Water Quality Analysis and Prediction at Mtendeli

April 2018

Bachelor Thesis, University of Toronto

Image: Rohingya Refugee Camp. **Source**: https://news.un.org/en/story/2017/11/636832-unicef-warns-contaminated-drinking-water-camps-rohingya-refugees

Background

Location of Mtendeli Camp, Tanzania

Q. How do we know when water is still safe to drink?

Water tapstand at Nyarugusu camp

Research Gap

- Statistical models developed in below studies are inaccurate at times.
- Authors admit that model parameters can change considerably even for a small improvement in R^2.

Objective & Impact

- Address current research gaps.
  - Develop a more robust model
  - Develop a more accurate model in predicting water quality
    - Water quality is measured by chlorine concentration [mg/L]
- Accurate prediction of future chlorine concentration will significantly improve, or maintain, water security measures by controlling the onset of waterborne diseases (e.g. cholera).

Dataset

- Data is provided by a water specialist from Médecins Sans Frontières (MSF).
- Extremely small dataset
  - ~50 samples in total
- Scarce data points within each sample
  - Only two measurements in time

Only two measurements in a time-series-type data!

	A	C	D	E	Y	Z
1	Sample Number	Tap Stand ID	Time event 1	FRC (mg/l)	Time event 2	FRC (mg/l)
2	1	54	9:40	0.65	8:30	0.1
3	2	54	10:00	0.69	8:45	0.14
4	3	54	10:25	0.68	9:00	0.09
5	4	40	10:59	0.6	9:15	0.48
6	5	40	11:40	0.6	9:30	0.14
7	6	40	15:30	0.69	9:48	0.63

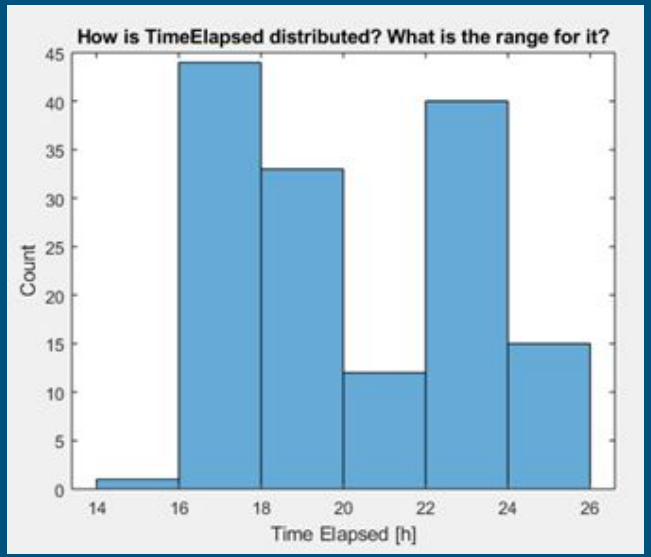
Foreseeable Challenges

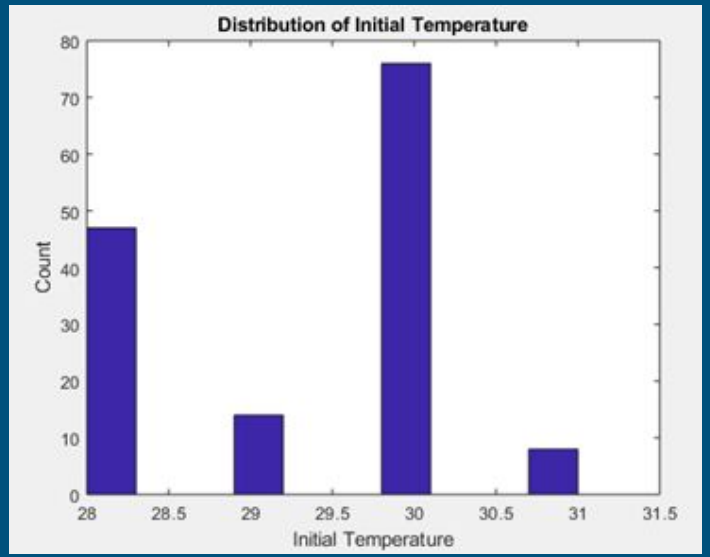
- Machine learning typically relies on having access to a large training data.
- Accurate (time series) modeling relies on many data points collected over time for *each* sample.

Both criteria NOT fulfilled in this project… Significant limitations are expected.

Exploratory Data Analysis

- Descriptive analytics -- some interesting facts



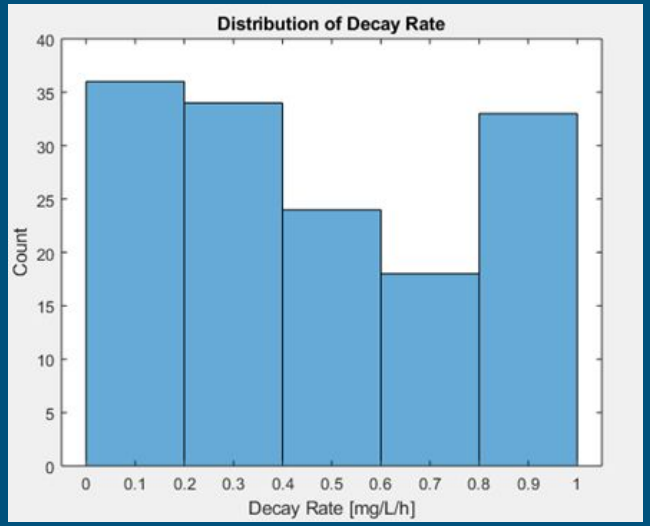


Exploratory Data Analysis

- Time elapsed between the initial measurement of water quality and the second varies widely.
  - Max: 26 hours
  - Min: 14 hours
  - Adds to the difficulty of the problem as the time of the second measurement is far from consistent!
- Most common initial temperature of a water sample at the refugee camp is 30 degree Celsius.
  - Max: 31 C
  - Min: 28 C
  - Difficult to say whether one degree difference will be significant or not.

