Water Quality Analysis and Prediction at Mtendeli Refugee Camp in Tanzania

by

Jangwon Park

Supervisor: Chi-Guhn Lee
April 2018

Abstract

Well-known first- and second-order decay models lack robustness and sometimes demonstrate poor performance in accurately predicting future chlorine concentrations of water samples in a similar refugee camp setting to Mtendeli. A two-stage custom algorithm was developed to address these issues. In the first stage, hierarchical clustering analysis divides the entire data into three clusters. Each cluster contains only similar water samples that are likely to decay at very similar rates, thereby creating three distinct decay rate classes: low-, medium-, and high-decay. The optimal parameters to hierarchical clustering are determined by feature selection algorithms and a heuristic optimality score based on cophenetic correlation coefficient and cluster size variance. In the second stage, a custom ensemble regression model is developed by averaging predictions from random forest and gradient boosted trees. Predictions are made in each of the three clusters separately. The results demonstrate that the model is effective for predicting future chlorine concentrations of water samples in the "low-decay" cluster with an average R^2 of 0.88 over 100 random trials, suggesting some promise of machine learning in resolving the issues introduced by traditional statistical models. However, in the other two clusters, the model performance is rather unsatisfactory. Significant improvements can be made in both clustering analysis and machine learning predictions by incorporating data on new features such as tap stand conditions, temperature differential, total organic carbon, and hours of direct exposure to sunlight, and by including more data both in general as well as within each water sample.

Acknowledgements

In developing this thesis report of the past eight months of research, I would like to express wholehearted gratitude to my mentor and supervisor, Professor Chi-Guhn Lee, for the inspirational thesis topic and for his in-person guidance in every single meeting from the start to finish of this journey. I have gained a lot of knowledge from this research, and the experience has provided me with a strong motivation and clear outlook for future endeavors. I would also like to express gratitude to Syed Imran Ali, an affiliate researcher of Doctors Without Borders, for providing the data for the project, past studies and reports, and finally for his guidance in the beginning of the project. Finally, I would like to thank the entire Engineering Science community including all the faculty staff, coordinators, communication consultants, and fellow students who allowed this work to be as intellectually stimulating as it was.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 Background and Research Gap

Many developing countries in Africa, including Tanzania, lack the infrastructure for the kind of piped water supply systems often seen in developed countries and therefore rely heavily on centralized batch chlorination [1]. Under this setting, refugees at Mtendeli camp must fetch water manually, enough to last a day or two at a time, from various tap stands where treated water is distributed. However, it is overwhelmingly difficult to accurately predict when this water may become unsafe to consume due to varying climate conditions every day and many opportunities of water contamination in an often-unsanitary refugee camp environment. Predicting an acceptable water quality – measured by the concentration of free residual chlorine (FRC) [mg/L] – is therefore crucial to the refugees' health and livelihood.

Two studies in the past have tackled the same problem using statistical models in a similar refugee camp setting in Africa [1, 2]. These models are well-known first- or second-order decay models which describe the decay rate of a water sample i.e. the rate at which the concentration of FRC decreases. However, two issues exist at large: 1) the statistical models are sometimes highly inaccurate, and 2) the models lack robustness as the author admits that the model parameters may change vastly even for a small improvement in R^2. The implication of an inaccurate, highly sensitive model is an unreliable and possibly incorrect recommendation of initial chlorine dose to water batches. As such, new approaches must be considered to fill this research gap.

1.2 Objectives

The objective of this project is to build a robust model which accurately predicts future chlorine (FRC) concentrations using machine learning techniques. The advantage of using machine learning for prediction is that it does not make assumptions on how chlorine will decay prior to prediction as statistical models do; this is particularly reliable in a refugee camp setting as it is an open, dynamic system where samples that are identical in all aspects may still undergo different rates of chlorine decay on different days purely due to weather conditions or unanticipated human factors.

2. Literature Review

Predicting a numerical, continuous target is called regression in the context of machine learning. Studies on predicting water quality using machine learning are quite scarce and therefore only a few works are discussed here. As a result, this section will also concentrate on building the conceptual foundation on some specific methods of machine learning regression which may prove effective for this project.

2.1 Clustering Analysis

Clustering is a method by which water samples that are inherently similar are grouped together into the same cluster prior to any prediction of chlorine concentration. It achieves this by minimizing the intra-cluster distances and/or maximizing inter-cluster distances accordingly. Countless number of works have used clustering analysis to label and structure their data set in numerous ways. For this project, the idea is that if clustering analysis can discover and classify only the very similar samples together, then they must follow the same model and decay at (more or less) the same rate of change. Predictions on these samples, then, could be much more accurate.

Two of the most common clustering methods are discussed: k-means and hierarchical clustering. The following table summarizes the pros and cons of each method in general.

Table 1. Pros and cons of k-means and hierarchical clustering methods *[3]*

K-means Clustering		Hierarchical Clustering	
Pro	**Con**	**Pro**	**Con**
Generally more efficient than hierarchical clustering	The number of clusters¸ *k*, is often difficult to determine without expert opinion.	Far greater number of options to specify regarding the combination of method and metric when forming clusters within MATLAB	Requires significant tuning and experimentation
	Far fewer options to specify regarding the combination of method and metric	Can visualize merging of clusters with dendrograms	Difficult to identify the best cutoff point to limit the number of clusters

2.1.1 Measures of the Quality of Clusters

With using either k-means or hierarchical clustering, a major challenge is to evaluate the quality of the clusters – how good is the current clustering analysis? In this section, two formal measures are proposed to assist in answering that question. However, the project will also develop a heuristic measure of the quality of clusters based on how many data points each cluster has. For example, to ensure that the training set in each cluster is sufficiently large when building machine learning models, it is desirable to have evenly balanced clusters. Therefore, cluster size variance will be a good heuristic measure to consider in addition to the following.

2.1.1.1 Cophenetic Correlation Coefficient (CCC)

In hierarchical clustering, it is desirable for inter- and intra-cluster distances to be proportional. One way to verify this is by computing the cophenetic correlation coefficient which, if close to 1, suggests that the data is well suited for clustering [3]. Alternatively, cophenetic correlation coefficient can be thought as the correlation between the dissimilarity between any two clusters with their inter-cluster distance [4, 5]; the greater the degree of dissimilarity, the greater the inter-cluster distance. A clustering algorithm would ideally preserve this notion all throughout the algorithm.

CCC is particularly useful when comparing all possible combinations of methods and metrics in hierarchical clustering against each other; the combination with the highest value produces the best clustering analysis on the data. [4] does exactly this to figure out the best clustering method for its unique data set. It compares four different methods – Single, Average, Complete, and Ward – with six different metrics including Euclidean, Standard Euclidean, Minkowski, Mahalanobis, Manhattan, and Cosine. Among the 24 different combinations, the Ward-Cosine combination proved optimal. However, the paper suggests that the optimal combination varies from data set to data set, and each project must implement its own to discover the optimal combination for themselves. [5] performs the same analysis and again chooses its own optimal combination based on the highest CCC produced among 63 different combinations. This provides further confidence in the application of CCC to determine the optimal clustering method for this project.

2.1.1.2 *Inconsistency Coefficient*

A common stopping criterion in merging clusters is to specify the maximum number of clusters. A more algorithmic approach to this is to use the **inconsistency coefficient**. For cluster A and B about to be merged, the coefficient is calculated as follows [3]:

$$i(A, B) = \frac{d(A, B) - \mu}{\sigma}$$

The distance metric *d(A, B)* is substituted by an appropriate metric i.e. Euclidean, Manhattan, etc. The parameters μ and σ are the mean and standard deviation of the distances in the previous merges. The equation effectively standardizes the merge about to be formed to the historical inter-cluster distances. Therefore, the greater the value of *i*(*A*, *B*), the worse, or the more "inconsistent", the merge is. A good clustering analysis often involves identifying the threshold on the inconsistency coefficient so that any potential merges whose inconsistency coefficients exceed this threshold are not created.

[4] again does exactly this analysis to determine the "right" number of clusters. Its approach is to iterate through multiple computations of the inconsistency coefficient on the same data set, each time with a different "depth" parameter. The depth parameter specifies how much historical inter-cluster distances one would consider when computing the inconsistency coefficient. The greater the depth, the higher the inconsistency coefficient will be for the merge in question since earlier merges were always formed between much closer clusters. The heuristic approach in [4] equates the depth parameter as the number of clusters to use, and graphically observes the trend or pattern in inconsistency coefficients with subsequent merges at each depth parameter. The general rule is to find the depth at which the graph appears in a more "compact" form and the maximum inconsistency coefficient does not rise as much, or slows down significantly. Many works relate to this particular paper to leverage its heuristic approach in their applications, and likewise, this project may perform a similar analysis.

2.2 Regression Trees

Unlike clustering analysis which is an unsupervised machine learning method, regression trees are a supervised, completely data-driven method of predicting the values of the target variable using decision trees. A decision tree uses a number of rules to divide the feature space into rectangles in a way that identifies regions having the most homogeneous responses to the features, capturing complex nonlinearities between them and the target variable [6]. There are largely two different algorithms which make use of decision trees for prediction: random forest and gradient boosted trees.

2.2.1 Random Forest

Random forest is an ensemble (collection) of many decision trees. Random forest typically makes use of a bagging algorithm. In a bagging algorithm, a subset of observations is randomly sampled with replacement from the full training set with equal weights [7]. By default, random forests develop deep trees, which are grown independently and in parallel [7]. Each tree's prediction ultimately casts a "vote" to decide on the final prediction based on the majority rule. Because each tree only uses a subset of the training set, random forests avoid overfitting the training data.

