CSE/ISyE/MGT 6748

Practicum Final Report

Esri: Water Point Data Analysis and Prediction

Team 7: Kim Criel[1], Pengfei Mei[2], Minseok Song[3]

Submission Date: 1 December 2021

[1] kcriel3, kcriel@gatech.edu
[2] pmei8, himeipengfei@gatech.edu
[3] msong309, msong@gatech.edu

Table of Contents

1. Introduction

1.1. Project Context

In this practicum, sponsored by Esri (Environmental Systems Research Institute, a leading supplier of geographic information software), we want to analyze water point data in Uganda.

Uganda is a developing, landlocked nation in East Africa with a population of 43 million and major challenges when it comes to water accessibility. Only 32% of the population has access to a basic water supply[4], and 63% do not have access to improved sanitation facilities[5]. Furthermore, more than three quarter of the population[6] lives in rural areas, for which this problem of access is even more outspoken.

In 2010, the United Nations General Assembly explicitly recognized water and sanitation as a basic need for all human beings[7], as poor access or lack of affordability has tremendous negative effects on the development of emergent economies.

1.2. Project Goals and Challenges

Given its importance, we want to understand the current situation regarding water sources and build a model to predict the status of the water points in Uganda. The insights and predictions can help local governments and international organizations direct their resources in a more effective manner to help those most in need.

In terms of challenges, we consider three major elements:

- Spatial component of the analysis and modeling: in this project we use geospatial data as well as “simple” tabular data. Spatial analysis requires different thinking and approaches than conventional analysis. Furthermore, our team has no background in spatial statistics.
- Quality of data: we rely on multiple datasets of very different natures (semi-structured information collected on the ground to reference data on a global scale), which requires extensive cleaning and imputation prior to manipulation.
- Learning curve tooling: Esri’s geospatial software called ArcGIS Pro has a vast and extensive toolbox available and requires an important time investment to get up to speed.

[4] USAID Global Waters: https://www.globalwaters.org/wherewework/africa/uganda

[5] Water.org: https://water.org/our-impact/where-we-work/uganda/

[6] The World Bank, 2020 data, https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS?locations=UG

[7] United Nations Resolution 64/292: https://www.un.org/waterforlifedecade/human_right_to_water.shtml

2. Data Exploration and Preparation

2.1. Overview

The exploratory data analysis was done with various datasets prepared by the Esri team. It provided a solid base to understand the current situation in Uganda. The descriptions of datasets are as follows:

- Uganda population data for 2016; i.e., point representation of the number of persons living in a raster around that point; there are 106,076 points in the dataset for a total population of 35.88 million.
- Water point data: it consists of 118,110 data points with 52 attributes including latitude, longitude and the current status of each water point.
- Administrative data about Uganda (2016 snapshot): namely administrative districts and extent of urban zones

2.2. Attributes

We can summarize the main attributes as follows:

- Geolocation: longitude and latitude
- Status: the target variable which contains three values (yes, no, and unknown in terms of whether the water point is functioning - 'yes' means that a water point is functioning and 'no' means that a water point is not functioning).
- Several dates: reported, created, updated
- Water point characteristics in a broad sense: source type (e.g., boreholes), technology type (e.g., tap stand) management type (e.g., private), installer, installation year, paying or not, district information, fecal coliform values and subjective quality

The first inspection showed that 21 attributes were in fact not adding any additional value as they had equivalent cleaned up versions or fully redundant data (e.g., longitude and latitude both had two other representations and nearly all characteristics listed above have a raw version).

A second major observation is that the water point data has three classes for what we initially thought to be binary class problem. Out of the 118,100 water points, we have:

- 2.6% with unknown status
- 18.1% with non-functioning status
- 79.3% with functioning status

In our initial exploration, we included all three classes, but for our modelling we excluded datapoints with unknown status, given the relatively small number compared to the overall dataset

When it comes to the district information, we used the given dataset with 135 unique districts, but we must note that it does not match the current district boundary. Uganda is further decentralizing and dividing districts since 2015 which is causing the boundary to change continuously. E.g., district Kaabong in the upper right part of the country got split into two 2019.

2.3. Geospatial Analysis and Summary Statistics

Given the geospatial nature of our dataset, this allows us to provide interesting visual representations. In this section we'll explore several exploratory questions to gain insights:

- What is the distribution across districts of non-functioning water points?
- What are the top causes of non-functioning water points?
- What is the distribution of water point installations over time?
- Is there a relation between age of water point and the status of the water point?
- What part of the rural population has access to water per district?

2.3.1. Water Points and Rural Population

Let's first start with giving an appreciation of the water points and rural population in Uganda:

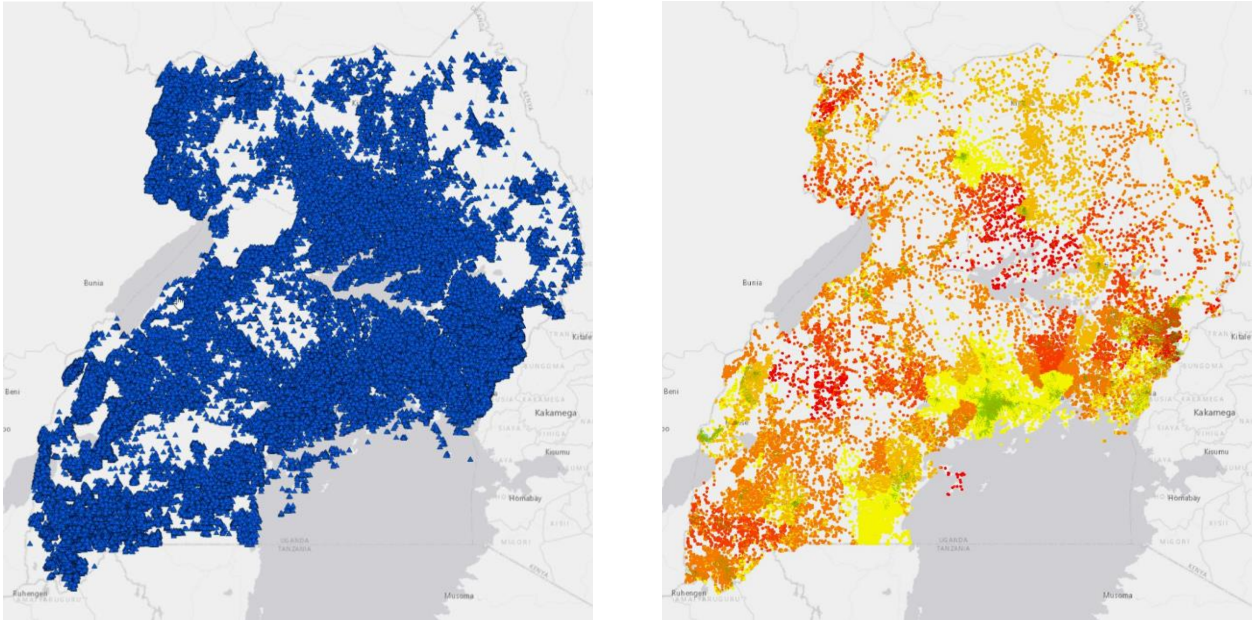

Figure 1: Side-by-side of water points and population density

From the water points plot, we see that the water points are located all over the country, but the density varies depending on the area. The population density plot shows the population in each area; Red means it's highly populated and yellow means that it's less populated. Combining two plots together, we see that the distribution of water points coincides with the distribution of population.

