# Pump it Up
# Data Mining for
# Tanzanian Water Crisis


## Team Poseidon

Karishma Agrawal

Rohit Sarda

Mahesh Gajwani


**Executive Summary**

Water is critical to a country's development, as it is not only used in agriculture but also for industrial development. Though Tanzania has access to a lot of water, the country still faces the dilemmas of many African countries where many areas have no reliable access to water. In a household where money is scarce, families have to often spend several hours each day walking to get water from water pumps. We are looking at the dataset of water pumps in Tanzania to predict the operating condition of a water point. By finding which water pumps are functional, functional needs repairs, and non functional, the Tanzanian Ministry of Water can improve the maintenance operations of the water pumps and make sure that clean, potable water is available to communities across Tanzania. While we weren't able to identify all the pumps that need repair, our confidence in the ones we did is high and we expect this to aid the maintenance process.

**Data**

We are using the data from Taarifa and Tanzanian Ministry of Water to predict which water pumps are functional, functional needs repairs, and non functional. The data was collected using handheld sensor, paper reports, and user feedback via cellular phones. The dataset has features such as the location of the pump, water quality, source type, extraction technique used, and population demographics of pump location. The training set has 59,401 rows and 40 features including an output column. The output column specifies the status of the water pump in the category of functional, functional needs repairs, or non functional. Out of the 40 features in the data, we have 31 categorical variables, 7 numerical variables, and 2 date variable.

**Objective**

Our motivation for the project is to identify the water pumps that are functional but need repair. To date, a total of $1.42 billion has been donated and spent to fix the water access crisis in Tanzania [3]. Even though there were many water pumps constructed with these donations, these water pumps are not well maintained. Many of the water pumps that were built with the donations are now in danger of failing across communities [3]. We want to help the Tanzanian Ministry of Water in identifying these water pumps that are functional but need repair so that an immediate action can be taken to keep them running in a healthy state. By fixing these water pumps early, the people of Tanzania could have improved and continuous access to running water.

**Risks and Challenges Identified**

Through our project, we encountered many challenges and risks that we tried to mitigate. Here are the risks and challenges we faced in our project:

1. Missing values:
   Our dataset contains many missing values in the features associated with the water pump. Since the data of the water pump was collected using handheld sensor, paper reports, and user feedback via cellular phone, many of the feature values were not accurate. For example, the construction_year feature of the water pump contained 20,709 missing rows making it hard for us to create a decay rate for the water pumps or add dynamic weather data. We also had 21,381 rows that contained missing population data. We mitigated the missing values risk by trying to find on the Tanzanian government web site if there are datasets that can be incorporated in our dataset. For example, we were able to get the population data from the Tanzanian National Bureau of Statistics to solve our missing population data.
2. Class Imbalance:

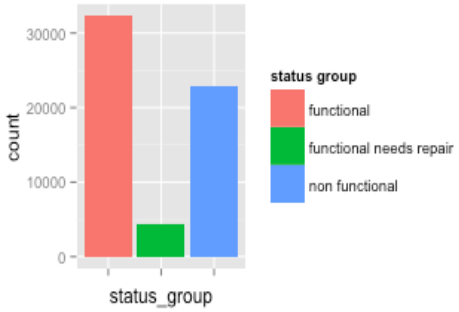


*Figure 1. Class Imbalance*

   Our dataset has severe class imbalance, with 32,259 data points for functional water pumps, 4,317 data points for functional water pumps but needs repair, and 22,824 data points for non functional water pumps as seen in Figure 1. If the goal was simply to classify the pumps and the cost for mistakes was same for all the classes, this situation poses less issue. However our focus is on functional pumps that need repair and can be fixed. In this case, we want to increase the true positive rate that would essentially result in effective maintenance. At the same time, we want to reduce false positives, i.e. misclassification of functional or nonfunctional pumps into functional needs repair. This would reduce the unnecessary expense. To mitigate the issue of class imbalance, we used cost based approaches to rank the classifier performance or various sampling methods like oversampling and under-sampling. The challenge we faced with the class imbalance was that our class improved classification for one or the other at all times.
3. High Arity:
   Our dataset contains many categorical features. There are 31 categorical features in our data and many of them have high arities. These high arity features include funders, installer, water point names, sub village, ward, and scheme name. These features contain over 1000 different values that causes huge memory overhead when running algorithms such as Random Forest on our machines. For example, the funder contains

