Clean Water: How the AI Community can Contribute to Accessing Water Sources in Developing Countries

Karthik Dusi[1], Thilanka Munasinghe[2]
[1]Department of Industrial and Systems Engineering
[2]Department of Information Technology and Web Science (ITWS)
Rensselaer Polytechnic Institute
Troy, NY

Abstract

Access to water is one of the fundamental human rights. Clean water is an issue plaguing many countries worldwide and is one of the world's largest health concerns. The poor are those who suffer significantly from access to improved water sources and often contract other infectious diseases from unsafe water. This paper examines how the AI community can further research into clean water data that is available and investigate the socio-economic factors that prevent some communities from gaining access to safe water sources. Preliminary and Exploratory Data Analysis were done on the UN data to understand the patterns, relations, and trends between related variables. Key correlations were investigated between different socioeconomic factors such as GDP, Corruption, and Infrastructure to understand what has the greatest effect on access to improved water sources. To do so, visualizations were built using Python and the Seaborn package, as well as using the Pandas package to curate the data.

Introduction

Over 1.1 billion people in the world lack access to a general water source (World Wildlife Organization 2020). According to Worldwildlife.org, 2.7 billion people suffer from water scarcity at least one month a year, and 2.4 billion people are victims to clean water inadequately. These numbers have been on the rise and continue to be as scientists predict that by 2025, over two-thirds of the world's population will face water shortage issues (World Wildlife Organization 2020).

Several scholars and politicians have called for clean water to be recognized as a human right. Germany and Spain put forward a resolution at the UN to recognize clean water as a fundamental human right. However, the US, Russia, and Canada rejected this resolution in favor of examining issues affecting access to safe drinking water and sanitation (Editors et al. 2009). Three main reasons are cited as to why water should be a human right. One that ensuring access to clean water will significantly reduce the number of people affected by diseases (Editors et al. 2009). Two, the privatization of water does not ensure everyone has equal access (Editors et al. 2009). Three, the world's resources are being exploited to a point where our current water supply quality is threatened and must be improved upon (Editors et al. 2009). These reasons explain why water is a human right, but other factors affect the lack of access to clean water.

Literature Review

Various socioeconomic factors may affect who has access to improved water sources, such as whether they live in an urban area, a rural area, a country's GDP, infrastructure, corruption, and government effectiveness. Countries classified as developing countries, like Afghanistan, Albania, Iran, and India, are considered 'developing countries' because of the rate at which its GDP per capita grows and the infrastructure it has to support necessary elements for human life clean water (investopedia 2019).

In rural areas, drinking contaminated water can lead to diarrheal illnesses, enteropathy, and other serious diseases. In a paper investigating water quality in a South African rural community, at least a third of the population perceived the water as unsafe and felt they could get sick from it (Edokpayi et al. 2018). The system used to supply water to the community did not test positive for containing contaminants, but the system does not reach all community residents and is subject to frequent shutdowns (Edokpayi et al. 2018). Additionally, due to increased amounts of available water in the monsoon season, the research shows that there is more treated water in the region and more people feel comfortable drinking the water in the monsoons (Edokpayi et al. 2018).

In an opinion raised by sustainability experts, they express that current Sustainability development goals (SDG) are based on the assumption that access to safe water sources includes sources with good quality water. However, there is an important distinction between safe water and qual-

AAAI Fall 2020 Symposium on AI for Social Good.

ity water. Over 1.8 billion people were exposed to water sources contaminated by fecal matter and were overlooked by the misguided SDG statistics reported in 2012. The article written on ”Current opinion in environmental sustainability” suggests that the number of populations reported are in lack of access to safe drinking water was underestimated. (Tortajada and Biswas 2018)

Introduction to Dataset Explored

The UN provides datasets that they have collected as well as datasets from related organizations like the WHO at data.un.org. Other sites like ourworldindata.org also has relevant data towards understanding the problems behind lack of access to clean water. To understand basic correlations and present ideas for the reader, a dataset with information on populations using improved water sources was explored(World Health Organization 2014). This dataset has a percentage of a population using improved water source for 192 countries, and further divides the percentages into whether they live in rural or urban areas. The rural areas are defined as areas not part of major metropolitan areas, which are defined by population density and distance from the metropolitan city, and the rural data reflects data collected on those areas. Similarly, the urban data reflects data collected in areas part of major metropolitan areas. There is also historical data ranging from 1990 up to 2012 for these countries, giving ample data to explore and analyze. The figure 1 shown below outline the general project work-flow.

