Exploration of Water Sanitation Improvement

Roman Nett

Data Analytics 4000 Level
Fall 2022

Abstract:

Water sanitation and access to uncontaminated drinking water is an extremely important issue and is vital to well-being and health[1]. Many studies have been conducted showing that restricted access to clean water has devastating effects. It can stunt a child's growth[2] and increase infant mortality[3]. It can also facilitate the spread of debilitating conditions such as diarrheal diseases and parasites[4]. Much work has gone into improving access to sanitary drinking water, and it is one of the main forms of humanitarian aid. In the past, this has mostly been done through policy and direct intervention - but because each country has unique conditions, it has been a slow and difficult process[5]. What if we could use the sanitation data we have to help inform future investments in improving sanitation levels? This is the analysis I aimed to perform. If I look at all countries together, however, the patterns I observe could be too broad to be applied to any specific country. Because of this I first wanted to cluster countries together, and then analyze the patterns within each cluster. I hypothesize that there will be clusters of countries with unique factors that are the best predictors of sanitation level, and those can be used to aid future humanitarian relief.

[1] Howard, Guy, et al. "Domestic water quantity, service level and health."
[2] Spears, Dean. "How much international variation in child height can sanitation explain?."
[3] Hathi, Payal, et al. "Place and child health: the interaction of population density and sanitation in developing countries."
[4] Prüss, Annette, et al. "Estimating the burden of disease from water, sanitation, and hygiene at a global level."
[5] Hutton, Guy, et al. *Evaluation of the costs and benefits of water and sanitation improvements at the global level*

Data Description:

My goal was to make predictions about water sanitation service level, so I wanted to ensure the source of this data was reputable. WASH[6] - short for water, sanitation and hygiene - is a project managed by the Joint Monitoring Program, a collaboration between the World Health Organization and UNICEF. This dataset includes information about drinking water, sanitation, and hygiene, split up by country or region from 2020. For my analysis, I focused solely on the drinking water service level split up by country. The raw dataset included a country code column, a percent of the population column, and then the service level accessible by that percentage of the population. This meant that for each country, there were between one and six different associated rows, with different percentages of the population having access to different levels. Even multinomial classification would not succeed at fitting these values as the response variable is split into multiple columns over multiple rows. Because of this, I aggregated the levels using a numerical scale weighted by the percentages. After doing this, I can still compare the aggregated values between countries as each one uses the same scale and the same original levels. With six distinct levels, the numerical scale I used spanned from 0 to 1 at intervals of 0.2.

$$score_{country} = \sum_{i,\, i \in country} percent_i * (level_{\#} - 1) * 0.2 \quad (1)$$

In equation (1), $level_{\#}$ represents the encoded level, a number from one to six with one being the lowest level and six being the highest. For my independent variables, I wanted to cast a wide net. With only around 200 countries, my row data was very precious, so I wanted a dataset with many columns so I wouldn't have to discard much row data to fit my models. I decided to use the NASA Environmental Performance Index dataset from 2020[7]. It includes a vast variety of

[6] WHO, UNICEF. *Joint Monitoring Program WASH Data*.
[7] Wendling, Z.A., J.W. Emerson, A. de Sherbinin, D.C. Etsy, et al. *Environmental Performance Index 2020*.

environmental data from 180 countries around the globe. It is also split up by country code meaning I could easily join it to my WASH data. The rest of its columns are grouped hierarchically, columns being the weighted aggregation of other columns. EPI is the weighted aggregation of two Policy Objectives - Environmental Health (40%) and Ecosystem Vitality (60%). Each of these objectives is the weighted aggregation of different Issue Categories which are themselves the aggregations of different Indicators. By using all of these columns as predictors, we can not only glean what indicates a high score but also at what hierarchical level it is. There are 46 columns in this dataset so I will only highlight the important ones. EPI is obviously an important column as it is the main focus of the dataset and an aggregation of all columns. Under the Environmental Health objective and contained within the Air Quality issue category is HAD. HAD measures household air pollution from solid fuels. Also falling under Environmental Health is the Sanitation and Drinking Water issue category. This category contains USD (unsafe sanitation) and UWD (unsafe drinking water). The rest of the columns I will highlight fall under the Ecosystem Vitality objective. TCL measures the tree cover loss over a five-year moving average and is under the Ecosystem Services issue category. FSH is the issue category measuring fisheries and is the aggregation of fish stock status, marine trophic index, and fish caught by trawling. CCH is the issue category pertaining to climate change and is the aggregation of CHA (methane emission growth), GIB (greenhouse gas intensity growth), GHP (greenhouse gas emissions per capita), and 5 others. Finally, I will highlight SDA, which is the sulfur dioxide emission growth rate and is under the Pollution Emissions issue category.