arity of 1900 different values. We tried multiple techniques to reduce the arity such as generation of synthetic levels and exploring dimension reduction techniques such as principal component analysis (PCA).

4. Repeated Values:
   Our dataset contains many features that contain similar representation of data presented in different grains. The group of features of (extraction_type, extraction_type_group, extraction_type_class), (payment, payment_type), (water_quality, quality_group), (source, source_class), (subvillage, region, region_code, district_code, lga, ward), and (waterpoint_type, waterpoint_type_group) all contain similar representation of data in different grains. Hence, we risk overfitting our data during training by including all the features in our analysis. We tried to avoid this risk by identifying features in each group that contained the finer grain which held more information in the analysis or looked at the correlation analysis across the features to see which one is a better fit.

**Preliminary Data Analysis**

On eyeballing the data, we identified few features that seemed discriminative based on our human intuition. According to us, amount_tsh (amount of water available to water point), gps_height, basin, installer, population, scheme_management, construction year, extraction_type, management_group, water_quality, payment type, source, and waterpoint_type seemed like they could be extremely important in identifying the pump status. We plotted the distribution of these features for all the classes to see if any feature was determinant for any class. As indicated in Figure 2, for most of the features there was no significant pattern. In the case of water quantity, shown in Figure 3, we did find that most of the dry water points were non functional as it was expected.

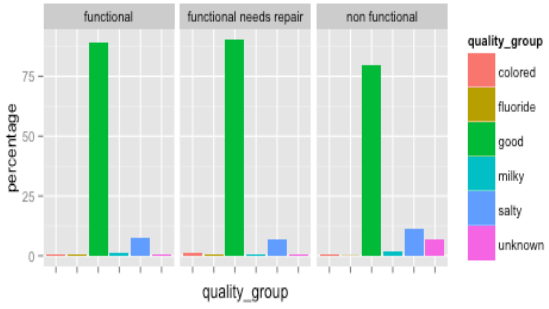


*Figure 2. Distribution of quality per class*

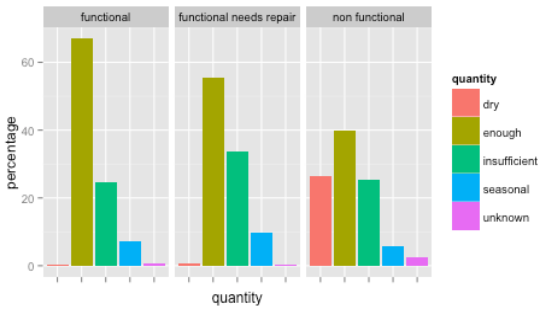


*Figure 3. Distribution of quantity per class*

Next, we decided to conduct a correlation analysis to check if there was any correlation between the features. Using Chi-square for categorical data, anova for categorical - numerical data, and Pearson correlation for numerical data, we checked the correlation scores for an

exhaustive combination of the feature points. The significance test showed that the p-values for all these cases ranged between the orders of e-9 to e-16. Surprisingly, we could not find any meaningful correlation between the features.

Since our dataset contained latitude and longitude for each water pumps, we wanted to identify any regions that contain high concentration of functional; functional needs repair, and non-functional water pumps. Below is our geographic map of Tanzania that contains all the data points of water pumps and the regional concentration of functional, functional needs repair, and non-functional water pumps.