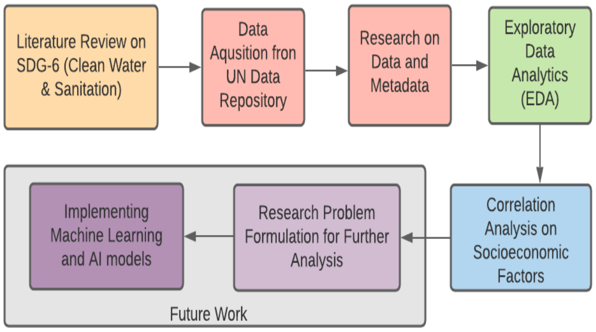
Literature Review on SDG-6 (Clean Water & Sanitation)
Data Aqusition fron UN Data Repository
Research on Data and Metadata
Exploratory Data Analytics (EDA)
Correlation Analysis on Socioeconomic Factors
Research Problem Formulation for Further Analysis
Implementing Machine Learning and AI models
Future Work

Figure 1: Data Acquisition and Project Work Flow

Exploratory Data Analysis (EDA)

Using the Pandas (Pandas NumFOCUS 2020) and Seaborn (Michael Waskom 2020) packages in Python, EDA was done on the collected dataset to understand correlations between GDP per capita on percentage of total population’s access to improved water source, as well as understanding the correlations between GDP and percentage of urban populations and percentage of rural populations’ access to the improved water sources. First, rows with empty data were dropped to make sure that only rows with usable data were present. Next, boxplots were generated to see the distribution and also to detect outliers.

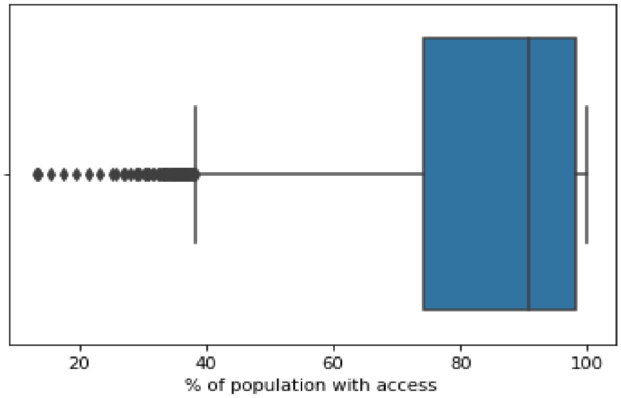

Figure 2: Percent of Total Population with Access

The boxplot in figure 2 shows the percentages of total populations with access to improved water sources; there are many outliers towards the lower-end of the plot. This means many outliers points are lower than the minimum, which was calculated to be 62.3%. The median of this data is 90.9% , and the IQR is 24%. This means that 50% of the percentage values fall within 24% of the median. Instead of merely deleting the outliers here, since they are the countries with lower percentages of people having access to improved water sources, a new data frame could be made to contain the outliers data and then compare the boxplot for the ’outlier’ data frame to the original data frame.

Similarly, the percentages of urban populations and rural populations having access to improved water sources were explored using boxplots to see if there are any outliers. The majority of the rural populations fell within the Inter-Quartile Range (IQR), with a minimal number of outliers, but the urban populations had a large number of outliers. The boxplots with outliers are further reinforced by histograms of the same variables.

