Determining NDMA Formation During Disinfection Using Treatment Parameters

Introduction Water disinfection was one of the biggest turning points for human health in the past two centuries. Adding chlorine to water has drastically reduced the cases of water-borne illnesses and increased the standard of living, however, water chlorination produces harmful (and likely carcinogenic) disinfection byproducts (DBPs). Today, there are many different known types of byproducts that form during disinfection, the most toxic of which include compound classes such as nitrosamines, haloacetonitriles, haloacetamides, haloacetaldehydes, trihalomethanes, and haloacetic acids. Of these classes, the environmental protection agency (EPA) enforces maximum contaminant limits on nitrosamines, trihalomethanes, and haloacetic acids, the most toxic of which on a volume basis are nitrosamines. For these three classes of compounds, utilities throughout the United States are required to report when, and the magnitude to which, any limits are exceeded. The lowest limit (implying the most toxic compound) set by the EPA is for *Nitrosodimethylamine* (NDMA). In the past, water plants have been shut down when NDMA limits had been exceeded in order to protect public health.

Most DBPs, and especially NDMA, are linked to the constituents of the water that is being chlorinated and can vary drastically based on water quality parameters, method of disinfection, and water treatment plant characteristics. While NDMA formation mechanisms are mostly understood, the external factors, such as the manner of chlorine addition are unknown. However, NDMA formation could be modelled based on disinfection parameters, which are a lot cheaper to determine than the actual nitrosamine concentrations post disinfection. Similarly, NDMA analyses of water can take up to 24 hours to complete (depending on analysis), and by the time that a water utility would know that limits have been exceeded, the water would have already been used by consumers. Therefore, if accurate results could be achieved consistently, it would be in a utilities and consumers best interest to predict the concentration of NDMA that could form during disinfection. The goal of this project was to lay the foundation for future work in NDMA prediction. Namely, the available drinking water treatment plant parameters were used as inputs to predict whether NDMA would form in treatment plants, and if so, the concentration of NDMA that is expected to form.

Related Work Even though many researchers have used limited data sets to predict regulated DBP formation, their attempts have been largely disjointed, and have not considered plant factors that could influence formation, such as treatment plant size (Chowdhury et al., 2009). In most papers in the DBP field, researchers look at a class of DBPs formed from a limited number of water sources (for example, several sampling points along a river), and use various statistical measures to correlate the data with known parameters (Hong et al., 2007). Only one research group thus far has compiled multiple data sets available in the literature in a machine learning context (Singh et al., 2012). Here, the authors used data compiled from 63 different papers, with parameters such as dissolved carbon, bromide concentration, pH, temperature and chlorine contact time. They used an artificial neural network (ANN), support vector machine with a Gaussian kernel (SVM), and gene expression programming (GEP) to predict trihalomethane formation. The average mean squared error associated the training set was 0.1%, 0.5% and 9.4% for the ANN, SVM, and GEP respectively. For the test set error, the ANN SVM and GEP average mean squared errors were 16.4%, 13.4%, and 13.2% respectively. The low training set errors could be a sign of model overfitting. At the same time, the data set compiled by the authors was very limited in scope: here, trihalomethanes were formed in lab, rather than at real water treatment plants, and the authors could not consider how DBP formation would change with, for example, plant size. Similarly, the focus of their research was trihalomethanes. NDMA is much more toxic than trihalomethanes, and predictive methods for their formation are necessary. The following analysis is intended to focus on the effect of various treatment plant characteristics on NDMA formation using treatment plant data throughout the United States.

While Hong et al. is currently the only study that focuses on DBPs in water, there are several relevant studies that seek to predict the mutagenicity in water samples. Even though mutagenic compounds and DBPs are not known to be correlated, they occur in similar concentrations in water (Hertzberg et al., 2000) and, by extension, modelling methods used in mutagenicity studies could be effective for NDMA prediction. Of mutagenicity studies, the most relevant one to this project is Zheng et al., where a Gaussian kernel SVM was used to predict total mutagenic compounds in an environmental water (unlike studies that use water prepared in a lab that does not mimic relevant NDMA formation conditions) (Zheng et al, 2013). Here, the best fit resulted in a 9.7% average mean squared error for the training set, and a 10.2% average mean squared error in the test set. While the data set in this study is also small (48 samples for the training set and 12 samples for the test set), the reported errors are a benchmark for models that could be used to predict NDMA formation.