Analysis:

First I analyzed the distribution of my dependent variable - score. The goal of this is to ensure that the patterns and trends make sense with the original source data and to inform the models I will apply.

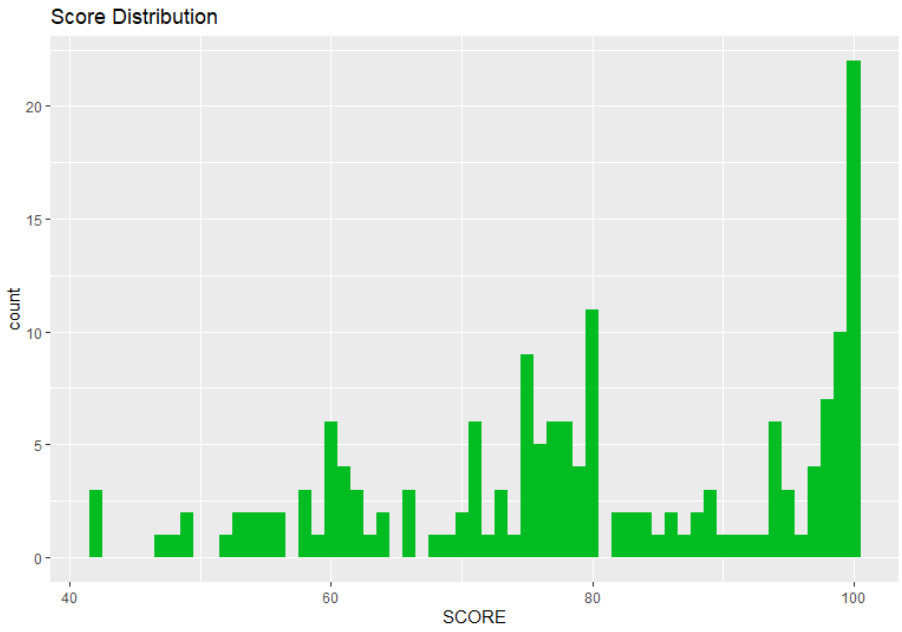

Figure 1: Here we see the distribution of the Score values over the 167 countries in my joined EPI-WASH dataset. There are large spikes at 60, 80, and 100 - this makes sense with respect to the original data, as countries with 100% (or almost 100%) of the population at certain levels will have a score corresponding very closely to that level.

Another way I validated my score transformation is by comparing it to similar variables. The Unsafe Sanitation and Unsafe Drinking Water data comes from the Institute for Health Metrics and Evaluation, meaning if my score and these metrics have a high correlation the aggregation that I performed did not meaningfully corrupt the data. Unsafe sanitation and unsafe drinking water had correlation coefficients of 0.867 and 0.876 respectively, giving me confidence that my methods were sound. I then removed these columns from the dataset, as learning that either of them affects score would not give much useful information. To prepare for my first model, I needed to determine how I was going to cluster my data. My end goal was to find patterns in

score within each cluster, so I needed to ensure a clustering that was not affected by it. To do this, I found the variables that had the lowest correlation with score. I chose Fisheries (FSH), Tree Cover Loss (TCL), and Greenhouse Gas Intensity Growth (GIB). They had correlations of -0.009, 0.029, and 0.033 meaning they were orthogonal to a country's score, and clusters based on them should have a representative distribution. Unfortunately, there was not complete data for all of these columns so I removed some rows from my final analysis. I experimented with data imputation techniques, but by filling in missing data with data from other countries, the clusters became less distinct. For example, using mean imputation resulted in each cluster having identical results, meaning the clustering did not actually provide us with any new information.

