Progress Presentation:
Predicting Chlorine Measurements in Mtendeli, Tanzania

Name: Jangwon Park

Supervisor: Chi-Guhn Lee

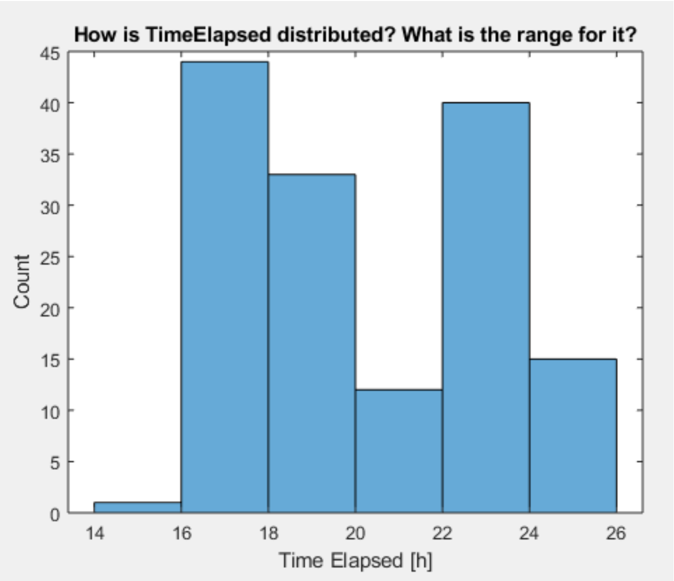
Date: March 26, 2018

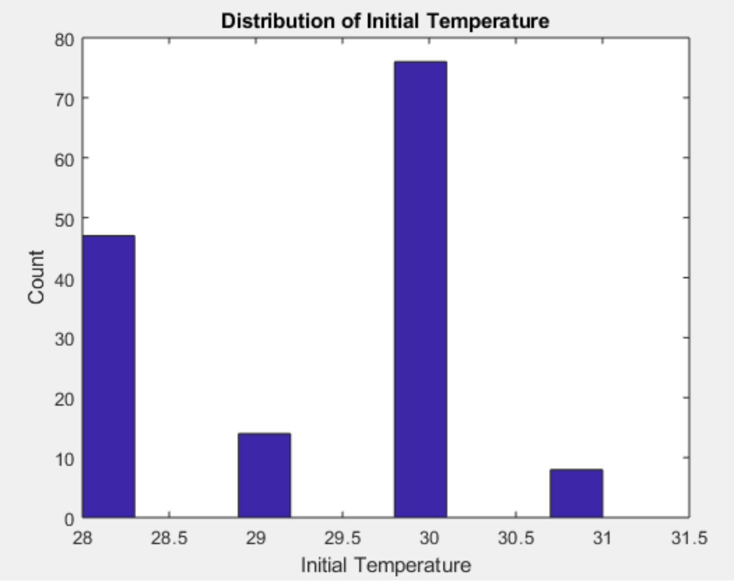
Agenda

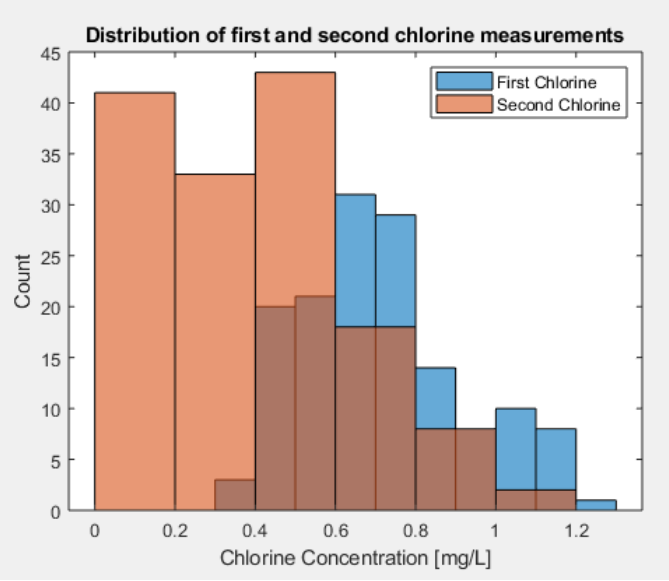
1. Results from exploratory data analysis (EDA)
2. Clustering algorithm (motivation, results)
3. Conjectures derived from cluster-specific data analysis and predictions
4. Future improvements

EDA 1: Distributions

Assumption: all samples in the data set can be described by the same model (decay curve)

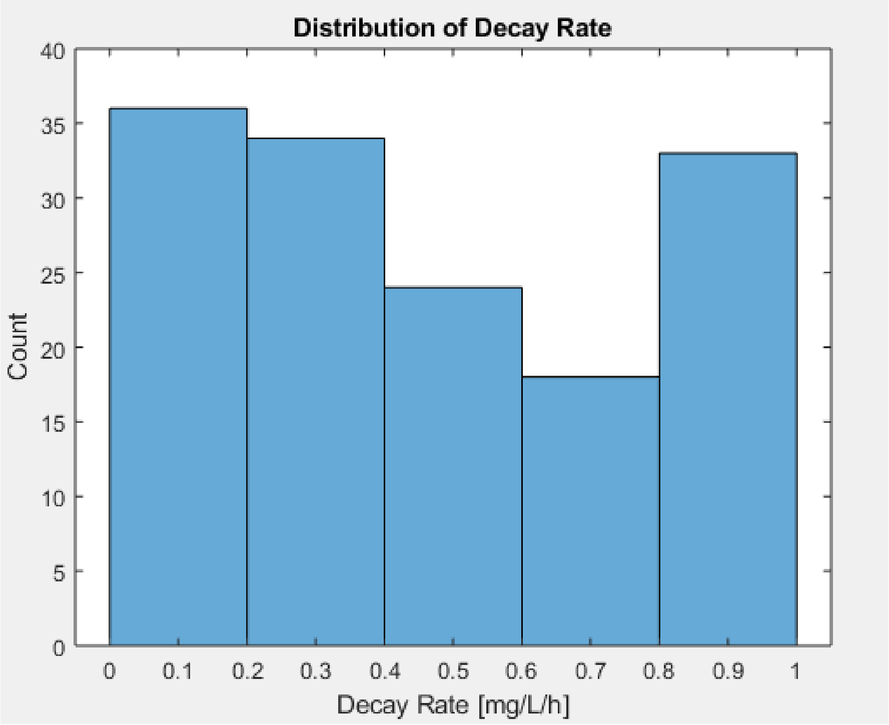






	Time Elapsed [h]	Initial Temperature [C]	Initial Chlorine [mg/L]	Final Chlorine [mg/L]
Min	15.8	28	1.27	1.12
Max	25.9	31	0.32	0

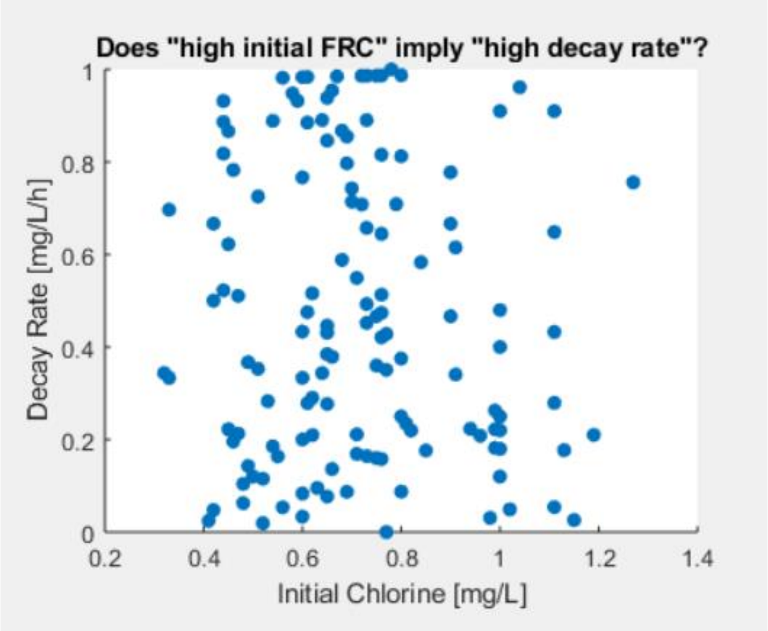
EDA 1: Distributions Cont’d.