[8] uses random forest to predict the degree of nitrate pollution in groundwater in Southern Spain. Nitrates typically originate from inorganic fertilizers in farming and can pollute water when discovered in excess amounts. Though this is in a different context, the application is similar to this project in that both aim to evaluate the quality of water by using random forest and a bagging algorithm for regression. The authors conclude that random forest is a very promising predictive method for water research and performs better than linear regression across many criteria including mean squared error. However, the discussion is rather inconclusive since the only benchmark is linear regression and because it will be difficult to generalize this result to a non-agricultural context. This indicates a need to carry this research forward to a different context and compare random forest to a broader variety of other predictive models.

2.2.2 Gradient Boosted Trees

Gradient boosted trees are similar to random forests with one notable difference: it makes use of a boosting algorithm which assigns different weights to the observations when it randomly samples a subset of the training set. Rather than growing trees independently and in parallel, gradient boosted trees are grown sequentially. At each sequence, a new tree focuses on observations that the model did not predict well in the previous sequence by adjusting the weights [6]. The goodness of predictions is evaluated by a loss function with the most typical one being sum of squared errors (SSE). The goal of gradient boosted trees algorithm is thus to keep adding trees, whose predictions are ultimately averaged for the same observation, to minimize the loss function iteratively. Gradient boosting consists of weak learners, which are “shallow” trees in this case.

A specific study for using gradient boosted trees for water quality prediction does not exist. It is reasonably assumed that as they are promising since random forest, which is based on decision trees as well, was found to be promising for water quality predictions. With no prominent research in leveraging this technique for water quality prediction yet, this project may provide meaningful ground for the validity of its performance.

2.3 Support Vector Regression

Support vector regression (SVR) is a regression technique based on statistical learning theory and a structural risk minimization principle, which had great successes in nonlinear modeling [9]. It attempts to discover a predictive function based on data by minimizing the generalized error bound rather than the observed error as statistical models do.

[9] applies SVR to predict water quality in river crab habitats over time as measured partly by dissolved oxygen and partly by temperature. It achieves great success in accurately predicting and describing the changes in water quality with time in comparison to back-propagation (BP) neural networks. River crab habitats are open, nonlinear, dynamic and complex settings which makes this study particularly relevant to the context of this project. Additionally, [10] applies least squares support vector regression in likewise nonlinear, dynamic, and complex Liuxi River in Guangzhou, China and predicts water quality very accurately in comparison ARIMA models or BP neural networks, and further highlights the strength of SVR in generalizing an accurate function to describe often uncontrolled target values. Generalization of the findings of these two studies is not practical as the research was conducted in a very specific environment. Therefore, application of SVR in this project may contribute to the current state of the field.

2.4 K-Nearest Neighbor

K-nearest neighbor (KNN) is known to be one of the simplest algorithms in terms of implementation which also proved very successful in many applications over the years. Traditionally, it is a non-parametric method (i.e. makes no assumptions on the function or the underlying distribution of data) for pattern classification based on the estimates of *k* nearest neighbors about the observation in question [11]. For regression, it works very similarly by averaging the estimates of the *k* nearest neighbors.

Research on using KNN for regression is relatively scarce compared to classification. Though it has been applied in various contexts from basal area diameter prediction to lithium-ion battery capacity prediction, regression applications for water quality prediction is extremely rare. [12] shares a limited insight into the performance of KNN for water quality classification. First, the setting is non-open, controlled water resources systems. Secondly, the quality prediction is a multi-class classification problem, which may be vastly different from regression in nature. Thirdly, thought KNN was found to work to some extent, its performance as measured by error rate was poorer compared to support vector machines or neural networks [12]. However, in consideration of the limitations and potential improvements KNN may achieve in a completely different setting such as the Mtendeli refugee camp, it may be sought as a candidate algorithm for making predictions.

3. Description of the Data

The data used in the project is provided by Syed Imran Ali, an affiliate researcher at Doctors Without Borders. The entire data set contains 145 observations with a total of 22 features including initial chlorine concentration, initial temperature, type of container, etc. Each observation contains two data points – initial chlorine concentration recorded at time 0 and a final chlorine concentration recorded anywhere between 15 to 25 hours post-distribution. The data is expected to propose two major challenges for accurate prediction due to its small size:

- The regression model may not be robust because smaller data sets usually produce inconsistent results with each trial as the impact of data partitioning is much greater.
- Having only two data points may render some machine learning techniques ineffective such as SVR since the generalized model could likely be a straight line connecting the two points. This would be far from accurately describing chlorine decay, which usually exhibits a first- or second-order exponential decay.

4. Methodology

The custom algorithm developed for predicting the second (also the final) chlorine concentration in each water sample largely consists of two sequential stages: hierarchical clustering and custom ensemble regression model using average predictions from random forest

(RF) and gradient boosted trees (GBT). In the first stage, hierarchical clustering requires three parameters – method, metric, and a subset of features. The first two parameters shall be determined using a heuristic optimality score computed based on cophenetic correlation score and cluster size variance while the subset of features is determined by three different feature selection algorithms: neighborhood component analysis (NCA), lasso regularization, and correlation criteria. Detailed discussion on this is presented in the next section. In the second stage, R^2 and regression error characteristic curves (RECC) are used to evaluate and select the best predictive model.

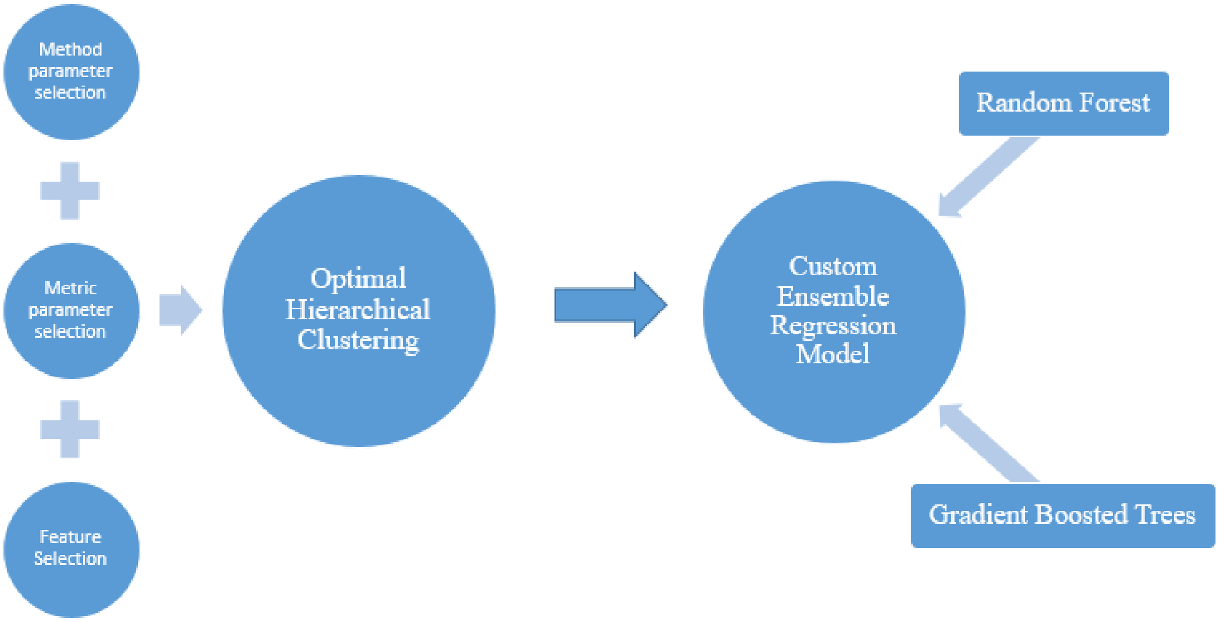

Method parameter selection
Metric parameter selection
Feature Selection
Optimal Hierarchical Clustering
Custom Ensemble Regression Model
Random Forest
Gradient Boosted Trees

Figure 1. Intuitive diagram of the custom algorithm developed for the project

4.1 Feature Selection

The objective of feature selection is often three-fold: improving prediction accuracy of predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data [11]. The purpose of feature selection in the context of this project lies most heavily in the third objective where feature selection may assist hierarchical clustering in maximizing its ability to identify consistent patterns only among water samples with similar decay behavior; here, the “underlying process” shall refer to the unique decay rate that each observation undergoes. Though expert domain knowledge may suffice if present, more algorithmic approaches can help select significant features, or more appropriately, exclude undesirable features that are highly correlated with others and/or add noise or bias to predictions. For robustness, three different algorithms are considered: neighborhood component analysis (NCA), lasso regularization, and correlation criteria.

NCA is an embedded method which minimizes prediction error of an underlying predictive algorithm with an additional regularization term [12]. Functionally, it is equal to k-nearest neighbors (KNN) and makes use of stochastic nearest neighbors, and the regularization term represents the sum of squared weights of each feature and therefore is conceptually equal to ridge regression [12]. Therefore, NCA can extract features that contribute to accurate predictions of chlorine concentration. In this project, only the feature weights that are greater than 0.01 are selected for the hierarchical clustering stage.