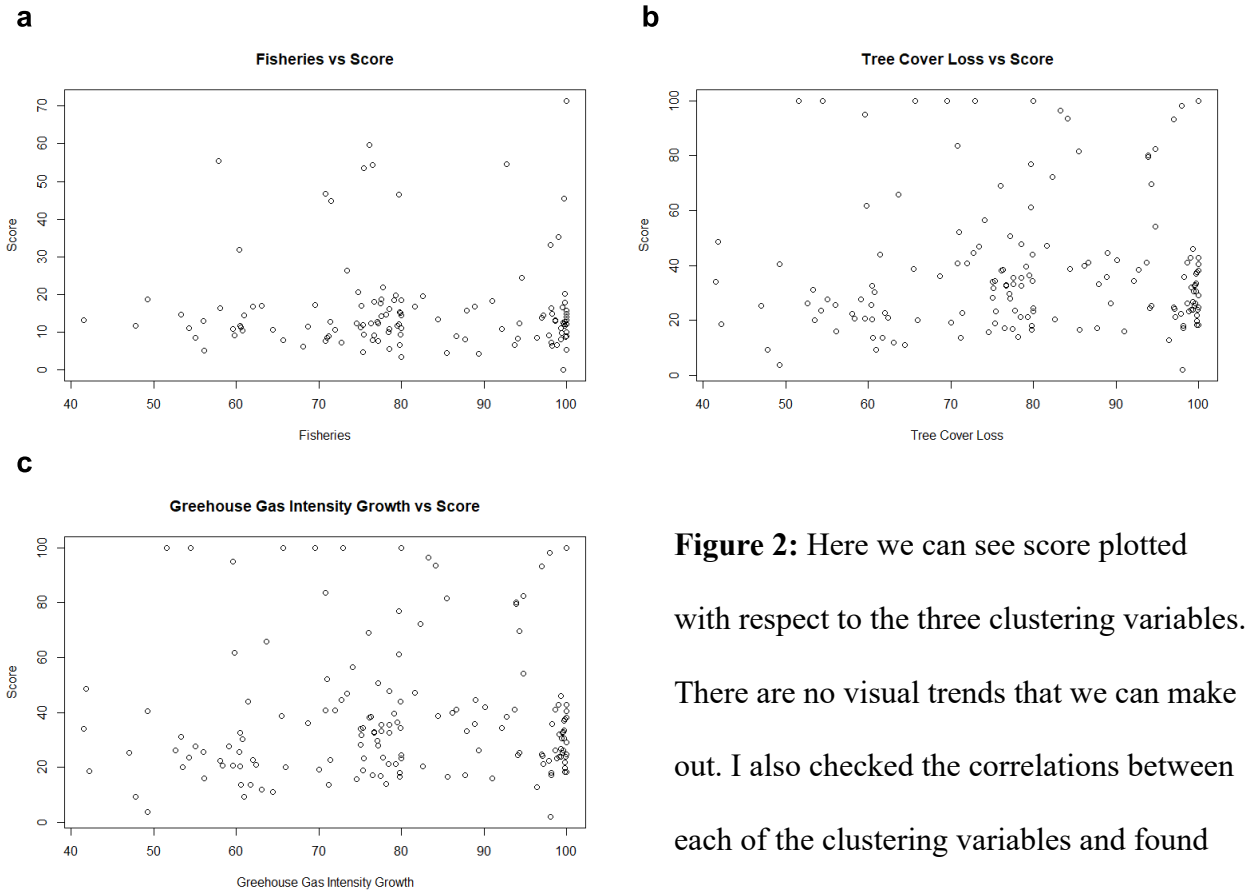

Figure 2: Here we can see score plotted with respect to the three clustering variables. There are no visual trends that we can make out. I also checked the correlations between each of the clustering variables and found that not only are they orthogonal to score, but also each other.

After completing my exploratory data analysis, I have identified a few uncertainties and potential sources of error. The first is that many of the variables contained within the EPI data have “bad” connotations. This means if any of the data is self-reported by a country, there may be motivation to lie about it. Another is that outliers may affect the final outcome of the analysis due to the small sample size. Removing the outliers would result in too much data loss, so countries at extreme values may affect variable importance measurements in my regression models. However, these will hopefully not affect the usefulness of the results, as countries that also lie about their data will then be clustered correctly and countries that also fall on the extremes should have similar characteristics.

Model application:

To cluster my data I used the unsupervised clustering technique kmeans. It uses moving centroids to arrive at a local optimum where each datapoint in a cluster is closest to its cluster’s centroid, and each centroid is positioned at the mean of its cluster’s datapoints. Choosing how many clusters to make is a difficult task because it is unsupervised there is no ground truth we can compare to and no accuracy to measure. To select the optimal number of clusters I used the “elbow technique”, which is a heuristic method that looks at the within-cluster sum of squares at each cluster number. The point at which the curve bends like an elbow is the number of clusters to use.