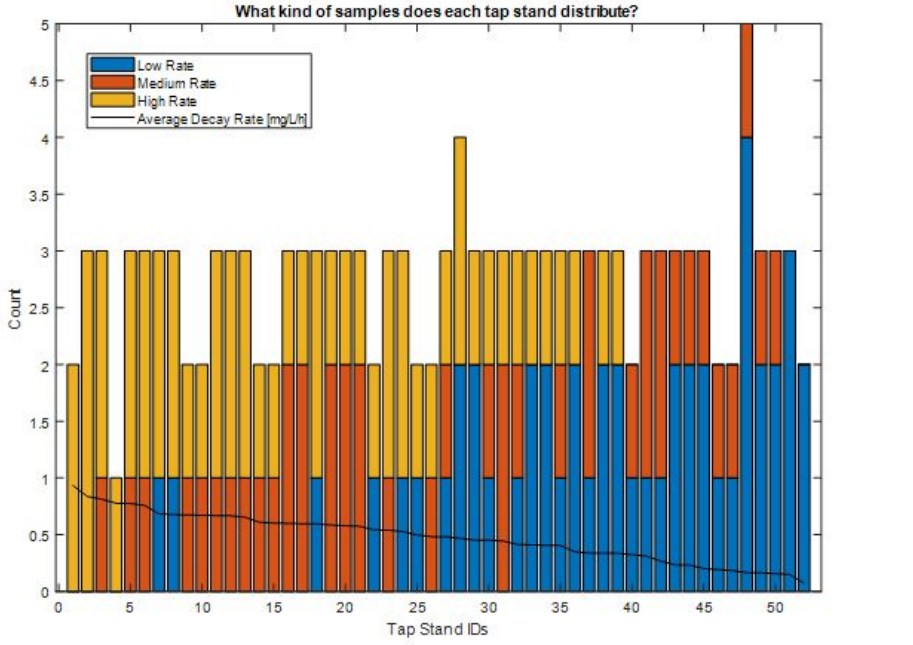
Data Analysis I

- Does knowing which tap stand the water came from an important information?
- Let us approximate decay rate using the only two measurements in time => linear



- Linear approximation is quite crude because we know water quality likely decays exponentially.
  - But this is the only option for us.
- **Decay rates vary significantly.**
- Possible suggestion that decay rate may depend on which tap stand the water came from.

Data Analysis I

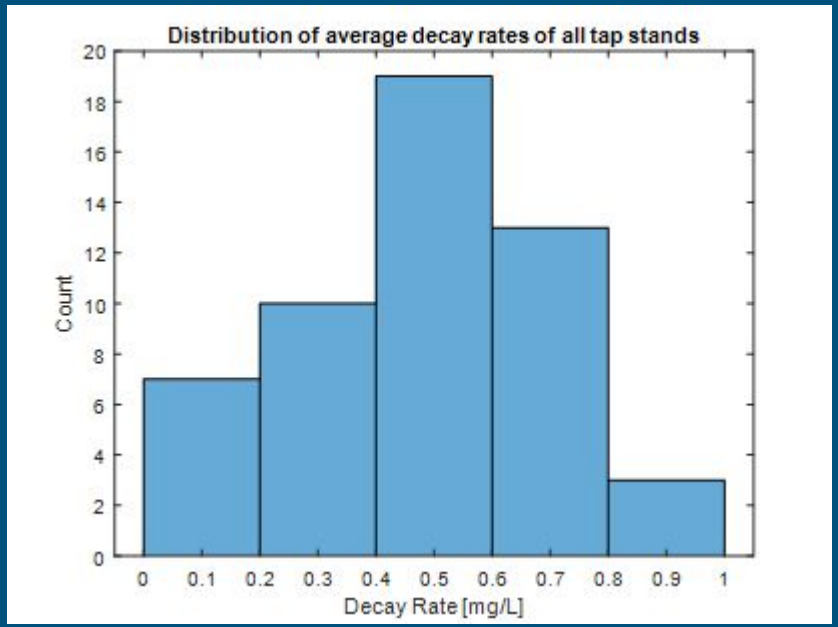


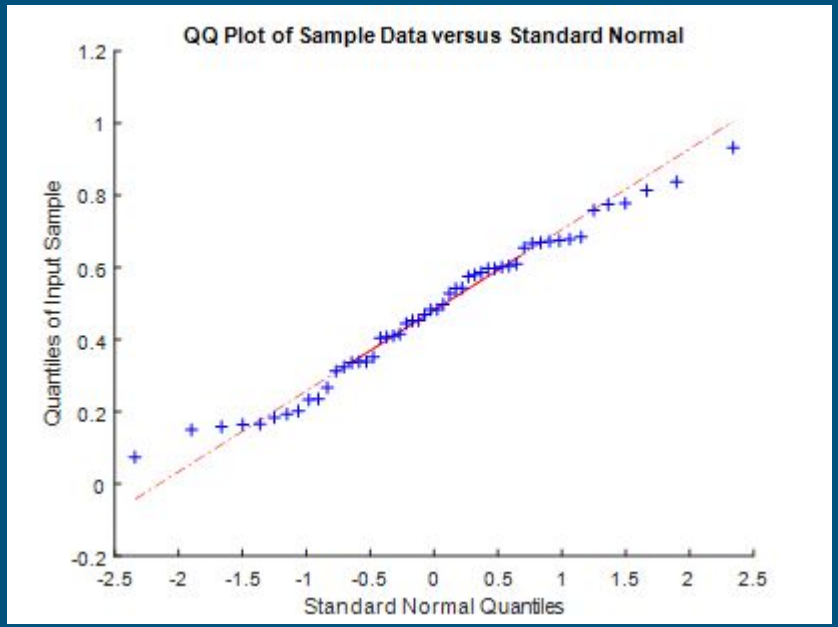
- 52 unique tap stands (*x*-axis)
- Each tap stand has anywhere from one to five samples in the dataset.
- When organized in decreasing order of “average decay rate”, the result shows:
  - Most samples labeled “high-decay” are concentrated on the left 33%.
  - Most samples labeled “low-decay” are concentrated on the right 33%.