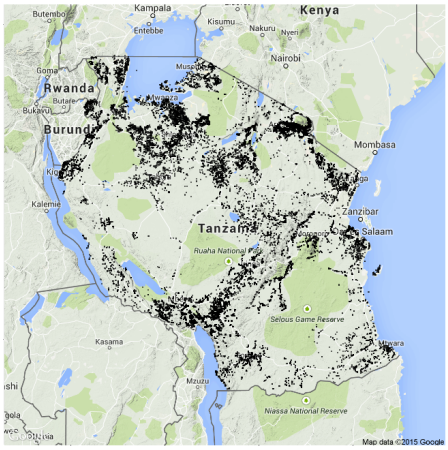


*Figure 4. Water Pump Data Points Across Tanzania*

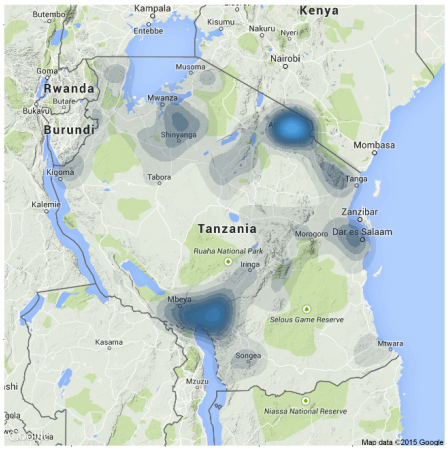


*Figure 5. Functional Water Pump Map*

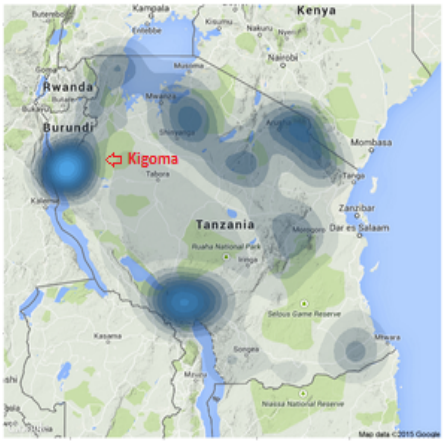


*Figure 6. Functional Needs Repair Water Pump Map*

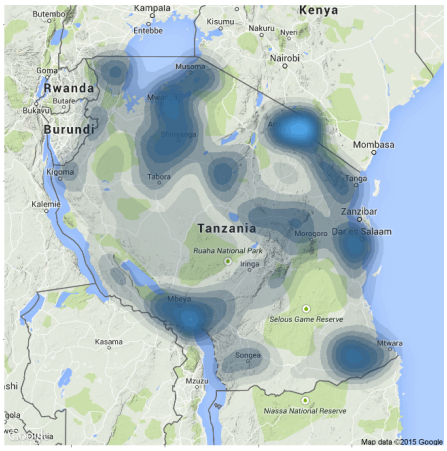


*Figure 7. Non Functional Water Pump Map*

Though most of the water pumps status is concentrated in the same regions most of the time, we can notice two distinct patterns from the maps. As Figures 6 and 7 suggest, the concentration for water pumps that are functional needs repair and non functional are more spread out throughout Tanzania, especially among the rural areas compared to functional

water pumps. We can see this map validated as only 44% of the rural areas in Tanzania get proper supply of water compared to 77.9% of the urban areas according to WHO/UNICEF [1].

The second pattern we see is the high concentration of water pumps that are functional needs repair in the Kigoma region. We can see this map validated, as Kigoma city is one of the cities that get the poorest supply of water in Tanzania. The Tanzanian Ministry of Water reported that Kigoma city only serves 31% of city's water demand with only 5 hours of running water every day [2]. The Tanzanian Ministry of Water can look at these maps and concentrate their water relief efforts in the areas where there are high concentration of functional needs repair and non functional water pumps but inadequate coverage by the functional water pumps. For example, by fixing the water pumps that needs repair in the Kigoma city, the Tanzanian Ministry of Water can potentially help relieve water crisis for 144,853 people by providing them with adequate supply of water.