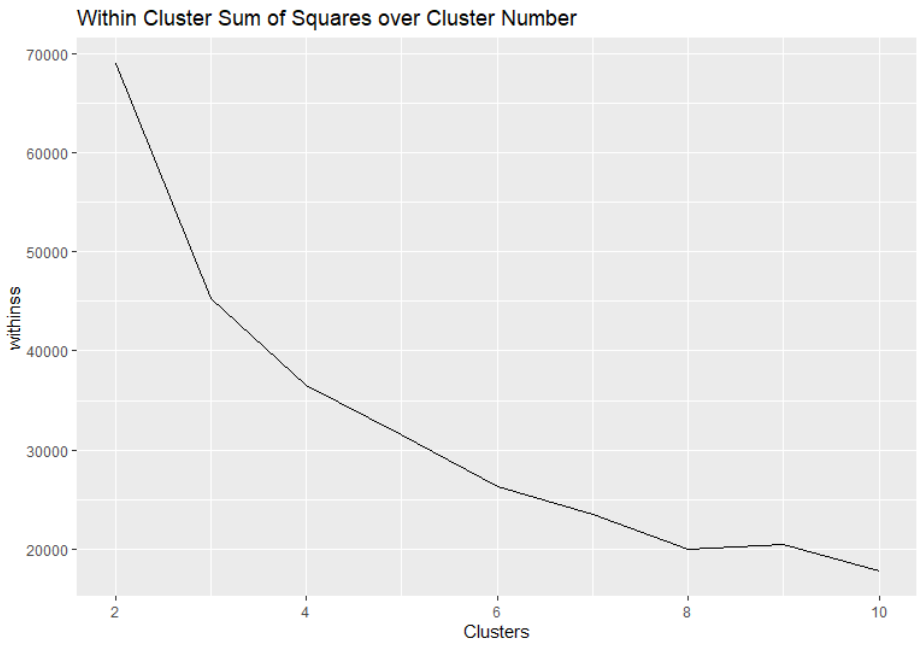

Figure 3: Here we see the total within-cluster sum of squares with respect to the number of clusters. At three clusters we see the slope of the plot begin to level out, so I chose three clusters.

After clustering, I did some further validation to ensure my results were good. I wanted to make certain that the score was not affected by the clustering, and that each cluster represented a certain class of countries. I first compared the mean and standard deviation of the score within each cluster.

Table 1

	Full Dataset	Cluster 1	Cluster 2	Cluster 3
Mean	80.99	80.26 (-0.73)	81.71 (+0.72)	79.76 (-1.23)
Standard Deviation	14.95	10.35 (-4.60)	15.90 (+0.95)	16.41 (+1.46)

As you can see in **Table 1**, the means and standard deviations of the score are very similar to the full dataset, showing that my clustering did not introduce bias in how score was distributed. We can see this as well in **Figure 4b**.