2.3.2. Non-functioning Water Points Per District

The plot of the fraction of non-functioning water points of each district is shown as below:

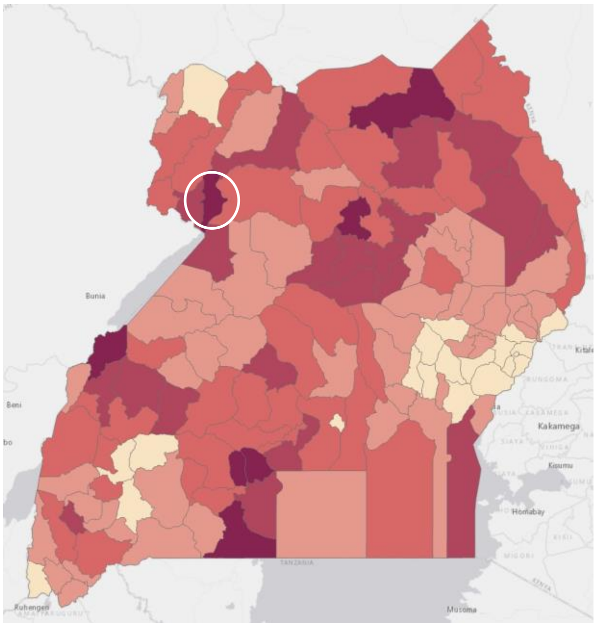

Figure 2: Non-functioning water points per district

There are five classes on the plot; The redder the area is, the higher the fraction is. The district with the highest rate is Pakwach which is located in the northwestern part of the country, right above Lake Albert with 44% of water points not functioning. On a side note, the water points with unknown status were heavily concentrated in 4 districts in the eastern part of the country which implies that there are some problems collecting the data in that area.

2.3.3. Top Causes of Non-functioning Water Points

To investigate this relation, we created pivot tables between the status of water points and the different predictors. Then, we performed chi-squared tests between our nominal variables and our target variable, which showed statistically significant relationships. Some of the most important insights are as follows:

- In terms of water source, one interesting outlier are shallow wells: twice as many are not functioning compared to the rest: 25% versus 14%.
- Installation year: once we go before 2007, we have a higher proportion of non-functioning water points.
- Installer: installations by local or central government go hand in hand with higher non-functioning proportions.
- Pay value: water points tend to be more functioning when the water committee collects the fees.
- Subjective quality: when looking at the table in detail, we see that bad quality goes hand in hand with non-functioning.
- We found a statistically significant relationship for water technology and facility types as well.

2.3.4. Distribution of Installation of Water Points Over Time

The bulk of water points were installed between 2000 and 2009. In 2007, the most water points were installed in Uganda with the number of 7,348. As of 2010, installations of new water points dramatically decrease.

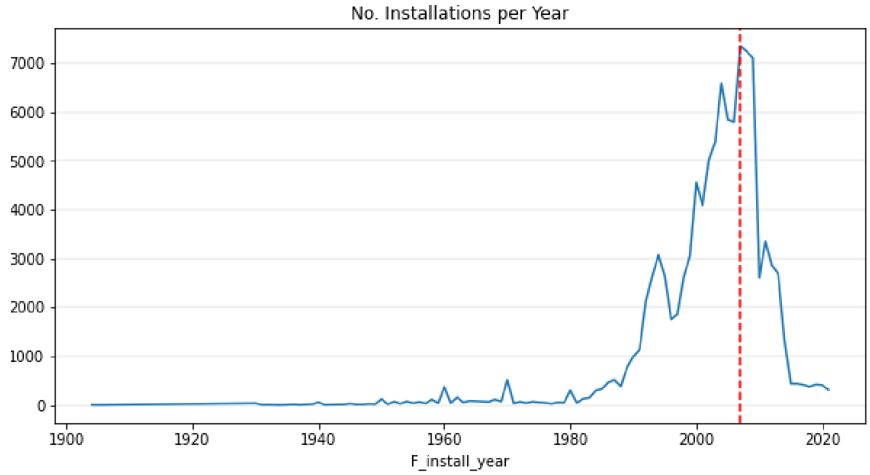

Figure 3: Installation of Water Points Over Time

2.3.5. Relationship Between Installation Year and Status

The below box plot for functioning and non-functioning both show right-skew. Furthermore, functioning water points have a median age that is slightly lower than non-functioning ones.

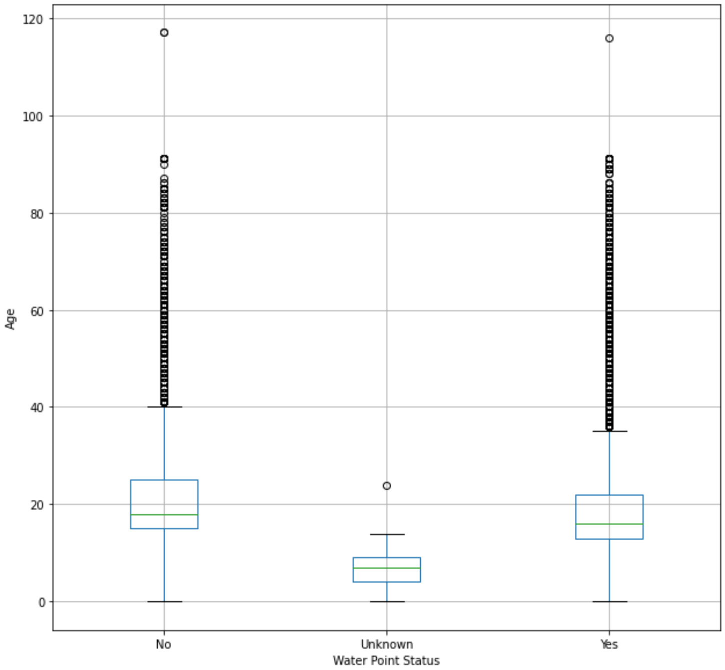

Figure 4: Box Plot of Age and Water Point Status

2.3.6. Measurement of Rural Water Access Per District

This piece of analysis shows the power of geospatial analysis. In a nutshell, we performed the following steps:

- Exclude urban population based on the contours of urban zones from our 2016 population data and spatially join with district information
- Select functioning water points and create a buffer of 1000 meter around each point
- Identify which rural population points are within those buffers
- Perform the necessary manipulations to group and aggregate to obtain the fraction of rural population with access to total rural population per district

This analysis leads to the below plot, in which we see 5 classes of 20% increments. In other words, the darkest color on this heatmap means that more than 80% of the rural population in the district has no access to a water point. To say the least this is dramatic.

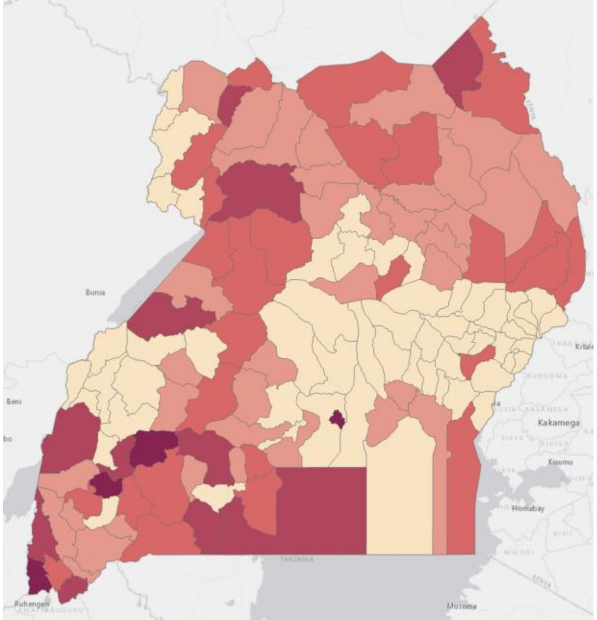

Figure 5: Heatmap of Rural Water Access Per District

2.3.7. Measurement of Water Point Pressure Rating and Crucialness Per District

We performed an additional piece of analysis that once again shows the power of geospatial analysis:

- For all functioning water points, we attributed a fraction of the sum of the local population around it in a 1000-meter radius: distance between a water point and a population point served as the weight. The complexity is that you can have multiple overlapping water points and populations points and this required a few steps of aggregation.
- In a second step, we take that attributed population and divided by the total population around the water point (again 1000-meter radius), this leads to a so-called crucialness score. I.e., this value shows the redundancy of the system and can be illustrated by an example: suppose you

have only one water point around a population point: crucialness would be 100%. With two water points at equal distance, this would be 50% each.

- Next, we repeated this analysis for each non-functioning water point (one at a time and filtered the local area to reduce computation time).
- A final step comprised calculating pressure ratings. Each type of water point source/technology has a typical usage capacity (in terms of persons served, e.g., 1000 persons for a mechanized well [1]). Taking the fraction of served population to capacity results in the pressure rating. Anything above 100% means that there is a mismatch in people served versus water point capacity.

The below plots contain the crucialness and pressure results per district. Especially the pressure rating plot is worrying: compared to the standard guidelines on capacity, we see that we have 51 districts have an average pressure rating above 100%. The caveat to this analysis is that we base ourselves on a 1000-meter radius. The reality is that the people in Uganda will need to walk much farther than 1000 meters (as that constraint meant excluding nearly 60% of the water points).

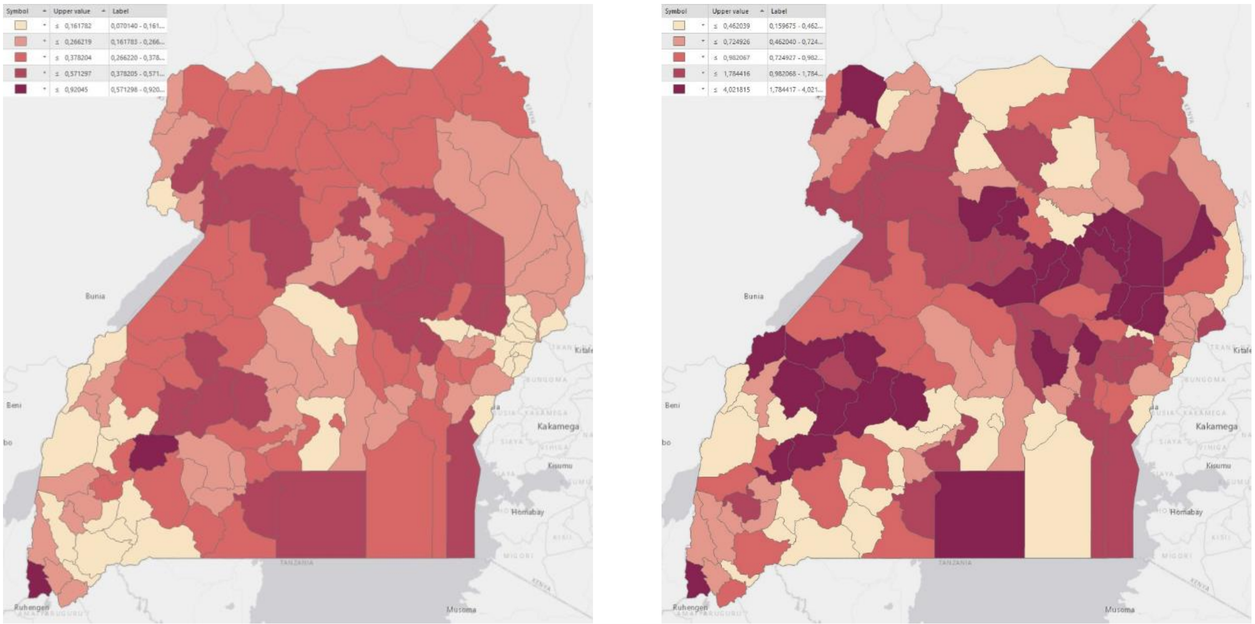

Figure 6: Crucialness and Pressure Plot Per District

2.4. Clustering Analysis

In our spatial statistics toolbox, we have a number of very powerful tools at our disposal. In this section, we'll illustrate a number of applications and the analyses we performed. We'll look into density-based clustering for both functioning water points and rural population. The involved technique HDBSCAN (or in full Hierarchical Density-based Spatial Clustering of Applications with Noise), which identifies clusters based on their spatial distribution, filters out noise and allows varying distances [2].