Data Analysis I

- The previous result is significant because it tells us that **different tap stands may distribute samples with largely varying decay rates, and that a tap stand may consistently distribute samples of one decay rate category**.
  - Therefore, knowing something about a tap stand can help us **predict the decay rate** of a water sample collected from it.

Data Analysis I





Data Analysis I

- Average decay rate of each tap stand is approx. normally distributed.
- Mean = 0.48 mg/L/h
- Standard deviation = 0.21 mg/L/h
- **Coefficient of variation = 44%**
- Conclusion: there is significant dispersion in the data.
    - **Which tap stand a water sample came from is important to know** in predicting its future water quality.

Data Analysis II: Clustering

- Since different tap stands exhibit different decay rates, let us **cluster** the tap stands into **three different classes of decay rates**.
  - Low-decay: water samples that exhibit slow water quality decay.
  - Medium-decay
  - High-decay
- Alternatively, five classes could be used instead to be more precise.

Data Analysis II: Clustering

- **Motivation:** we should be able to cluster similar tap stands together in terms of how fast water quality decays when it came from these tap stands.
- Two well known clustering algorithms are below. Any can be pursued.
  - K-means clustering
  - Hierarchical clustering
- Objective: minimize intra-cluster distance while maximizing inter-cluster distance.
- *Cophenetic clustering coefficient* = “correlation between dissimilarity between two clusters and their inter-cluster distance” [1, 2].
  - Higher the better

Data Analysis II: Clustering

Evaluation Criteria	Result of Clustering		
	Cluster 1	Cluster 2	Cluster 3
Number of samples	49	43	53
Range of time elapsed [h]	8.4	8.5	8.8
Mean initial chlorine [mg/L]	0.7	0.8	0.7
Mean decay rate [mg/L/h]	-0.33	-0.46	-0.61

- Hypothesis testing (equal medians, equal means) indicates that all three clusters are statistically different at the 5% significance level.
- Shows some promise that I could use existing features to categorize all tap stands into distinct decay rate classes.

Data Analysis II: Clustering

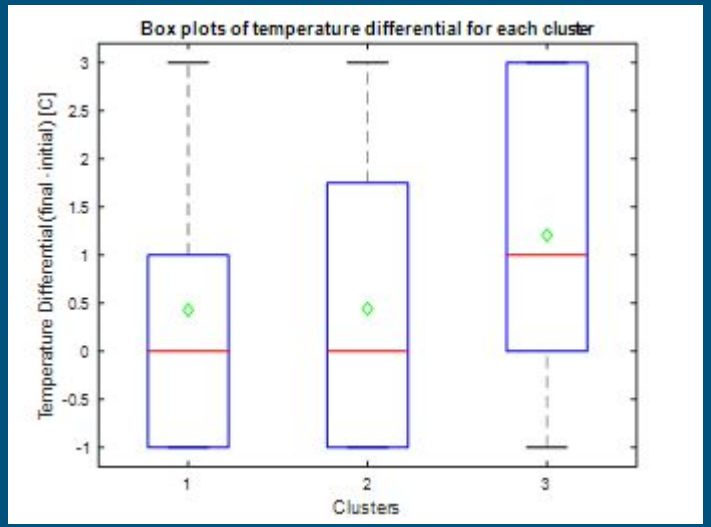
- So what have I shown?
- Water samples can potentially be grouped into distinct categories because which tap stand it came from may help me predict how fast its quality will decay.
- If water samples can be precisely clustered in this way, prediction of future water quality will be done in each category separately, and this will be much more accurate.

Data Analysis II: Clustering

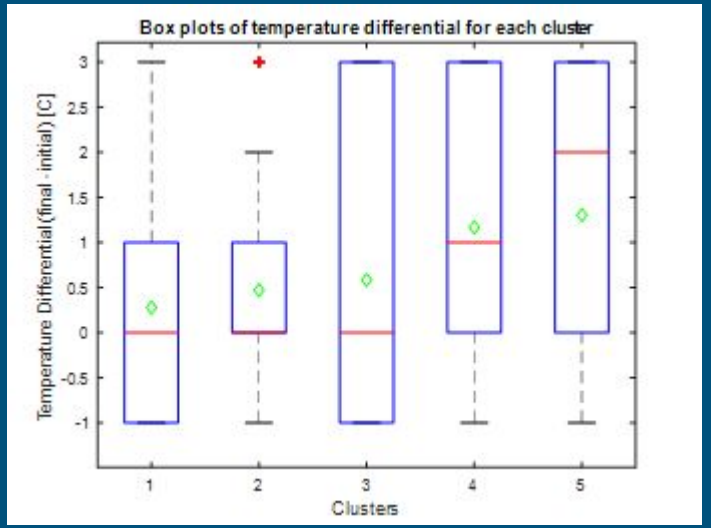
- How can clustering be improved (produce more distinct clusters)?
- In short, new features need to be introduced:
  - Information on tap stands (age, level of dirtiness, etc.)
  - Information on the day's weather patterns (how temperature changed throughout the day)
  - More informative measure of the cleanness of the water itself (total organic carbon)
  - Information on hours of exposure to sunlight.
    - Sunlight is known to accelerate water quality decay

Data Analysis III

- Temperature differential (initial - final temperatures) is positively correlated with decay rate.