**Baseline Performance**

Our main focus is on detecting the functional needs repair pumps. We noticed high overlap in the water pumps of all three classes, which was apparent in the false positive rates for functional water pumps and non-functional water pumps. We decided that we can let go of some discriminative information for the irrelevant classes, i.e. functional and non-functional by clubbing them together, to favor the classification of functional needs repair. Going forward, all our analysis referred is conducted in the functional needs repair vs. others setup. Since our dataset has substantially less positive examples, in construction of the ROC plots, we have performed the log transformation of false positive rates.

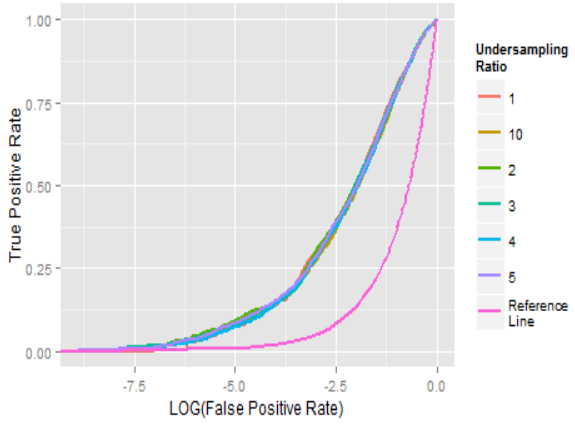


*Figure 8. ROC plot for Naive Bayes with different under sampling ratios*

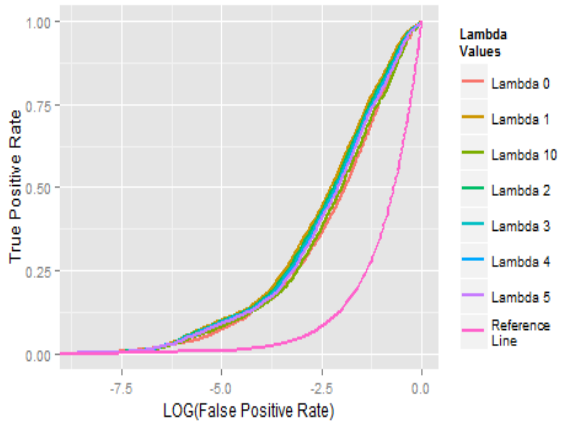


*Figure 9. ROC plot for Naive Bayes with varying Laplace smoothing parameter*

To establish our baseline, we tested our raw dataset with Naive Bayes classifier and tested different Laplace smoothing parameter, lambda, and found that while the overall performance was good, 87.5%, the false positive rate for functional needs repair was high, 18%. This lead us

to believe that perhaps the presence of the "functional needs repair" class, which formed 7.8% of our data, is under represented. We decided that under sampling the "others" class could help us improve our results and decrease the false positives.

We carried out under sampling with different ratios, varying from 1:1 to 1:10, where the latter represents the proportion of "others" in the mixture. For performing adequate comparison using ROC plot, we considered only the common data points from all the under sampled datasets. As indicated in Figure 9 we could not find any significant difference in using the raw data and the naive approach for classification combined with under sampling.