Similarly, lasso regularization is an embedded method which minimizes the sum of squared errors between observed data and predictions made by linear regression [13]. In lasso regression, the beta coefficients of the features are encouraged to be reduced to zero and thus is an effective method for feature selection [14].

Finally, correlation criteria describe Pearson correlation coefficient between each of the feature to the target variable i.e. second (final) chlorine concentration [11]. In this project, features whose absolute correlation coefficient are less than 0.15 are not considered as inputs to the hierarchical clustering algorithm.

4.2 Hierarchical Clustering

Hierarchical clustering is a promising and reliable approach to group “similar” observations. Similarity is a concept of distance – the closer the two observations are, the more similar they are. A proper distance function is uniquely defined by every method-metric combination supported in MATLAB.

Table 2. List of methods and metrics supported by MATLAB in hierarchical clustering

Method	Metric
Single	Euclidean
Complete	Standard Euclidean
Average	Squared Euclidean*
Centroid*	Mahalanois
Ward	Cityblock (Manhattan)*
Median	Chebychev
Weighted	Minkowski
	Hamming*
	Cosine*
	Correlation*
	Jaccard
	Spearman

*Method and metrics supported by MATLAB for k-means clustering.

When a group of water samples all exhibit a similar decay behavior, thus demonstrating a consistent pattern among them, accurately predicting chlorine concentrations may become significantly easier. Therefore, the **goal of the clustering analysis is to group the entire data into a few clusters such that each cluster may only have the samples will all likely undergo the same rate of decay**. Feature selection algorithms proposed in the previous section only exists to complement this processing by advising which set of features may be provided as inputs to the clustering algorithm. If successful, it can be shown that hierarchical clustering is a simple, but powerful data-structuring pre-analysis to complement predictive algorithms.

The procedure of hierarchical clustering analysis in this project can be summarized in the following steps:

1. Input features selected from the feature selection algorithms and a specific method-metric parameter into the hierarchical clustering algorithm. Specify the maximum number of clusters to three.
2. Compute the score for the provided method-metric parameter by averaging the cophenetic correlation coefficient (CCC) of the algorithm and normalized cluster size variance. Store this score.
3. Repeat steps 1 and 2 for all possible method-metric combinations.
4. Choose optimal method-metric combination corresponding to the highest score from step 4.

Relying solely on CCC may sometimes produce high unbalanced clusters where only one or two observations are found in a cluster while the rest are concentrated into another. Therefore, an arithmetic average is calculated between CCC and cluster size variance to acquire good quality clusters while also ensuring roughly evenly sized clusters.

Finally, while k-means clustering can be much quicker to implement, hierarchical clustering presents more combinations of methods and metrics to experiment with in MATLAB to discover the optimal approach for the specific data in question. This outweighs the advantage of k-means clustering as the accuracy of the algorithm is ultimately prioritized over its efficiency.

4.3 Custom Ensemble Regression Model

4.3.1 Data Preprocessing

Prior to predicting chlorine concentrations with the custom ensemble regression model, some data preprocessing is beneficial to improve the performance as much as possible. Two topics are discussed here – data partitioning/validation and feature engineering approaches.

4.3.1.1 Data Partitioning/Validation

70% of the entire data set is generally used for training and the rest for testing. However, small datasets are particularly sensitive to how the data is partitioned between the training and testing sets and consequently may yield significantly varying estimations of model accuracy [13]. To make the model more robust and help generalize it for any data set, 5-fold cross validation is implemented. In a K-fold cross validation, the training set is partitioned in to K equal portions, exactly one of which will always serve as a validation set when training the model; this is repeated K times each with a different validation set [13]. Cross validation is therefore an effective approach which allows one to make the best use of scarce data.

4.3.1.2 Feature Engineering

Feature engineering approaches can be supplementary tools to boost model performance. Two of the most common methods are feature scaling, which normalizes all numerical features to place their relative importance/influence on predictions on a common scale, and principal component analysis (PCA) to address undesirable correlation and noise among features which may harm model performance [15].

PCA is essentially a change of basis technique which computes the most meaningful basis to re-express a potentially noisy data [15]. It achieves this by projecting the original feature space to lower dimensions which are effectively a set of orthogonal directions, or principal components, which explain the most variability of the data. In this project, by selecting the first several principal components which cover 99% of the variability of the data, the essence of the data is retained while a completely new set of "features" are derived.

4.3.2 Candidate Algorithms

There are several candidate algorithms. In this project, the following are considered: k-nearest neighbors (KNN), support vector regression (SVR), random forest (RF), gradient boosted trees (GBT), and a custom ensemble model which averages predictions of any combinations of the techniques. The custom ensemble model eventually achieves the highest performance and is therefore pursued. The details of how each algorithm works is discussed in the literature review section. In this section, only the strengths and limitations are briefly discussed.

Table 3. Outline of each predictive algorithm's strengths and limitations

	KNN	**SVR**	**RF**	**GBT**
Strengths	Non-parametric	Structural risk minimization More robust than statistical models	Non-parametric Handles outliers and overfitting well	Non-parametric
Limitations	Influenced heavily by binary variables Sensitive to outliers	Relies on having many observations for accuracy	Generally slower run-time performance Easy to lose interpretability with large numbers of trees	May not handle overfitting well Easy to lose interpretability with large numbers of trees

4.3.3 Selection Criteria

The following two common quantitative measures exist for selecting the "best" predictive model which be used to evaluate each model in the previous section: R^2 and regression error characteristic curve. R^2 is intuitively the measure of performance of a model relative to a simple average line for predicting chlorine concentrations. By default, the R^2 of this benchmark is 0 and any model whose accuracy is positive indicates that it performs better than the benchmark; however, the goal is to get as close to 1 as possible. Since R^2 is a very well-known measure of performance in regression, only the regression error characteristic curve will be discussed.

4.3.4 Regression error characteristic curve (RECC)

For classification problems, receiver operating characteristic (ROC) is a widely used, powerful visualization and comparing the predictive power of classification models. Regression error characteristic curves (RECC) generalize ROC to regression and therefore serve the same purpose for comparing regression models, which has gained popularity in many studies over the years [18]. On the *x*-axis of a RECC, error tolerance is plotted often computed by squared errors while on the *y*-axis the accuracy of the regression model is plotted from 0 to 1. Naturally, as the error tolerance increases, the accurate of the model nears and ultimately reaches 1. Qualitatively, the RECC of a good model tends towards the top left corner of the graph; quantitatively, this means it achieves a high area under the curve (AUC). A great AUC is highly correlated with a high R^2 value and therefore both will be used to evaluate model goodness.

5. Results and Discussions

5.1 Feature Selection

The following summarizes the significant features as identified by each of the three feature selection algorithms:

Features	NCA	Lasso Regularization	Correlation Criteria
Initial FRC concentration*	✓	✓	✓
Initial TRC concentration*	✓	✓	✓
Initial temperature*	✓	✓	✓
Jerrycan	✓	✓	✓
Bucket	✓		✓
Other container			
White		✓	✓
Green			
Blue	✓		
Yellow	✓		
Orange			
Pink			
Red			
Container opacity	✓		✓
Container covering	✓		
Container cleanness			

Same container	✓		
Same water	✓		
Container fullness*	✓	✓	✓
Method of drawing	✓	✓	✓
Container outside			
Time elapsed*			

*numerical variables (the rest are binary)

There are fundamental limitations with each of these algorithms. For instance, the underlying predictive model in NCA is KNN while for lasso regularization, it is linear regression. As these may not be effective models for making accurate predictions in this project, the results here are only lightly considered as suggestions. With some trial and errors, the following features are selected as inputs to the clustering algorithm, which are also highlighted in the table above: initial FRC concentration, initial temperature, jerrycan, bucket, container covering, and container cleanness.

5.2 Hierarchical Clustering

With these input features, and each of the 84 method-metric combination from Table 3, the hierarchical clustering is ready to be evaluated. As mentioned, CCC and cluster size variance are averaged to output a final score for each combination.