2.4.1. Density-based Clustering of Functioning Water Points and Rural Population

In below plots, we've captured such a density-based clustering method[8] for our two main data tables of interest. These plots show clusters of functioning water points and rural population around urban sprawl, which coincides to some extent with the rural-urban transformation[9] ongoing in Uganda:

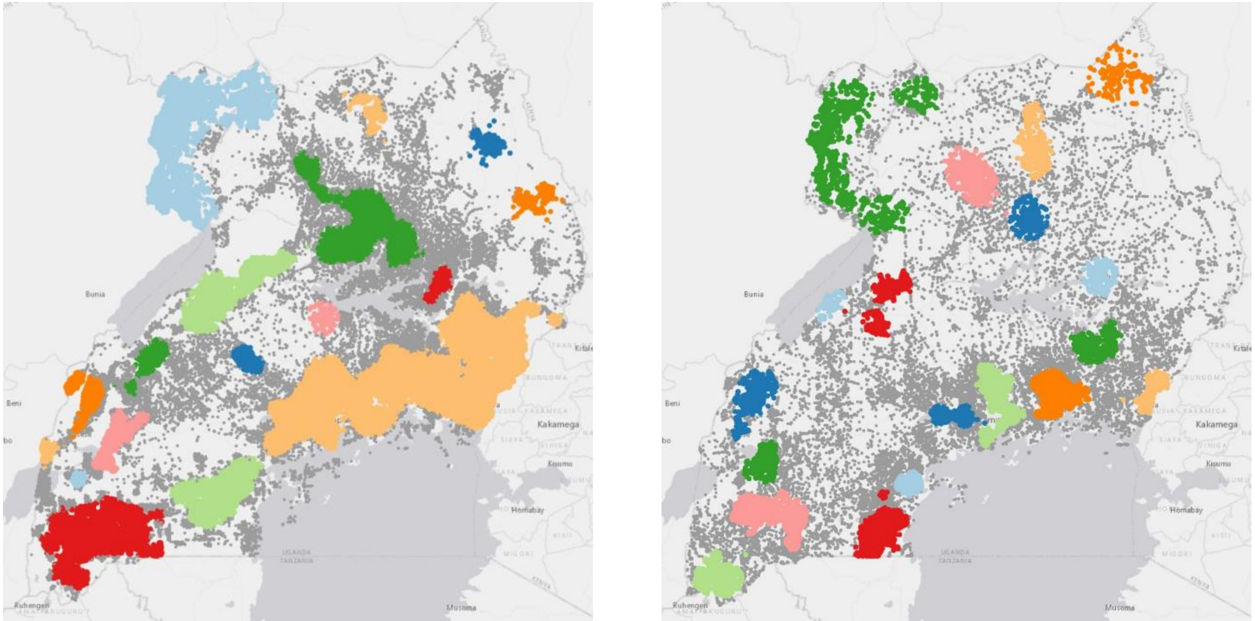

Figure 7: Density-based Clusters of Functioning Water Points and Rural Population

2.4.2. Hot Spot and Cluster Outlier Analysis

In below side-by-side plot, we show two different kinds of outlier analysis for the percentage rural water access in a district:

- Hot spot analysis based on the Getis-Ord Gi* statistic [3]; it determines whether the feature in a certain neighborhood is statistically significantly different from the area we are studying. We'll tag the feature in that neighborhood as hot spot if significantly higher or cold spot if significantly lower.
- Cluster outliers based on Anselin Local Moran's I [4]. Contrary to a hot spot analysis, it removes the feature in the neighborhood being studied and then it checks whether that feature is statistically different from its neighborhood.

In the hot spot analysis, red and blue mean hot and cold spot, respectively. One large cluster of hotspots is found in the eastern region, whereas in the southwest we have a concentration of cold spots. So, what's happening? In the eastern part, we actually have a region in which the rural water access is very high, except for two districts. It's also high compared to the map average. In the southwest, the inverse reasoning can be made.

[8] Self-adjusting density-based clustering with 300 and 350 minimum features per cluster respectively

[9] Rural-Urban Transformation in Uganda, IFPRI and University of Ghana conference, 10 May 2011

In the cluster outlier analysis, there are three dark red districts in the southwest area. These are in fact districts with high rural water access, whereas the rest of the southwest region has low rural water accessibility. The opposite result is found in the eastern region. There are two blue districts with low rural water accessibility in a region where the rest has a high rural water accessibility.