Data Set and Features The data set used to predict NDMA formation in drinking water treatment plants was compiled from 10 data sets available through the EPA (USEPA). The final data set consisted of over 20 defining characteristics, such as facility size, EPA processing number, type of water disinfected, point of sampling, disinfectant used in the facility, and date that the water sample was collected. From this set, 10 features were extracted: irrelevant characteristics, such as the date that a water sample was collected and EPA processing number were omitted such that only potentially relevant parameters, such as facility size and disinfectant used in the facility remained. Further, for each feature, the data included was preprocessed for ease of use in Matlab, for instance, categorical variables such as plant size were assigned numerical values ranging from 1 to 5 rather than string values of 'XL' through 'S' as present in the original data set. Continuous features were given standard scores to aide normalizing the data.

The total dataset consisted of 108,604 examples. Of these examples, in 1906 samples, there was NDMA detected (concentration was greater than 0.2 ng/L) and in the remaining samples (106698 of the total samples) the NDMA concentration was lower than the accepted detection limit of 0.2 ng/L (counted as zero/non-detect samples). This detected to non-detected ratio was maintained by using stratified sampling when dividing the set into test and training data. 76023 examples (~70% of the full dataset) comprised the training set, and 32582 examples (~30% of the full dataset) comprised the test set.

Methods Two classes of algorithms were used in this project: in order to predict whether or not NDMA would form, classification algorithms were used, and in order to predict how much NDMA would form, regression algorithms were used. In terms of classification algorithms, Logistic Regression, Naïve Bayes, SVM (Gaussian and Linear kernel), and Classification Decision Trees were used. In terms of regression algorithms, Linear Regression and Regression Decision Trees were used. Built in Matlab functions were used for all implementations.

In the case of logistic regression, for the feature vector $x^{(i)}$ and a labelling $y^{(i)}$, the regression finds parameter θ that fits the hypothesis $h_\theta(x) = P(y = 1|x;\theta)$, where the hypothesis is the sigmoid function. θ is fit to minimize the cost function $J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$. For Naïve Bayes, the classifier assumes that all the $x^{(i)}$'s are conditionally independent on y and uses Bayes theorem and joint probability to label a given example (using the MAP decision rule). For SVM, a hyperplane is used to classify data by separating positive and negative examples with a hyperplane. Even though SVM separates data using linear classification, kernels can map inputs to higher dimensional feature spaces. Here, SVM classification was performed using a Linear and Gaussian kernel. The final classification algorithm used was a classification decision tree. Here, the data is subdivided using decision nodes, where for each feature, an optimal threshold is used to label the example.

In terms of regression algorithms used, Linear Regression was used to find θ such that a least-squares cost function is minimized. Regression Decision Trees were used in a similar manner to classification decision trees, however, the final labels were treated as numbers rather than categories.

Results and Discussion In my approach, NDMA prediction was broken up into two sub problems: one, where the main goal was to predict whether NDMA would form given the available features in my dataset (classification), and the second, where in the case that NDMA is detected, whether it would be possible to predict how much NDMA would form. For classification, given that the main concern of treatment utilities is whether NDMA is detected, and not when it isn't detected, for classification performance was measured based precision and recall, as well as the AUC for both the training and test data. In this case, the precision was taken as the probability that samples that were indeed detect samples were labelled as detects by the classification algorithm, and the recall was defined as the probability that samples labelled as Non-detect samples did not have detectable NDMA concentrations. Similarly, the AUC metric (or area under the Receiver Operating Characteristic curve) was appropriate for this dataset, as the AUC should not be sensitive to imbalanced classification: in the case of the AUC as applied here, the closer the resulting AUC value is to 1, the better the algorithm (with AUC's closer to 0.5 indicating a very poor algorithm) (Huang et al., 2005). For regression, the performance was assessed based on the average mean squared error, with lower values indicating better performance.