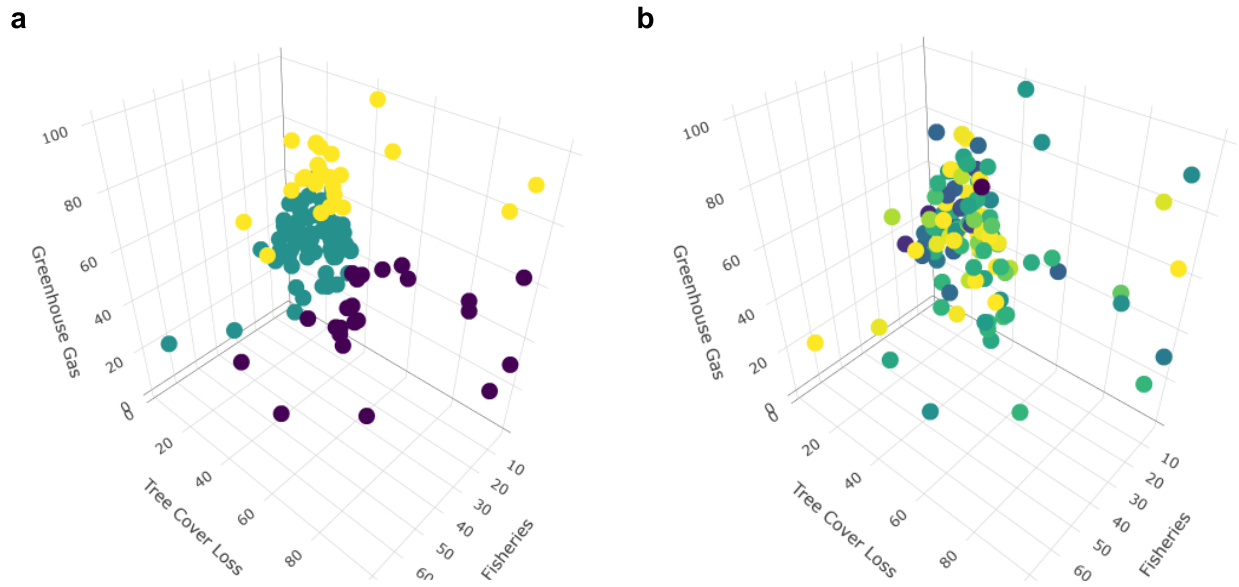

Figure 4: In **a**, we see the data split plotted in 3D space colored by cluster. In **b**, we see the same datapoints however now they are colored by their score. The brighter the color, the higher the score. The colors seem to be randomly distributed, implying that our clustering was independent of score.

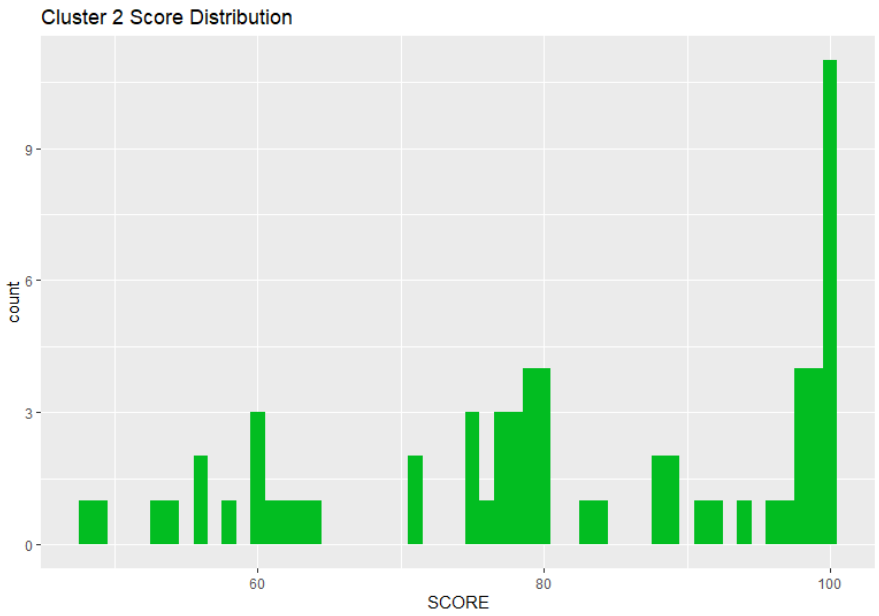

Figure 5: Here we see the distribution of scores in one of the clusters. The distribution looks almost identical to the distribution of the full dataset, giving even more credence to the fact that the clustering was done independently of score.

Now that I had my clusters, I wanted to describe their characteristics with the clustering variables that I chose. I achieved this by looking at how their distributions differed between clusters.