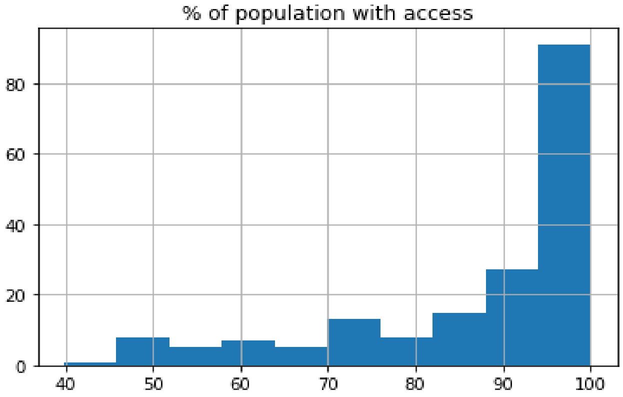

Figure 3: Histogram of Percent of Total Population with Access

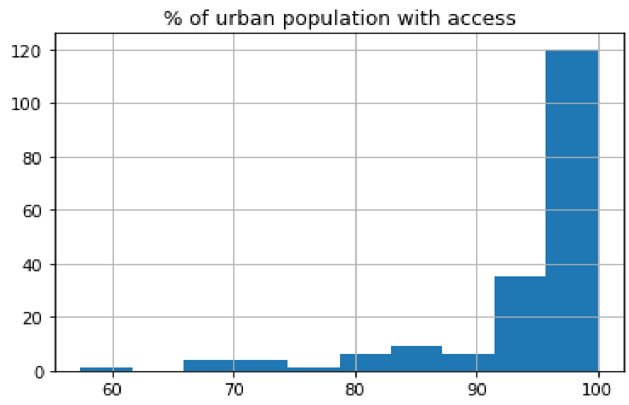

Figure 4: Histogram of Percent of Urban Population with Access

The same method mentioned above to deal with the total population's outliers could be used here to further explore the rural populations and in which countries exactly rural populations are suffering more.

By looking at heatmaps, we can understand the correlations between each variable better. In this case, we want to look at the relation between GDP per capita and the percentages of populations with improved water sources access. If we look at the total populations' heatmap in Figure 5, we can see a 0.49 correlation.

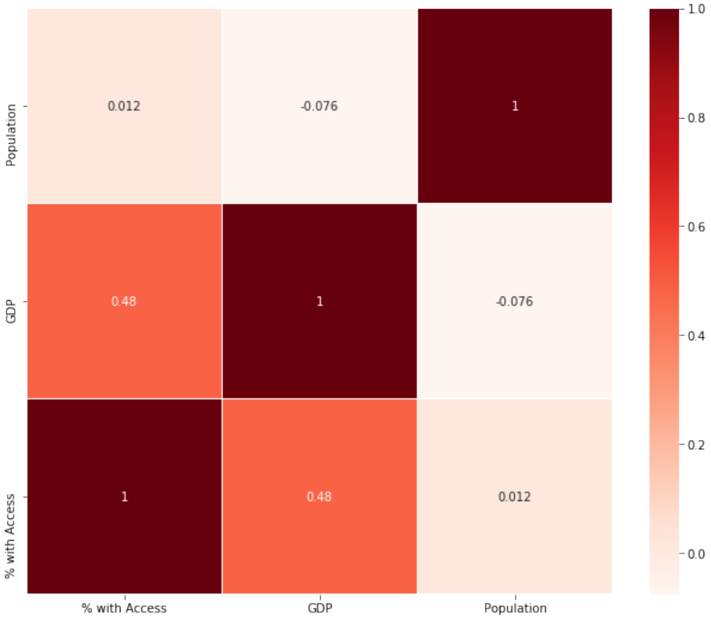

Figure 5: Correlation Matrix with GDP and Population

This implies a moderate level correlation here that could be worked with further if the outliers are removed. We also see that the country's total population does not correlate with the percentage of the population with access to improved water sources, with a correlation of 0.017. This means that socioeconomic factors are worth looking at since we see that GDP per capita is worth analyzing further; other factors like infrastructure, corruption, and effectiveness can be included in the dataset to build a predictive models.

As we see, this dataset can be built open further by including other socioeconomic factors and also utilizing the historical data that is provided from 1990 to 2012. More recent data is also provided by ourworldindata.org, which could be used to verify a predictive model if developed (Ritchie, Max Roser 2019).

According to a paper by economists (Gomez, Perdiguero, and Sanz 2019) investigating factors affecting water access in rural areas of developing countries, they cite gross national income, female primary completion rate, agriculture, growth of rural population, and governance indicators as the main socio-economic factors affecting access to improved water sources for rural populations. By governance indicators, they refer to political stability, control of corruption, and regulatory quality as examples. They also recognize that the water source itself and income of the group are two things that should influence the selection of factors being looked at and include other indicators of 'good' governance such as infrastructure, taxation, etc.

Combining this initial dataset with other indicators provided by the World Bank (The World Bank 2018, 2020) resulted in variables measuring Government Effectiveness, Overall Infrastructure, and the Corruption Perception Index. The initial dataset ranged from 1990 to 2012, but the World Bank dataset had data from 1995 to 2012. For preliminary purposes, the following analyses were done on data collected on the year 2012. Looking at only 124 countries in 2012, the following heatmap in Figure 6 to investigate correlations was generated.

In Figure 6, we can see a strong blue color means a higher correlation between the two variables. We see a medium to a strong correlation between the corruption perception index (Corruption Perceptions Index 2020) and percent of the rural population with access. This can be perceived as certain rural populations not having access to improved water sources because of a higher corruption perception index.