**Data Transformation**

Next, we looked into different classifiers. We used Random Forest on our raw data, but faced issues due to low memory. While the heap size was extended to 4GB, because few features had high arity e.g. for funders the arity was 1897, we could not build a Random Forest model with 50 trees and maximum depth of 10.

| Model | TPR | FPR |
|---|---|---|
| Naïve Bayes – lambda 1 | 65.2 % | 18.2 % |
| Random Forest – reduced levels | 25.7 % | 1.2 % |
| Random Forest - funder PCA only | 1.4 % | 0.15 % |
| Random Forest - Installer + funder PCA only | 0.35 % | 0.17% |
| ***Random Forest – Augmented funder PCA*** | ***28.2 %*** | ***1.3 %*** |
| Random Forest – Augmented funder and installer PCA | 21.6 % | 1.05 % |
| Random Forest – no dry data | 28.3 % | 1.4 % |

*Table 1. True positive rates and false positive rates for different models*

Our approach to mitigate this issue was to conduct some data transformations. We looked at features that were representing the same information in different grains, such as extraction_type, extraction_type_class and extraction_type_group, and retained only one of them with the finer grain. Next for features with high arity, we calculated the top 10 values, based on frequency and assigned all the remaining values to 11th synthetic value as "others". Additionally, as most of our population values were missing, we used population data available from National Bureau of Statistics of Tanzania to get reliable population data by wards [4].

Using this new data with 20 features, we generated a random forest model and found 94.24% accuracy. On further examination of the true positive rate (TPR) and false positive rate (FPR), shown in Table 1, we found that the concerned class "functional needs repair" had relatively low TPR.

While reducing the levels and eliminating features seemed reasonable, we did not wish to make any assumptions with respect to the dataset. This led us to explore different methods to deal with high arity. In case of water point name, we simply dropped the feature since it was completely irrelevant. However, in case of funders we binarized the feature and generated a sparse matrix of 1897 x 1897 dimension. Principal component analysis (PCA) was then conducted and the top 10 principal components were used to replace the actual funder feature.

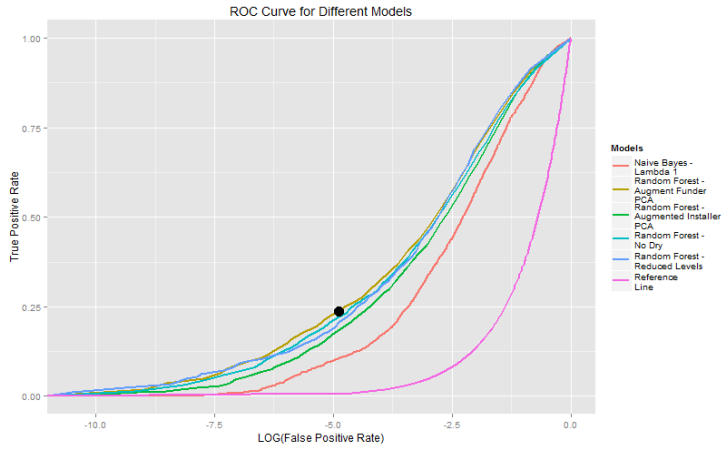


*Figure 10. ROC plots for different models*

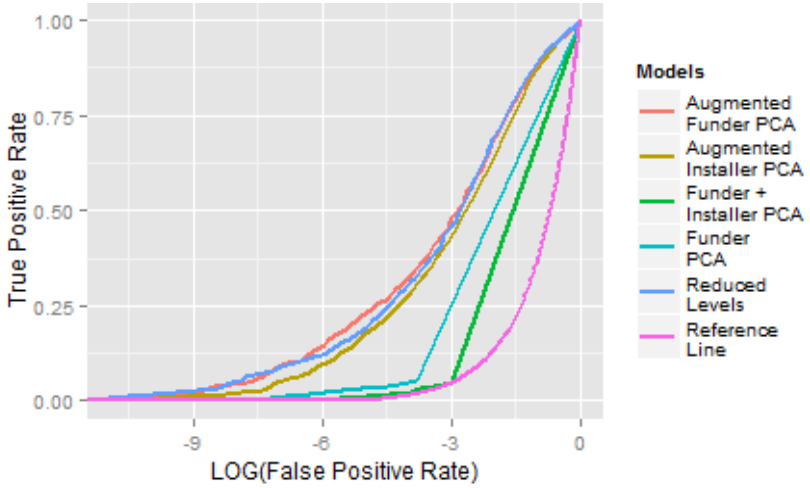


*Figure 11. ROC plots to study effect of Principal Component Analysis*

The principal components essentially grouped the funder based on various attributes, such as their spending patterns. We tested the performance of the principal components in isolation, shown in Figure 11, and found that they are only useful when augmented with the raw data. Using this augmented data and Random Forest, we found the accuracy was 93.6% and saw improvement in TPR, while the FPR did not change.