Three-cluster approach



Five-cluster approach

Data Analysis IV

- Container type has important information regarding decay rates.
- “Jerry cans” are found more often in the low-decay cluster, implying that they may provide better protection against chlorine decay than “buckets”.

3-Cluster Approach	Low-decay Cluster	Medium-decay Cluster	High-decay Cluster
Jerry Can : Bucket Ratio	3.1	1.3	0.6

5-Cluster Approach	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Jerry Can : Bucket Ratio	3.5	1.6	1.2	0.8	0.4

Data Analysis V

- Method of drawing water has important information regarding decay rates.
- “Pouring out” are found more often in the low-decay cluster, implying that they may provide better protection against chlorine decay than “dipping glass”.

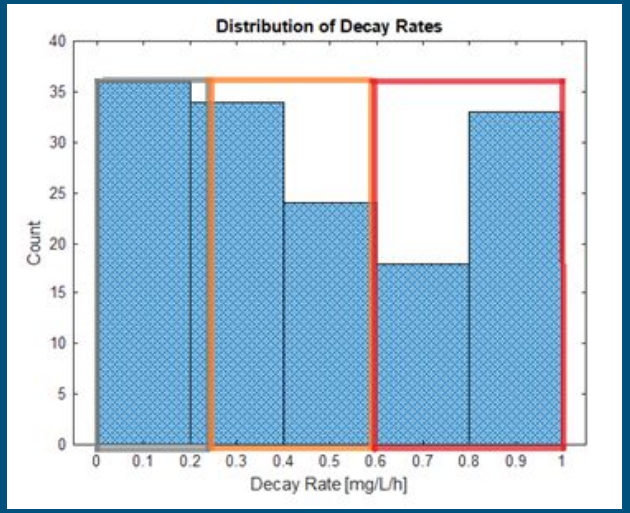
3-Cluster Approach	Low-decay Cluster	Medium-decay Cluster	High-decay Cluster
Pour Out : Dip Glass Ratio	3.5	1.4	0.7

5-Cluster Approach	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Pour Out : Dip Glass Ratio	3.5	2.4	1	0.8	0.6

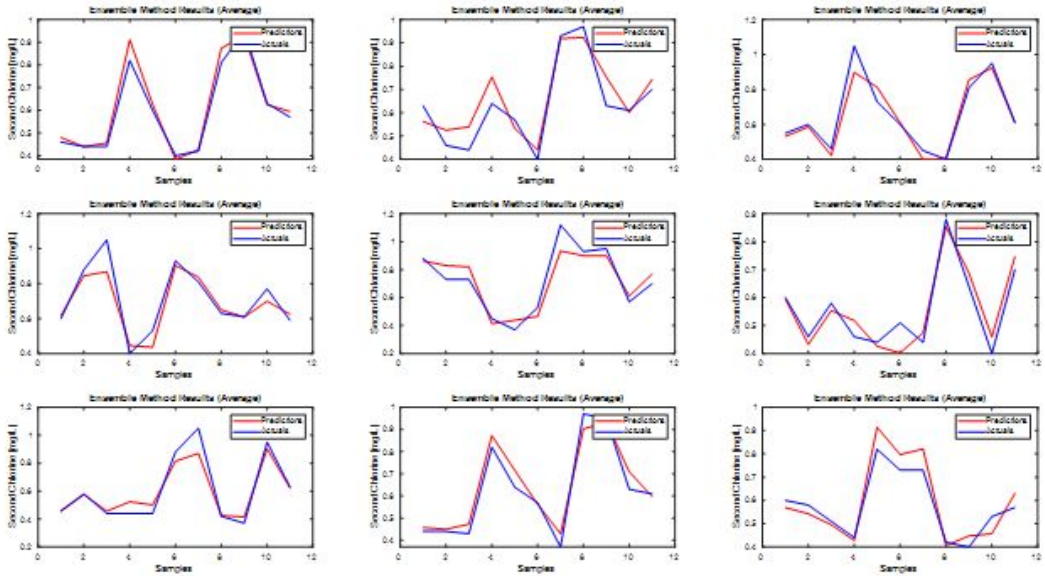
Regression

- Assume water samples have been “perfectly” clustered.
- Let us now predict water quality in each cluster.