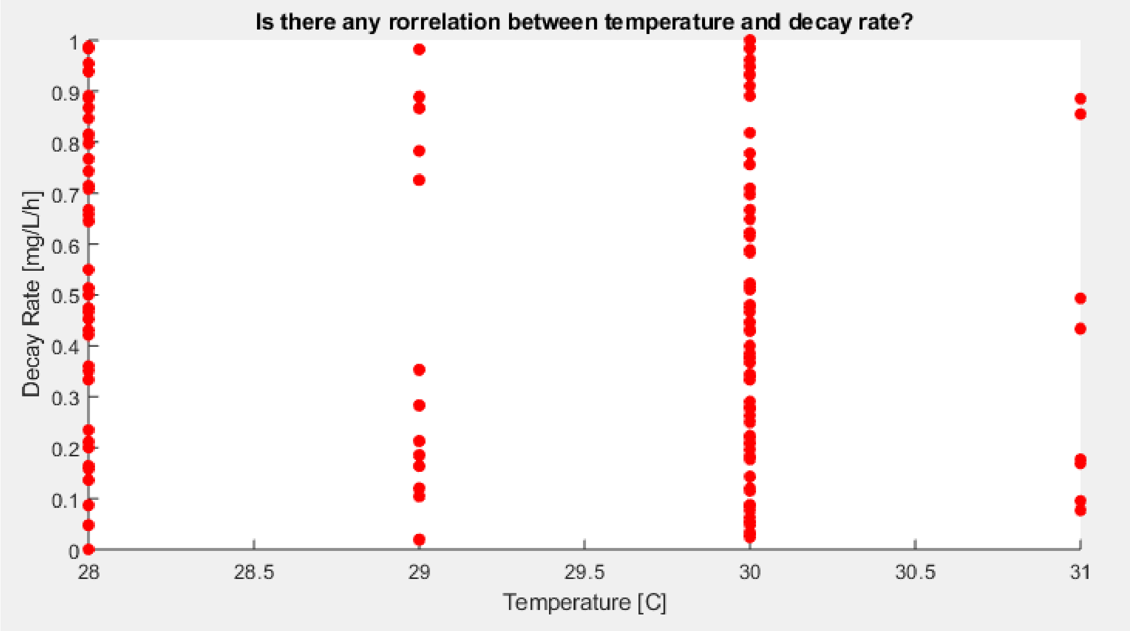
- Decay rates are **linearly** approximated
- Min: 0
- Max: 1
- Higher number indicates higher decay rate
- Decay rate is (arguably) the most important feature to look at because:
  - One of the parameters that make a model unique
  - It appears in all orders of reaction (decay constant)
  - Conceptually, it is what we are trying to predict (i.e. predict future chlorine concentration at 12-hour post-distribution mark)

EDA 2: Correlations

Q. Is there a correlation between initial chlorine measurements and decay rates?



Q. Is there a correlation between temperature and decay rates?

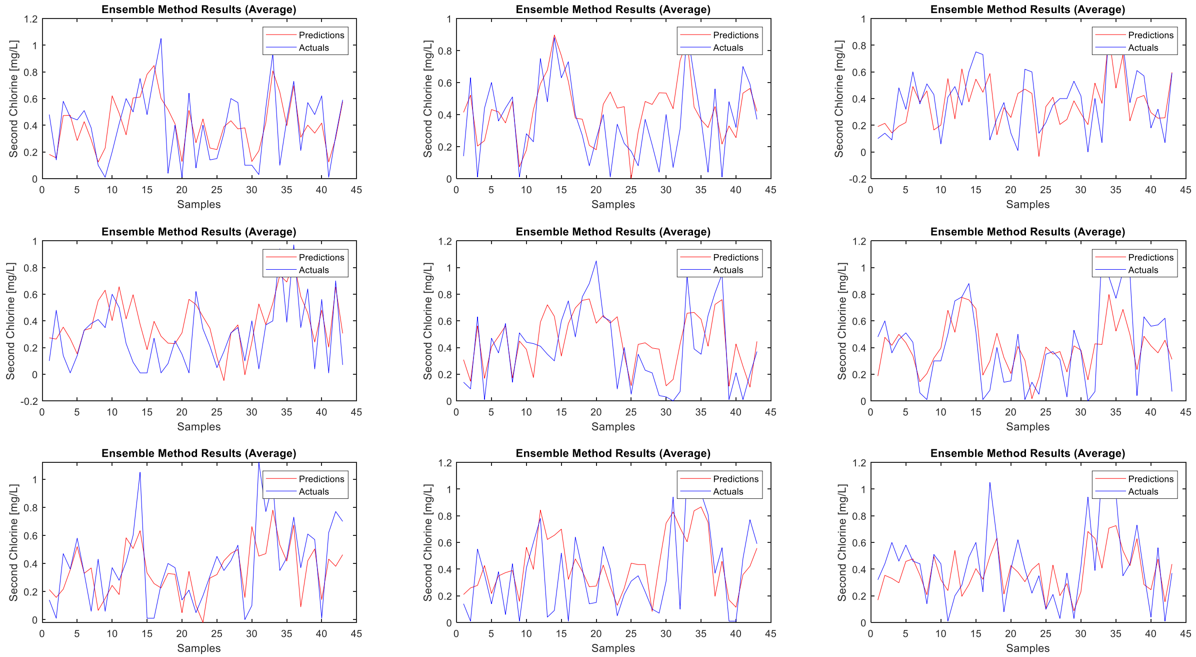


Conclusion: There are no intuitive correlations among the current data set

EDA 3: Machine Learning Predictions

- Predictive machine learning algorithms were applied to predict the final chlorine measurements regardless to evaluate its current performance.

- The key measure of performance is R^2.



- Average $R^2 = 0.4$
- Red: predictions
- Blue: actuals

- Conclusion: machine learning algorithms cannot predict final chlorine measurements accurately

Clustering Algorithm

- **Motivation:** the previous assumption appears to be wrong and the data set contains samples that are inherently dissimilar, which are meant to be analyzed separately. Clustering algorithms can help us group samples that behave similarly.

Well-known clustering approaches:

<u>1. K-means clustering</u>

- Randomly initializes *k* clusters and place them optimally through many iterations by minimizing the distance from every point to its nearest centroid.

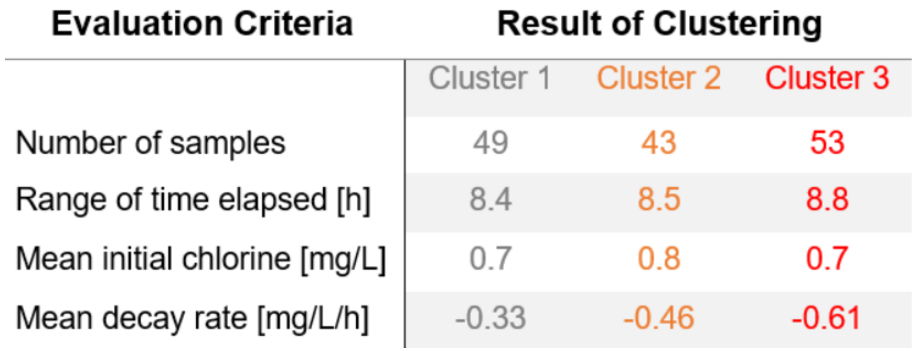
<u>2. Hierarchical clustering</u>

- Assumes each observation is a cluster on its own, and sequentially and hierarchically merges closest clusters together until there are only *k* clusters left.

