Predicting Tanzania Water Pumps Maintenance

Silvia Maione

Springboard Data Science Career Track

Outline

- ➢ Motivation
- ➢ Dataset and methodology
- ➢ Key findings
- ➢ Modeling
- ➢ Summary, next steps
- ➢ References

Motivation

- Predicting the functionality and aging of water pumps in the state of Tanzania in order to timely suggest maintenance
- Allowing continuous access to clean and fresh water
- Scheduling needed maintenance before loss of functionality in order to save on cost

Benefit to the population, the government of Tanzania, private companies and public agencies involved

Dataset

- Originally made available by the Government of Tanzania and Taarifa
- Used in competitions, typically as an example of classification problem
- About 60k entries, each representing a water pump
- 40 attributes, mainly categorical, many geographical, population and amount of water
- Each water pump classified as “functional”, “non-functional”, “needs repair”

Dataset (cont.)

- The data is sparse and has high cardinality
- Most of the attributes are categorical
- Some entries look wrong
- Categories are label-encoded
- Missing values are filled using transformation of other attributes

```
Percentage of missing values in funder: 7.4
Percentage of missing values in installer: 7.5
Percentage of missing values in wpt_name: 6.0
Percentage of missing values in subvillage: 0.6
Percentage of missing values in public_meeting: 5.6
Percentage of missing values in scheme_management: 6.5
Percentage of missing values in scheme_name: 47.4
Percentage of missing values in permit: 5.1


Number of columns of float64 type: 3
 ['amount_tsh', 'longitude', 'latitude']

Number of columns of str type: 29
 ['date_recorded', 'funder', 'installer', 'wpt_name', 'basin', 'subvillage', 'region', 'lga', 'war
_management', 'scheme_name', 'extraction_type', 'extraction_type_group', 'extraction_type_class',
group', 'payment', 'payment_type', 'water_quality', 'quality_group', 'quantity', 'quantity_group',
'source_class', 'waterpoint_type', 'waterpoint_type_group', 'status_group']

Number of columns of int64 type: 6
 ['gps_height', 'num_private', 'region_code', 'district_code', 'population', 'construction_year']

Number of columns of bool type: 2
 ['public_meeting', 'permit']
```

```
gps_height
Mean: 669.0 Median: 370.0  Std: 693.0  Q25: 0.0  Q75: 1320.0

district_code
Mean: 6.0 Median: 3.0  Std: 10.0  Q25: 2.0  Q75: 5.0

population
Mean: 180.0 Median: 25.0  Std: 472.0  Q25: 0.0  Q75: 215.0

construction_year
Mean: 1301.0 Median: 1986.0  Std: 951.0  Q25: 0.0  Q75: 2004.0

recorded_year
Mean: 2012.0 Median: 2012.0  Std: 1.0  Q25: 2011.0  Q75: 2013.0

amount_tsh
Mean: 318.0 Median: 0.0  Std: 2998.0  Q25: 0.0  Q75: 20.0

longitude
Mean: 34.0 Median: 34.91031805  Std: 7.0  Q25: 33.0951871375  Q75: 37.179490449999996

latitude
Mean: -6.0 Median: -5.023822095  Std: 3.0  Q25: -8.54190396  Q75: -3.32691784
```

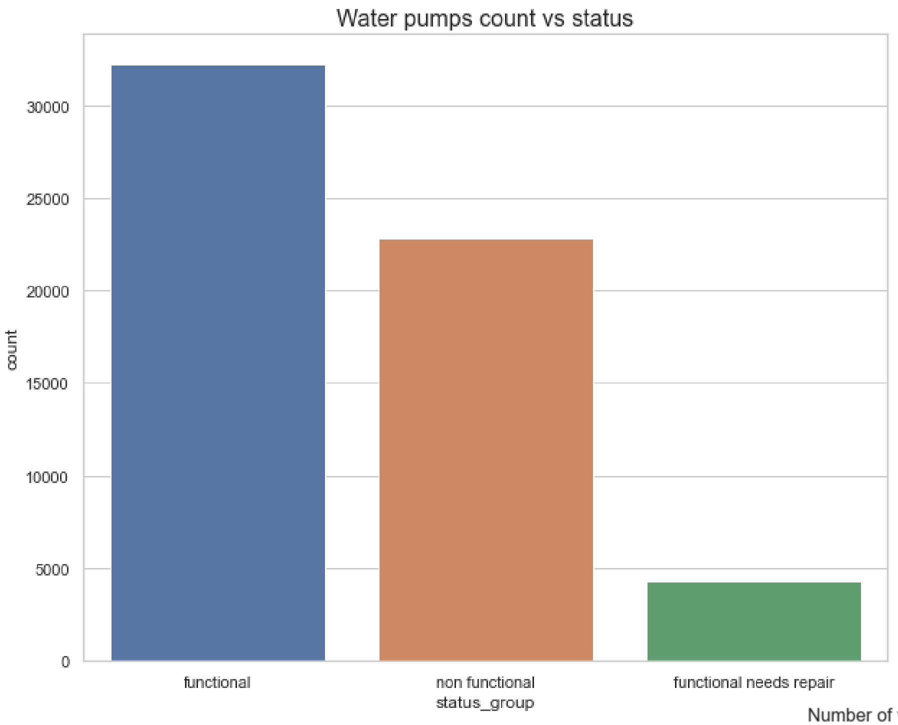
```
funder 1898
installer 2146
wpt_name 37400
basin 9
subvillage 19288
region 21
lga 125
ward 2092
scheme_management 13
extraction_type 18
extraction_type_group 13
extraction_type_class 7
management 12
payment 7
payment_type 7
quantity 5
quantity_group 5
source 10
source_type 7
waterpoint_type 7
waterpoint_type_group 6
status_group 3
```