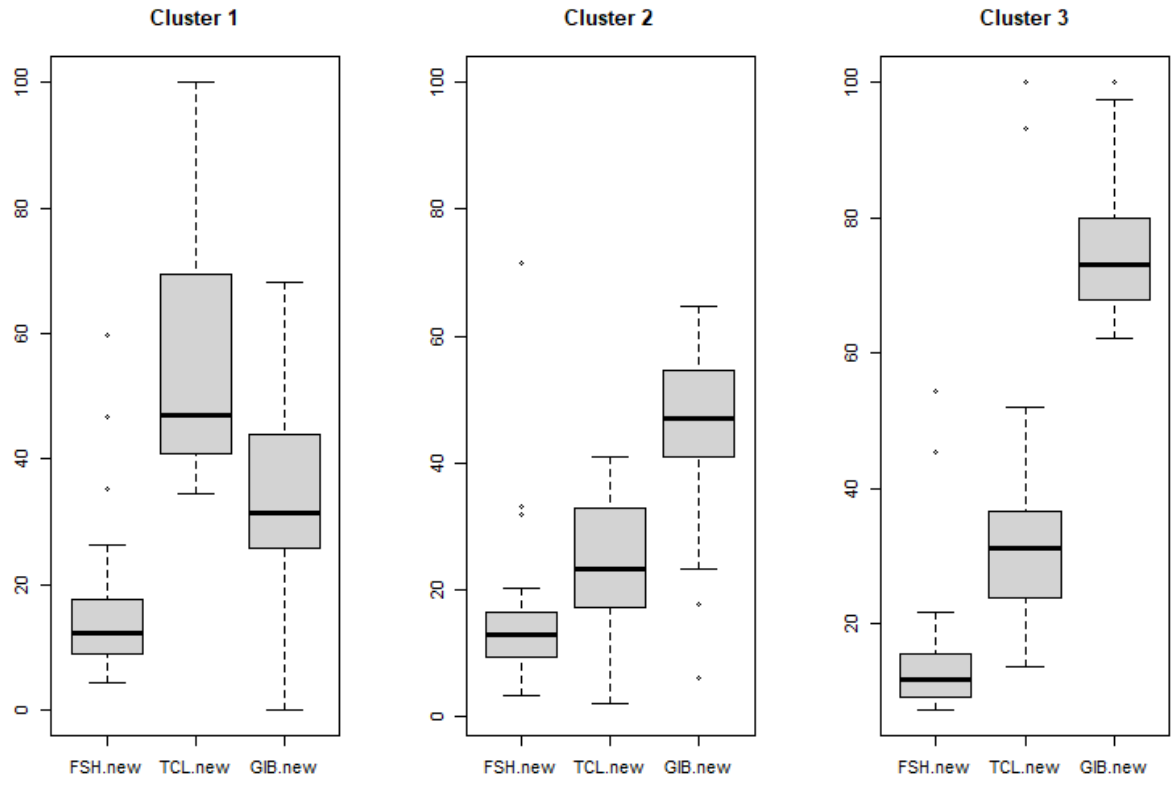

Figure 6: Here we see the distributions of the orthogonal variables for each cluster. This illustrates how each cluster differs from the other two. The first cluster has the highest number of fisheries, the highest amount of tree cover loss, and the lowest average greenhouse gas emission. The second cluster has a medium amount of fisheries, a medium to low amount of tree cover loss, and a medium mean greenhouse gas emission. The final cluster has a low amount of fisheries, a medium to high amount of tree cover loss, and a high mean greenhouse gas emission.

The whiskers of these boxplots cover most of the range of what a potential out-of-bag value might be, meaning given a new country to consider we could easily classify it into one of the three clusters without having to have access to the specific model that created the clusters.

In order to analyze the variables for each cluster, I needed to fit a regression model. I chose random forest for a few reasons. The first is that for random forest it is very easy to determine variable importance which is what my entire end goal is. The second is that random forests are very stable due to the large number of smaller models that are used internally, so with the small amount of data I have I can be more confident in my results. Finally, my data is not guaranteed to be linearly fittable, and decision trees can make use of important benchmarks and thresholds to fit spikes or curves. There were two hyperparameters that I needed to tune for my model, mtry and ntree. Mtry specifies the number of variables that the model tries at each split, and ntree specifies the number of trees that make up the model. In order to find the optimal values I generated random forests using a set of possible values and compared their mean squared errors. First I tuned mtry using values from three up to the max, where the max is the total number of variables in the cluster. I then recorded the mtry that created the forest with the lowest MSE. However, due to the stochastic nature of the model generation, this value could vary, so I ran this 10 times, and took the mtry value that resulted in the lowest MSE the most times. If there was a tie, I ran trials until one of the tied values appeared. I used a similar method when determining the optimal ntree. Using the mtry that I found previously, I generated random forests using 100, 200, 300, 400, 500, 600, and 700 trees and recorded the ntree that resulted in the lowest error. Again I ran these trials 10 times, with additional trials for tiebreakers. I then used the tuned hyperparameters for each cluster for my final models. For cluster 1 the optimal mtry and ntree were 30 and 400, for cluster 2 they were 19 and 400 and for cluster 3 they were

26 and 600. Cluster 1 had 23 datapoints, cluster 2 had 64 datapoints, and cluster 3 had 24 datapoints. With clusters of this size, I couldn't make large test and training sets so to validate my models and get a good estimate of the out-of-bag error, I used leave-one-out cross validation. This method is normally quite expensive, but with less than 100 points in each cluster, we could afford to generate a model for each datapoint. I found the squared error for each left-out datapoint when predicted by the model generated without it and then found the mean over all points.