- **Conclusion:** use hierarchical clustering to find the optimal metric-method combination

Clustering Algorithm Cont'd.

- Optimal method: "Weighted"
- Optimal metric: "Squared Euclidean"
- Optimality is evaluated based on the average of two measures: **cophenetic correlation coefficient** (CCC) and **cluster size variance**. CCC is intuitively understood as the *correlation between the dissimilarity between two clusters and their inter-cluster distance* [1,2].



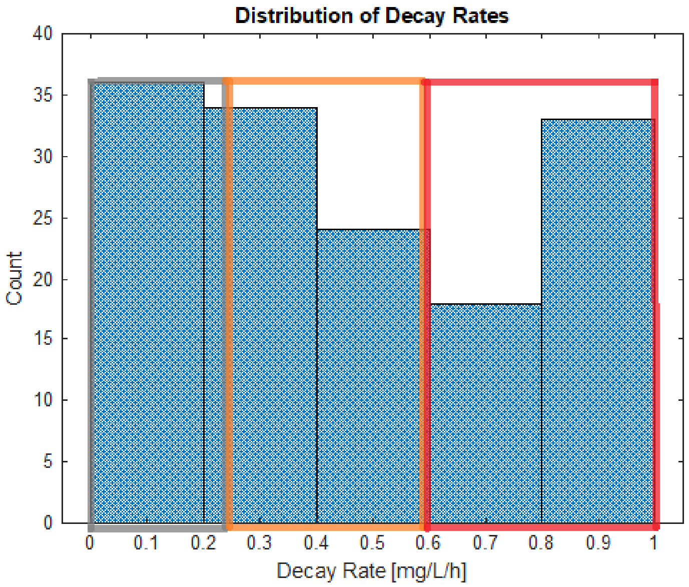
Evaluation Criteria	Result of Clustering		
	Cluster 1	Cluster 2	Cluster 3
Number of samples	49	43	53
Range of time elapsed [h]	8.4	8.5	8.8
Mean initial chlorine [mg/L]	0.7	0.8	0.7
Mean decay rate [mg/L/h]	-0.33	-0.46	-0.61

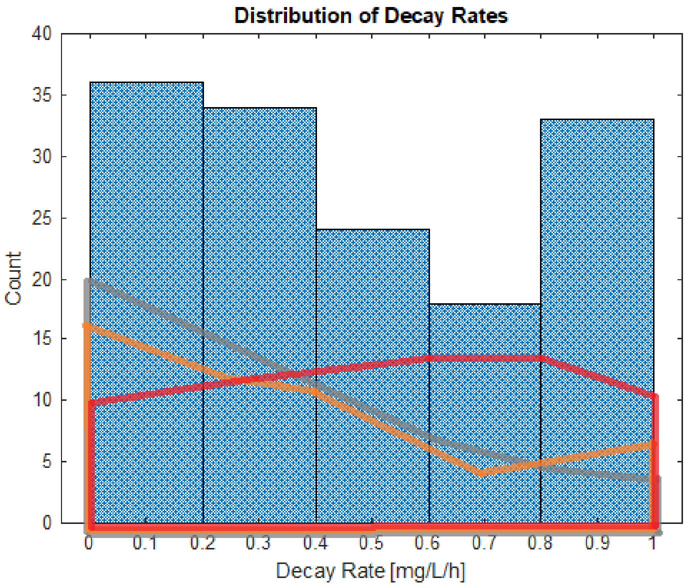
- While other features are not considerably different from each other, **"mean decay rate" appears to be the most distinct nature** that distinguishes a cluster from another.
- Hypothesis testing (equal medians, equal means) indicates that all three clusters are statistically different based on decay rate at the 5% significance level.

Clustering Algorithm Cont'd.

- Comparison to "perfect" clustering where the clustering algorithm is applied solely on the decay rate feature:

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





*This is a rough sketch done manually for visualization purposes only

Clustering Algorithm Cont'd.

Conclusion: by hypothesis testing and visualizing the clusters, there exists some promise that clustering algorithms can group samples according to their distinct decay rate category i.e. low-, medium-, and high-decay clusters.

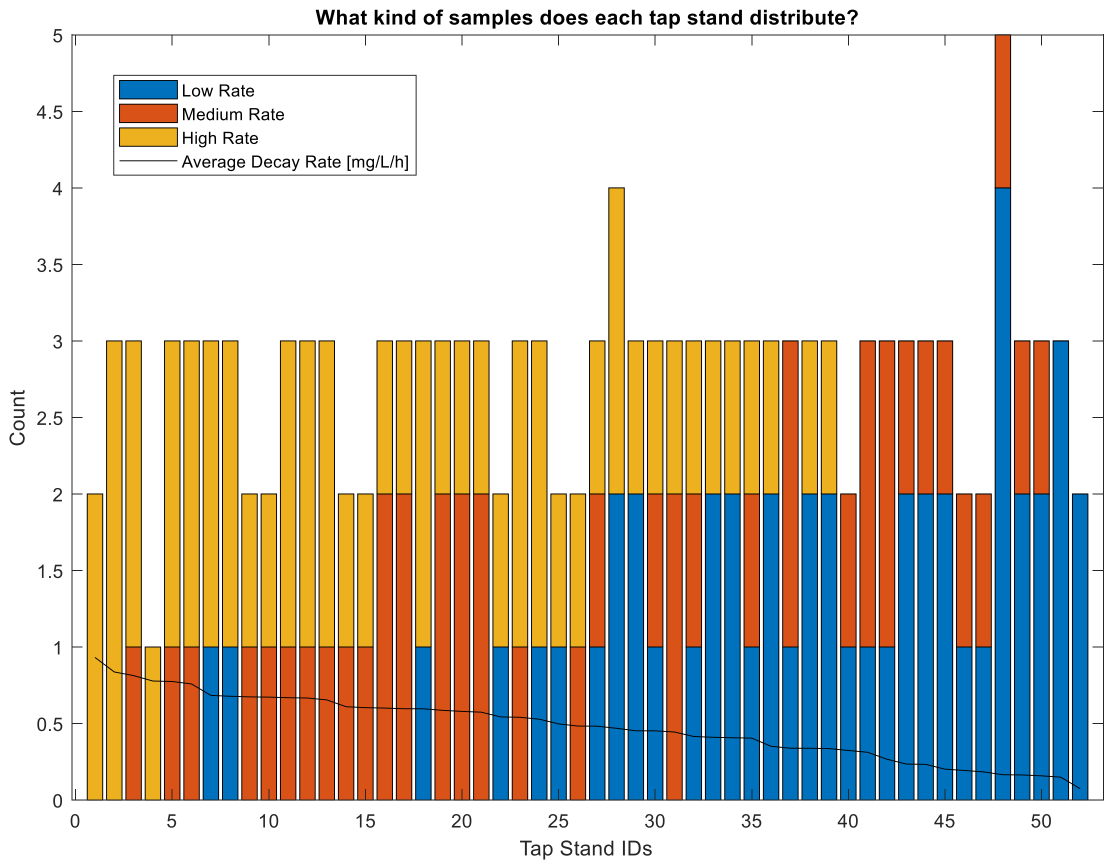
However, significant improvements are required for precise clustering by incorporating new features, which are outlined below and discussed in detail in the next few slides:

- Information on the tap stands
- Information on temperature differentials throughout a day
- Information on the cleanness of the water itself e.g. total organic carbon
- Information on hours of exposure to direct sunlight

Conjectures

- For now, it is assumed that the data set was clustered precisely according to three decay rate categories – *low-*, *medium-*, and *high-decay* clusters.
- Conjectures are made by analyzing each cluster separately and comparing the results.
- There isn’t sufficient data to fully validate the conjectures, but they still help us gain some insight.