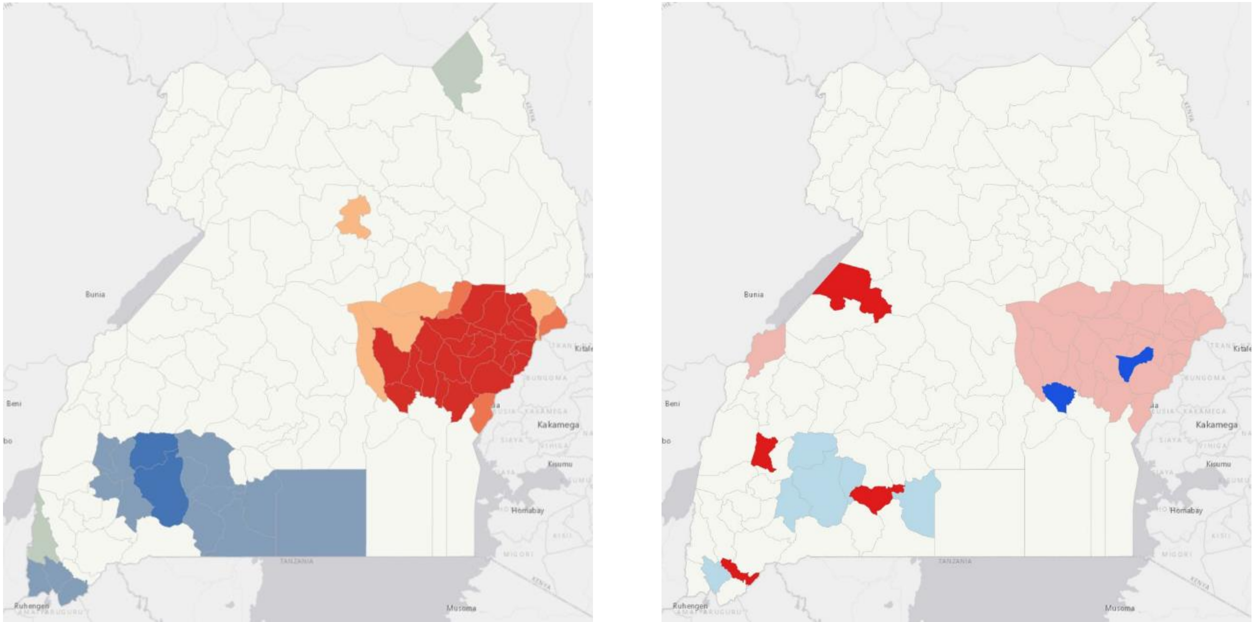

Figure 8: Hot Spot and Cluster Outlier Analysis for Rural Water Access

2.5. Class Imbalance

As we saw in the exploratory data analysis section, there was a class imbalance in the dataset which could pose a problem. Our concern turned out to be true. We tried out methods with minimal feature engineering and found out that methods such as logistic regression and Naïve Bayes were simply predicting the majority class, which still yielded the accuracy around 80%. Models with this accuracy is not meaningful when a majority class accounts for 80% of the dataset.

Class imbalance is not a new problem in data science and still an active field of interest [5] [6]. Typically, the problem is tackled from three angles:

- Consider and use the right metrics for the problem under consideration
- Evaluate the implication on the methods that were selected (next to varying the methods)
- Address the imbalance in the data itself

In this section, we will focus the third approach; metrics & methods will be covered in a later section. The three main methods for addressing the class imbalance in the data itself are as below:

- Oversample the minority class, in which we randomly sample additional data points of the minority class with replacement
- Undersample the majority class, in which we randomly remove records with the majority class
- Generate additional data for the minority class synthetically via SMOTE (Synthetic Minority Oversampling TEchnique) or ADASYN (adaptive version)

Given the size of our dataset, we opted for undersampling of the majority class to construct a balanced dataset. The dataset has 42,702 data points which is enough number of datapoints for modelling. We used this dataset as well as the original one when evaluating the different methods.

2.6. Feature Enrichment

One of the powerful features of the software package we used for this project, is that it allows for data enrichment: either through built-in methods or workflows that allow relatively easy imports. The main challenge in that enrichment is finding the right data sources and of course have these data sources support an initial hypothesis on potential relevance.

In below list, the hypotheses, the data and the relevance to the model. The detailed explanation will be discussed in the later section.

Data	Initial Hypothesis	Modeling Relevance
Uganda Roads	Distance to nearby road infrastructure could help in repairs	No, only used with geospatial random forest
Distance to Urban Areas	Ease of repair and/or access to parts, for more remote areas that would be slightly more complicated	Yes
Soil Data: General, Hydric, Chemistry and Bulk Density	Soil characteristics influence infiltration, permeability and the capacity to store water	Yes, but we had equivalently powerful predictors (i.e., swapping the one for the other made no difference)
Recalculation of rural population within 1000 meters	Our dataset for water points had an attribute on rural and urban population within 1000 meters, as we are modeling rural access, we recalculated	Yes
Elevation	Elevation and distance to the water table show a large correlation [7]	Yes
Demographic data for Uganda	We looked at education level, gender balance, purchasing power parity (PPP), household size, and population density: we thought that this might affect maintenance of water points	Yes, for PPP and population density, but ultimately did not retain
GDP Per District	More developed districts (as measured by GDP) have better infrastructure	No

For the demographic enrichment, we had to consider the cost related to this paying Esri GeoEnrichment service. Given the size of our dataset and the associated cost, we first ran experiments on a reduced dataset. We found two predictors of interest, but ultimately did not retain as the trade-off between cost and the performance improvement were not worth it.

3. Methodology

3.1. Feature Engineering and Selection

In this section, we arrive at the place where art meets science. We had two problems. First, values in some predictors were not structured. Second, there were too many predictors. We'll detail our approach to deal with the problems as follows.

3.1.1. Feature Engineering of Unstructured Data

In our water point dataset, we had three predictors that seemed relevant but had too many values:

- Payment: 119 unique values
- Raw Status: 481 unique values
- Subjective quality: 294 unique values

We performed a manual remapping of the values of these predictors to 4, 6, and 7 values respectively. For example, there were 55 values in subjective quality that mentioned the word smell or some variant such as stinks, smelling, smelly and so on. These values were remapped to a single value 'Smell'.

3.1.2. Feature Selection

We spent significant amount of time on performing feature selection to reduce the complexity of the model. The steps on enrichment, engineering and selection were done through multiple iterations.

To perform feature selection, we used the following methods:

- Feature importance in random forest
- Feature importance in a gradient booster
- Feature importance via chi-square
- Feature importance via mutual information gain
- In a final step, we added feature importance via Boruta, which, in a nutshell, is a wrapper around a random forest, duplicates & shuffles the values in order to determine z-scores for relevance, and at each iteration it gets rid of the non-performing features [8]

We looked at the intersection of these methods on which predictors carried the most relevance. Nevertheless, there was crucial threshold where we let the model performance speak for itself. There were predictors for which we could argue they did not carry enough weight, but once removed model performances dropped significantly (in other words the art of feature selection came into play here).

3.2. Metrics

As mentioned before, we had an imbalanced dataset and our initial modeling showed some potential problems with the standard metrics. That is to say, the simple accuracy metric can be misleading. To tackle this problem, we need to consider other metrics besides accuracy to evaluate our models. Workhorses of choice are precision and recall as well as the F1-score, which combines both [9] [10].

Precision tells us to which extent we can trust the model when giving us a positive prediction (i.e., the ratio of true positives to the sum of true and false positives). Recall, on the other hand, paints the picture of predicted relevant items with respect to all existing relevant items (i.e., the ratio of true positives to the sum of true positives and false negatives). With these two metrics, we have functions that span the columns and rows of a confusion matrix. Finally, the F1-score is the harmonic mean of precision and recall.

In our performance evaluation, we used the unweighted version of precision, recall and F1-score for our two classes as this will allow easier comparison across the balanced and full datasets. Accuracy, on the other hand in below results is weighted across the two classes.

3.3. Methods

As this a geospatial problem next to a binary class problem, we approached the problem from two angles: conventional methods and a geospatial method. The methods we investigate are as below:

- Logistic Regression
- Naïve Bayes
- K-Nearest-Neighbors
- Support Vector Machines
- Neural Network
- Random Forest
- Gradient Boosting
- Geospatial Random Forest

All these methods are supervised learning methods. Our main objective is to strike the right balance between precision and recall, which we assess via the F1-score. A secondary objective is to minimize computation time.