Table 2

	Cluster 1	Cluster 2	Cluster 3
MSE for LOOCV	40.9	53.3	82.1

As you can see in **Table 2** the estimated out-of-bag error is quite low! Given these mean squared errors, we estimate a predicted score will be off by about 6.4, 7.3, and 9.1 for clusters 1, 2, and 3 respectively. We see this illustrated in **Figure 7**.

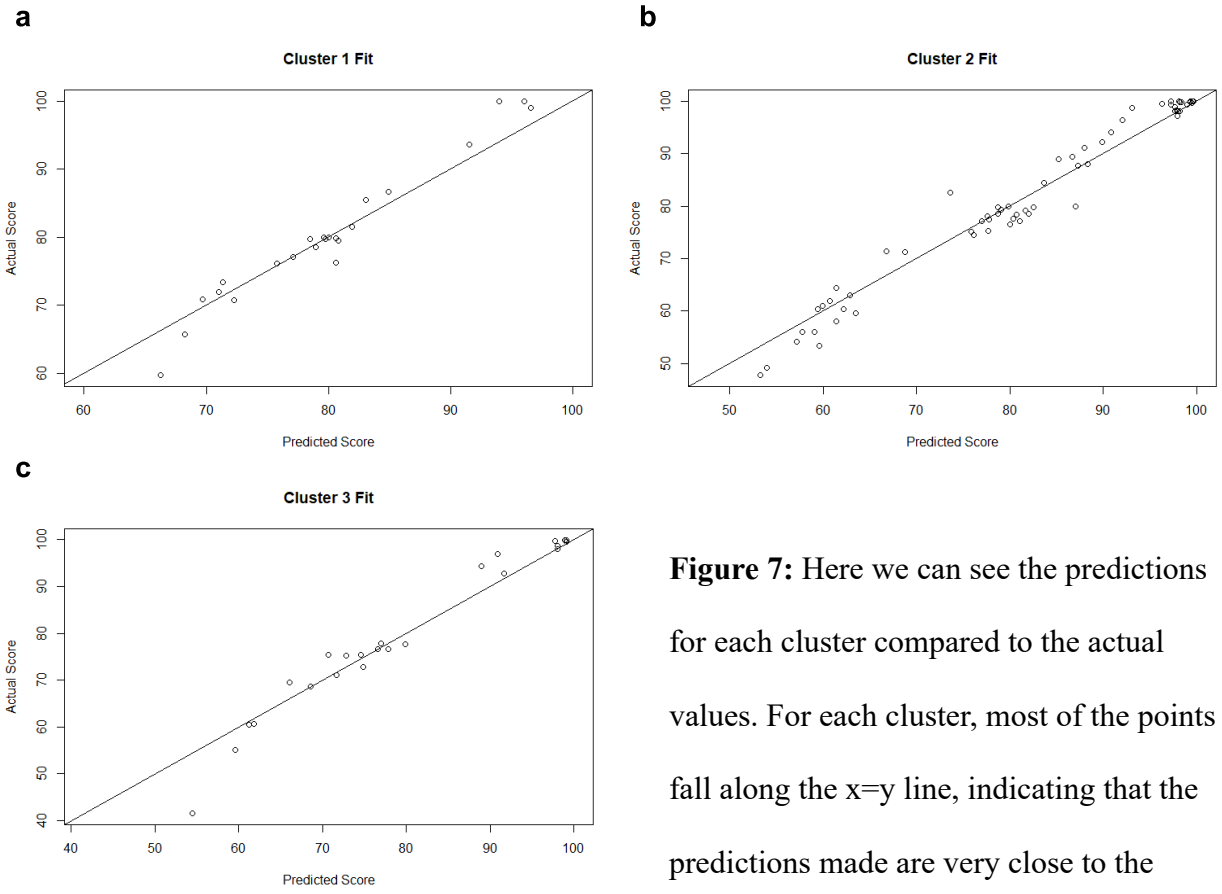

Figure 7: Here we can see the predictions for each cluster compared to the actual values. For each cluster, most of the points fall along the x=y line, indicating that the predictions made are very close to the ground truth.