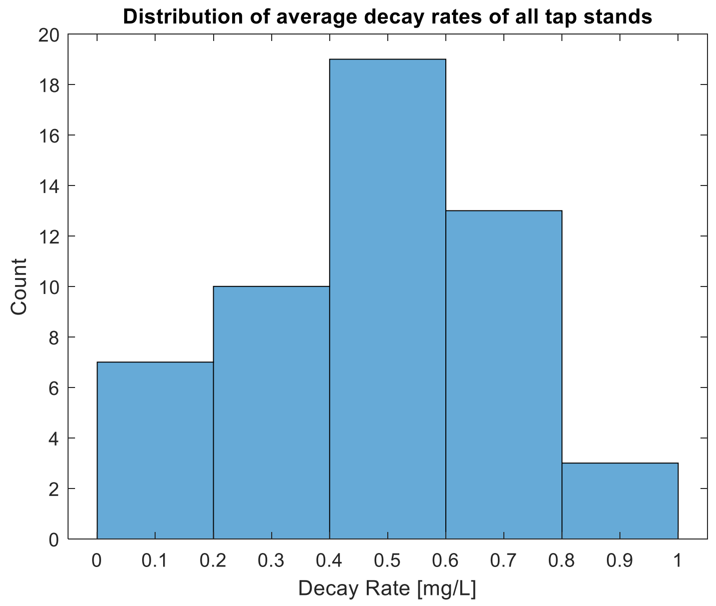
Conjecture 1: Tap stands tend to distribute water samples of similar decay rates

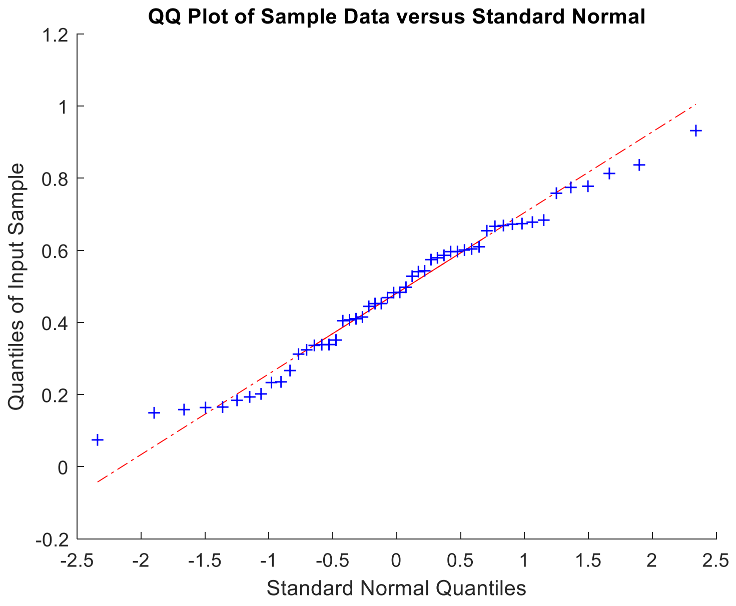


- 52 unique tap stands.
- Each tap stand has anywhere from one to five samples in the data set.
- When organized in *decreasing* order of "average decay rate", the result shows that:
  - Most samples labeled as "high-decay" are concentrated on the left 33%.
  - Most samples labeled as "low-decay" are concentrated on the right 33%.

Conjecture 1: Tap stands tend to distribute water samples of similar decay rates

- 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 (which improves clustering AND machine learning prediction accuracy).



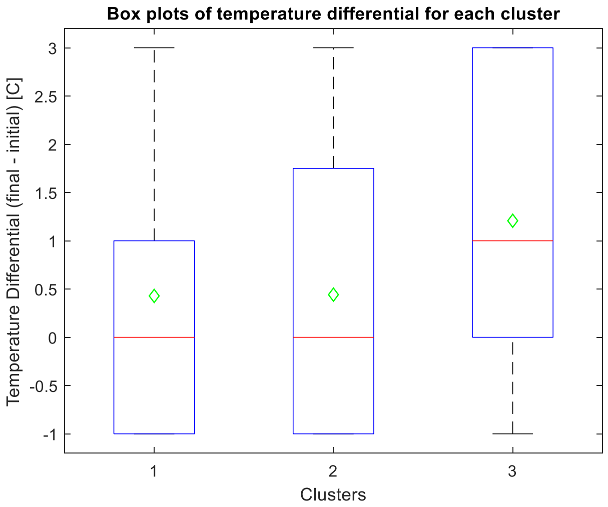


- Approx. *normally* distributed (more data can validate this)
- Mean = 0.48 mg/L/h
- Std. dev = 0.21 mg/L/h
- **Coefficient of variance (CV) = 44%**
- Intuitively, there is significant dispersion in the data to support the claim.

Conjecture 1: Tap stands tend to distribute water samples of similar decay rates

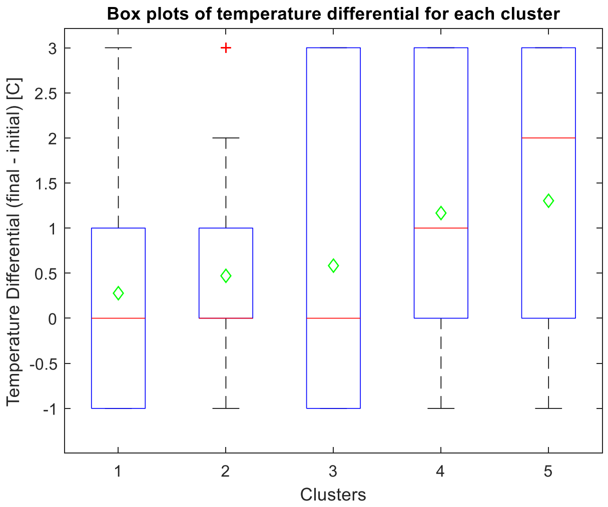
- Information on the tap stands can be incorporated as a new feature by gathering data on their conditions e.g. age, cleanness, etc.
- At the least, if such data is hard to collect, their geographical locations can be coded into the data as a categorical variable.
- **Conclusion:** information on the tap stands can help improve the precision of clustering as well as the prediction accuracy of machine learning algorithm because they may hint something about a water sample’s decay rate.

Conjecture 2: Temperature differential correlates positively with decay rate



3-Cluster Approach:

- From left to right are low-, medium-, and high-decay clusters.
- Mean temperature differential (final – initial):
  - Low-decay cluster: 0.43
  - Medium-decay cluster: 0.44
  - High-decay cluster: 1.21



5-Cluster Approach:

- From left to right are clusters with increasing average decay rate.
- Mean temperature differential (final – initial):
  - Cluster 1: 0.28
  - Cluster 2: 0.47
  - Cluster 3: 0.58
  - Cluster 4: 1.17
  - Cluster 5: 1.30

Conjecture 2: Temperature differential correlates positively with decay rate

- Initial temperature alone does not exhibit this increasing trend. Therefore, it isn’t enough by itself to help describe decay rate of a water sample.
- Information on temperature differentials can be collected by using the weather forecast of the day’s high or low temperature.
- However, there isn’t enough data to validate the correlation between temperature differential and decay rate.