Evaluation Criteria	“Perfect” Clustering			Result of Clustering		
	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
Number of samples	49	43	53	49	43	53
Range of time elapsed [h]	8.0	9.1	9.9	8.4	8.5	8.8
Mean initial chlorine [mg/L]	0.7	0.7	0.7	0.7	0.8	0.7
Mean decay rate [mg/L/h]	-0.14	-0.40	-0.84	-0.33	-0.46	-0.61

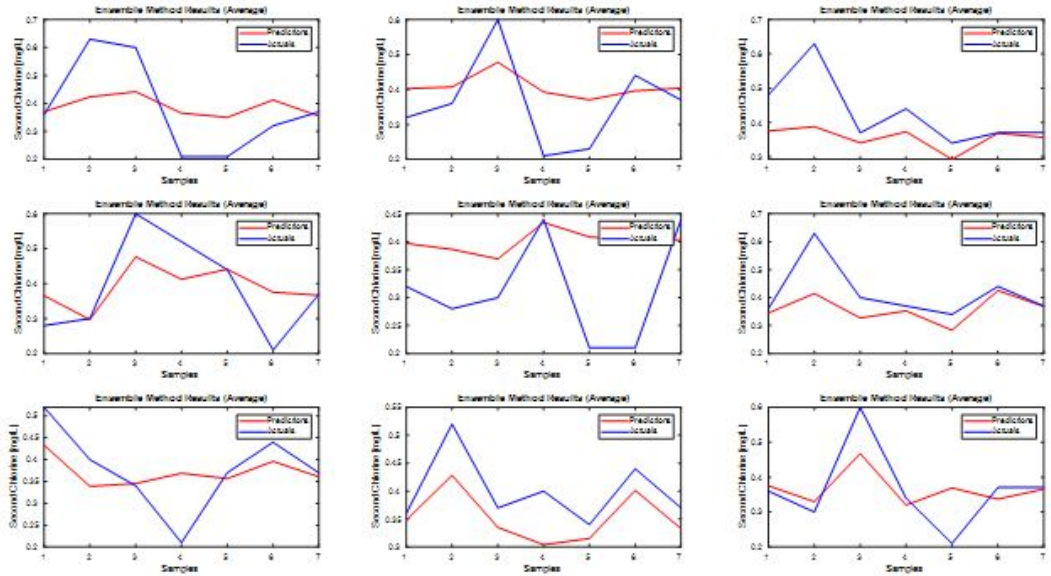


Regression: “Low-Decay” Cluster



- Average R^2 = 0.88
- Red = predictions
- Blue = Actual

Regression: “High-Decay” Cluster



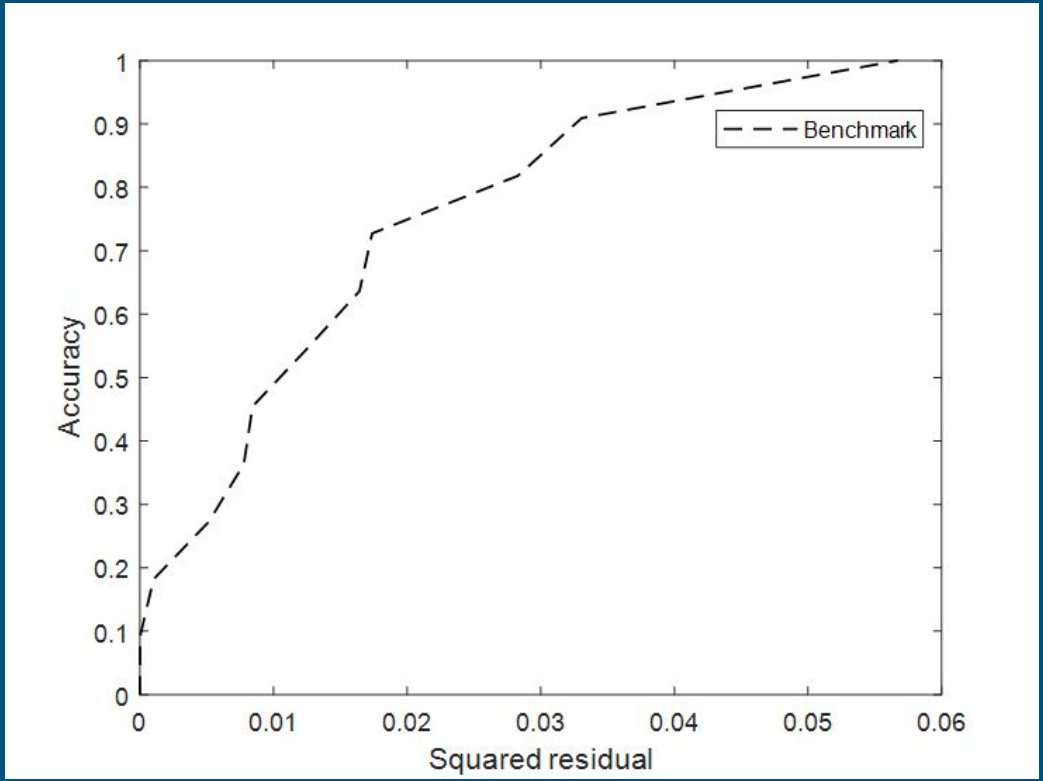
- Average R^2 = 0.08
- Red = predictions
- Blue = Actual

Regression

- Why choose R^2 as a measure of performance?
- R^2 measure the “degree to which predictions are correlated with actuals”.
  - As the type of the data in question is time-series, measuring the correlation between predicted and actual water quality seems reasonable.
- How is the model selected?
  - **Regression error characteristic curve** (RECC): measure of the accuracy of each machine learning algorithm given increasing squared error tolerance. If the prediction lies within this tolerance, then the prediction is assumed to be correct.