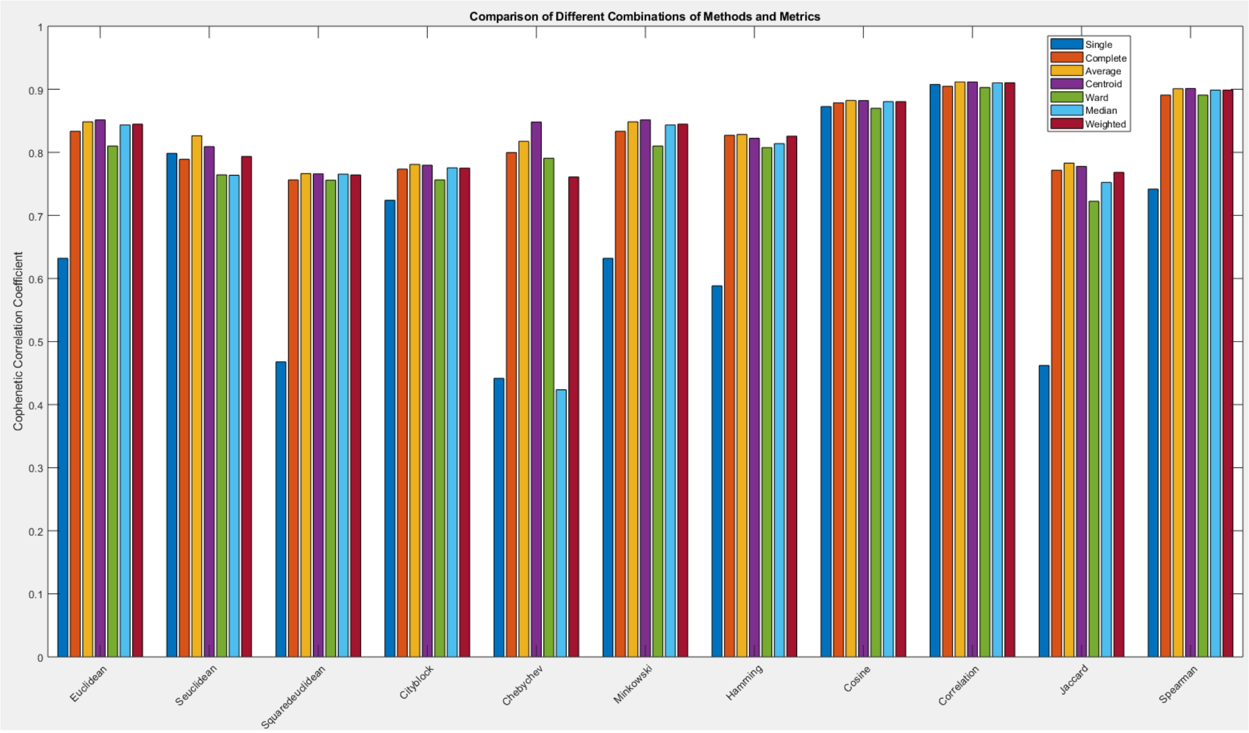


Figure 2. CCC scores of each method-metric combination.

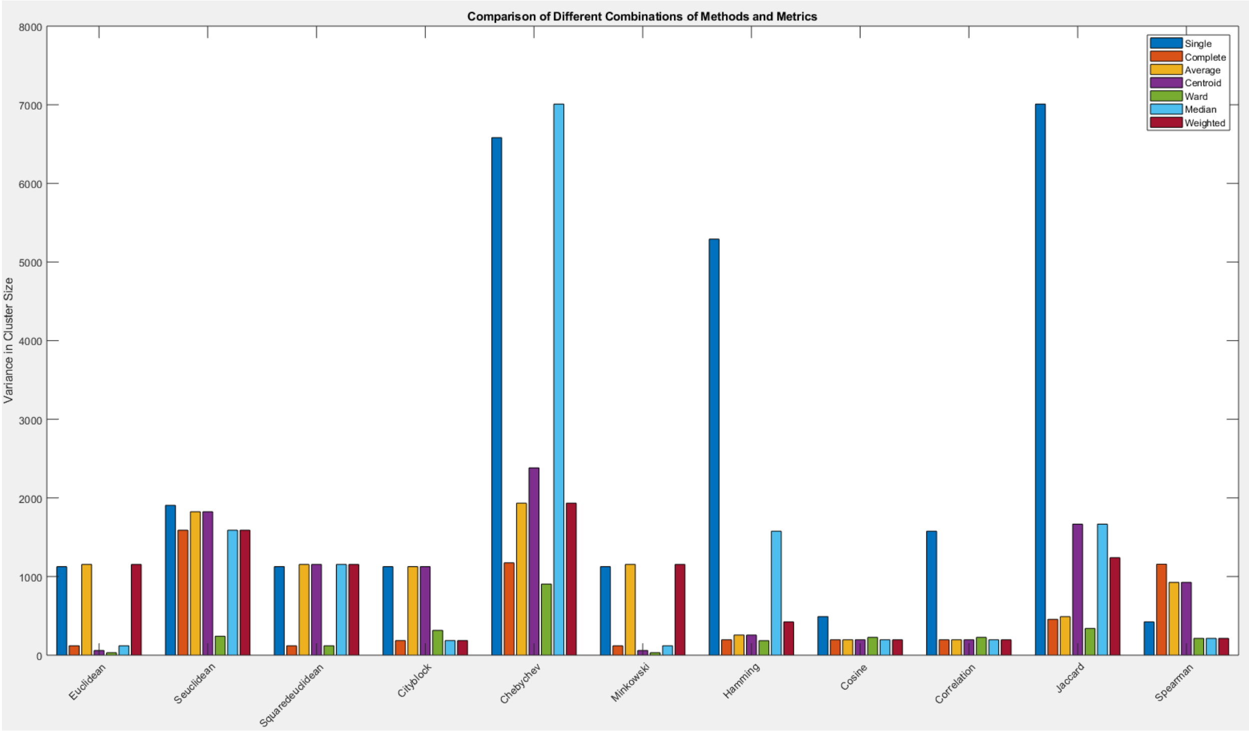
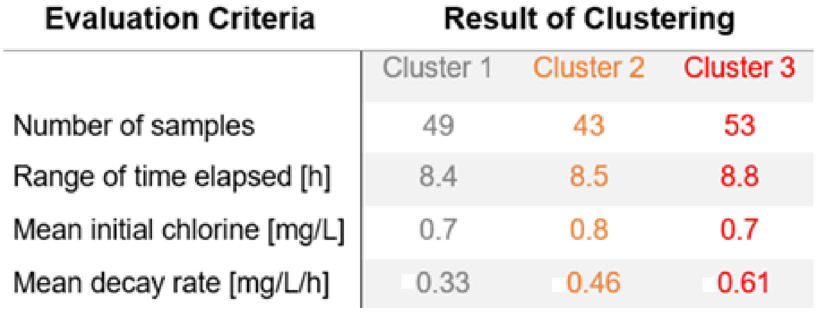


Figure 3. Cluster size variance for each method-metric combination

CCC scores do not generally differ as significantly as cluster size variance does across all method-metric combinations. Therefore, cluster size variance scores have a lot of impact on determining the final, optimal method-metric parameters for the data. Prior to averaging both scores, the cluster size variances are first normalized by dividing by the largest observed variance and then inverted to allow the smallest variance to have the highest score (lower the variance, the better). At the end of this analysis, it is discovered that the **Ward** method with the **Euclidean** metric are optimal.

Table 4. Result of Ward-Euclidean hierarchical clustering analysis



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 the first three evaluation criteria are not distinct enough to distinguish a cluster from another, the mean decay rates of the clusters are sufficiently different. Hypothesis testing on equal means and medians using the two-sample t-test and Wilcoxon rank sum test reveals that the clusters are indeed statistically different based on the decay rates at the 5% significance level. This is an important finding in this project as it is a working, supporting proof of the objective of the clustering analysis to some extent. In conclusion, this supports the idea that **the data set can possibly be divided based on decay rates into three clusters based on three distinct decay rate categories – namely low-, medium-, and high-decay clusters**.

However, the clustering analysis needs to be improved significantly to allow for more precise clustering to produce more distinct clusters i.e. their mean decay rates should be further apart than now. Table 4 presents the best result that cannot be improved as of now. This implies that the current set of features are not informative enough to cluster the data more precisely. However, new features that contain suggesting information about a water sample's decay rate can potentially improve clustering analysis if available. Some insightful conjectures are proposed below to support this claim.

Conjecture 1: Tap stands distribute water samples with significantly varying decay rates, but the same tap stand tends to distribute samples with consistently similar decay rates

The current data set has 52 unique tap stands based on their unique ID's. Each tap stand contributes anywhere from one to five samples/observations to the data set. No features on tap stand conditions such as age or cleanness exist. As a result, it is currently assumed that 1) all tap stands are equally capable of distributing water samples of all kinds of decay rates from low- to high-decay, and 2) tap stands do not significantly vary from each other in affecting decay rates. However, an analysis into the mean decay rates of the samples from each tap stand hints otherwise.