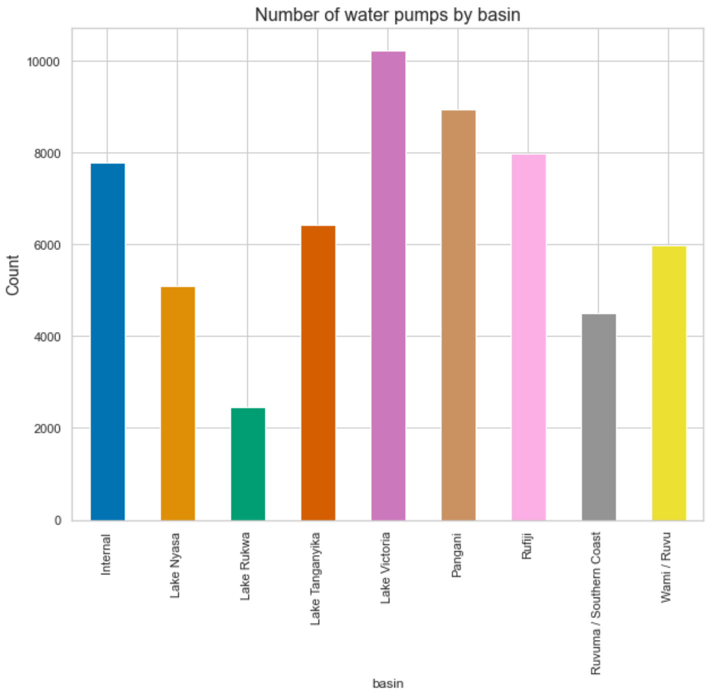
Methodology

- Data wrangling, clean-up and visualizations
- Using water pumps' status as the event, calculating the age of water pumps from construction year and recorded date (to be used as a timeline/time to event), both needed for survival analyses
- Applying various algorithms (Kaplan Meyer, Cox, Support Vector Machine and Random Forest Survival) to make predictions and compare the results using the C-index (a metric that has the same interpretation as AUC of ROC in classification problems)

Water pumps count

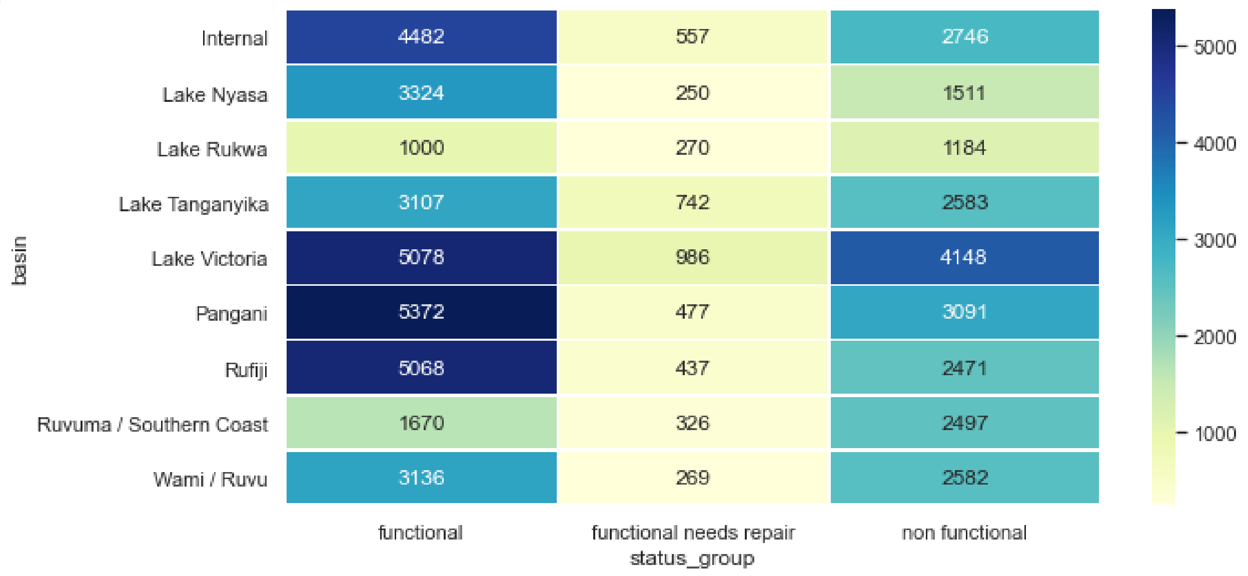
➢ The minority classes in the status group will be combined for the purpose of modeling