We applied the methods on four dataset variants:

- One-hot encoded and full dataset
- Label encoded and full dataset
- One-hot encoded and balanced dataset
- Label encoded and balanced dataset

We included both one-hot encoding and label encoding of our dataset because of the following reasons. First of all, we have methods that don't perform very well with label-encoded datasets as they try to infer meaning from a nominal variable (e.g., Logistic Regression, Naïve Bayes, k-Nearest Neighbors, SVM). Secondly, we have predictors with many levels, which leads to a blow-up of the feature space in one-hot encoding and increases training time

3.4. Hyperparameter Tuning

We have added the details of our grid search approach on hyperparameter tuning in the appendix. We can summarize our approach as follows (which we applied for each encoding/dataset variant):

- 80%-20% split of dataset into training and testing.
- We used the training data alone to tune the hyperparameters using 5-fold cross validation.

- From this tuning, we selected one set of parameters for each method that was then used to train on the full training set and, subsequently, to assess performance on the test set.

We should note that the tuning was a computationally expensive task requiring a few iterations, which forced us to restrict the search space.

4. Results

4.1. Feature Selection

As a result of the steps outlined in the feature selection section, we ended up with 19 predictors in our conventional methods. The geospatial method was further extended with spatial datasets on road data and urban area, but we also had to remove some predictors as they were not compatible with the implementation of the method (e.g., too many categories). Including the spatial extensions, we have 15 predictors for this method.

The below two plots show the feature importance for the geospatial random forest and how the importance changes across ten runs (i.e., box plot for each variable):

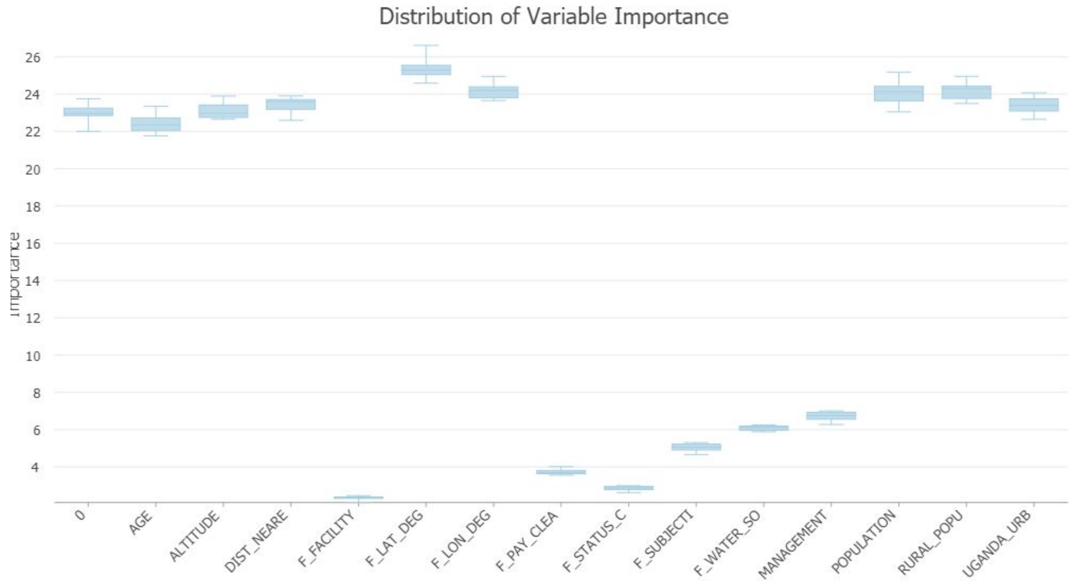

Figure 9: Feature Importance Plot Geospatial Random Forest Full Dataset

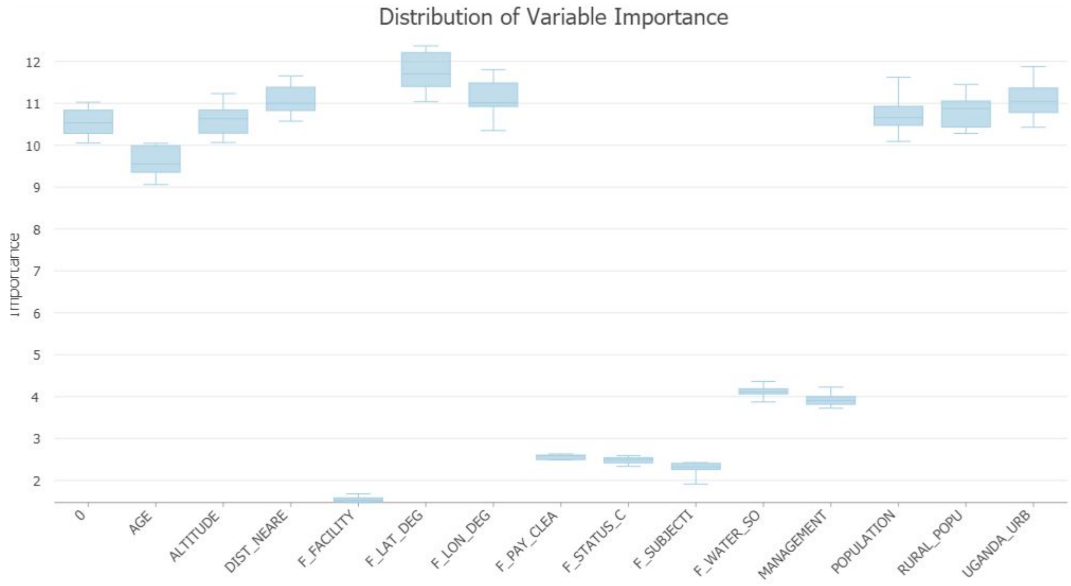

Figure 10: Feature Importance Plot Geospatial Random Forest Balanced Dataset

4.2. Model Performance

The below tables show the performance of models from four different approaches: one-hot encoding versus label encoding and the full dataset versus the balanced dataset. We should note that Geospatial Random Forest and Naïve Bayes were only run on the balanced dataset due to limitations of each method.

As mentioned before, The tables below show the results of all the models. The precision and the recall below are unweighted averages of the two classes and accuracy, on the other hand, is a weighted metric.