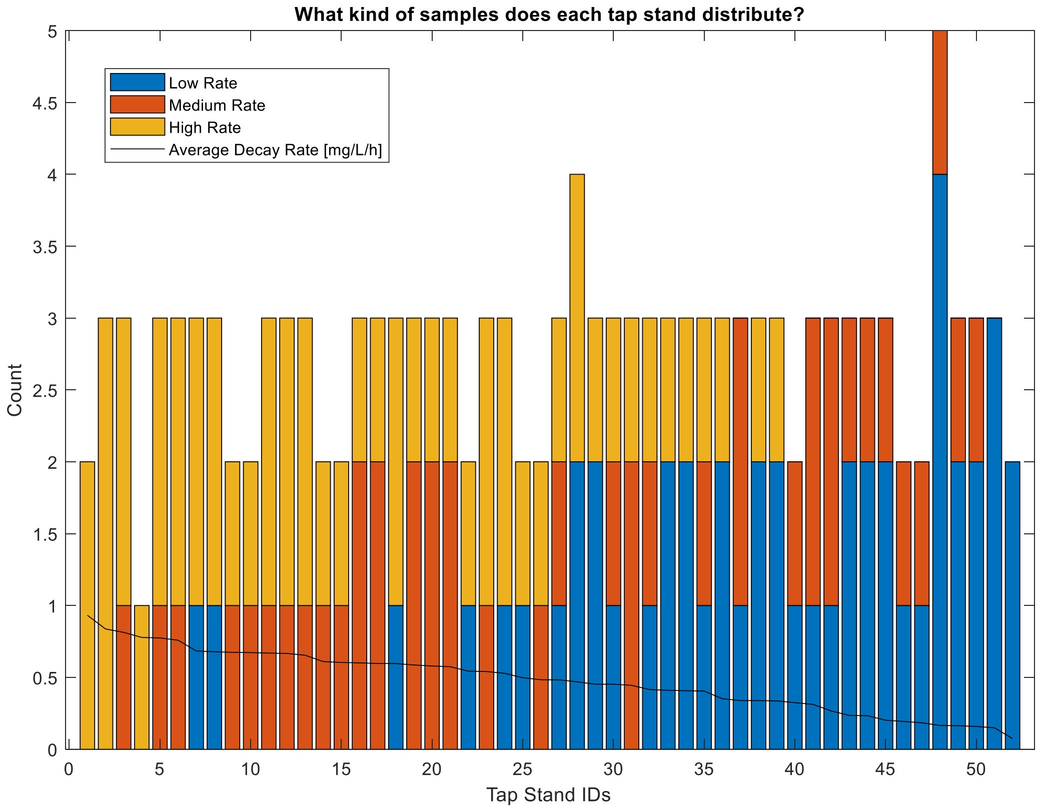


Figure 4. Mean decay rates of each tap stand in decreasing order

The height of each bar represents how many samples are recorded from that tap stand in the current data set, and the black trendline is the mean decay rates of all samples from that tap stand. When arranged in the decreasing order of this mean decay rate, the figure shows most of the tap stand only tends to distribute water samples of one kind of decay rate. For example, all three observations from tap stand 2 are labelled “high rate” while four of five observations from tap stand 48 are labelled “low rate”. Therefore, it can be expected that the next sample collected from these two tap stands are very likely to exhibit high decay and low-decay respectively. Furthermore, contrary to the current assumption, tap stands do vary significantly from each other in what kind of decay rates they each tend to produce.

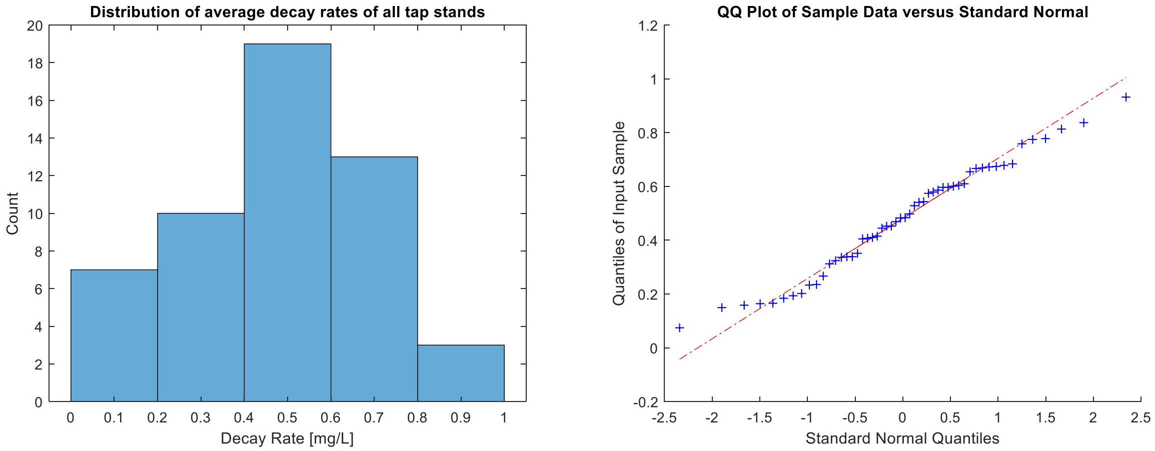


Figure 5. Distribution of mean decay rates of the tap stands (left) and normality check of the distribution using (right)

In fact, the distribution of the mean decay rates of the tap stands tend to follow a very close normal distribution as presented in figure 4 with the mean of 0.48 mg/L/h and standard deviation 0.21 mg/L/h. Coefficient of variation can be used as an effective descriptive statistic to indicate the dispersion of the data, which in this case is 44%. In general, this is a very large dispersion and thus helps dispel the assumption that tap stands do not vary significantly from one another in affecting decay rates. In conclusion, this conjecture reveals that **tap stands may distribute observations with largely varying decay rates, and that a tap stand may consistently distribute samples of one decay rate category.**

Conjecture 2: Temperature differential correlates positively with decay rates

Temperature is widely known to accelerate decay rate when higher. However, the temperature distribution in the current data set is not diverse and is in fact discrete.

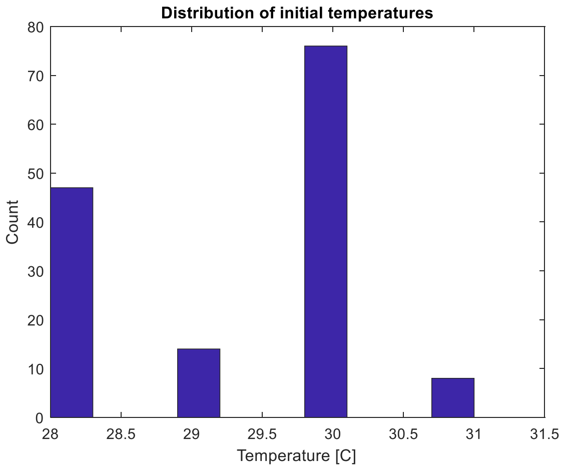


Figure 6. Distribution of initial temperatures of all observations

Therefore, the effect of temperature on decay rate cannot be studied effectively. In addition, the positive effect of higher temperatures on decay rates is reasonable in closed systems where the temperature is maintained. In an open, dynamic setting such as the Mtendeli refugee camp, where the temperature can fluctuate throughout the day, the initial temperature alone may not be sufficient even if the distribution of the temperatures is continuous. An alternative measure is temperature differential – the difference between the temperatures at various points in the future and the initial temperature. Fortunately, the current data set has recorded a final temperature for each observation. These, however, are not used for clustering or predictions since they are not known in advance.

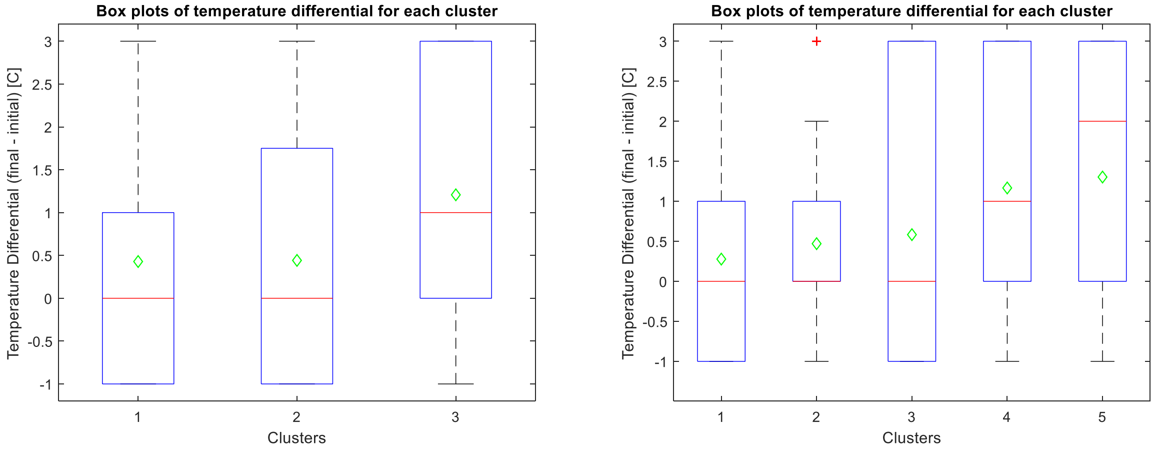


Figure 7. Box plots of temperature differentials of each cluster for the 3-cluster approach (left) and for the 5-cluster approach (right). The clusters are arranged in increasing order of mean decay rate.

The green diamond indicates the mean temperature differential of each cluster. With both the 3-cluster and 5-cluster approach, it can be shown that the mean temperature differential correlates positively with the mean decay rate of each cluster. This is reasonable since an observation will have decayed at a faster pace if it experiences a higher change in temperature. Such an analysis with the initial temperature alone does not reveal the same correlation.

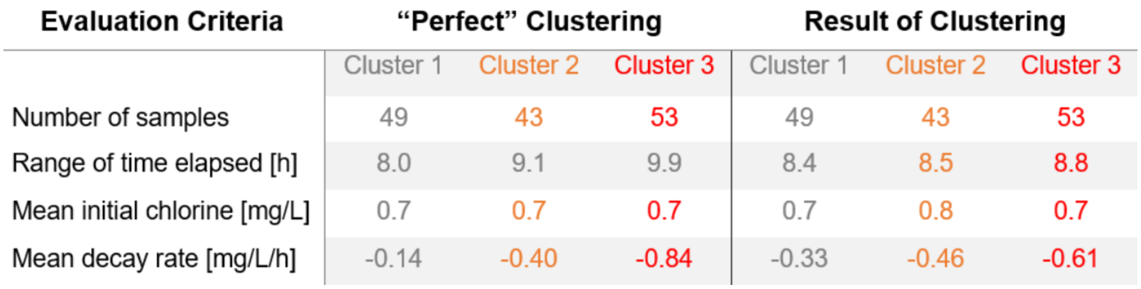
In addition to conjectures 1 and 2, other features which are closely related to decay rates are total organic carbon (TOC) and hours of direct exposure to sunlight. TOC is an estimate of the cleanness of water itself which reacts with chlorine. Therefore, it is expected to accelerate decay rate when found in greater amount. concentration. As it may require a specific equipment to measure TOC, some papers have suggested a linear relationship between TOC and chlorine demand, thereby possibly allowing one to compute TOC without measuring it [19]. As for hours of direct exposure to sunlight, it is intuitively to understand that it will accelerate decay rate. The current data set has a feature "container outside" which attempts to capture the same information. However, it overlooks the fact that each day experiences varying degrees of cloudiness and rain which renders this feature ineffective.

