Predicting the Condition of Waterpoints in Tanzania

James Toop

Data Science, Module 3 – Final Project

Purpose – A Global Water Crisis

Water is an essential of life yet, according to a WHO report from 2019 :

- 2.2 billion people globally do not have safely managed drinking water services.
- 4.2 billion people do not have safely managed sanitation services.
- 3 billion lack basic hand washing facilities.
- One of the most common causes of death in the developing world is drinking dirty and diseased water.

Sources: United Nations Sustainable Development Goals and World Health Organisation

Purpose – The Situation in Tanzania

Tanzania has a water and sanitation crisis :

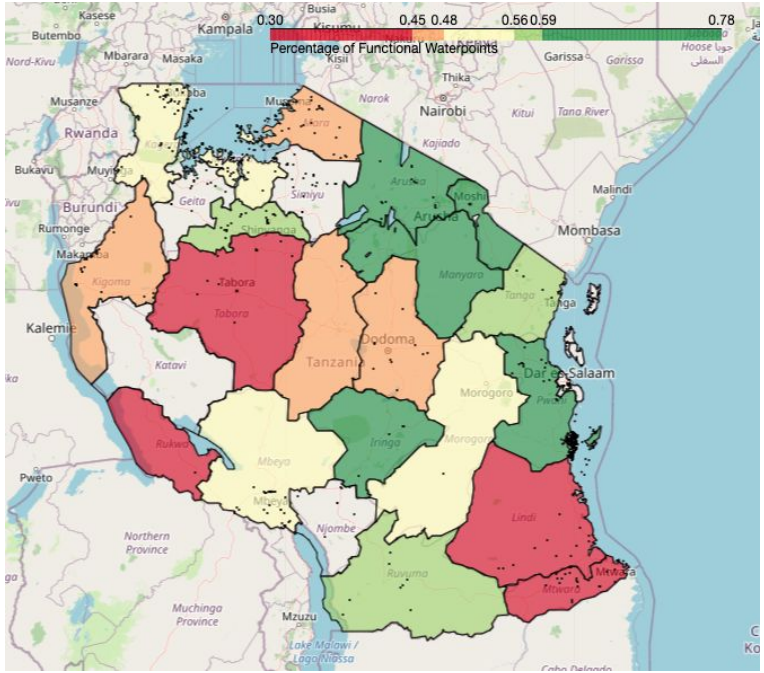
- Only 60% of the population of 61 million have access to an improved (protected from contamination) source of safe water.
- Since 2007, the government of Tanzania has been implementing a Water Sector Development Programme at a cost of around $300 billion USD.
- Non-functioning waterpoints force communities to rely on unsafe sources of water, often several kilometres away, affecting not just health but education too.

Source: wateraid.org

Objective –

Create a classifier that predicts which waterpoints are functional, which need some repairs, and which don't work at all.

Observations – Regional Variations



Almost half of all waterpoints are in need of repair or not functioning at all.

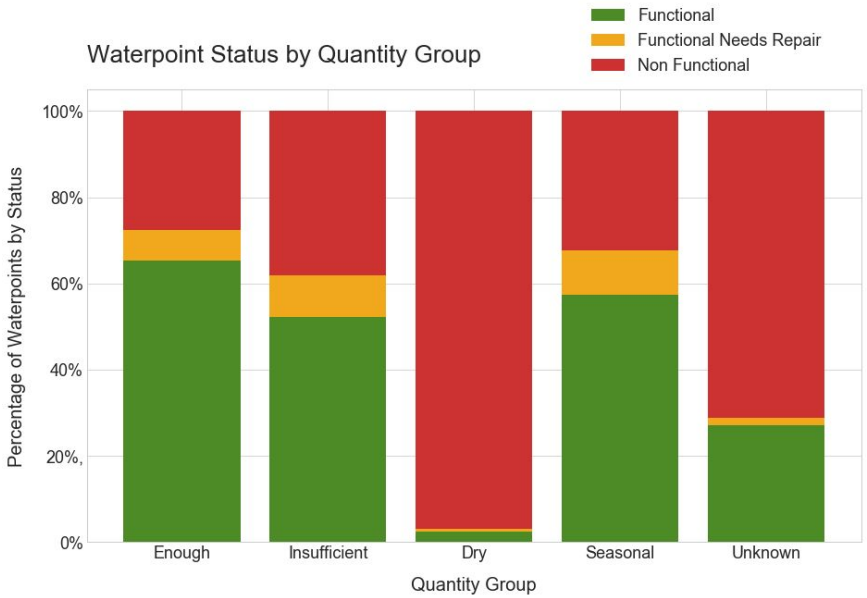
- 38% of waterpoints are non functional.
- Some regions exhibit a much lower percentage of functional waterpoints than others.
- Inconsistent regional coding with some regions with no waterpoints?