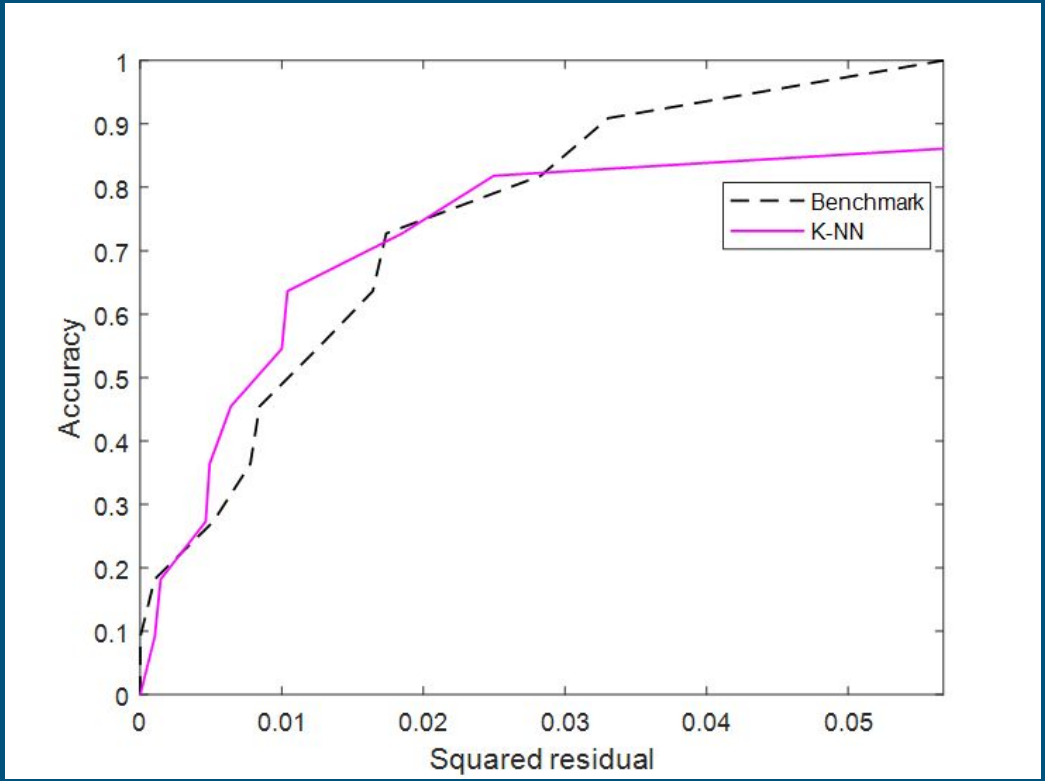
Regression: Model Selection

Benchmark: simple average line (constant)



