Fluorescence data and Stabilizing Variable Selection for Analysis of Water Treatment

Claire Cui

B00787480

Supervised by Dr.Hong Gu and Paul Bjorndahl

Submitted in partial fulfillment of the requirements for the Degree of Bachelor of Science: Concentrated Honour in Statistics

Dalhousie University

Halifax, Nova Scotia

May 2, 2022

Contents

1 Acknowledgements

In the process of writing this paper, I have received support and help from many professors, friends, and classmates, especially my thesis supervisor, Dr. Hong Gu, and her Ph.D. student Paul Bjorndahl. From the beginning of the topic selection, data collection, and modification, to the final draft, I have received their careful guidance. Weekly meetings with Dr.Gu and Paul to discuss areas for improvement and how to address them have generated tremendous support in the process. Especially in the process of revising the paper, Paul took great pains to give professional guidance not only on the specific content of the paper but also on the format and reference method of the paper.

I am pleasured to pursue my undergraduate degree in the Department of Mathematics and Statistics at Dalhousie University, and I am especially grateful to the professors who have taught me during my undergraduate phase. Their guidance and knowledge shared in my study make me want to pursue my master's degree in Statistics. In addition, I would like to thank my parents for their unwavering support and love.

2 Introduction

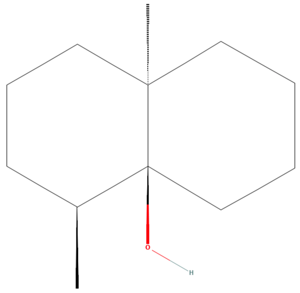

Figure 1: Structure of geosmin[1]

Geosmin [trans-1, 10-dimethyl-trans-9-decalol], a terpenoid alcohol (fig 1), is one of the primary taste and odor-causing compounds in in surface drinking water supplies[2]. Previous work by R.Srinivasan and G.A.Sorial indicates that the main source of geosmin in water is cyanobacteria, the blue-green algae, and can also be the result of the presence of certain types of filamentous bacteria or actinomycetes[3]. Even though geosmin has not been connected with any human health effects, the presence of its earthy and musty odor[4] results in a negative public experience and decreased consumer trust in water[3].

As a result of low human odor threshold of concentration between 1 and $10ng \cdot L^{-1}$, even low levels of geosmin in water are a source of consumer complaints[5]. Therefore, it is critical that drinking water utilities need effective surveillance strategies to respond to geosmin outbreaks. Since water utilities generally do not have direct access to analytical equipment for measuring geosmin, typically samples are sent to external specialized laboratories for analysis. The high cost of these methods and the long turnaround time to obtain data on geosmin concentrations hinders the decision-making of water utilities. To speed up testing and decision-making, a way to reduce analytical complexity and generate predictive tools is needed. One possible approach is to explore the co-occurrence of water quality features, using one measurement as a surrogate for the other. This technique is ideal when indicator properties are easier and more cost-effective to monitor in the field.

The objective of this thesis is to use model-based predictions to develop relatively simple and effective screening methods to improve the assessment of geosmin outbreaks. In this study, stabilized regression analysis was applied for variable selection with fluorescence excitation-emission matrices(FEEM) of wavelength pairs to predict geosmin in water from the JD Kline water supply plant in Nova Scotia, Canada. Our aim is to discuss whether fluorescence spectroscopy can act as a quick and convenient indicator for water quality monitoring assessment in geosmin outbreaks and further to find which specific excitation and emission wavelengths (or peak regions), such as humic-like, fulvic-like, protein-like, and microbial product regions are associated with geosmin.

3 Methods

3.1 Data collection

3.1.1 Water treatment data and pre-preparation

Water samples were collected from the direct filtration system of JD Kline Water Supply Plant in Pockwock Lake, near Halifax, Nova Scotia, Canada. Further detailed information about this full-scale system is described by Vadasarukkai, Yamuna S., et al. in their arcticle[6]. Ambient geosmin was measured in Pockwock Lake water and 20 $ng \cdot L^{-1}$ of geosmin was added to samples prior to advanced oxidation process (AOP) exposure (a water treatment method that effects geosmin) to ensure a 1-log reduction in geosmin was beneath human detection limits.