Source: Taarifa Waterpoints Dashboard and the Tanzanian Ministry of Water

Observations – Quantity

As quantity decreases, the proportion of non functional waterpoints increases & the “dry” waterpoint paradox.

- A large proportion of waterpoints that have been recorded as “dry” are also non functional.
- It is unclear from the documentation if these waterpoints are non functional because they are dry or for some other reason (e.g. the extraction mechanism has broken down).

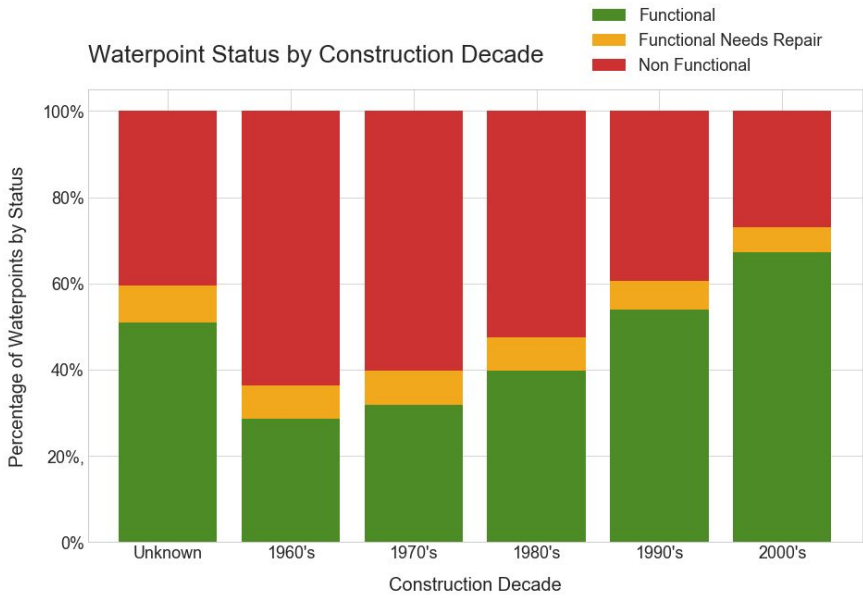


Source: Taarifa Waterpoints Dashboard and the Tanzanian Ministry of Water

Observations – Age

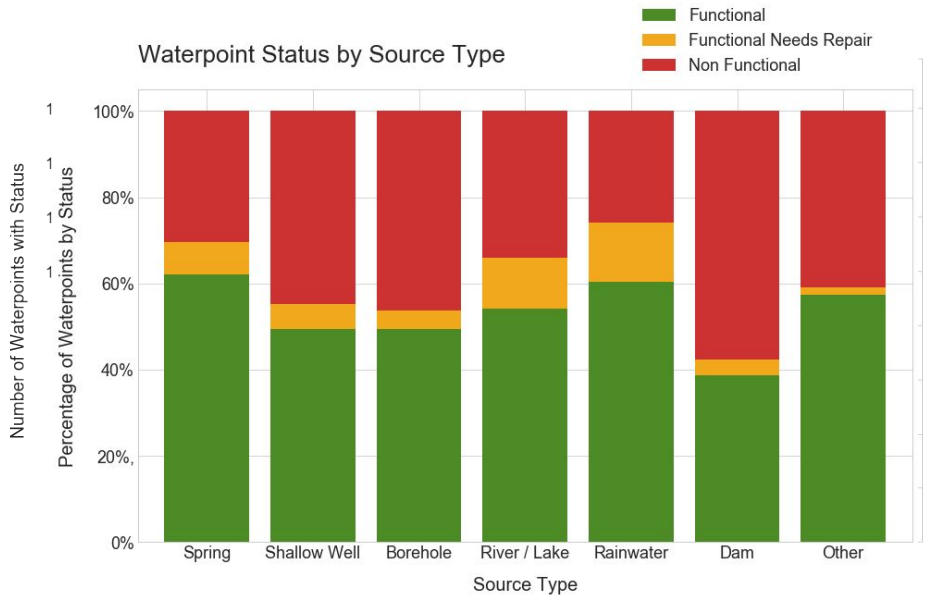
The older the waterpoint, the greater the proportion of non functional waterpoints.

- Rather obviously, the age of the waterpoint is a factor BUT
- The number of waterpoints constructed has increased significantly since 2000 with double the number of waterpoints constructed than in the 1990's.