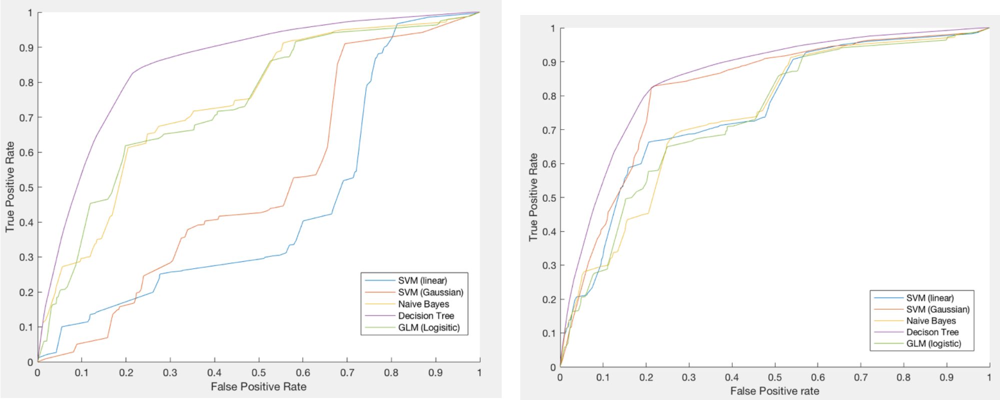


Figure 1. (Left) ROC curves for the unaltered data set and (Right) ROC curves for the most optimal Non-detect: Detect data deletion ratio.

For each sub problem, the goal was to determine which algorithms would perform best given the data. For classification, the skew in the data set posed the greatest Figure 1(left panel). challenge to applying algorithms to the data directly without further processing. As can be seen in for the unaltered data set, based on the ROC curves the algorithm performance varies widely. SVM (both using a linear and Gaussian kernel) are outperformed by Logistic Regression, Naïve Bayes, and the Classification Decision Tree. However, while the decision tree is the best performing option out of these algorithms based on the ROC curves, it is important to note that for classification decision trees the ROC is based on a lower number of points than the remaining algorithms, and the performance may be overstated through the ROC.

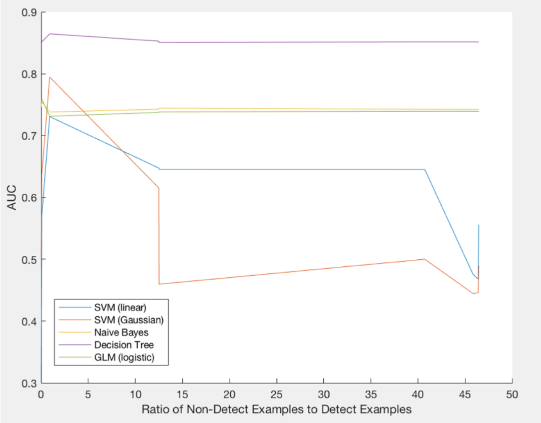
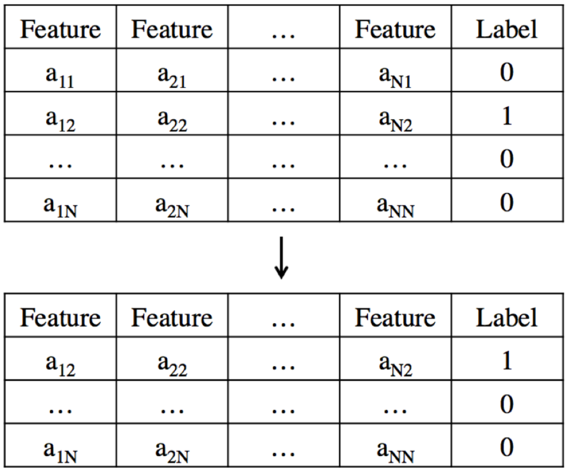


Figure 4. AUC for data deletion using various Detect: Non-Detect ratios.