Though the hierarchical clustering results lack precision now, further analysis found promising results and potentials for their improvement. To verify the idea that clustering based on the decay rate is indeed complementary to machine learning prediction, the rest of this paper will assume that the algorithm is improved and therefore have clustered directly on decay rates for further illustrations.

5.3 Custom Ensemble Regression Model

Directly clustering on the decay rate leads to three precisely divided data sets:

Table 5. Comparison of results between "perfect" clustering and empirical clustering



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

By the same method presented in the previous section, it was discovered that the **Weighted** method with the **Squared Euclidean** metric performed the best among all other options. The decay rate used in the "perfect" clustering scenario was **linearly approximated**. The only reason for the linear approximation is because each observation in fact has only two data points; attempting to fit first- or second-order decay models despite this fact led to a highly concentration distribution of decay rates as shown:

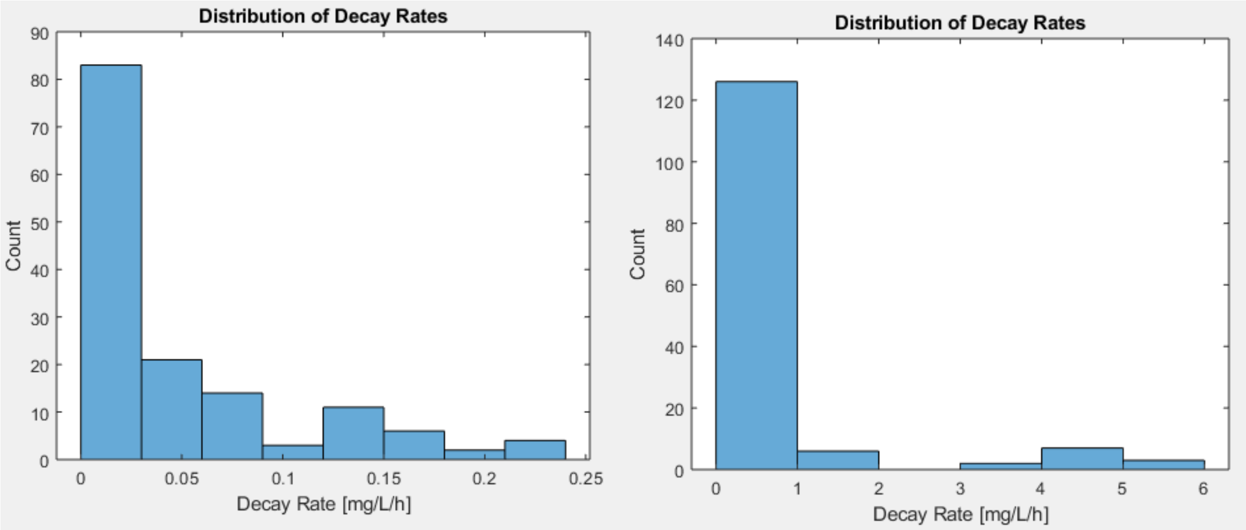


Figure 8. Distribution of decay rates as approximated by first-order (left) and second-order (right) models

The fact that the vast majority of the observations have more or less the same decay rate makes it impractical to cluster the data into three groups based on three distinct decay rate categories. Again, this is due to lack of data points in each observation, and therefore linear approximation was pursued.

Selection of the "best" regression model for predicting chlorine concentrations in each cluster is decided by each candidate model's performance as measured by R^2 and RECC. For convenience, only the results for the "low-decay" cluster are presented.

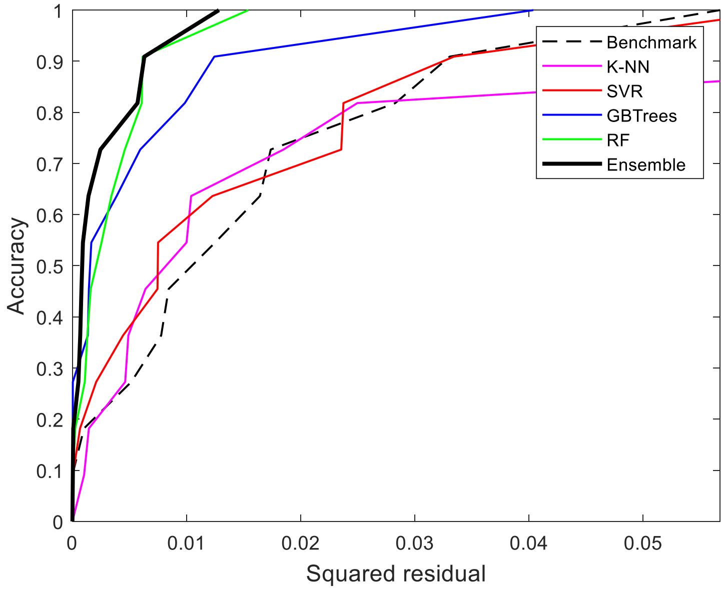


Figure 9. Regression error characteristic curves of different models for the low-decay cluster

Table 6. Average R^2 over 100 trials for each model for the low-decay cluster

	KNN	SVR	GBT	RF	Ensemble
R^2	-0.2	-0.08	0.82	0.87	**0.88**

The benchmark is simply a horizontal, average line to predict chlorine concentrations and therefore has an R^2 of 0. The ensemble model, which averages the predictions of GBT and RF, not only achieves the highest AUC in the RECC, but it also has the highest R^2 over 100 trials. For these reasons, the ensemble model is selected as the optimal model. Nine trials are shown below for demonstration purposes:

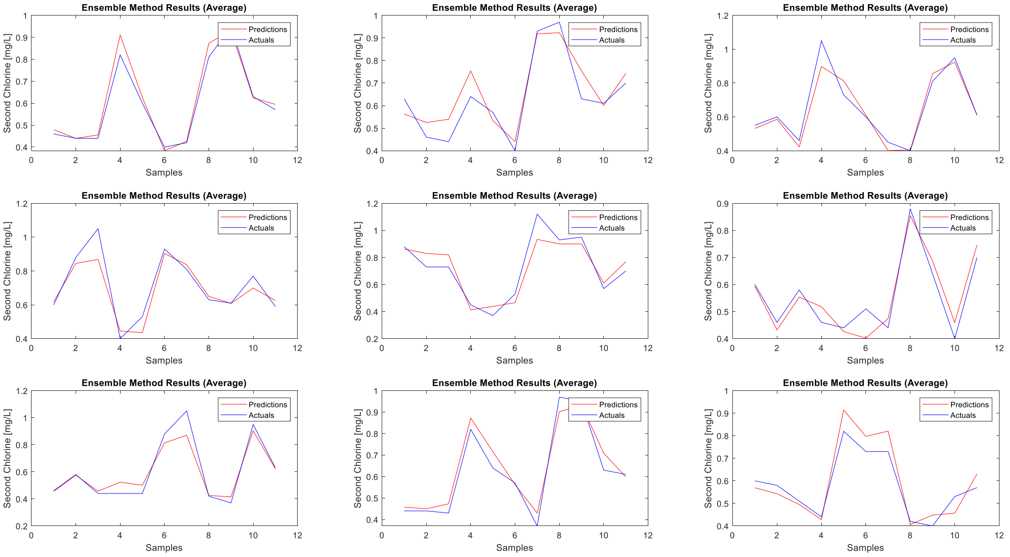


Figure 10. Predicted (red) vs. actual (blue) chlorine concentration for the low-decay cluster

Data preprocessing approaches such as 5-fold cross validation, feature scaling, principal component analysis led to a moderate improvement in model performance after implementation.

The same degree of accuracy, however, are not seen in the other two clusters. While the average R^2 for the medium-decay cluster was 0.71, it was 0.08 for the high decay cluster over 100 trials, which is only slightly better than the benchmark.

Table 7. Average R^2 over 100 trials using the custom ensemble regression model

	Low-Decay Cluster	**Medium-Decay Cluster**	**High-Decay Cluster**
R^2	0.88	0.71	0.08

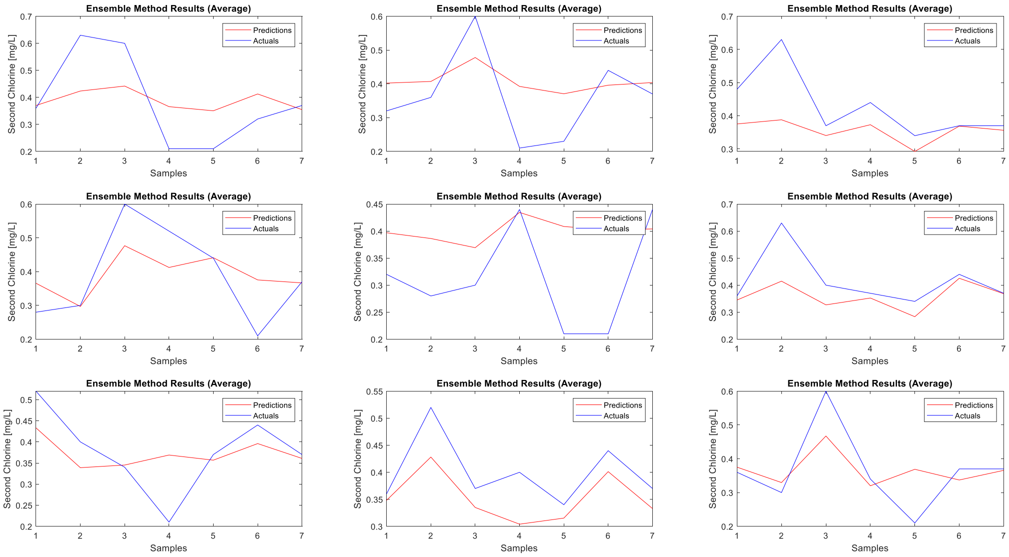


Figure 11. Predicted (red) vs. actual (blue) chlorine concentration for the high-decay cluster

This is likely due to the fundamental limitation imposed by the linear approximation of the decay rate. In reality, chlorine in water does not decay linearly; it undergoes a form of exponential decay which cannot possibly be observed accurately with a crude linear approximation. This can be avoided by collecting more data points in each observation. Once there are at least four or five data points, the decay rate can be estimated much more accurately for that observation. Clustering analysis and machine learning prediction with these observations are then expected to improve dramatically.