Using these fitted models, I could then analyze the variable importance to determine which columns were most influential for each cluster. Each variable is analyzed based on its mean decrease Gini index, which measures the decrease in accuracy when the column is left out. For the first cluster, the most important variable was CHA, the methane intensity trend. The second and third most important variables were HLT (overall environment health policy objective) and CCH (climate change issue category) respectively. For cluster 2, the most influential variable was HAD, the household air pollution from solid fuels, followed by HLT and then overall Environmental Performance Index. Finally, for cluster 3 the most important column was EPI

with over twice the importance of the next variable. Grouped in second place were SDA (sulfur dioxide emission), HAD, and GHP (greenhouse gas emissions per capita).

Conclusions:

Using these results, we can learn a lot about how different environmental variables are associated with different sanitation levels in different countries. Looking at the different clusters, we can interpret the variable trends to determine the country type. The first variable, fisheries, measures the amount of fishing and trawling done by that country, if a country is landlocked, this value is going to be very low, as they are restricted to just their fish stock. Comparatively, an island country will have a high value for fisheries, as much of its food and trade comes from the sea. The second variable, tree cover loss, represents the amount of forest cover lost by that country. Arid, desert countries are likely to have low tree cover loss values, as they often have no tree cover to start with. On the other hand, tropical and temperate countries may have a high TCL due to logging or other forms of deforestation. The last variable, greenhouse gas intensity, measures the greenhouse emission rate normalized by economic status. A highly industrialized country will have a large GIB, while a sparsely populated or underdeveloped country may have a lower one. With these classifications, we can describe our three clusters. Cluster 1 contains countries that are likely to have ocean or sea access and tend to have higher tree cover. They also are more likely to have low greenhouse emissions, so probably have a smaller population and less industry. Countries in cluster 2 are likely to have partial fishing access, are more likely to be arid or non-forested, and have a medium amount of greenhouse emissions. Cluster 3 countries have the lowest amount of fisheries, so are more likely to be landlocked, and are slightly more forested than the countries in cluster 2. They also have the highest greenhouse emissions of the countries in all groups meaning they are likely densely populated and developed. For cluster 1

the most important variables focused on climate change and greenhouse gas emission rates. This means for countries with already lower-than-average emissions, keeping the value low is an indicator of a higher score. For cluster 2 the most important variable was household air pollution followed by greenhouse gas emissions. It is possible that household solid fuels have a larger impact on sanitation level when a country has lower tree cover. For the final cluster, the most important variable was overall environmental performance index, followed by various emissions variables. For landlocked and highly industrialized countries, the total environmental performance is very important, possibly because with so much development, pollution and contamination can start to affect the water supply.

These results give us a view into what environmental variables are important in predicting sanitation level, but are not fully validated. In future work, more models can be applied to do this. By using a different clustering technique such as mean shift, we can get a different set of clusters. Then, we can compare to the original clusters using the adjusted rand index to determine how stable the clustering is. In the same vein, we can compare the regression done by random forest with other regression models such as support vector machines or xboost. We can then not only compare the performance of these models but also their respective variable importances. Another way to validate these results is by changing the response variable. In this study, I transformed it by aggregating it into a continuous variable, but other transformations can be done. If we define certain cutoffs we can change the score into an ordinal variable. For example, we can classify each country into a level if at least 50% of the population has access to that level. We can then replace our regression models with classification models and repeat the variable importance analysis. If we arrive at the same or similar variable importances, we validate the original transformation and subsequent regressors.

Bibliography

1. Hathi, Payal, et al. "Place and child health: the interaction of population density and sanitation in developing countries." *Demography* 54.1 (2017): 337-360.
2. Howard, Guy, et al. "Domestic water quantity, service level and health." (2003).
3. Hutton, Guy, et al. *Evaluation of the costs and benefits of water and sanitation improvements at the global level*. No. WHO/SDE/WSH/04.04. World Health Organization, 2004.
4. Prüss, Annette, et al. "Estimating the burden of disease from water, sanitation, and hygiene at a global level." *Environmental health perspectives* 110.5 (2002): 537-542.
5. Spears, Dean. "How much international variation in child height can sanitation explain?." *World Bank policy research working paper* 6351 (2013).
6. Wendling, Z.A., J.W. Emerson, A. de Sherbinin, D.C. Etsy, et al. *Environmental Performance Index 2020*. New Haven, CT: Yale Center for Environmental Law and Policy. https://doi.org/10.13140/RG.2.2.21182.51529 (2020).
7. WHO, UNICEF. *Joint Monitoring Program WASH Data*. (2000). https://washdata.org/data/household#!/table?geo0=region&geo1=sdg

Github: https://github.com/TheNamor/DataAnalytics2022_Roman_Nett