We can also observe that there is a strong relationship between government effectiveness and percentage value of rural population that has access to improved water sources, which makes sense given that more effective governments are able to provide water sources to all parts of the country.

There is a medium correlation between infrastructure rating and percentage of the total population with access to improved water sources. This could be because this infrastructure rating considers all infrastructure in the country, and it may be more prudent just to observe water-related infrastructure, like drainage basins, sewers, reservoirs, etc.

Further Analysis and how AI Community can help

The AI community can help leverage this data and turn it into a usable tool for governments and relief organizations

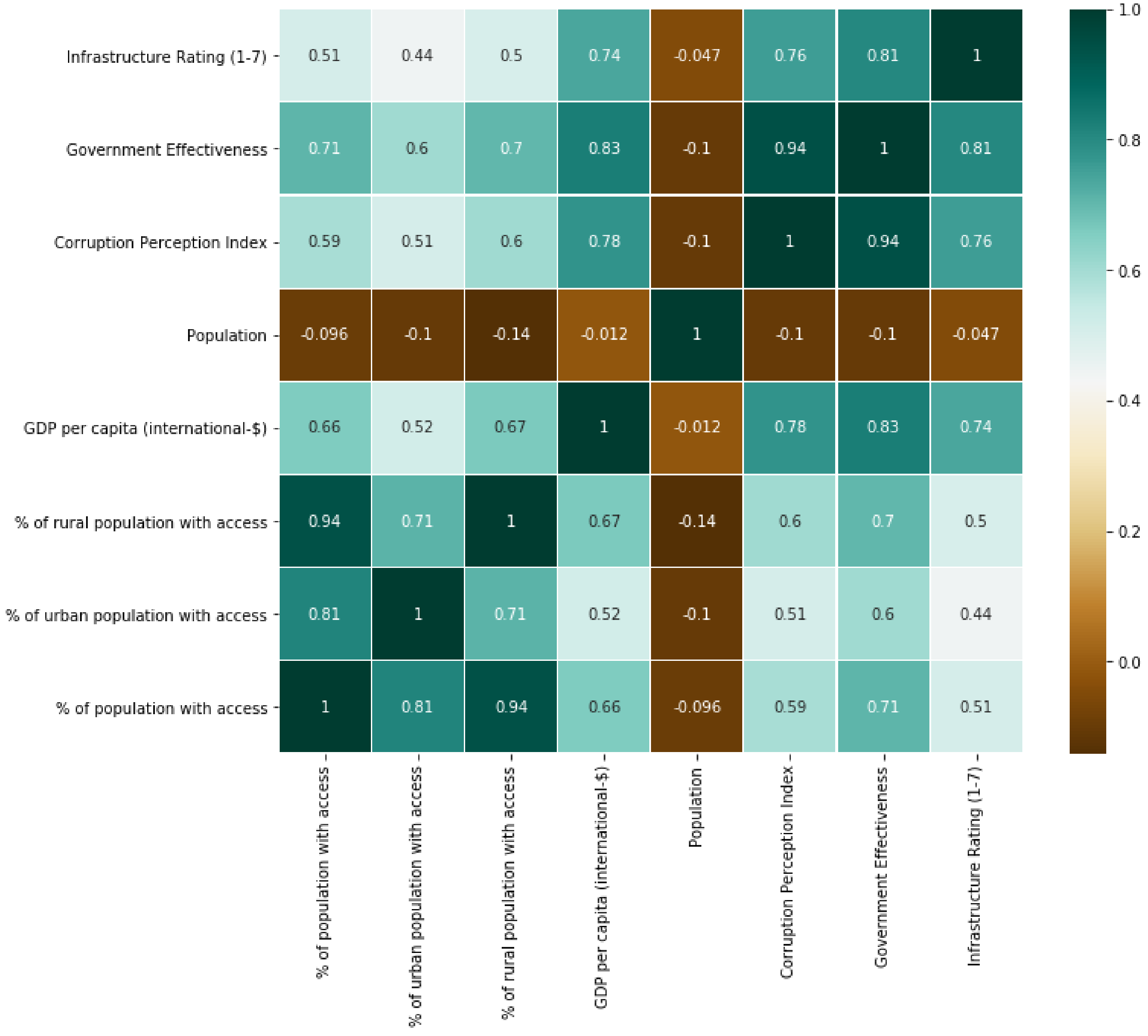

Figure 6: Correlation Matrix with Other Socioeconomic Factors

by helping them predict where resources must be allocated first to enhance access to improved water sources. Using it as a model to predict where clean water sources will deplete given trends in GDP, infrastructure, and other socioeconomic factors would be very useful as several scholars assert that by 2025, two-thirds of the world's population will face water shortage.
(World Wildlife Organization 2020).