3.1.2 Bench-scale advanced oxidant-influenced FEEM data

Water samples were treated with either hydrogen peroxide or ozone and exposed to Ultraviolet (UV) fluences via a bench-scale collimated beam unit to degrade natural organic matter (NOM) containing geosmin in water samples. For the chemical oxidants, hydrogen peroxide was added at concentrations of 1 $mg \cdot L^{-1}$ and ozone was added at 10 $mg \cdot L^{-1}$. For the photo-oxidants, fluences of either 100 or 1000 $mJ \cdot cm^{-2}$ were used in both chemical treatment type. All FEEM data was collected immediately after AOP exposure and stored in amber vials. Analysis for both geosmin and FEEM was performed within one week to minimize NOM profile changes and geosmin volatilization.

3.2 Statistical methods

3.2.1 Fluorescence excitation emission matrix (FEEM) analysis

FEEM provided a snapshot of the overall changes in fluorescent NOM, which contains non-fluorescent geosmin, before and after exposure to AOP, including fluorophore components associated with microorganisms that have been shown to be suitable indicators of changes in water quality. The FEEM samples were divided into 12 treatment types and, in total, 273 FEEM samples were analyzed. Table 1 describes the number of observations corresponding to each treatment type. For each sample, there were 121 wavelengths from 240 nm to 600 nm and 125 emission wavelengths from 213.652 nm to 620.467 nm, hence there is a total of $121 \times 125 = 15125$ excitation-emission wavelength pairs. However, wavelength pairs that had zero intensity values over all samples were removed from the dataset, so in total, 6990 wavelength pairs were ultimately used. To visualize FEEM, one sample matrix (i.e. sample number 22 of 273) was selected and shown in Figure 2.

3.2.2 Geosmin analysis

Geosmin samples were assayed as described by Wright, E., et al.[7], it was measured by GC-MS in parallel with FEEM in order to understand the removal efficiency of geosmin in natural water matrices when exposed to various AOPs. The samples were bottled immediately after AOP treatment and stored at $6°C$ in a 1L headspace amber bottle and all of the processing was completed within 7 days after collection. There were 12 types of treatment representing 12 different environments of varied geosmin concentrations. These observed environments can be thought of as perturbations or