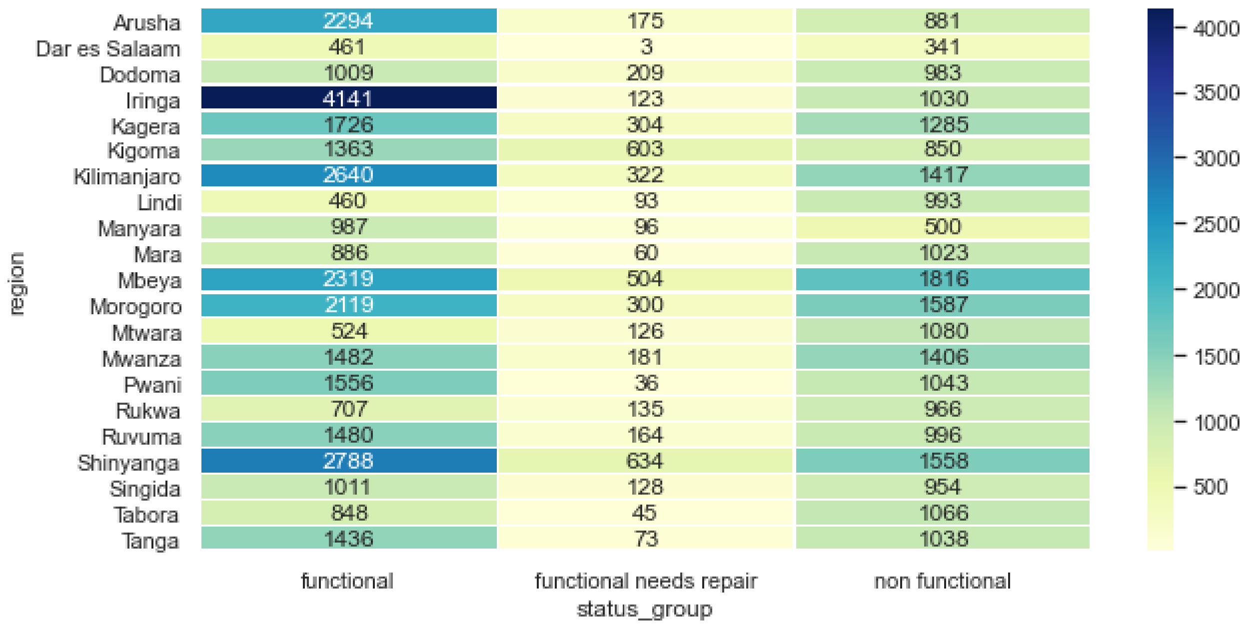




Water pumps functionality

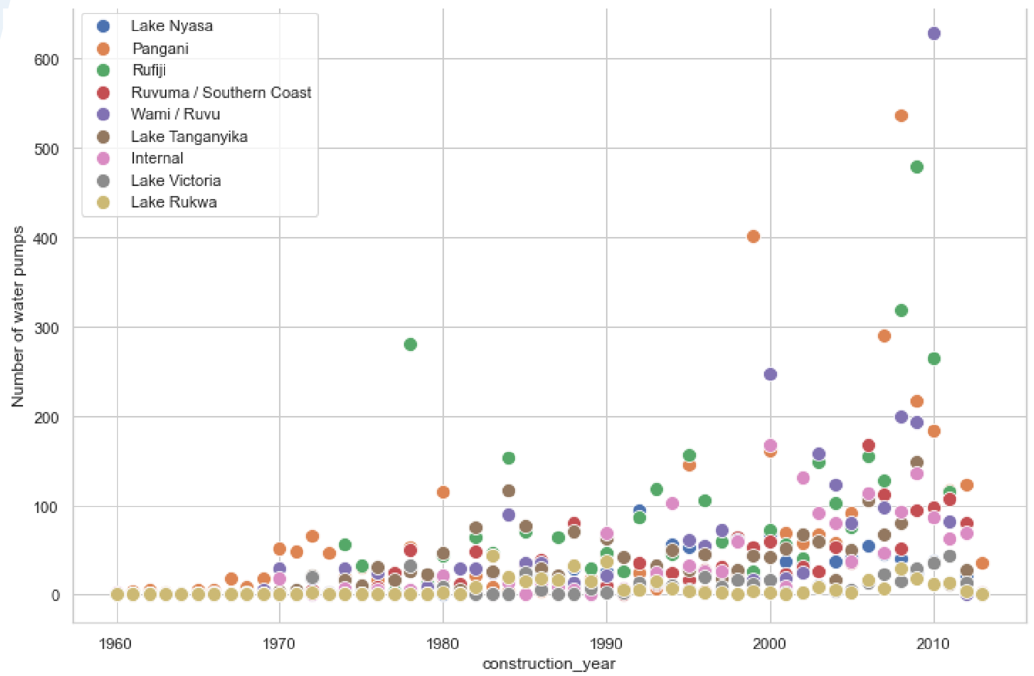
➢ The number of pumps in each group is similar across the various basins, while there's more variation if looking at regions