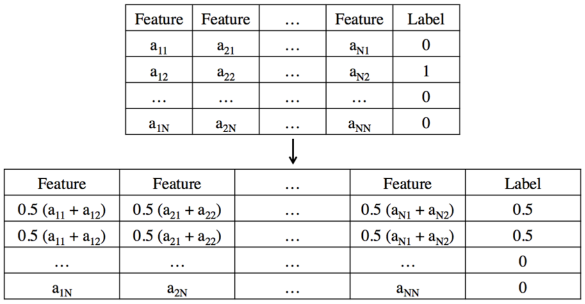


Feature	Feature	…	Feature	Label
a_{11}	a_{21}	…	a_{N1}	0
a_{12}	a_{22}	…	a_{N2}	1
…	…	…	…	0
a_{1N}	a_{2N}	…	a_{NN}	0

↓

Feature	Feature	…	Feature	Label
a_{12}	a_{22}	…	a_{N2}	1
…	…	…	…	0
a_{1N}	a_{2N}	…	a_{NN}	0

Figure 2. Illustration of random data deletion preprocessing.



Feature	Feature	…	Feature	Label
a_{11}	a_{21}	…	a_{N1}	0
a_{12}	a_{22}	…	a_{N2}	1
…	…	…	…	0
a_{1N}	a_{2N}	…	a_{NN}	0

↓

Feature	Feature	…	Feature	Label
$0.5\ (a_{11} + a_{12})$	$0.5\ (a_{21} + a_{22})$	…	$0.5\ (a_{N1} + a_{N2})$	0.5
$0.5\ (a_{11} + a_{12})$	$0.5\ (a_{21} + a_{22})$	…	$0.5\ (a_{N1} + a_{N2})$	0.5
…	…	…	…	0
a_{1N}	a_{2N}	…	a_{NN}	0

Figure 3. Illustration of data averaging preprocessing.

In order to reduce the data skew and improve performance, two data preprocessing methods were used: one which involved random deletion of non-detect samples (Figure 2) and a second, where data points were averaged at random (Figure 3). For random data point deletion, in the training set, samples labelled as non-detects were deleted at random. As can be seen in Figure 4, based on the AUC metric, the best performance was achieved by all algorithms in training when the ratio of the detect to non-detect samples was around 1:2. Both the Linear and Gaussian kernel SVMs were the most affected by point deletion, while the improvement in results for Logistic Regression, Naïve Bayes, and Classification Decision Trees was incremental. Figure 1 (right panel) shows the ROC curves for the most optimal detect to non-detect ratio of samples, and Table 1 shows the corresponding AUC values for the training sets on each algorithm. For data point averaging in the training set, two examples were chosen at random and the values of their features and labels were averaged (for categorical features, the category closest to the midpoint category was chosen for the resulting example). Here, the best results were achieved when the data was averaged 10 times (averaging 2, 10 and 100 times was tested), and the resulting AUC values are shown in Table 2. Here, Linear and Decision Tree regression algorithms were used, and predictions were labeled as detects if the predicted label was greater than 0.5 and as non-detects if the predicted label was less than or equal to 0.5. Overall, the improvement in AUC values from data point averaging was negligible.

Fitting Method	Train All Data	Train With Random Deletion
Generalized Linear Model (logistic)	0.739	0.718
Naive Bayes	0.742	0.735
SVM (linear)	0.556	0.749
SVM (gaussian)	0.511	0.802
Decision Tree	0.851	0.855

Table 1. Training AUC results comparing algorithms used on the original data set to algorithms used along with data deletion with the optimal Non-detect to detect ratios of samples.

Fitting method	All data	With augmentation
Generalized Linear Model	0.664	0.638
Decision Tree	0.533	0.564

Table 2. Training AUC results comparing algorithms used on the original data set to algorithms used along with the optimal data averaging.

Fitting method	Precision (%)	Recall (%)
Generalized Linear Model (logistic)	0.6	18.7
Naive Bayes	0.7	14.5
SVM (linear)	0.75	20.7
SVM (gaussian)	2.4	99
Decision Tree	23	77
Random	1	48

Table 3. Precision and Recall of the test data set using algorithms trained optimal random deletion of the data.

As can be seen in Tables 1 and 2, random point deletion outperformed data point averaging and applying algorithms to the original data set, and therefore, random point deletion was used to make predictions on the test set. The quantitative results thereof can be seen in Table 3. Precision and recall were used to evaluate the test data: unlike the AUC, these two metrics are more helpful in visualizing how useful a given algorithm is in water treatment utility applications. A water treatment utility may not be as concerned by false positives that are predicted, rather, the main concern would be that as many samples as possible that have detectable NDMA are correctly caught by the algorithm and labeled as positive. Here, in terms of recall, the decision tree outperforms the remaining algorithms, however, in terms of precision, the SVM using the Gaussian kernel has the best performance.