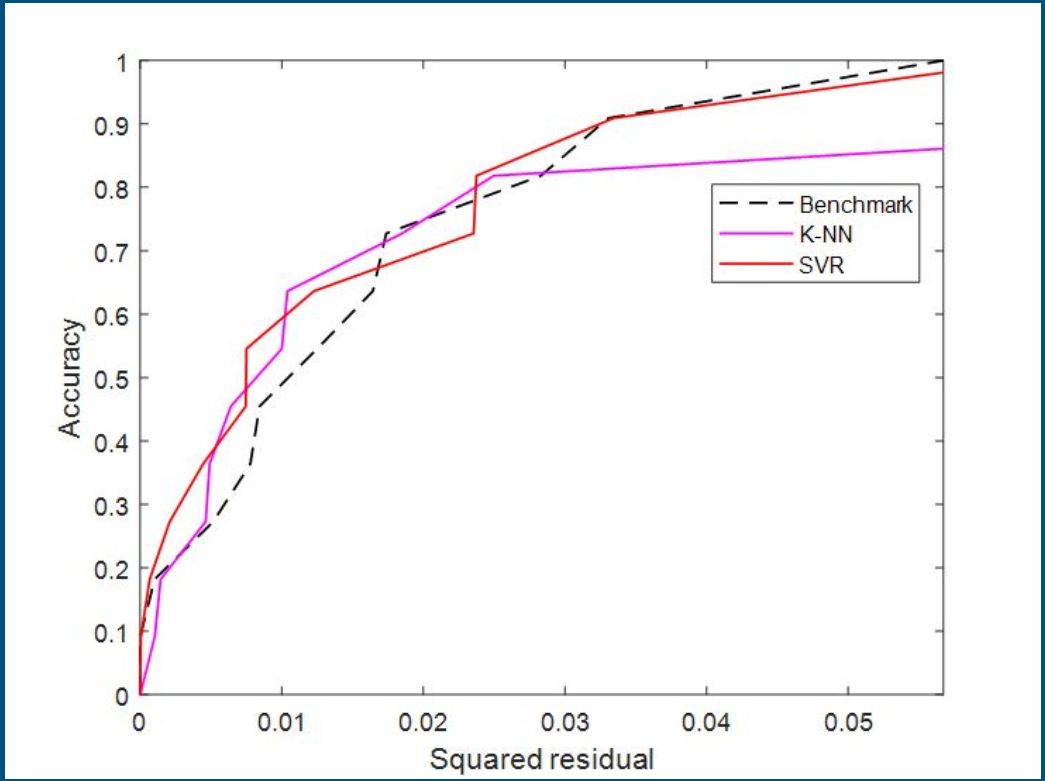
Regression: Model Selection

Benchmark: simple average line (constant)
KNN: K-nearest neighbors



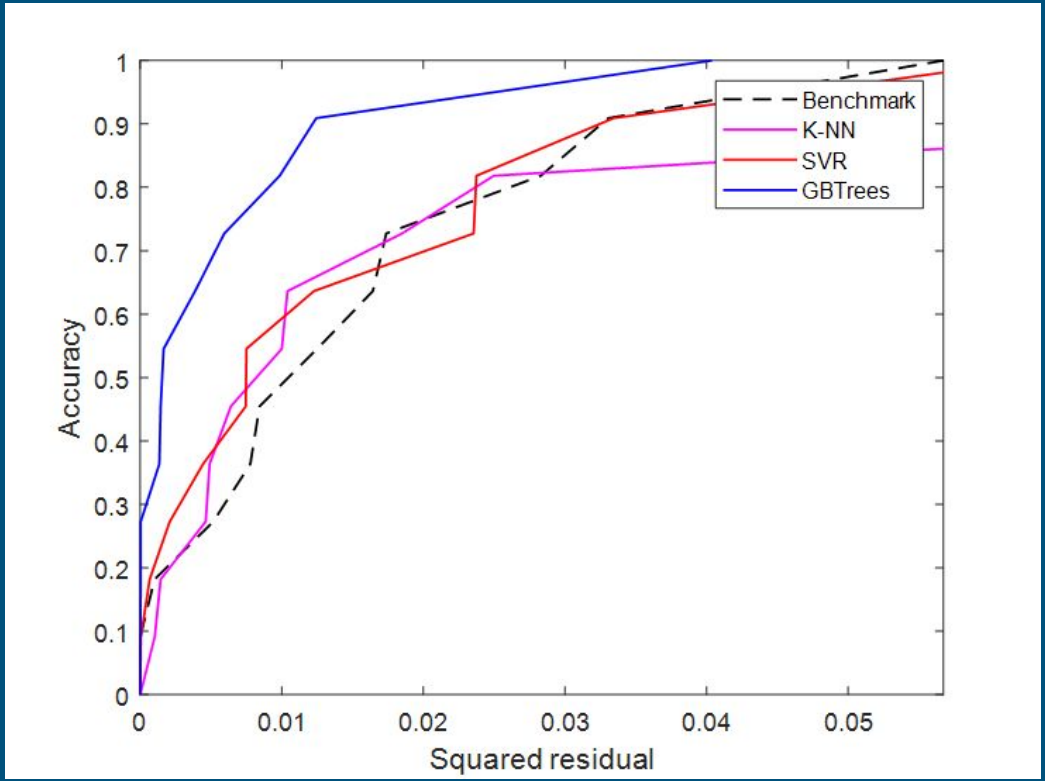
Regression: Model Selection

Benchmark: simple average line (constant)
KNN: K-nearest neighbors
SVR: support vector regression



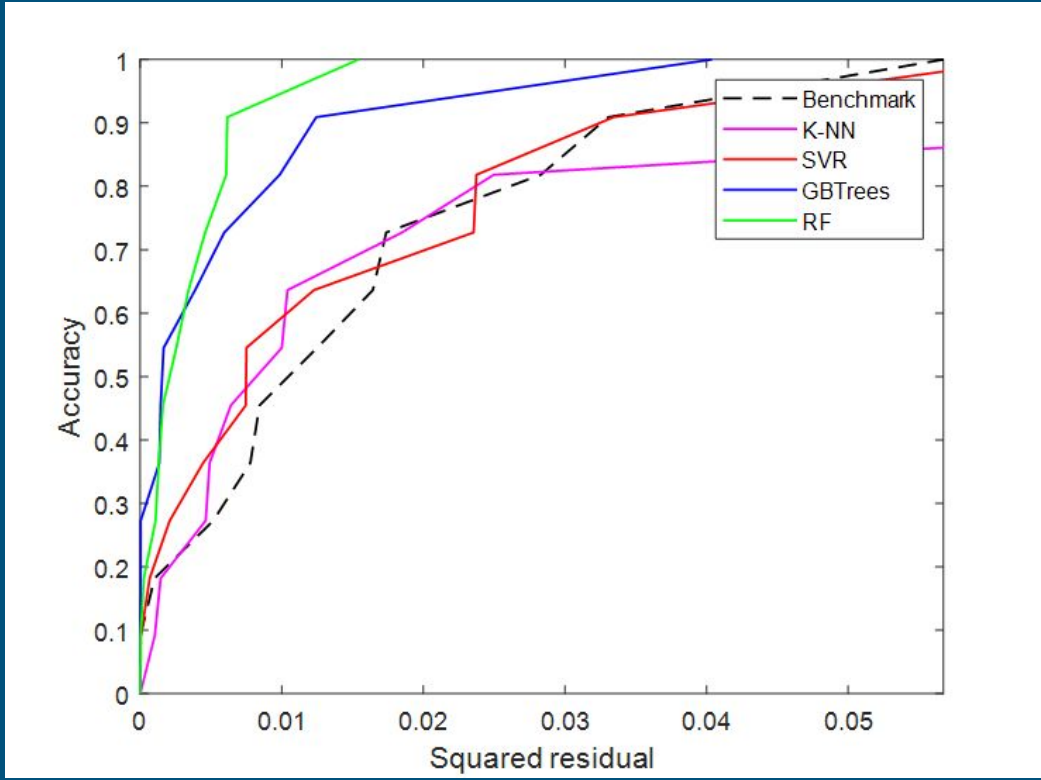
Regression: Model Selection

Benchmark: simple average line (constant)
KNN: K-nearest neighbors
SVR: support vector regression
GBTrees: gradient boosted trees