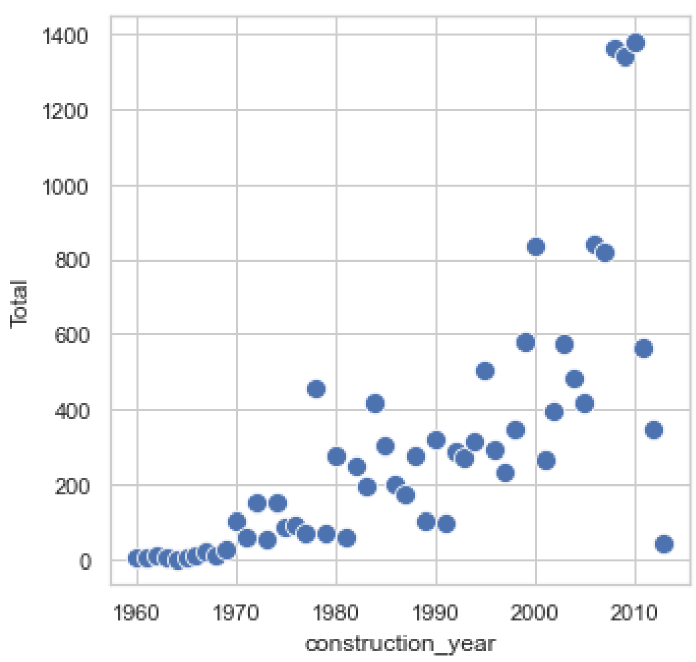


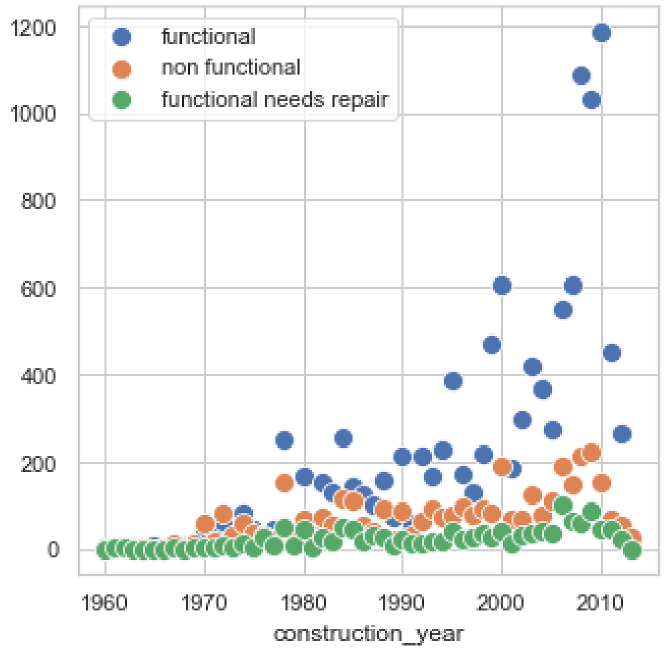


Water pumps over the years

- ➢ Some basins haven't seen a significant increase in the number of water pumps
- ➢ The number of pumps needing maintenance hasn't changed much

