Exploration of Water Sanitation Improvement

Roman Nett **nettr@rpi.edu**

Rensselaer Polytechnic Institute, Tetherless World Constellation, Troy, NY, United States

Abstract

Water sanitation and access to uncontaminated drinking water is an extremely important issue and is vital to well-being and health[2]. Many studies have been conducted showing that restricted access to clean water has devastating effects. It can stunt a child's growth[5] and increase infant mortality[1]. 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[3]. What if we could use the sanitation data we have to help inform future investments in improving sanitation level? 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.

Data

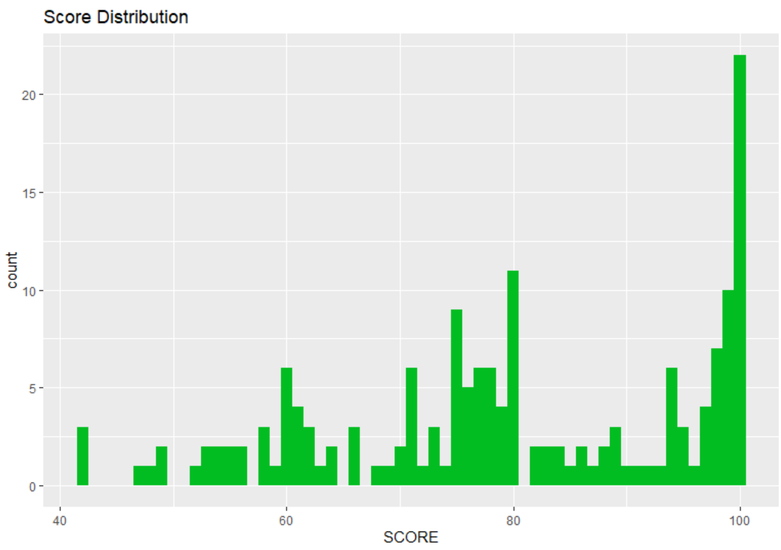
My goal was to make predictions about water sanitation service level, so I wanted to ensure the source of this data was reputable. WASH[7] - short for water, sanitation and hygiene - is a project managed by the Joint Monitoring Program, a collaboration between the World Health Organization and UNICEF. 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 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.

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

The data I used to predict the score is from the NASA Environmental Performance Index dataset from 2020[6]. It includes a vast variety of environmental data from 180 countries around the globe. Its columns are grouped hierarchically, columns being the weighted aggregation of other columns, with the top level being a country's overall Environmental Performance Index. There are two Policy Objective columns that make up EPI, each of which are aggregations of different Issue Categories. These in turn are aggregations of Indicators, for a total of 46 different columns.

Data cont.

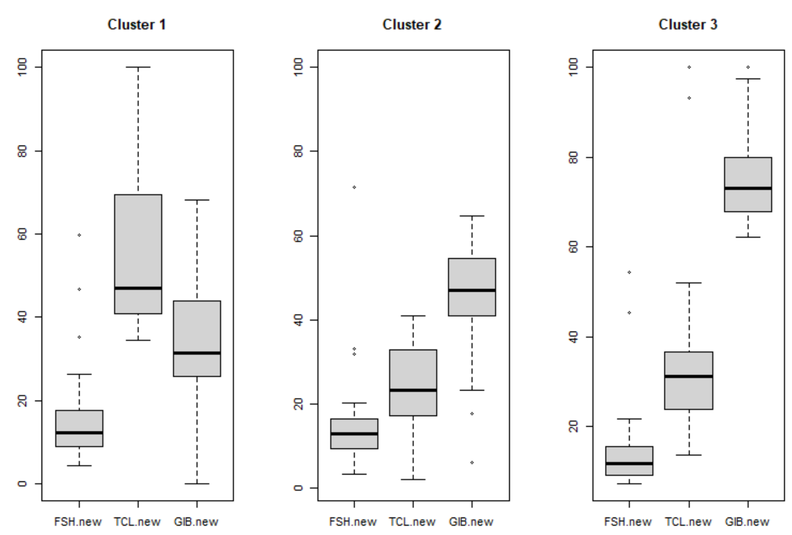
Here we see the distribution of the Score values over the 167 countries in my joined EPI-WASH dataset. There are large spikes at 60, 80, and 100 - this makes sense with respect to the original data, as countries with 100% (or almost 100%) of the population at certain levels will have a score corresponding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