Additionally, models can be used to investigate where water quality is low. With machines and water filters that continuously check whether the water is safe to drink or not, a data collection feature could be added and could provide data for data scientists to use in narrowing down where the water contamination is happening. Prototypes of devices that can detect whether water quality is low and can report the data to a database exist, and could be used for this application.

By using machine learning techniques and neural networks, this existing data coupled with other socio-economic datasets can be used for the further analysis and develop prediction models. Lack of clean water leads to many infectious diseases, such as deadly diarrheal diseases, cholera, and typhoid, and by using a model to see where there is no clean water available, medical professionals can help try to prevent the spread of infectious diseases in those areas utilizing those models. Stakeholders for this type of application would be public policy experts, healthcare professionals, and infrastructure professionals who could help provide data and insights regarding what sort of socioeconomic factors are most prevalent in prohibiting access to clean water.

Conclusion

This paper presents a preliminary understanding of what could be done to collect and explore the data to help solve access to improved water sources. Looking at correlations between key indicators and populations with access to water sources provides a basic understanding of what features to use in future models. Additionally, looking at rural populations over urban populations may be more productive since urban populations tend to be well developed and have good water sources. We plan to use the insights gained from this initial analysis to test out different hypotheses and research questions in the future. Obstacles that must be overcome are the lack of data for specific countries and certain yearly periods. Most countries have recent data, but only some go back up to 1995 and beyond. More emphasis needed to be done on adequate data collection. Organizations such as the World Bank, the United Nations should emphasize the importance of regular and thorough data collection from their member countries. As upstanding citizens of the world and with the new technologies available to us, the AI community must push themselves forward to develop and come up with tools that can be used in directing relief efforts in the right places where access to clean water is a problem.

References

Ritchie, Max Roser. 2019. Clean Water - Our world in data. Unsafe water is responsible for 1.2 million deaths each year, https://ourworldindata.org/water-access, Accessed on: September 24, 2020.

Corruption Perceptions Index. 2020. The corruption perceptions index Ranks of countries. , https://www.transparency.org/en/cpi/2019/results#, Accessed on: September 24, 2020.

Editors, P. M.; et al. 2009. Clean water should be recognized as a human right. *PLoS Med* 6(6): e1000102.

Edokpayi, J.; Rogawski, E.; Kahler, D.; Hill, C.; Reynolds, C.; Nyathi, E.; Smith, J.; Odiyo, J.; Samie, A.; Bessong, P.; et al. 2018. Challenges to sustainable safe drinking water: A case study ofwater quality and use across seasons in rural communities in Limpopo Province, South Africa, Water (Switzerland), 2018, 10: 1–18. *DOI* 10: w10020159.

Gomez, M.; Perdiguero, J.; and Sanz, A. 2019. Socioeconomic factors affecting water access in rural areas of low and middle income countries. *Water* 11(2): 202.

investopedia. 2019. Top 25 Developed and Developing Countries. , https://www.investopedia.com/updates/top-developing-countries/, Accessed on: September 24, 2020.

Michael Waskom. 2020. seaborn: statistical data visualization. Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics, https://seaborn.pydata.org/, Accessed on: September 24, 2020.

Pandas NumFOCUS. 2020. Pandas Library. , https://pandas.pydata.org/, Accessed on: September 24, 2020.

The World Bank. 2018. Government Effectiveness. Perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies, https://bit.ly/30c2MrT, Accessed on: September 24, 2020.

The World Bank. 2020. Quality of overall infrastructure. , https://bit.ly/331Wywr, Accessed on: September 24, 2020.

Tortajada, C.; and Biswas, A. K. 2018. Achieving universal access to clean water and sanitation in an era of water scarcity: strengthening contributions from academia. *Current opinion in environmental sustainability* 34: 21–25.

World Health Organization. 2014. Population using improved drinking-water sources . , https://data.un.org/Data.aspx?q=water&d=WHO&f=MEASURE_CODE%3aWHS5_122, Accessed on: September 24, 2020.

World Wildlife Organization. 2020. water-scarcity. , https://www.worldwildlife.org/threats/water-scarcity, Accessed on: September 24, 2020.