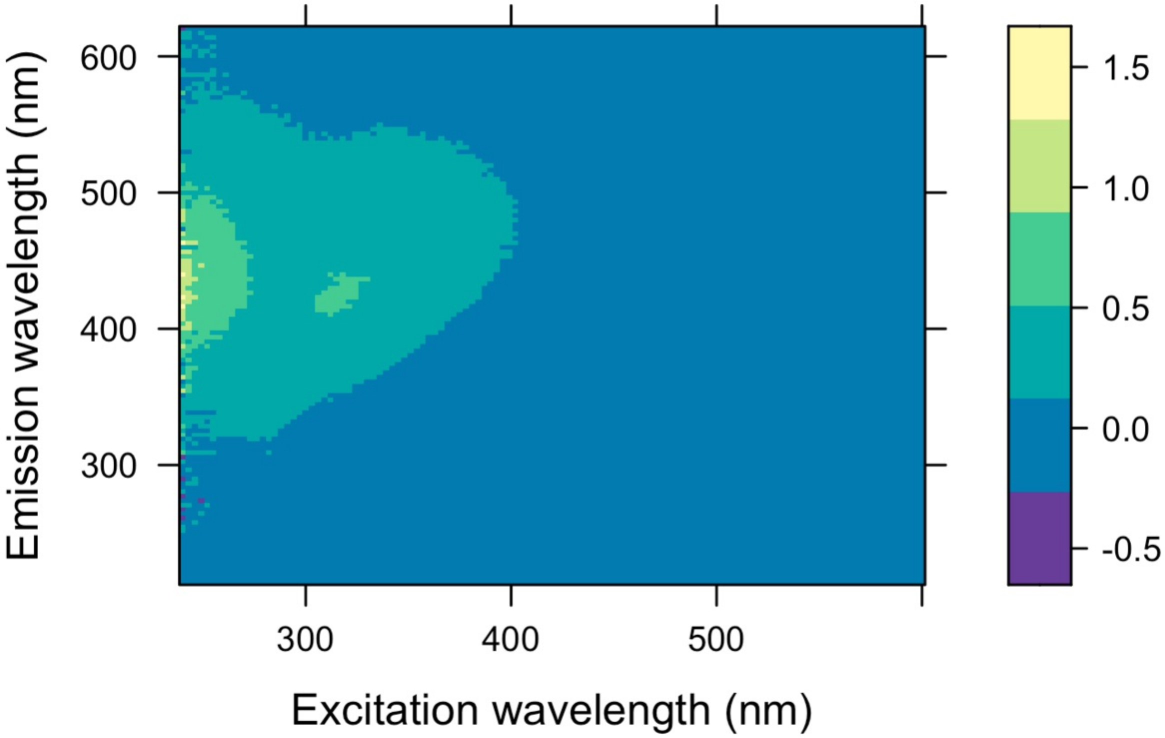

Figure 2: A FEEM from the bench-scale advanced oxidation experimental data

interventions of the natural water samples. Our goal is to make generalizable predictions about unobserved environments based on the 12 observed environments. Geosmin concentration values of the water samples for different treatment types are given in figure 3. Thus, it can be seen that the geosmin levels varied in different environments.

Table 1: Treatment types and corresponding number of observations

	Treatment types	**Number of observations**
1	Raw	10
2	Raw+Spike	12
3	UV High UV High	12
4	H2O2 High	12
5	UV High H2O2 Low	12
6	UV High O3 High	12
7	UV High O3 Low	11
8	UV Low	11
9	UV Low H2O2 High	11
10	UV Low H2O2 Low	11
11	UV Low O3 High	11
12	UV UV Low O3 Low	11

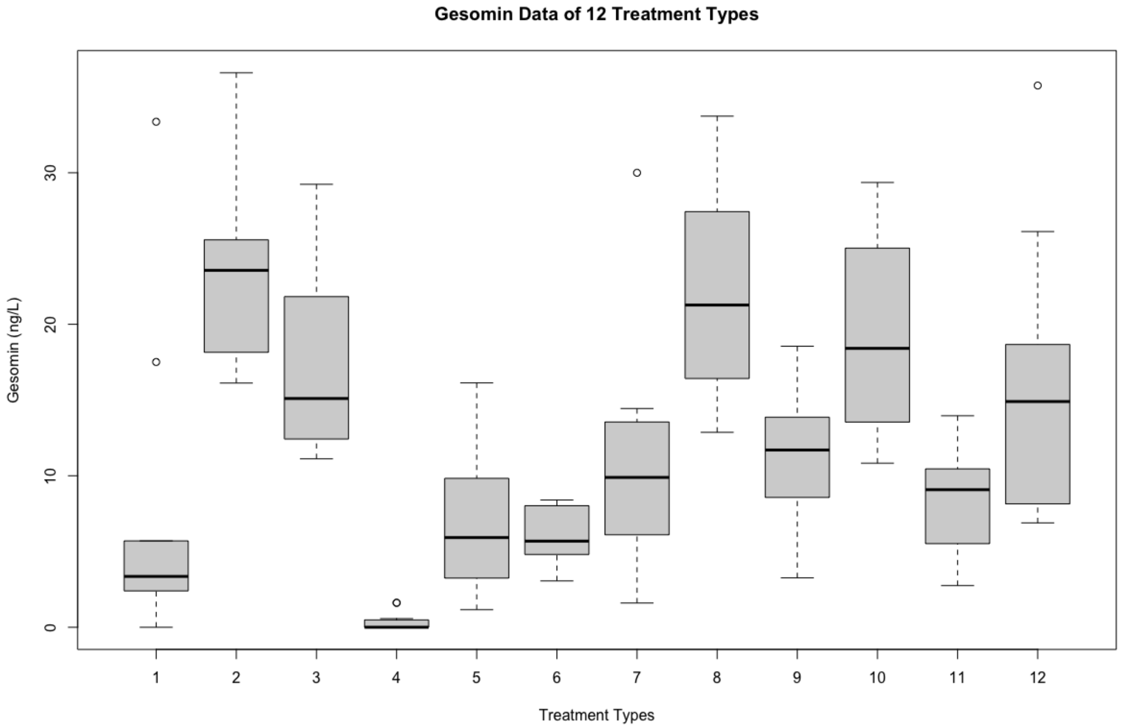

Figure 3: Gesomin Data of 12 Treatment Types. In the x-axis, the 12 numbers correspond to the 12 treatment types in table 1.