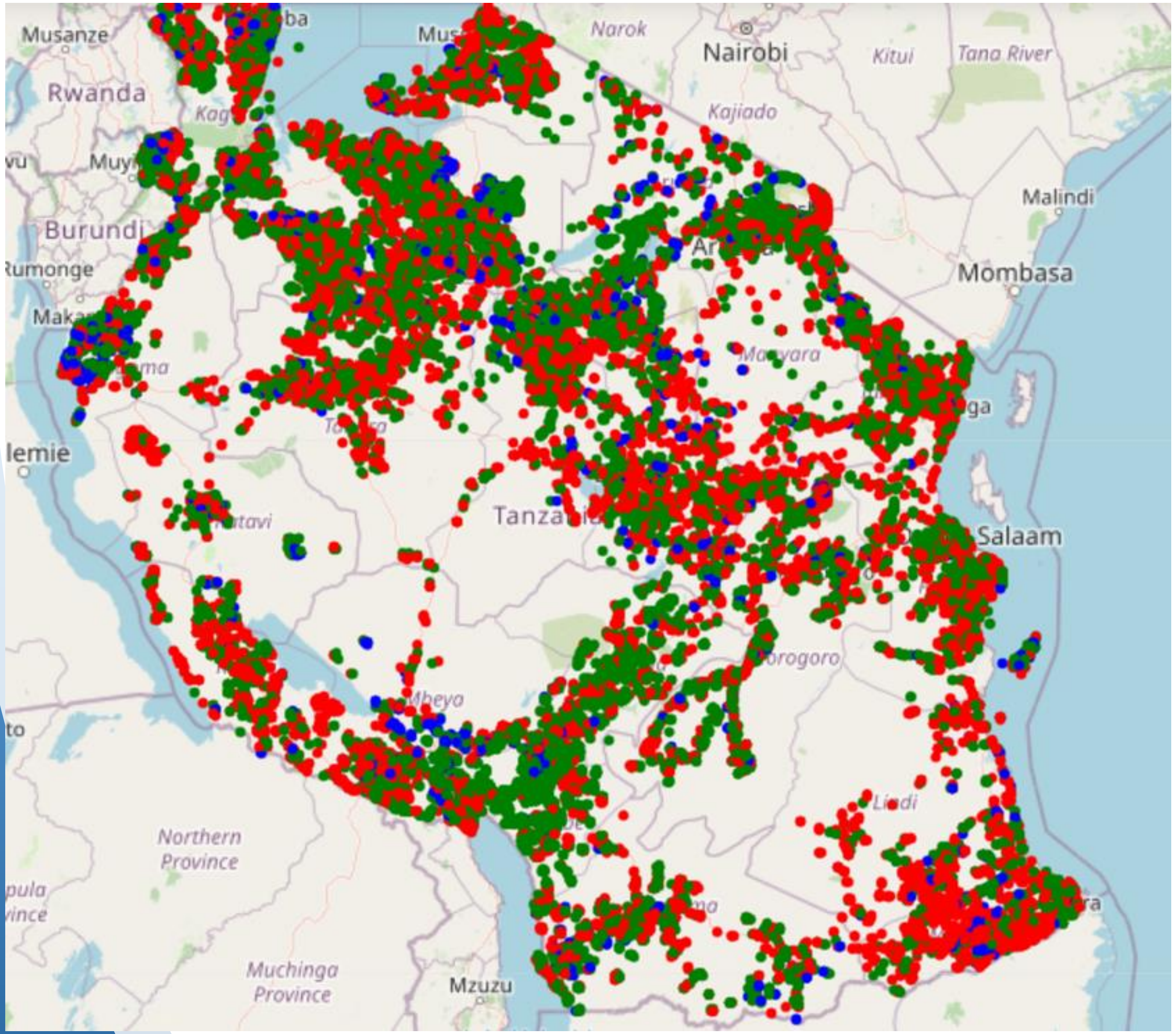






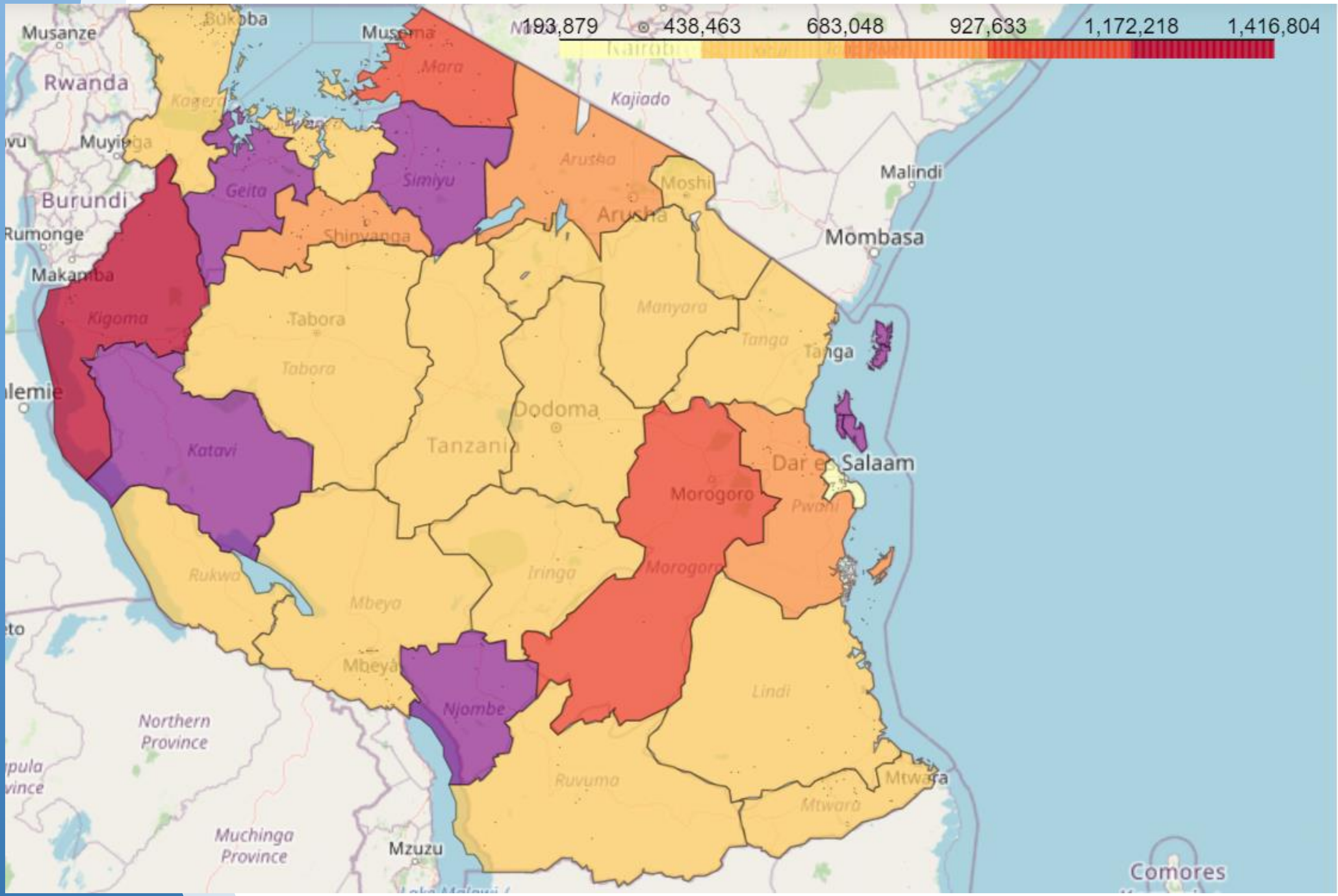
Water pumps by region

- Color coding corresponds to functionality



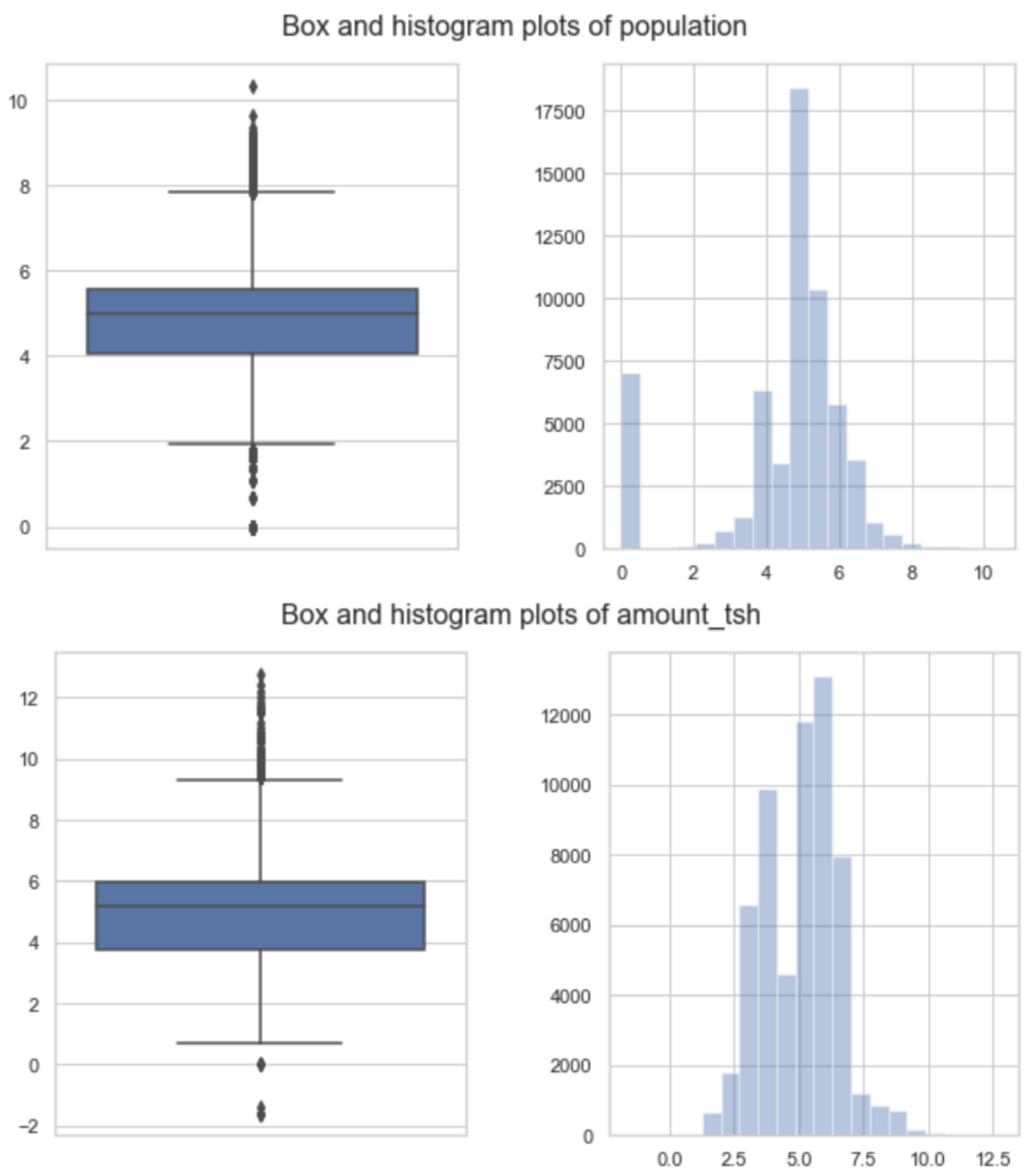
Population by region

- For observations with null population in the original data, the mean of the regions in the same basin has been calculated
- Regions in purple indicate no observation is available (not present in original data)