Conjecture 3: Clustering can be used to group samples into distinct decay rates

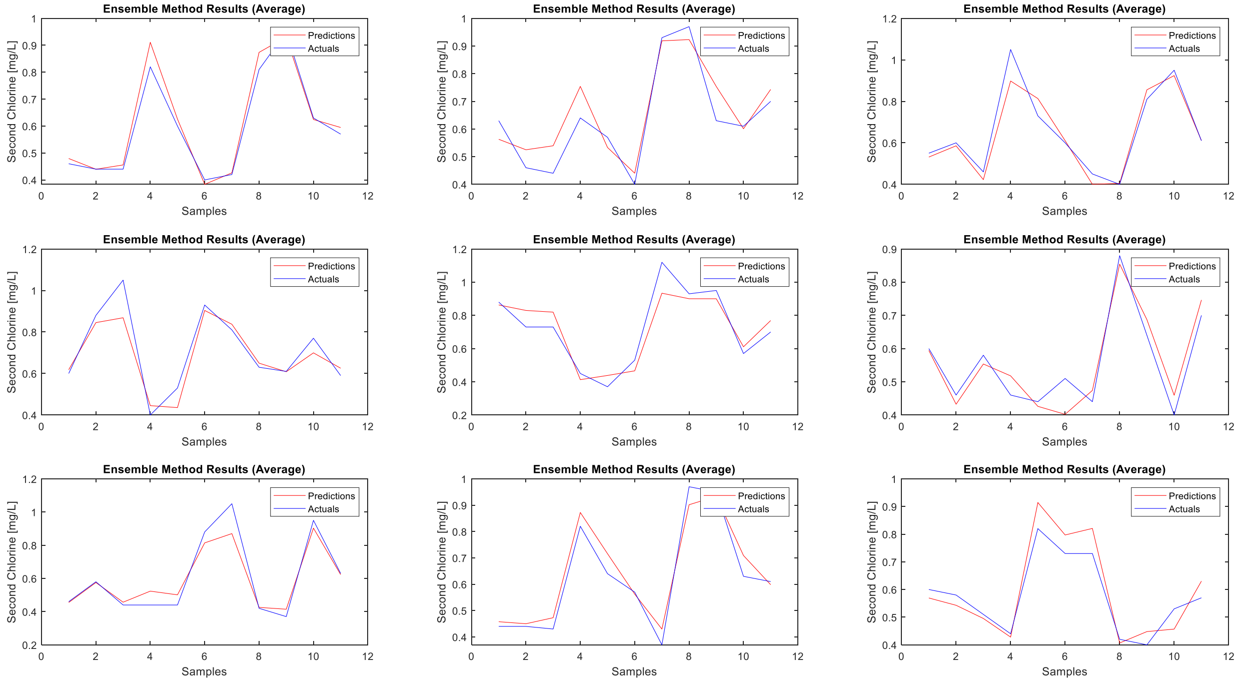
- Clustering based on a combination of features known in advance demonstrated some promise in dividing the data set into distinct decay rate categories.
- However, significant improvements are required to precisely cluster the samples.
- Clustering can be improved by incorporating conjecture 1 and 2.
- Other helpful features that are not reflected in the current data set include:

  Total Organic Carbon (TOC): an estimate of the cleanness of the water itself. Since it reacts with chlorine, it is expected to accelerate decay rate when found in higher amount. Some papers have suggested a linear relationship between TOC and chlorine demand, thereby possibly allowing us to calculate TOC without actually measuring it [3].

  Hours of exposure to direct sunlight: whether a container is outside or not may not be an accurate indication of this feature due to varying degrees of cloudiness or rain every day. This could be estimated by the weather forecast of the day's hours of sun and is known to accelerate decay rate if greater.

Conjecture 4: Machine learning is effective in predicting chlorine concentrations for the low-decay cluster

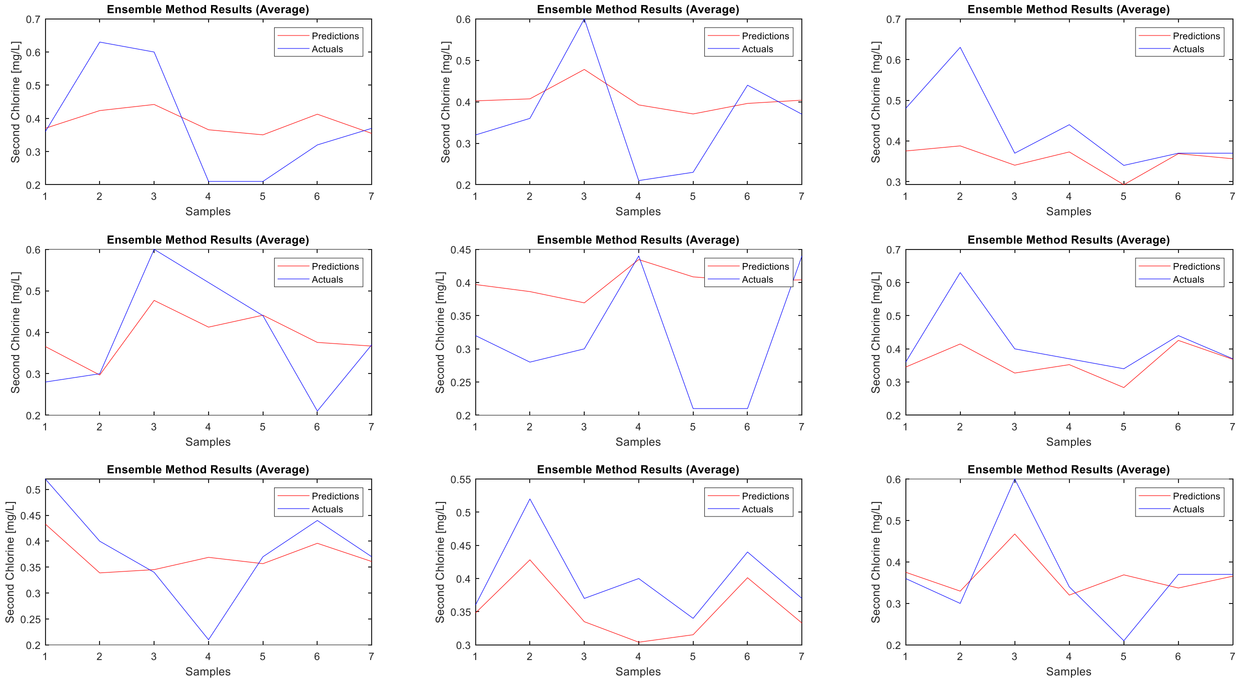
Only nine trials are shown, but 100 random trials were done to verify the performance of machine learning:



- **Average $R^2 = 0.88$**
- Red: predictions
- Blue: actuals

Conjecture 4: Machine learning is effective in predicting chlorine concentrations for the low-decay cluster

However, machine learning performs very poorly with the high-decay cluster:

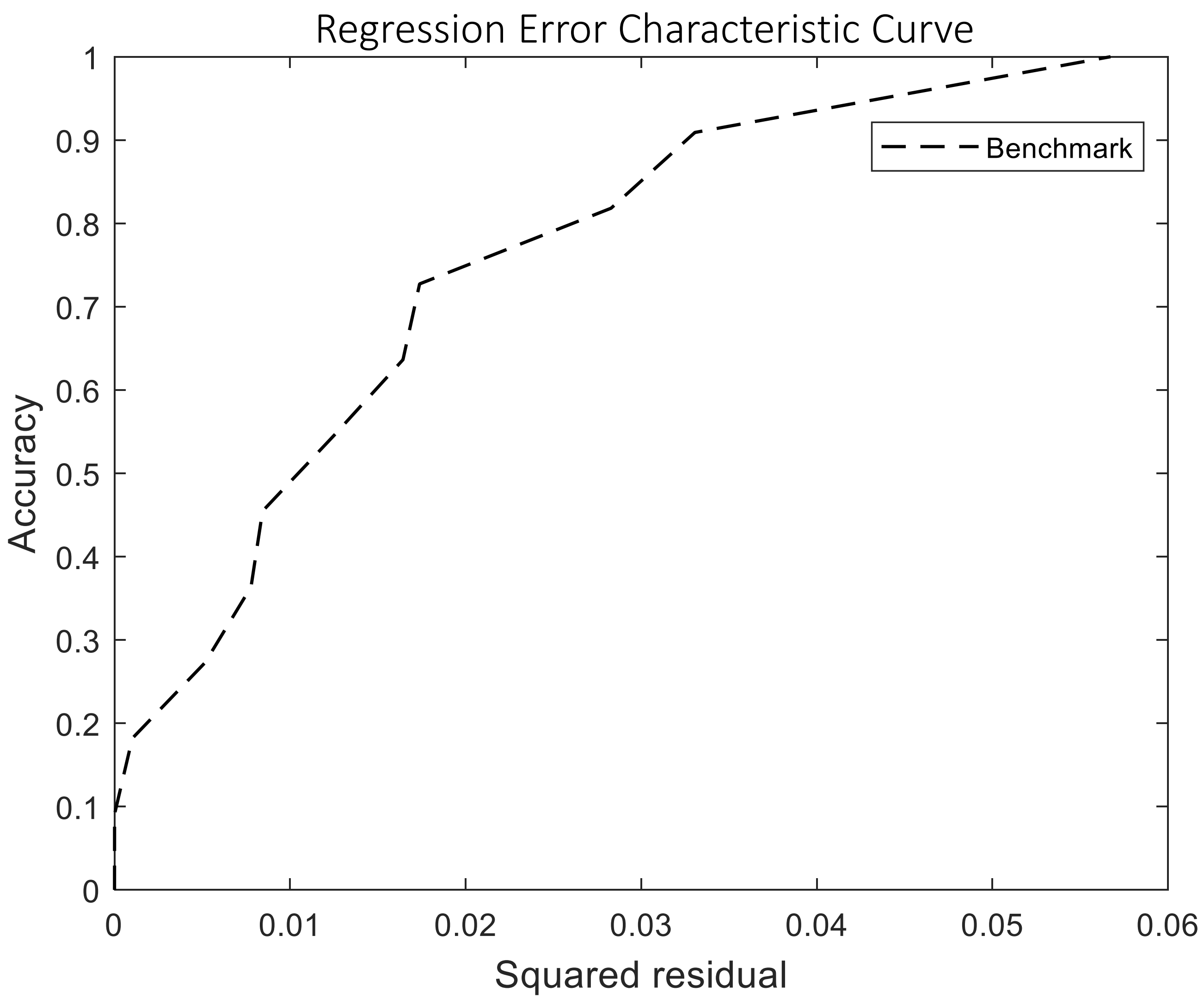


- **Average R^2 = 0.08**
- Red: predictions
- Blue: actuals

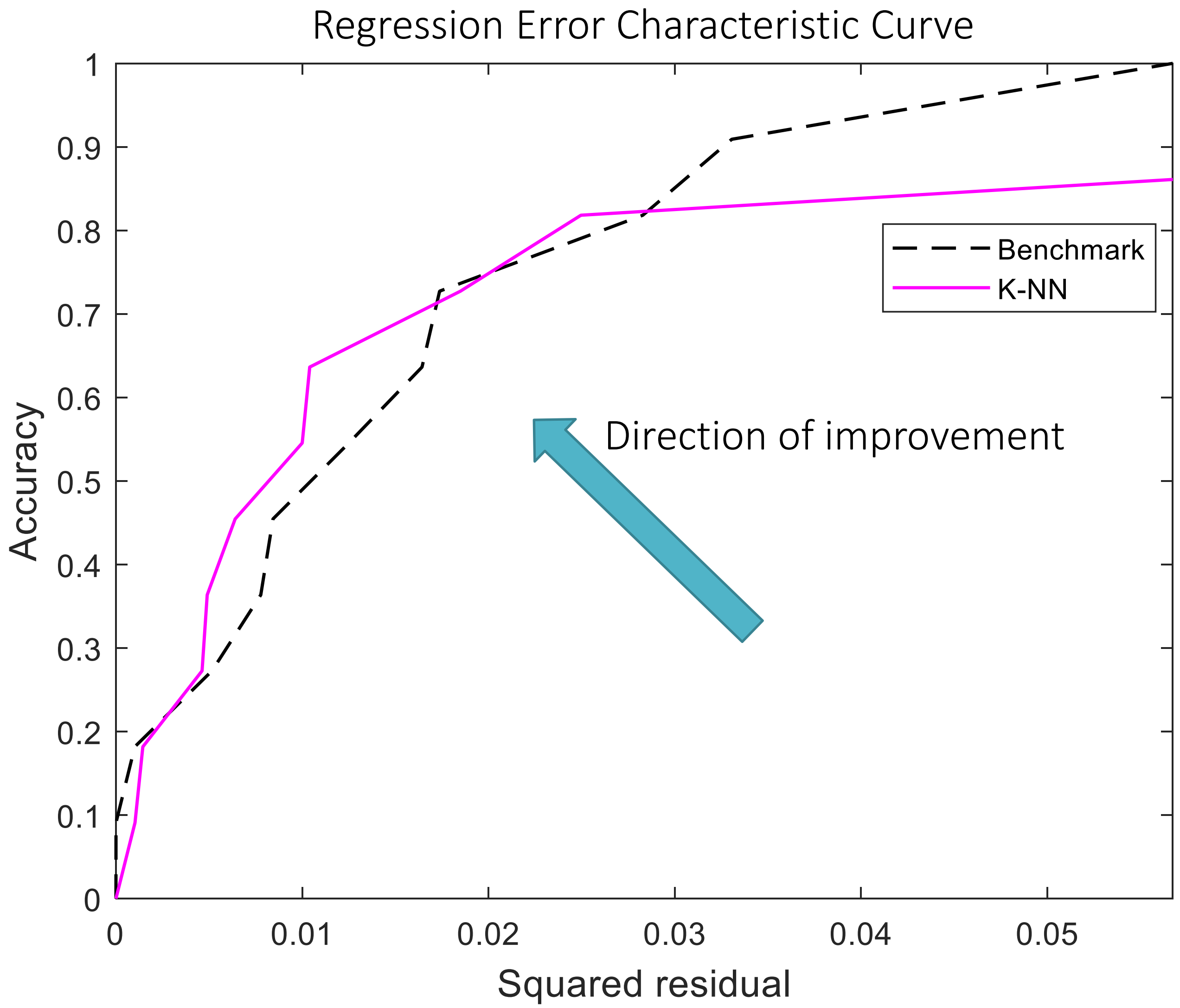
Conjecture 4: Machine learning is effective in predicting chlorine concentrations for the low-decay cluster

Approaches taken to improve the model:

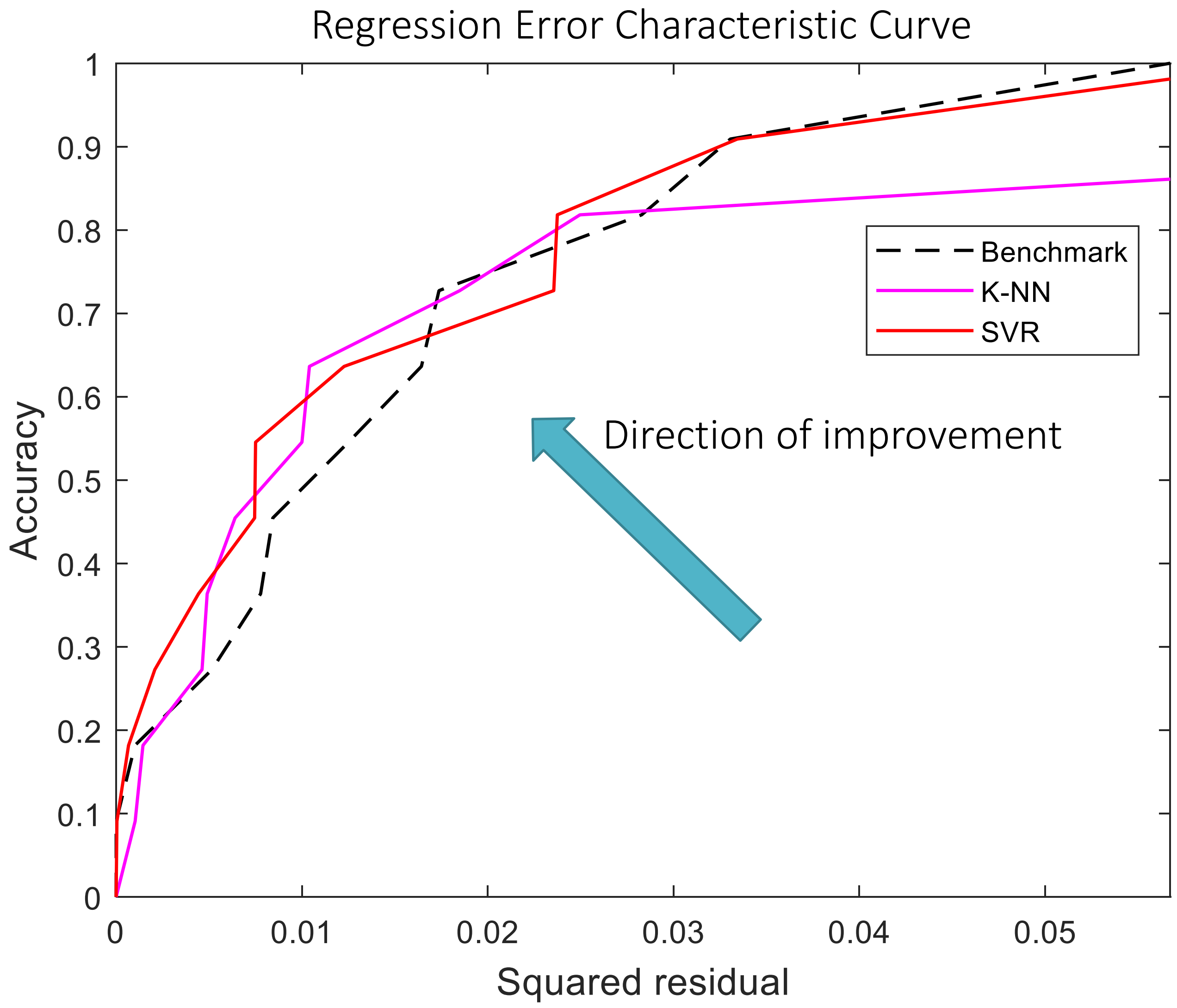
- Different predictive algorithms
  - K-nearest neighbors (K-NN), random forest, gradient-boosted trees, support vector regression (SVR), and the combinations thereof as an ensemble model were tested.
  - Different algorithms have different assumptions, strengths, and limitations.
  - Performance of each machine learning algorithm can be evaluated by:
    - R^2
    - *Regression error characteristic curve* (RECC), which measures 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.



- **Benchmark** = simple *average* line for predicting chlorine concentrations.

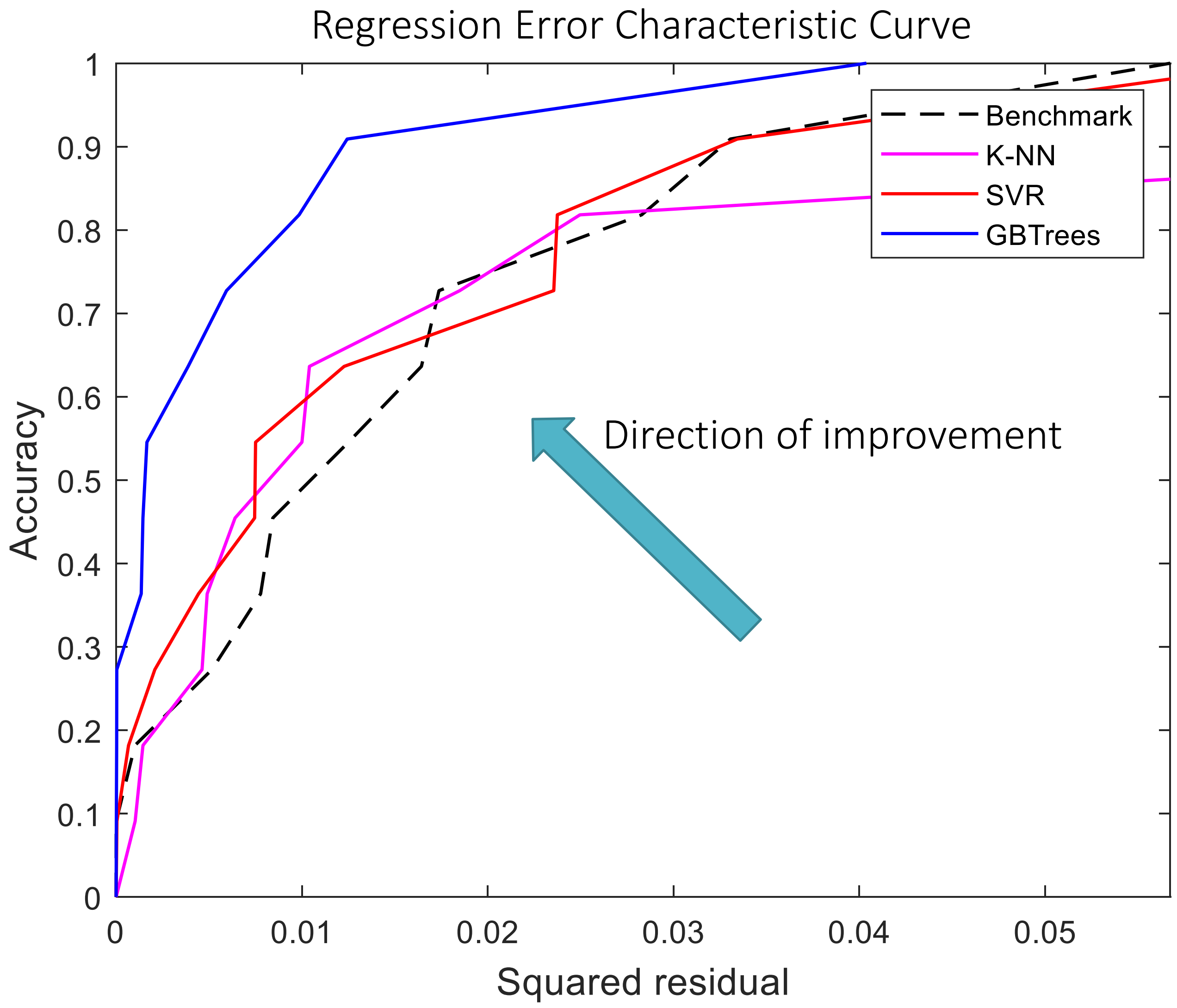


- **Benchmark** = simple *average* line for predicting chlorine concentrations.