Qualitatively, of the features available, only four significantly contributed to the predictive algorithms: namely, the treatment plant size, type of water disinfected, sampling point in the treatment train, and method of chlorine application. The most important parameter for all the algorithms tested was the method of applying chlorine to the water. Most the samples that were chloraminated were labeled as detect samples, while the majority of chlorinated samples were labeled as non-detect samples. This is consistent with accepted scientific knowledge, as the presence of ammonia in water is associated with NDMA formation.

Fitting method	Train Error	Test Error
Generalized Linear Model	10.9%	34.4%
Decision Regression Tree	15.2%	11.7%

Table 4. Average mean squared error for regression algorithms predicting the concentration of NDMA formed.

Regression algorithms were applied to samples in which NDMA was detected, and the average mean squared errors for a Generalized Linear Model and Regression Decision Tree are shown in Table 4. For the training set, the errors from both algorithms were comparable, however, the Decision Tree outperformed the Generalized Linear Model in the test set, implying that the GLM may have been overfitting the training data (the test set error was almost 3 times greater than the training set error for GLM). The average mean squared error for Regression Decision Trees is very close the accepted analytical error (10%) when NDMA is quantified in samples (physical quantification), and is similar to results achieved for other contaminants as discussed in the related work section. These results indicate that machine learning could be applicable to NDMA prediction.

Conclusion and Future Work As can be seen in the results, predicting NDMA from EPA data is possible, even though the results of the classification are not perfect, likely due to the skew in the data set obtained. With random non-detect data deletion in the training set, the more promising algorithms are the SVM that uses a Gaussian kernel and can achieve precision of around 23% and Classification Decision Trees which can achieve 99% recall. More promising results are achieved when predicting the amount of NDMA formed, where Decision Trees, with an average mean squared error of 11.7% can come close to the accepted analytical error when NDMA samples are physically run. The results achieved are therefore a good demonstration that machine learning algorithms could be applied to this specific problem.

However, in the future, before I run additional algorithms, I would like to compile a data set with more features. As applied in this project, only four features significantly contributed to the models. Even though the EPA has been monitoring NDMA formation throughout the US, they do not provide data for water quality treatment parameters, rather, they provide parameters such as disinfection technique used. While this is important information, it does not paint the full picture for NDMA formation. The best fix now would be to map the water quality parameters available online with the correct water treatment plants using GPS coordinates. Further processing of obtained data to reduce skew would also be appropriate, using techniques such as stratified sampling. Cross validation would also likely improve the results.

References

Chowdhury, Shakhawat, Pascale Champagne, and P. James McLellan. "Models for predicting disinfection byproduct (DBP) formation in drinking waters: a chronological review." *Science of the Total Environment* 407.14 (2009): 4189-4206.

Hertzberg, R., et al. "Supplementary guidance for conducting health risk assessment of chemical mixtures." *Washington, DC, Risk Assessment Forum Technical Panel*. 2000.

Hong, H. C., et al. "Modeling of trihalomethane (THM) formation via chlorination of the water from Dongjiang River (source water for Hong Kong's drinking water)." *Science of the Total Environment* 385.1 (2007): 48-54.

Huang, Jin, and Charles X. Ling. "Using AUC and accuracy in evaluating learning algorithms." *IEEE Transactions on knowledge and Data Engineering*17.3 (2005): 299-310.

Singh, Kunwar P., and Shikha Gupta. "Artificial intelligence based modeling for predicting the disinfection by-products in water." *Chemometrics and Intelligent Laboratory Systems* 114 (2012): 122-131.

“Third Unregulated Contaminant Monitoring Rule.” Environmental Protection Agency. Web. www.epa.org

Zheng, Weiwei, et al. "Support vector machine: classifying and predicting mutagenicity of complex mixtures based on pollution profiles." *Toxicology*313.2 (2013): 151-159.