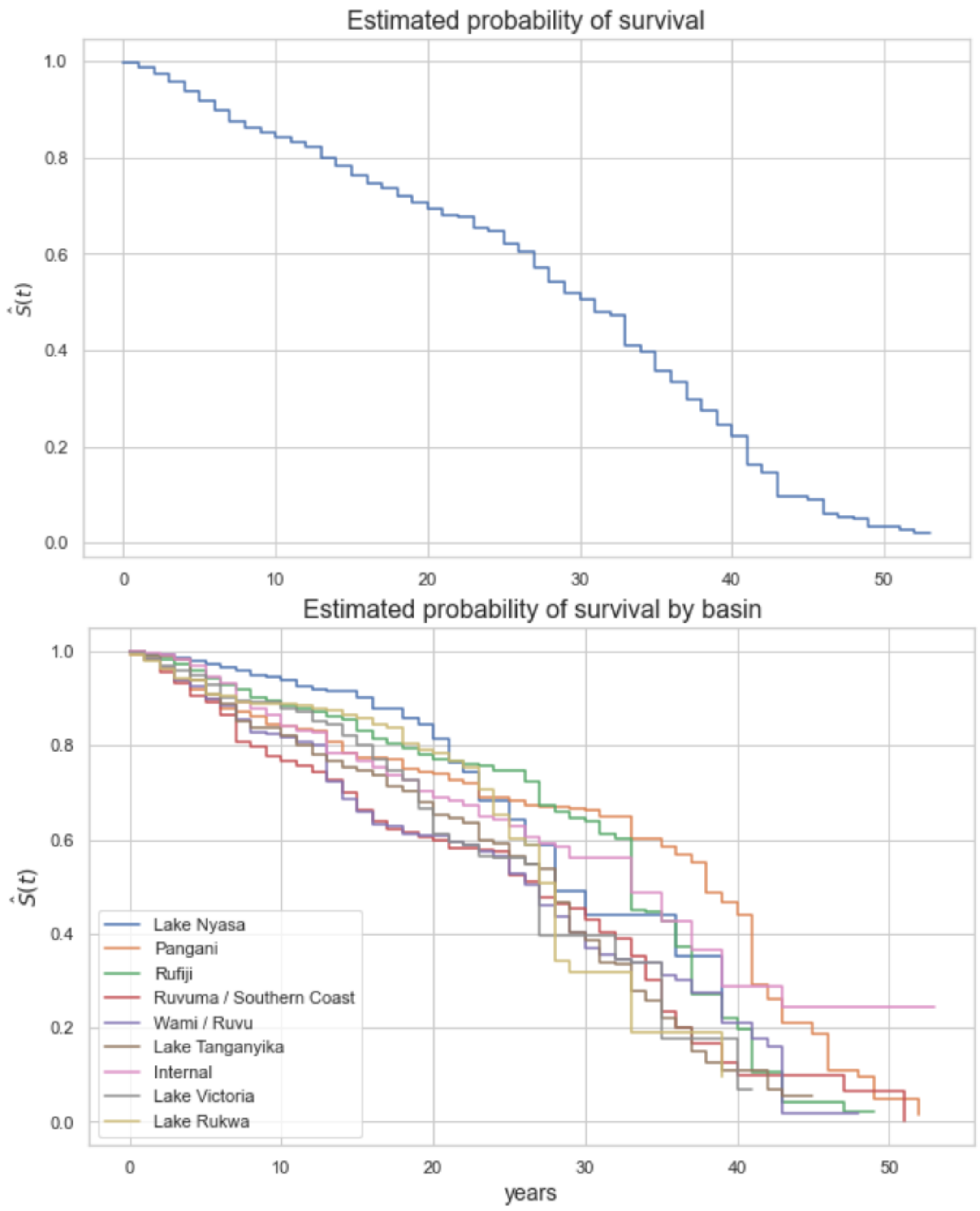
Population and amount of water

- Data is skewed
- Log scale is used in the plots



Kaplan Meier Estimate

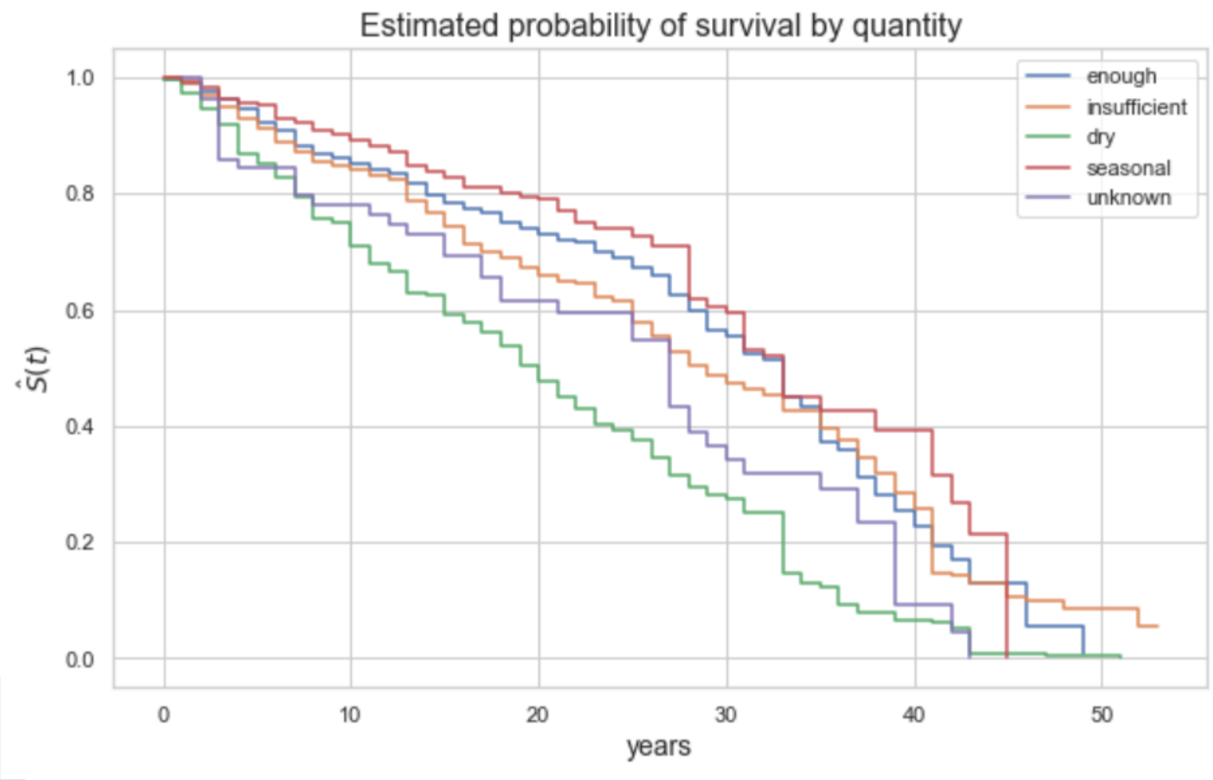
- Overall the rate of decrease is constant for about 25 years, then it becomes steeper and finally it flattens
- When looking at the single basins, the behavior is quite different



Kaplan Meier Estimate (cont.)