4.2.1. Performance of One-hot Encoding on Full Dataset

Method	Precision	Recall	F1-score	Accuracy
Logistic Regression	94.82%	96.13%	95.46%	97.23%
k-Nearest Neighbors	94.23%	94.23%	94.23%	96.53%
Radial SVM	94.77%	96.19%	95.46%	97.23%
Neural Networks	94.82%	96.21%	95.49%	97.25%
Random Forest	94.86%	96.07%	95.45%	97.23%
Gradient Boosting	94.81%	96.29%	95.53%	97.27%

4.2.2. Performance of Label Encoding on Full Dataset

Method	Precision	Recall	F1-score	Accuracy
Logistic Regression	92.57%	94.44%	93.47%	95.98%
Naïve Bayes	92.88%	90.86%	91.83%	95.20%
k-Nearest Neighbors	93.70%	92.78%	93.23%	95.97%
Radial SVM	40.78%	50.00%	44.92%	81.56%
Neural Networks	94.74%	95.88%	95.30%	97.14%
Random Forest	94.90%	96.58%	95.72%	97.37%
Gradient Boosting	95.00%	96.65%	95.80%	97.43%
Geospatial Random Forest	94.07%	96.46%	95.21%	97.03%

4.2.3. Performance of One-hot Encoding on Balanced Dataset

Method	Precision	Recall	F1-score	Accuracy
Logistic Regression	97.29%	97.20%	97.22%	97.23%
k-Nearest Neighbors	96.33%	96.26%	96.27%	96.28%
Radial SVM	97.22%	97.13%	97.15%	97.15%
Neural Networks	97.20%	97.12%	97.14%	97.14%
Random Forest	97.20%	97.11%	97.13%	97.13%
Gradient Boosting	97.31%	97.23%	97.25%	97.25%

4.2.4. Performance of Label Encoding on Balanced Dataset

Method	Precision	Recall	F1-score	Accuracy
Logistic Regression	94.10%	94.03%	94.05%	94.05%
Naïve Bayes	94.34%	94.23%	94.19%	94.19%
k-Nearest Neighbors	93.85%	93.79%	93.80%	93.81%
Radial SVM	88.62%	88.38%	88.40%	88.42%
Neural Networks	97.30%	97.18%	97.20%	97.20%
Random Forest	97.37%	97.30%	97.32%	97.32%
Gradient Boosting	97.23%	97.16%	97.18%	97.18%
Geospatial Random Forest	96.88%	96.89%	96.89%	96.89%

5. Conclusion

5.1. Discussion

Our proposed methods exhibit favorable performances overall. We have multiple methods that score more than 97% across the different metrics, which is a vast improvement over the initial models with around 85% accuracy.

As the F1-score is a combination of precision and recall which allows a simpler comparison of the balanced versus full dataset, we used this metric for our conclusions:

- Random Forest and Gradient Boosting come out on top across the 4 categories and a small performance improvement for the label-encoded dataset. Interestingly, these two methods achieve their results in a different way: gradient boosting by reducing bias and random forest by reducing variance. In any case, it leads to the same sweet spot with any of these two methods that could be crowned king.
- In general, the results on the balanced datasets are better by about 1% to 2% depending on which metric we use.
- We need to state the obvious: the performance differences are very small. If we take the top 10 best performances, then the min-max difference for F1-score and accuracy is around 1%.
- When it comes to encoding, label vs. one-hot doesn't matter for random forest and gradient boosting as expected, whereas for the other methods it does as it tries to derive meaning from ordinal values. Furthermore, the results on the balanced dataset with one-hot encoding are all extremely close to one another; the difference is in the third and fourth decimal places.
- Support Vector Machines with a radial kernel performed catastrophically on the label-encoded dataset. It is because SVM tries to derive meaning from an order that is not actually present in the dataset. One-hot encoding goes around this point and in that case performed just behind the winning methods.
- Logistic regression and k-Nearest-Neighbors showed decent results: mid-nineties across the performance metrics with peaks in the one-hot encoded balanced dataset. Still, it suffered from the class imbalance in the full datasets, as well as performing worse on the label-encoded datasets for the same reason as SVM.
- Naïve Bayes is at the bottom of the considered methods. Its approach is based on independent features, an assumption which more than likely is not correct for our dataset.
- Neural Networks performed very strongly across the board and only just behind the two winning methods. In retrospect, since we did not tune that many topologies, there might be some room for improvements. Still, this will come at the expense of rather lengthy training times.
- Final note goes to geospatial random forest, which performed strongly as well despite the removal of a few predictors.

Our two winning methods are ensemble methods. Although they performed fantastically, they have one major drawback. These methods are not locally explainable as are classic linear regression models. That being said, we have feature importance scores, that show how important each feature from an overall perspective. The trade-off of between interpretability and performance was inevitable in this case.

5.2. Future Work

We have several areas that we would like to explore in the future:

- Identifying which water points are going to fail is one thing, but determining which locations to build new ones is another. With the power of GIS, this could be a very interesting problem to tackle.
- In terms of tuning, we could enlarge the search space for other neural network topologies to squeeze out even more performance
- Although we spent a significant amount of time on our feature enrichment, it still seems there is much more potential: establishing an automated workflow for importing, data-wrangling, modeling and results generation would be an interesting setup to accelerate the exploration and incorporation of other datasets.

5.3. Lessons Learned

We tremendously enjoyed this practicum sponsored and supported by the great Esri team. We have a few takeaways from this project that will hopefully help us in the future:

- The learning curve on the geospatial dimension of this project was steep: not only does it require a new way of thinking, but we also had to work with a new software package that comes with a very extended toolbox. We were fortunate to go through an extended bootcamp and this helped to get the first baby steps right.
- As mentioned in the future work section, feature enrichment was not a walk in the park, in retrospect, we should perhaps have spent more time on building an automated pipeline to simplify our lives somewhat.
- We had the chance of working on an imbalanced dataset, this caused a few headaches at the start, but it was a great opportunity at the same time to really think through on results and which metrics make sense in which circumstances.
- Finally, with regards to tuning, though it is not as sophisticated, it is extremely time-consuming. We made a number of small errors such as copy-paste problems across the four variants, not enabling verbose logging or having a search space that was too large and for which our back-of-the-envelope calculation on how it would extrapolate from a subset of the data to the full dataset proved wrong. Small but important learnings that we will take with us for the next project.

6. Appendix

6.1. Source Code

We have uploaded all our code in a separate compressed file along with this write-up. This source code has been shared with the practicum sponsors Esri.