It is also concluded that the reason predictions are more accurate in clusters with lower mean decay rates is primarily because the observations in the low-decay cluster, for instance, have much smaller difference between their initial and final chlorine concentrations, which often make predictions easier.

6. Conclusion

A custom algorithm was implemented to pick the optimal parameters for hierarchical clustering that would best group the entire data into three clusters based on three distinct decay rate categories, namely low-, medium-, and high-decay rates. The parameter selection process consisted of three feature selection algorithms – NCA, lasso regularization, and correlation criteria – as well as score calculations using CCC and cluster size variance. It is expected that significant improvements to the precision of clustering can be achieved by incorporating data on new features such as tap stand conditions (age, cleanness, …), temperature differential, total organic carbon, and hours of direct exposure to sunlight.

A custom ensemble regression model which averages predictions from RF and GBT was selected to be the optimal model based on its R^2 over 100 trials and its performance on the RECC plot. Clustering analysis proved effective for the low-decay cluster, achieving an average R^2 of 0.88, but not for the other two clusters. It is concluded that this is most probably due to linearly – and quite crudely – approximating the decay rates. With more data points in each observation, decay rates can be estimated more accurately using well-known first- or second-order decay models. Not only will this improve machine learning predictions, it is also expected to improve the precision of the clustering analysis.

Other improvements include collecting more observations in general. The current data set is too small in the context of machine learning (145 observations). Larger training sets are bound to make the model more robust and potentially improve model accuracy as well.

7. References

[1] S. I. Ali, S. S. Ali and J.-F. Fesselet, "Effectiveness of emergency water treatment practices in refugee camps in South Sudan," *Bulletin of the World Health Organization,* 2015.

[2] S. I. Ali, "Study Report: Evidence Based FRC Targets for Centralized Chlorination in Emergencies," Medecins Sans Frontieres, 2017.

[3] T. Chan, *Clustering,* Toronto: University of Toronto, Department of Mechanical and Industrial Engienering, MIE465 Analytics in Action, 2018.

[4] B. Uragun and R. Rajan, "The discrimination of interaural level different sensitivity functions: Development of a taxonomic data template," *BMC Neuroscience,* vol. 14, pp. 114-134, 2013.

[5] S. Saracli, N. Dogan and I. Dogan, "Comparison of hierarchical cluster analysis methods by cophenetic correlation," *Journal of Inequalities and Applications,* vol. 1, pp. 203-211, 2013.

[6] J. Elith, J. Leathwick and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology,* vol. 77, pp. 802-813, 2008.

[7] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[8] V. Rodriguez-Galiano, M. P. Mendes, M. J. Garcia-Soldado, M. Chica-Olmo and R. Luis, "Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulunerability: A case study in an agricultural setting (Southern Spain)," *Science of the Total Environment,* vol. 476, pp. 189-206, 2014.

[9] S. Liu, H. Tai, Q. Ding, D. Li, L. Xu and Y. Wei, "A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction," *Mathematical and Computer Modelling,* vol. 58, pp. 458-465, 2013.

[10] X. Yunrong and J. Liangzhong, "Water Quality Prediction Using LS-SVM with Particle Swarm Optimization," in *Knowledge Discovery and Data Mining*, Moscow, 2009.

[11] L. E. Peterson, "K-nearest neighbor," Scholarpedia, 21 February 2009. [Online]. Available: http://scholarpedia.org/article/K-nearest_neighbor. [Accessed 8 April 2018].

[12] F. Modaresi and S. Araghinejad, "A Comparative Assessment of Support Vector Machines, Probabilistic Neural Networks, and K-Nearest Neighbor Algorithms for Water Quality Classification," *Water Resources Management,* vol. 28, no. 12, pp. 4095-4111, 2014.

[13] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research,* vol. 3, pp. 1157-1182, 2003.

[14] W. Yang, K. Wang and W. Zuo, "Neighborhood Component Feature Selection for High-Dimensional Data," *Journal of Computers,* vol. 7, no. 1, pp. 161-169, 2012.

[15] T. Chan, *Linear Regression,* Toronto: University of Toronto, Department of Mechanical and Industrial Engineering, MIE465 Analytics in Action, 2018.

[16] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* vol. 58, no. 1, pp. 267-288, 1996.

[17] J. Shlens, "A Tutorial on Principal Component Analysis," Cornell University, 2014.

[18] J. Bi and K. P. Bennett, "Regression Error Characteristic Curves," in *Proceedings of the Twentieth International Conference on Machine Learning,* Washington DC, 2003.

[19] L. Yee, M. Abdullah, S. Ata and B. Ishak, "Dissolved organic matter and its impact on the chlorine demand of treated water," *Malaysian Journal of Analytical Sciences,* vol. 10, no. 2, pp. 243-250, 2006.

8. Appendices

8.1 Appendix A: Exploratory data analysis I – basic statistics of select features

Some basic statistics of select features are presented here.

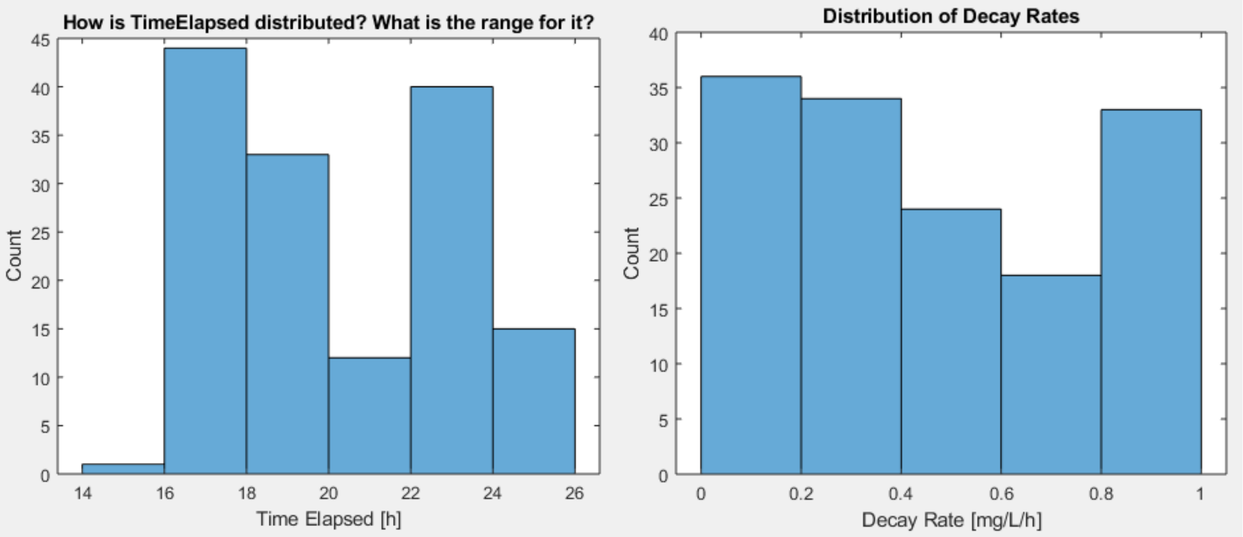


Figure 12. Distribution of time elapsed (left) and linearly approximated decay rate (right)

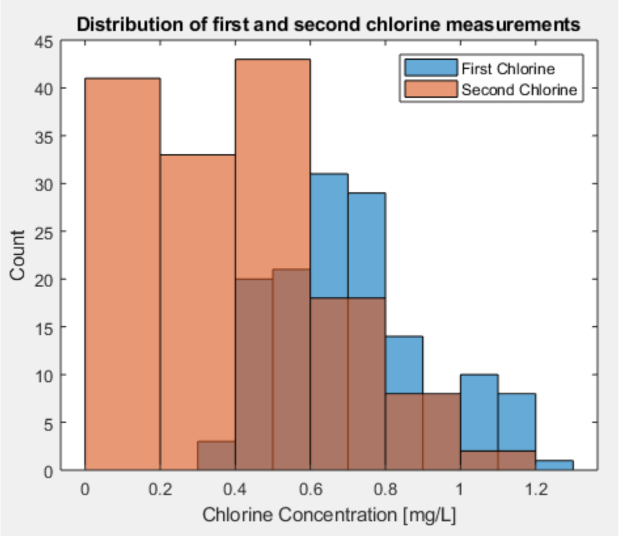


Figure 13. Distribution of initial and final chlorine (FRC) concentrations

Table 8. Basic statistics of time elapsed and linearly approximated decay rates

	Max	Mean	Min	Range
Time Elapsed [h]	25.9	20.3	15.8	10.1
Decay Rate [mg/L/h]	1	0.47	0	1
Initial chlorine [mg/L]	1.27	0.71	0.32	0.95
Final chlorine [mg/L]	1.12	0.38	0	1.12

8.2 Appendix B: Exploratory data analysis II – correlation analysis

Exploratory data analysis on the non-clustered (entire) data set reveals close to no correlations among the features, which partly motivated the pursuit of clustering analysis to begin with.