- Other features affect the curve, for example the water quantity

- Other models are considered



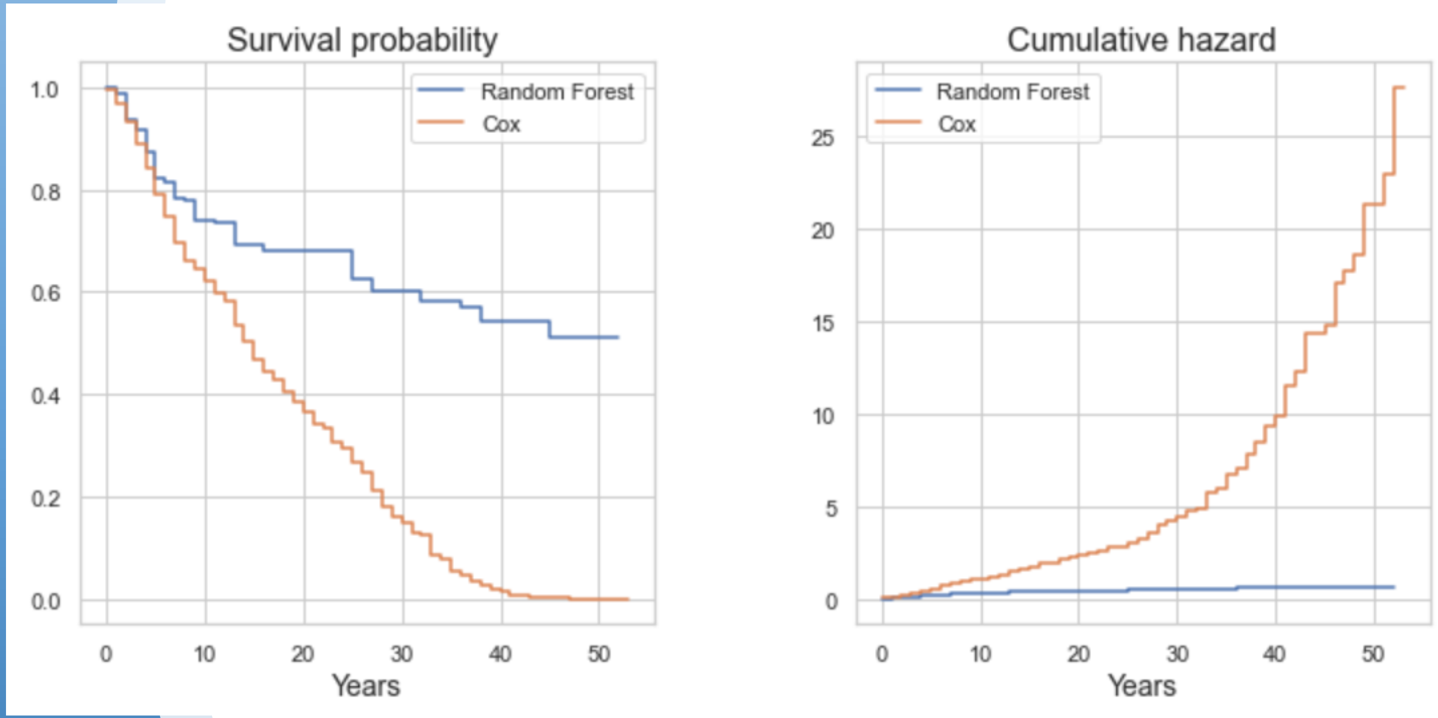
Models comparison

- Cox regression
- SVM
- Random Survival Forest
- C (concordance) index is similar to AUC - ROC

C- index			
	Cox	SMV	Random Forest
Training set	0.629	0.583	0.867
Test set	0.624	0.582	0.797

Prediction

- Select a water pump and plot survival probability and hazard for the Cox regression and Random Survival Forest model



Summary

- Random Survival Forest gave the best results in terms of concordance index
- Cox linear regression model can't learn the complexity of the data

Next Steps

- Look into feature selection
- Try SVM with non-linear kernel and other algorithms available in the scikit-survival package

References

1. https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/
2. https://github.com/aspds18/Springboard_capstone2
3. https://scikit-survival.readthedocs.io/en/latest/api.html
4. Pölsterl, S., Navab, N., and Katouzian, A., Fast Training of Support Vector Machines for Survival Analysis. Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, Lecture Notes in Computer Science, vol. 9285, pp. 243-259 (2015)

References

5. Pölsterl, S., Navab, N., and Katouzian, A., An Efficient Training Algorithm for Kernel Survival Support Vector Machines. 4th Workshop on Machine Learning in Life Sciences, 23 September 2016, Riva del Garda, Italy
6. Pölsterl, S., Gupta, P., Wang, L., Conjeti, S., Katouzian, A., and Navab, N., Heterogeneous ensembles for predicting survival of metastatic, castrate-resistant prostate cancer patients. F1000Research, vol. 5, no. 2676 (2016).