Source: Taarifa Waterpoints Dashboard and the Tanzanian Ministry of Water

Observations – Natural vs Engineered

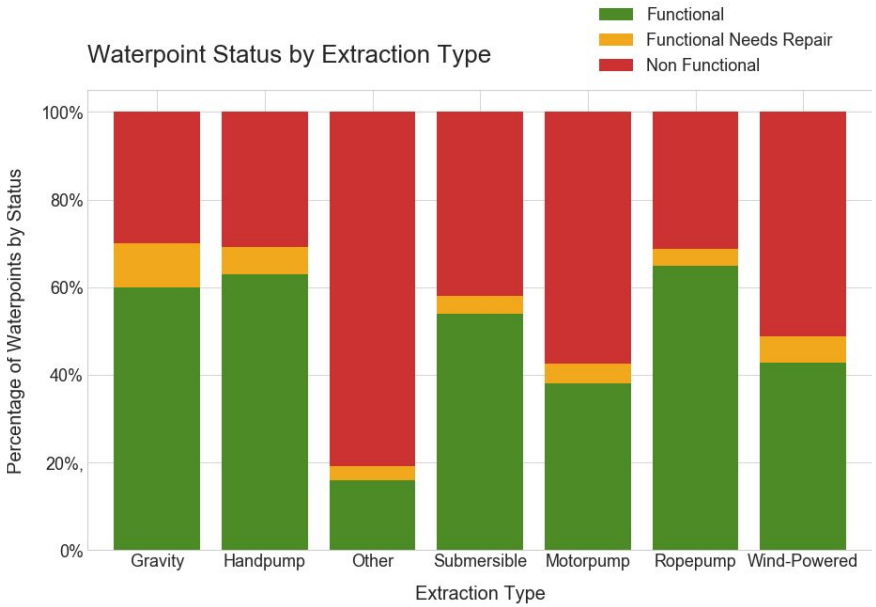


Sources with little or no engineering intervention are more likely to be functional.

- Waterpoints with a rainwater, spring or river / lake source have a higher proportion of functional waterpoints.
- But springs, shallow wells and boreholes account for a very large proportion of all waterpoints.

Source: Taarifa Waterpoints Dashboard and the Tanzanian Ministry of Water

Observations – Extraction Type



Waterpoints with motorised extraction types are less likely to be functional.

- The more complex the working parts of the waterpoint, the less likely the waterpoint is to be functional.
- Compare the proportion of non-functioning waterpoints that use a hand pump to those that use a motor pump.

Source: Taarifa Waterpoints Dashboard and the Tanzanian Ministry of Water

Modelling Approach – Accuracy

Building a classifier to predict the condition of the waterpoints and therefore be able to prioritise repairs efficiently.

- In a many instances, the data fields provided are either direct duplicates of other fields or provide the same information but in greater detail.
- Dropping the more detailed data fields, converting other key data into appropriate categorical variables and through an iterative process arrived a shortened list of predictor variables.
- The Random Forest classifier and Support Vector Machine (SVM) classifier produced the best results correctly classifying 77-78% of the “unseen” waterpoints, but there is a significant computational overhead when using the SVM classifier.

Results and Possible Next Steps

Results (so far) –

	Model	Classification Rate
1	Support Vector Machines	78%
2	Random Forest	77%
3	XGBoost	76%
4	Bagged Tree	71%
5	Logistic Regression	71%

Recommended next steps –

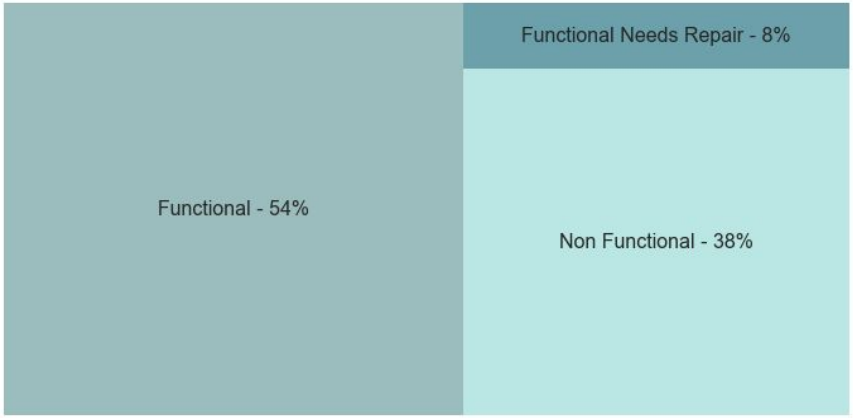
- Further re-engineering of the existing features could potentially improve the accuracy of the classifiers further.
- Source alternative datasets for cross reference and to improve the existing dataset, particularly in relation to geographical coding and the age of waterpoints.
- Improve data collection of population figures for each waterpoint or augment with additional datasets. Knowing how many people use each waterpoint will facilitate prioritisation for repair.

Thank you.

Any questions?

Appendix – Overall Condition

Almost half of all waterpoints are in need of repair or not functioning at all.



Source: Taarifa Waterpoints Dashboard and the Tanzanian Ministry of Water