a basic level of drinking water systems. The main objective of this research is to conduct risk analysis for drinking water facilities worldwide using a clustering approach with Python programming. The study aimed at solving four problems associated with the JMP dataset, specifically the units with unclear risk levels, ineffectiveness, lack of global context, and inconvenience. By solving the four problems of the dataset, this research helped better understand the problem of drinking water facilities in places around the world. However, it is important to note that rows with missing data value(s) in the CSV file of the online dataset were not accounted for in this research because it prevents the calculation of risk levels of specific areas and, thus, prevents better information about drinking water facilities around the world. Also, it is important to note that area(s) -- if any -- were not documented in the CSV file of the online dataset because it prevents the calculation of risk level(s) of the specific area(s) and, thus, prevents better information about drinking water facilities around the world. Furthermore, it is important to note that the criteria for each of the three risk level categories used could change because what constitutes a low, medium, and high-risk category per this research may vary. To further improve the results, the variability of the data before the year 2020 can be considered for better measurements of areas' risk levels.

5. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1832536 for the project, "Advancing Student Success by Utilizing Relevant Social-Cultural and Academic Experiences for Undergraduate Engineering, Computer Science Students (ASSURE-US).

6. REFERENCES

JMP. (2021, December 20). *SDG Indicator Metadata*. https://washdata.org/sites/default/files/2022-01/jmp-2021-metadata-sdg-611.pdf

Kluyver, T., Benjamin Ragan-Kelley, Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). *Project Jupyter | Home*. Jupyter Notebooks -- a Publishing Format for Reproducible Computational Workflows. https://jupyter.org/

Python Software Foundation. (2016). *Welcome to Python.org*. https://www.python.org/about/

Ritchie, H., & Roser, M. (2019). Clean water- Our World in Data. In *Our World in Data*.

WHO/ UNICEF. (2021). Progress on Household Drinking Water , Sanitation and Hygiene 2000-2020: five years into the SDGs. In *UNICEF journal*. https://www.eea.europa.eu/publications/industrial-waste-water-treatment-pressures%0Ahttp://files/558/Rapport EEA Industrial waste water treatment – pressures on Europe's environment.pdf

WHO/UNICEF. (2017). *Drinking water | JMP*. Https://Washdata.Org/Monitoring/Drinking-Water. https://washdata.org/monitoring/drinking-water

Wijesiri, B., & Hettiarachchi, A. (2021). How gender disparities in urban and rural areas influence access to safe drinking water. *Utilities Policy*, *68*, 101141. https://doi.org/10.1016/J.JUP.2020.101141