6.2. StoryMap and WebApp

We produced two other deliverables not in this report:

- To illustrate in a geospatial manner our results from the rehabilitation analysis, we've built a web application that allows filtering on the various water points and shows the calculation details
- We also created a StoryMap, which could be thought of as a fancier way of presenting slides

6.3. Data Sources

Data sources graciously prepared by Esri and made available at the start of this practicum:

- Raw water point data downloaded from Water Point Data Exchange (WPDx) on August 10, 2021, filtered for Uganda and enriched [11]
- Uganda administrative districts from the Humanitarian Data Exchange and dissolved boundary [12]
- Africapolis urban areas classification from the Africapolis dataset: this designates whether a water point is urban or rural [13]
- Uganda population density in 2016: show counts of people and urban/rural classification from Esri World Population Estimate 2016 and World Population Density Estimate 2016 [14] [15]

Data sources for enrichment:

- 2017 Uganda GDP per district from Frederick S. Pardee Center for International Futures at the University of Denver [16]
- World elevation from Global Multi-resolution Terrain Elevation Data [17]
- Soil Data: general, chemistry, hydric and bulk density [18] [19] [20] [21]
- Uganda Road Data [22]
- Demographics [23]

6.4. Hyperparameter Tuning

Below we have included a small discussion on the parameters we tuned via grid search for the different models:

- Logistic Regression was tuned on the following parameters:
- C: the inverse of the regularization strength, bigger values imply weaker regularization,
- Solver: we tried different solvers as not all have the same behavior (i.e., speed when it comes to dataset size
- Penalty: we looked at no penalty as well as L1, L2 and elastic net

- Naïve Bayes was tuned on its variance smoothing parameter (which influences the distribution's variance, in other words it changes the filtering)
- For k-nearest-neighbors, we performed a grid search on the number of neighbors (parameter k)
- For Support Vector Machines, we did a grid search on the cost parameter for a radial basis kernel. This cost parameter is the regularization term (or in other words the trade-off between margin width and misclassification). We did not tune on gamma (which is inversely related to the variance of the Gaussian kernel).
- Neural networks: we tried different topologies (3 layers, but different sizes) as well as different L2-penalty terms. We also looked at different solvers and decay rates (or learning rates). The topology helps for non-linearly separable problems and the learning rate directly impacts how much an update step influences the values of the weights.
- Random Forests: here we tuned the number of features to include in the split of each node as well as terminal samples. Other parameters were the depth of the trees and the number of trees (although we limited that to 100 or 200, the performance increase should very limited between these two choices [24]).
- Gradient Boosting was tuned on the following parameters (loss was deviance):
    - Learning rate (which determines the contribution of each tree)
    - Number of boosting stages (which is typically quite robust to overfitting)
    - Minimum number of samples to split a node
    - Maximum depth of the individual tree
    - Subsampling (we tried 0.8 which results into stochastic gradient boosting, which can be beneficial for variance reduction at the expense of bias as well as 1.0)

Bibliography

[1] Sphere Association, "The Sphere Handbook: Humanitarian Charter and Minimum Standards in Humanitarian Response," 2018.

[2] R. J. Campello, D. Moulavi and J. Sander, "Density-based clustering based on hierarchical density estimates," *Pacific-Asia Conference on Knowledge Discovery and Data Mining,* pp. 160-172, 2013.

[3] A. Getis and J. K. Ord, "The Analysis of Spatial Association by Use of Distance Statistics," *Geographical Analysis,* vol. 24, no. 3, 1992.

[4] L. Anselin, "Local Indicators of Spatial Association-LISA," *Geographical Analysis,* vol. 27, no. 2, pp. 93-115, 1995.

[5] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," in *Progress in Artificial Intelligence*, 2016.

[6] A. Fernández, S. García, M. Galar, R. .. Prati, R. C. Prati, B. Krawczyk and F. Herrera, Learning from imbalanced data sets, 2018.

[7] C. W. Fetter, Applied Hydrogeology (Fourth Edition), Pearson, 2000.

[8] M. B. Kursa and W. R. Rudnicki, "Feature Selection with the Boruta Package," *Journal of Statistical Software,* vol. 36, no. 11, pp. 1-13, 2010.

[9] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies,* vol. 2, no. 1, pp. 37-63, 2020.

[10] M. Grandini, E. Bagli and G. Visani, "Metrics for multi-class classification: An overview," 2014.

[11] "Water Point Data Exchange (WPDx)," [Online]. Available: https://www.waterpointdata.org/access-data/.

[12] Humanitarian Data Exchange, "Uganda administrative districts," [Online]. Available: https://data.humdata.org/dataset/uganda-administrative-boundaries-admin-1-admin-3.

[13] Africapolis, "Urban areas classification," [Online]. Available: https://africapolis.org/data.

[14] Esri, "World Population Estimate 2016," [Online]. Available: https://www.arcgis.com/home/item.html?id=92d3005feb84428a8f85160f2451ec63.

[15] Esri, "World Population Density Estimate 2016," [Online]. Available: https://www.arcgis.com/home/item.html?id=0f83177f15d640ed911bdcf6614810a5.

[16] M. Rafa, J. D. Moyer, X. Wang and P. Sutton, "Estimating District GDP in Uganda," 2017.

[17] Esri, "World elevation from the Global Multi-resolution Terrain Elevation Data 2010," [Online]. Available: https://www.arcgis.com/home/item.html?id=e393da08765940e49e27e30e1df02b58.

[18] Esri, "World Soils Harmonized World Soil Database - General," [Online]. Available: https://www.arcgis.com/home/item.html?id=af37c984900c48618b158352fb41da4d.

[19] Esri, "World Soils Harmonized World Soil Database - Chemistry," [Online]. Available: https://www.arcgis.com/home/item.html?id=0e71d0e63c494d75b2bc897b7515f89a.

[20] Esri, "World Soils Harmonized World Soil Database - Hydric," [Online]. Available: https://www.arcgis.com/home/item.html?id=233818f3e40a4bc39e4f8a942c19e6fb.

[21] Esri, "World Soils Harmonized World Soil Database - Bulk Density," [Online]. Available: https://www.arcgis.com/home/item.html?id=9b1cefacf7be47ab93c2dab2e2f24d68.

[22] RCMRD GeoPortal, "Uganda Roads," [Online]. Available: https://gmesgeoportal.rcmrd.org/datasets/africageoportal::uganda-roads/about.

[23] Esri, "Global demographic data," [Online]. Available: https://doc.arcgis.com/en/esri-demographics/.

[24] T. Oshiro, P. P. Perez and J. Baranauskas, "How many trees in a random forest?," in *Lecture Notes in Computer Science*, 2012.