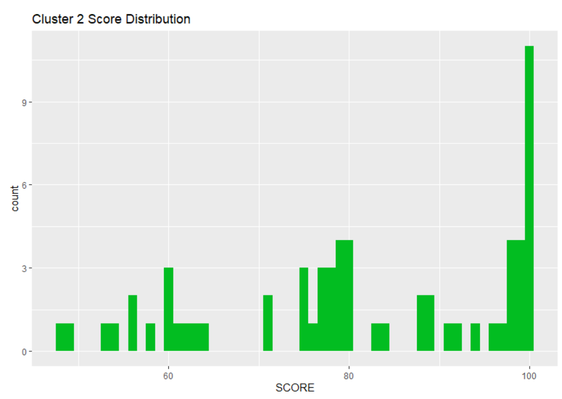
Score is also highly correlated with the Unsafe Sanitation and Unsafe Drinking Water columns from the EPI dataset – another indicator it was not meaningfully changed by the aggregation.

Clustering

To have a representative distribution of score within each cluster, I needed to organize using variables that had nothing to do with it. Fisheries, Tree Cover Loss, and Greenhouse Gas Intensity Growth had the lowest correlations with score: -0.009, 0.029, and 0.033 respectively.



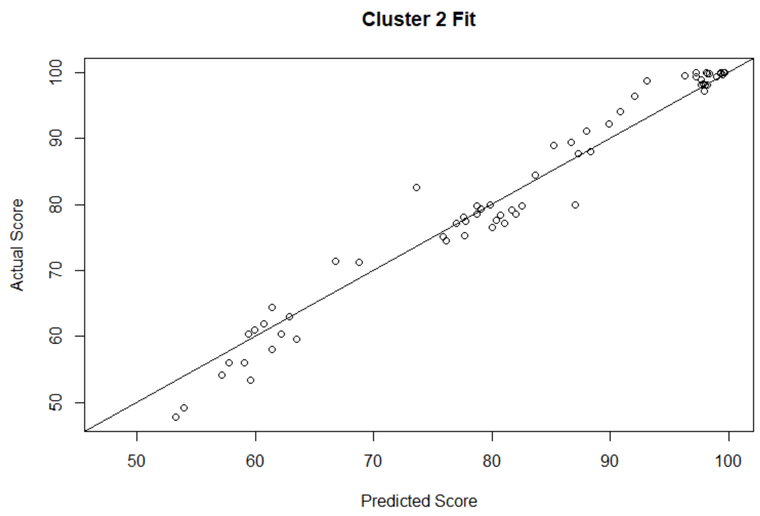
Here we see the distributions of the orthogonal variables for each cluster. The first cluster has the highest number of fisheries, the highest amount of tree cover loss, and the lowest average greenhouse gas emission. The second cluster has a medium amount of fisheries, a medium to low amount of tree 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.



Here we see the distribution of score for Cluster 2. It looks almost identical to the distribution for all the data.

Regression

Within each cluster, I then used Random Forests to predict the score. By doing this, I could then analyze which variables were most important.



Here we see the fit for Cluster 2, as you can see it predicts very accurately. In order to get a good idea of how the models might perform on new, never-before-seen data, I performed leave-one-out cross validation, where I trained a model on all the data except for one point, and then used the model to predict that point. I then found the average Mean Squared Error over all data points.

	Cluster 1	Cluster 2	Cluster 3
LOOCV MSE	40.9	53.3	82.1

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 - which is pretty good!

Conclusions

Using these results, we can learn a lot about how different environmental variables are associated with different sanitation levels in different countries. 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. So, 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.

Sponsors:

References:

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