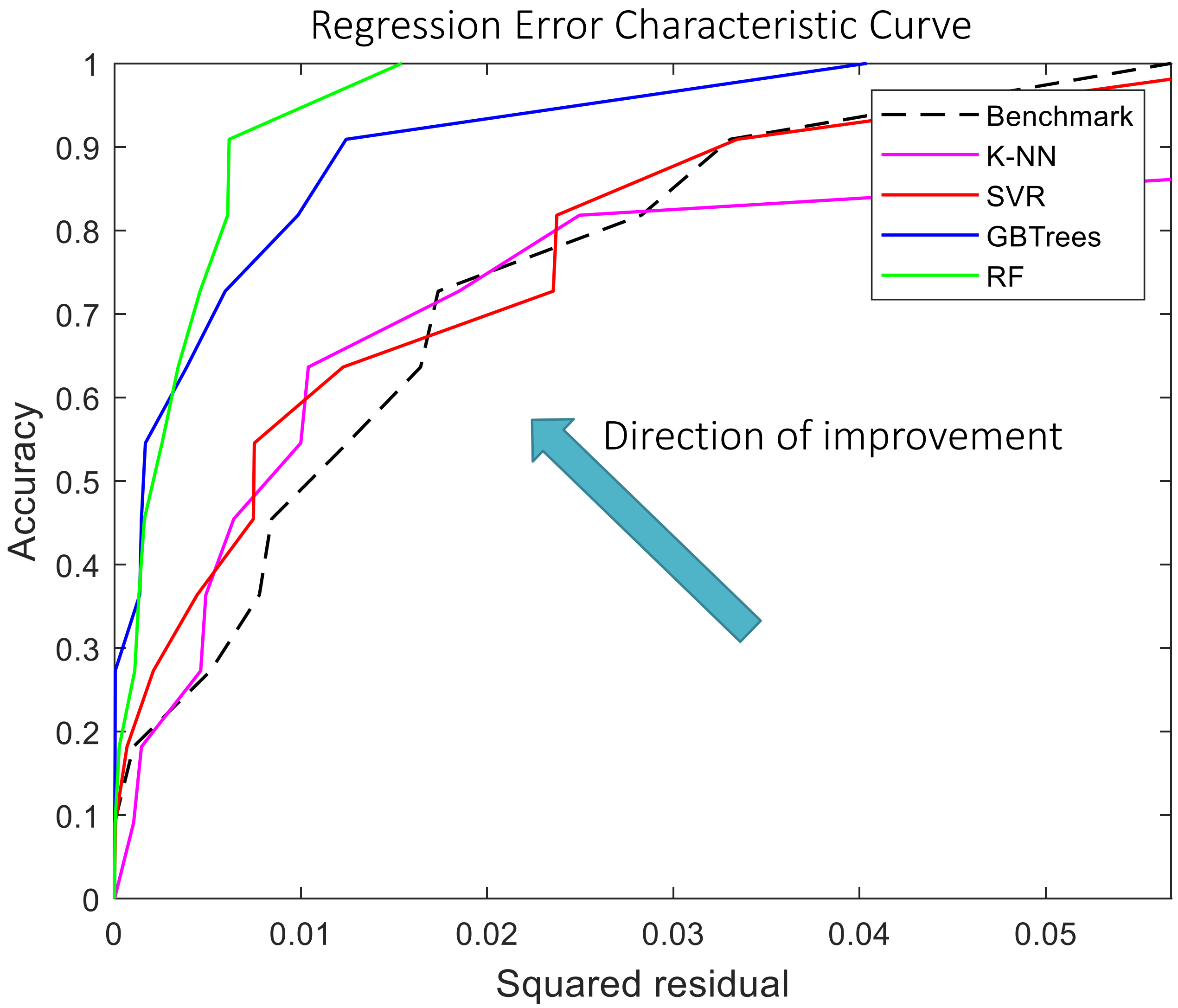


Direction of improvement

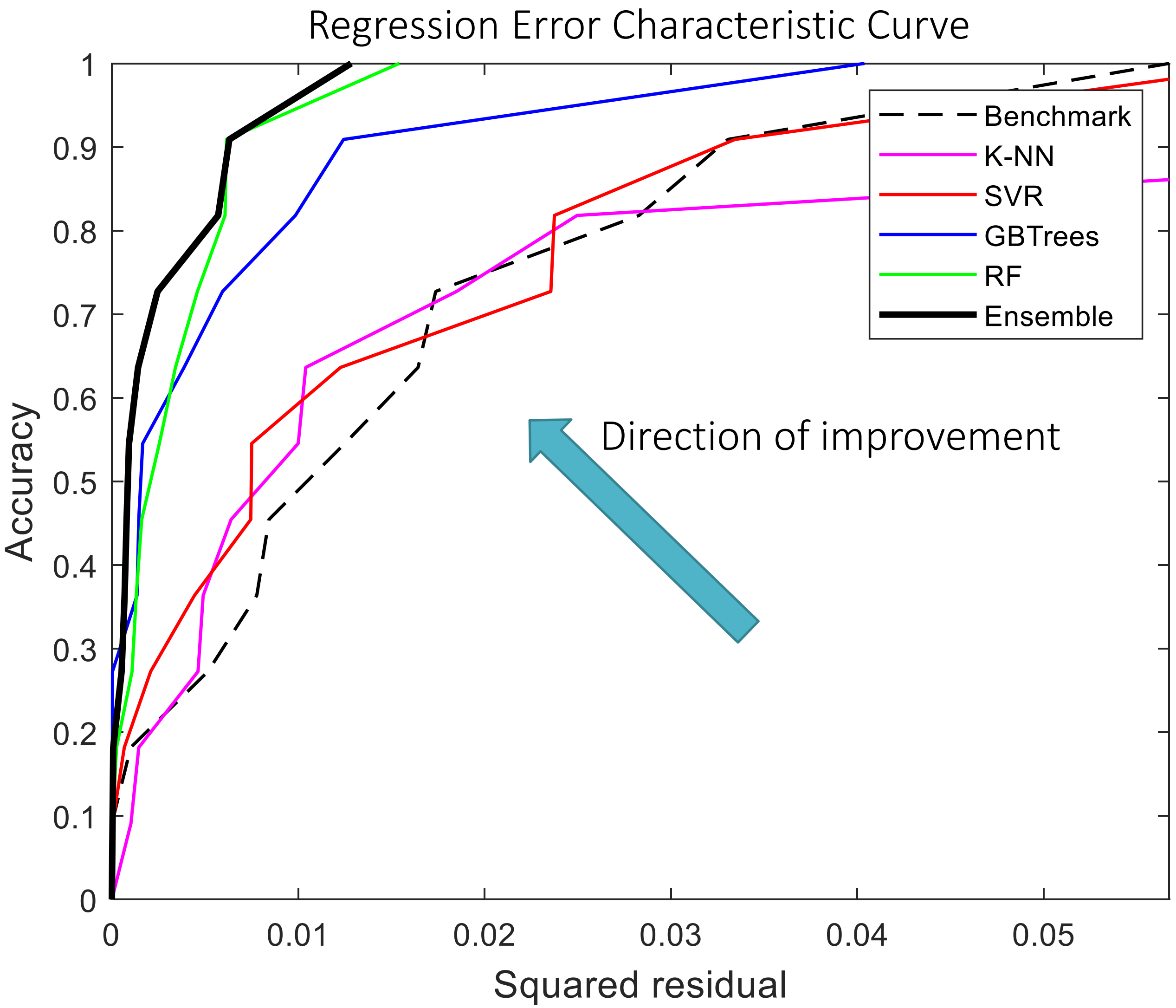
- **Benchmark** = simple *average* line for predicting chlorine concentrations.



- **Benchmark** = simple *average* line for predicting chlorine concentrations.



- **Benchmark** = simple *average* line for predicting chlorine concentrations.



- **Benchmark** = simple *average* line for predicting chlorine concentrations.
- **Ensemble** = average of GBTrees and RF (top two performing algorithms)

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

Conjecture 4: Machine learning is effective in predicting chlorine concentrations for the low-decay cluster

Approaches taken to improve the model cont'd.:

- 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.

Conjecture 4: Machine learning is effective in predicting chlorine concentrations for the low-decay cluster

Approaches taken to improve the model cont'd.:

- 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 R^2 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).

Other Conjectures

- Other than initial chlorine concentration and temperature, use of feature selection methods discovered that the following features are also important predictors:

Container type: 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

Method of drawing: "pouring out" are found more often in the low-decay cluster, implying that it provides 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

Feature selection methods included neighborhood component analysis (NCA), lasso regularization, and computing correlation score with the dependent variable (final chlorine measurement)

Future Improvements

There are largely two areas of improvements:

1. Quality of clustering

    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 rates

2. Accuracy of machine learning techniques, particularly for the high-decay cluster.

    Size of training set is too small (~50 samples, 35 of which are for training) which may affect accuracy

    Larger data set will make machine learning results more robust

    Changes in data collection may also improve machine learning performance:

        Min. time elapsed recorded is 15.8 h in the current data set.

        Decay rates may be more accurately captured if a sample is recorded earlier on.

Works Cited

[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.