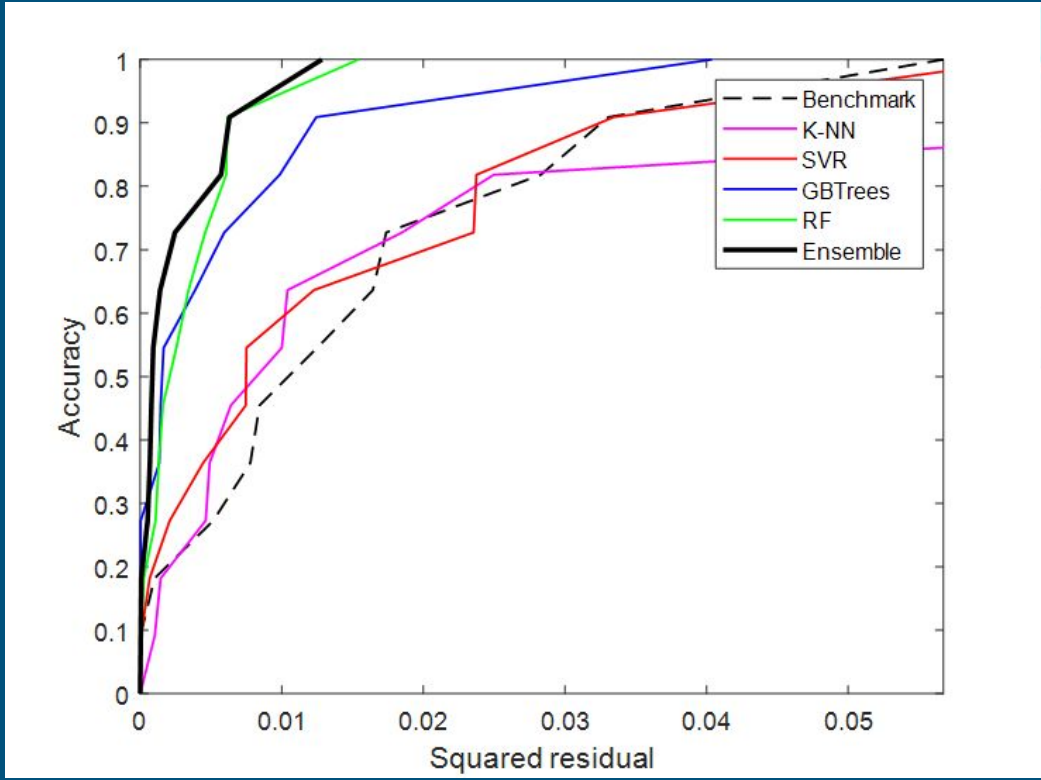
Regression: Model Selection

Benchmark: simple average line (constant)
KNN: K-nearest neighbors
SVR: support vector regression
GBTrees: gradient boosted trees
RF: random forest



Regression: Model Selection

Benchmark: simple average line (constant)
KNN: K-nearest neighbors
SVR: support vector regression
GBTrees: gradient boosted trees
RF: random forest
Ensemble: average of GBTrees and RF



Method	R^2
K-NN	-0.20
SVR	-0.08
GBTrees	+0.82
RF	+0.87
Ensemble	**+0.88**

Regression: Model Improvements

Adjusting the model parameters in machine learning algorithms

K-nearest neighbors (K-NN): varying values of K. Generally, K=2 or K=3 works best based on several trials, but tends to change frequently on every trial due to different partitioning of data.

Random forest and gradient-boosted trees: varying the number of trees. Any more than 100 trees do not result in any significant improvement.

Support vector regression: using different kernel functions. Generally, linear or gaussian functions work best.

Model training/validation

5-fold cross-validation: partitioning the training set into 5 equal portions and using each one at a time as a validation set to avoid overfitting and generalize the model. Improved R^2 by an average of 0.35 over 100 trials for the high-decay cluster.

Lasso regularization: minimizes least squares between predictions and actuals by forcing as many beta coefficients of the features to zero as possible with an extra regularization term to address overfitting and highly correlated features. After cross-validation, this improved R^2 by an average of 0.06 over 100 trials for the high-decay cluster.

Regression: Model Improvements

Feature engineering

Normalization of continuous features (aka. feature scaling): by normalizing data, an improvement in R^2 by 0.3 was observed in the high-decay cluster.

Principal component analysis (PCA) for dimension reduction:

- Defined new features (aka. “principal components”) and projected the data onto a lower dimensional space which still retained 99% of the variability within the data.
- Improved R2 by an average of 0.04 for the high-decay cluster on top of all approaches.

The R^2 for the high-decay cluster after all these approaches is 0.08 (average).

Future Improvements

There are two large areas of improvement.

1. Quality of clustering
    a. Data for new features (tap stand conditions, temperature differential, TOC, hours of direct sunlight), which may also help with improving the accuracy of machine learning since they are related to decay rate.
2. Accuracy of regression model, particularly for the high-decay cluster
    a. Size of training set is too small (~50 samples, 35 of which are for training) which may affect accuracy.
    b. Larger data set will make machine learning results more robust.
    c. Changes in data collection may also improve machine learning performance.
        i. Min. time elapsed recorded is 15.8 h in the current data set, but decay rates may be more accurately captured if a sample is recorded earlier on.

References

[1] S. Saraçli, N. Doğan and İ. Doğan, "Comparison of hierarchical cluster analysis methods by cophenetic correlation", Journal of Inequalities and Applications, vol. 2013, no. 1, p. 8, 2013.

[2] B. Uragun and R. Rajan, "The discrimination of interaural level difference sensitivity functions: development of a taxonomic data template for modelling", BMC Neuroscience, vol. 14, no. 1, p. 19, 2013.

[3] L. Yee, Md. Abdullah, S. Ata, and B. Ishak, “Dissolved organic matter and its impact on the chlorine demand of treated water”, The Malaysian Journal of Analytical Sciences, vol. 10, no. 2, p. 243-250, 2006.