The results obtained from the principal component analysis of funders prompted us to perform PCA on installer as well. In this case we had just selected the top 5 components. The resulting classifier essentially had the same FPR but the TPR had shown some reduction.

Our preliminary analysis had shown that the most of the dry water points were non functional. We decided to remove those data points and see if our classifier can learn a model for distinguishing the water pump status without that. This was motivated by our concern that a lot of functional needs repair water pumps were getting classified as non-functional. We found out that the results did not change with respect to "functional needs repair". On finer analysis, removing dry data points had improved the identification of non functional data points but increased the misclassification of functional points into non functional.

Finally, we decided to focus on the success metric. We looked at the average population per pump and per functional pump. Then we looked at our prediction and estimated the new average population per functional pump, if the pumps detected by us were indeed fixed. Merely looking at the population and number of pumps in in that region does not provide us any pattern. We saw few expected data points, such as places where less population and more number of pumps had more functional water pumps. This result was not significant enough or distinct enough to lay a clear trend, as indicated in Figure 12.

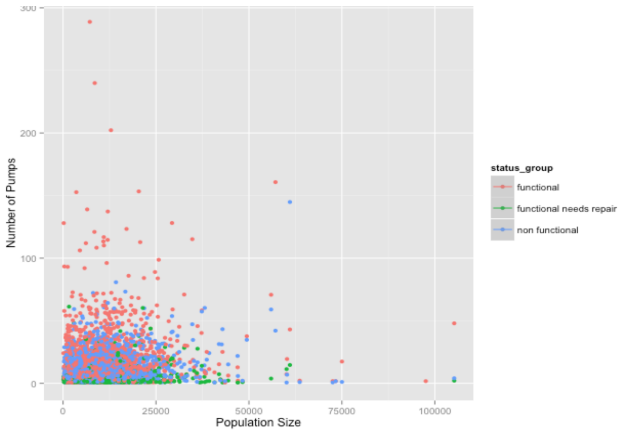


*Figure 12. Distribution of number of pumps wrt the population*

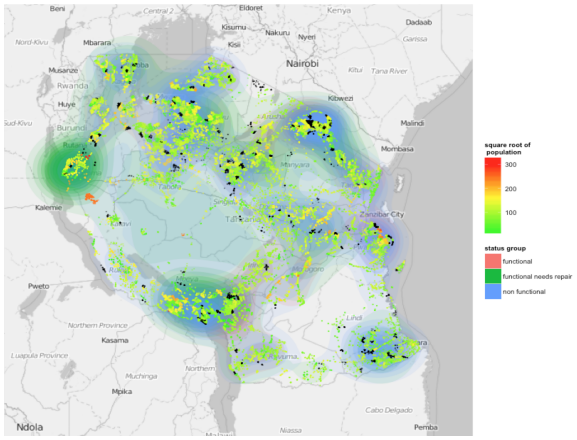


*Figure 13. Distribution of population wrt the pump density over Tanzania*

Further on, we also carried out the GIS mapping to see the distribution of the population vs the density of the pumps as per their status. True to our expectation, there wasn't any apparent useful trend.