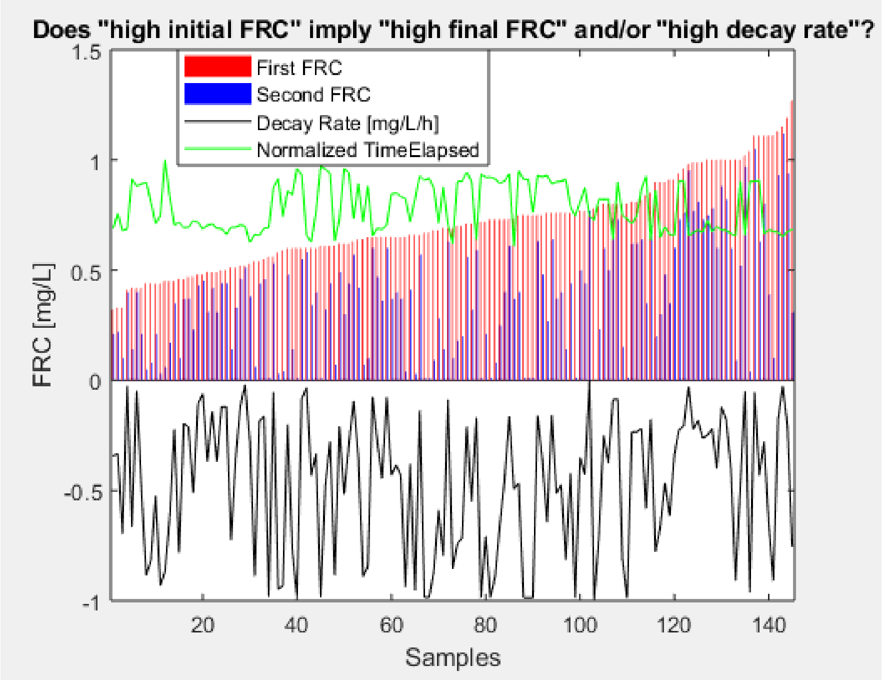


Figure 14. Initial chlorine concentrations are arranged in increasing order to find positive correlation with final chlorine concentrations

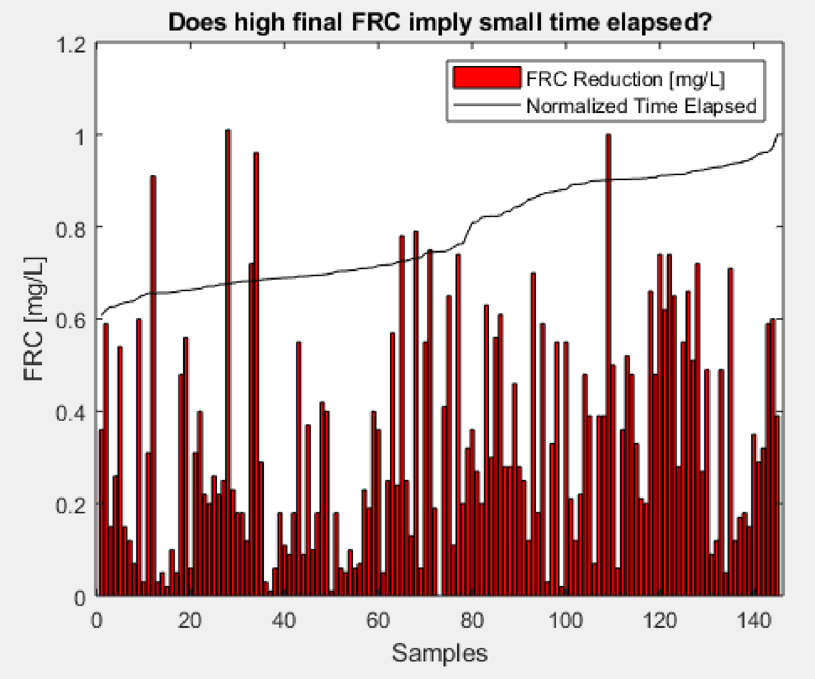


Figure 15. Time elapsed in increasing order to find relationships with total chlorine reduction

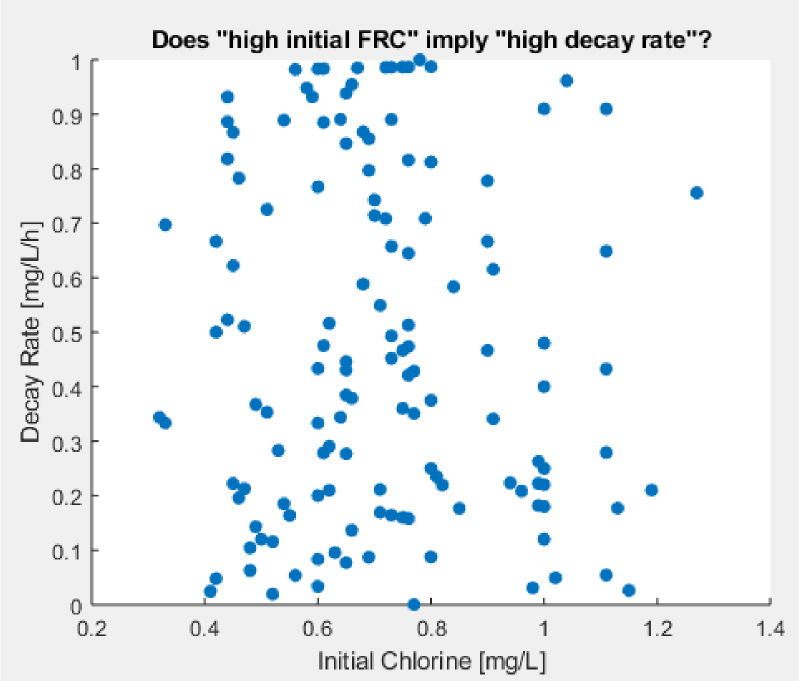


Figure 16. Correlation plot between initial chlorine and decay rate

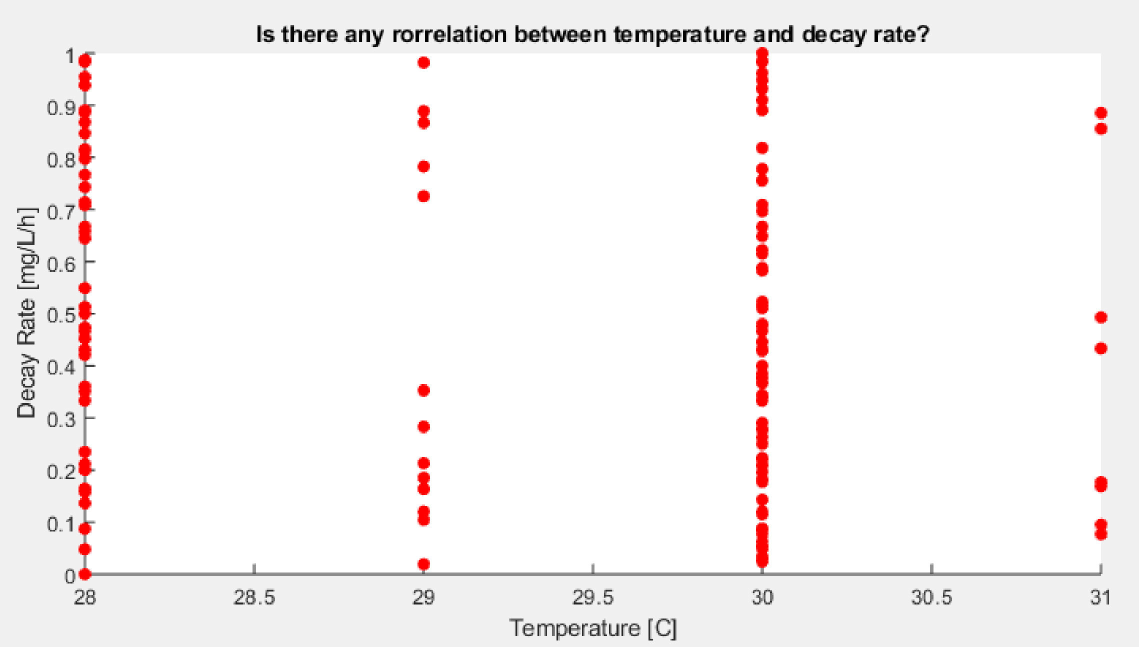


Figure 17. Correlation plot bewteen initial temperature and decay rate

One would expect, for instance, to observe high decay rate for a water sample which started out with a high initial chlorine concentration if all water samples truly followed the same model. However, this is not found to be true. In conclusion, no obvious or intuitive correlations are discovered among the features, which motivated clustering analysis.