**Final Results**

Our analysis leads us to believe that the dataset was difficult to classify. We focused our results based on the costs associated with fixing a pump and effort wasted in false identification. We prioritized the reduction of false positives, since these were either functional pumps that needed no repair or non-functional where maintenance cost was significantly higher being misclassified into functional needs repair. The cost of a standard pump ranges from $100 - $2000 [5]. Installing this pump requires drilling which can be anything between $1000- $3000. On the other hand maintaining the pump would only cost tens of dollars [6]. Here we have not considered the transportation costs. While we want to improve our true positive rates which would eventually be beneficial to the society, our major concern was to reduce effort in travelling to a non functional water pump or even worse to functional water pump. That is time which could have be effectively used elsewhere.

Our final classification model consisted of raw data augmented with top 10 principal components obtained from funder and Random Forest made up of 50 trees each with depth of 10. We then selected the cutoff threshold based on the ROC plot , shown in Figure 10 , as **0.6.** This resulted in our final model performing with 23.45% true positive rate and false positive rate had reduced to 0.75%.

Though we could not identify all the pumps that needed repair, we have high confidence in the ones which we did recognize. By figuring out the 23.45% of water pumps that are functional which needs repair, the Tanzanian Ministry of Water can set appropriate priority for their maintenance operations and help relieve water crisis for the communities in Tanzania.

**Future Work**

Since our project was only 6 weeks long and we were working with a particularly difficult dataset, we still have a list of tasks that we would like to try in the future.

1. Identify the lifetime of a pump
   We would like to identify the decay rate of a pump using the construction year and see how it correlates across different population and number of pumps across the area. Our hypothesis is that the pump will be more prone to being non-functional and functional needs repairs in areas where there is dense population and not enough coverage by the functional pumps. We would further work with Tanzanian Ministry of Water and Taarifa to provide us the appropriate construction year for the missing values in the 20,709 water pump data points.
2. Sparse Classification Methods
   We would like to fit a sparse classification model such as logistic regression with L1 penalty on our dataset to create a list of features that are important in determining the

status of the water pumps as being functional, functional needs repair, and non functional. Since our dataset contains 40 features with 31 categorical features, we believe that using sparse classification models will help determining important features that will significantly speed up our training time for models such as Random Forest. Additionally, we would like to try out few other methods to reduce the arity of the features such as clustering.

3. Improving the Success Metrics

   We would like to further improve our success metrics for the project by defining how many lives are we improving by fixing a water pump. Since there is a water crisis in Tanzania, a dashboard for Tanzanian Ministry of Water that specifies the functional but needs repair water pumps and the number of lives it will help improve by fixing the water pump will be important. As international donations can't fix all the water pumps in Tanzania, it will be important for the government to know which water pumps are crucial to help mitigate the water crisis concerns.

4. Density Cube

   To improve our overall performance we think another method worth exploring would be to construct a density cube using negative examples, and then testing using positive examples. This method might help us identify pattern in distribution of positive and negative examples.

**References**:

[1] Joint Monitoring Programme for Water Supply and Sanitation: Drinking water and sanitation coverage: country estimates by type of drinking water. WHO / UNICEF. http://www.wssinfo.org/data-estimates/tables/

[2] Water Sector Status Report 2009. Tanzania Ministry of Water and Irrigation. https://www.kfw-entwicklungsbank.de/migration/Entwicklungsbank-Startseite/Development-Finance/About-Us/Local-Offices/Sub-Saharan-Africa/Office-Tanzania/Activities-in-Tanzania/Water-Sector-Status-Report-2009.pdf

[3] How Tanzania failed to fix its water access problem. Tom Murphy. http://www.humanosphere.org/world-politics/2014/12/tanzania-failed-fix-water-access-problem/

[4] Population Distribution of Tanzania Regions by District, Ward and Village/Mtaa. National Bureau of Statistics. http://digitallibrary.ihi.or.tz/2168/

[5] MSABI aims at creating sustainable WASH services through a multilayered and multidisciplinary approach. http://msabi.org/sustainability/

[6] TRUE DOLLAR COSTS. http://www.simplepump.com/APPLICATIONS/Developing-Nations/True-Dollar-Costs.html