3.2.3 Stabilized regression analysis

Stabilized regression (SR) is a multi-environment regression technique, aimed at determining which set of variables leads to a more generalizable model. Let $X = (X_1, ..., X_d)$ be a random vector of predictor variables, let Y be a random response variable, let E_{tot} be a collection of all intervention environments and E_{obs} be a collection of observed environments and subset of E_{tot}. For each environment $e \in E_{tot}$, the variables X_e and Y_e have joint distribution P_e [8]. The goal is to make predictions based on a potentially unobserved environment $e \in E_{tot}$. Here, an assumption about invariance has been made: there exists a subset $S \subseteq \{1, ..., d\}$ such that

$$\mathbb{E}(Y_e | {X_e}^S = x^S) = \mathbb{E}(Y_h | {X_h}^S = x^S) \tag{1}$$

holds for all environments $e, h \in E_{tot}$ and all $x \in X$ [8]. By least squares method, we fit a regression function with a solution to minimize

$$\sum_{e \in E_{obs}} \frac{n_e}{n} \cdot \mathbb{E}((Y_e - f(x_e))^2)$$

where n_e is the number of observations in environment e and f is a function from $\mathbb{X}$ to $\mathbb{R}$ under the constraint that there exists a subset $S \subseteq 1, ..., d$ such that for all $e \in E_{tot}$ and all $x \in \mathbb{X}$

$$f(x) = \mathbb{E}(Y_e | {X_e}^S = x^S) \tag{2}$$

In order to find such a solution, Pfister, N., et al.[8] suggest to simply find the conditional mean based on X^S with a weighted average. They proposed to construct the weights as follows,

$$\hat{w_s} = \begin{cases} \frac{1}{|\hat{\mathbb{O}}|} & \text{if } S \in \hat{\mathbb{O}} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $\hat{\mathbb{O}}$ is a subset of the power set of $\{1, ..., d\}$ that estimates the collection of generalizable and regression optimal sets with respect to E_{obs}. The set $\hat{\mathbb{O}}$ is estimated by two scores calculated for each set $S \subseteq \{1, ..., d\}$. First, a stability score ($s_{stab}(S)$), measures how well the regression satisfies the invariance assumption (1) based on predictors in S. That is, the stability score measures the extent to which predictors in S are generalizable from E_{obs} to E_{tot}. Second, a prediction score ($s_{pred}(S)$), measures how predictive the regression is based on predictors in S. For $s_{stab}(S)$, we set it to be the p value of the test for the null hypothesis that set S satisfy equation (1). For $s_{pred}(S)$, we set it to be the negative mean squared prediction error. Finally, the sets of predictors which are both generalizable and predictive are selected, so the selected sets have both relatively high s_{stab} and relatively high s_{pred}.

With respect to our data of Pockwock Lake FEEM samples, we set the response variable Y to be the geosmin concentration, set predictor variables X to be 6990 excitation-emission wavelengths pairs, and set the environment variable E_{obs} to be the different treatment types. Our goal is to select sets of excitation-emission wavelengths pairs that are both generalizable and regression optimal for predicting geosmin level.

4 Results

4.1 SR analysis with original 12 environments

First, stabilized regression models were applied with all 12 treatment types as environments and results are shown in in figure 4. Based on this result, there were no wavelength pairs selected that were both generalizable and predictive. The wavelength pairs 336-509.648 and 294-590.154 (ex-em) have high selection probability (SR) but zero selection probability (SRdiff), which means these two predictor variables are generalizable but have low predictive performance. The four wavelength pairs shown on the x-axis with a selection probability around 0.65 are more predictive but not necessarily

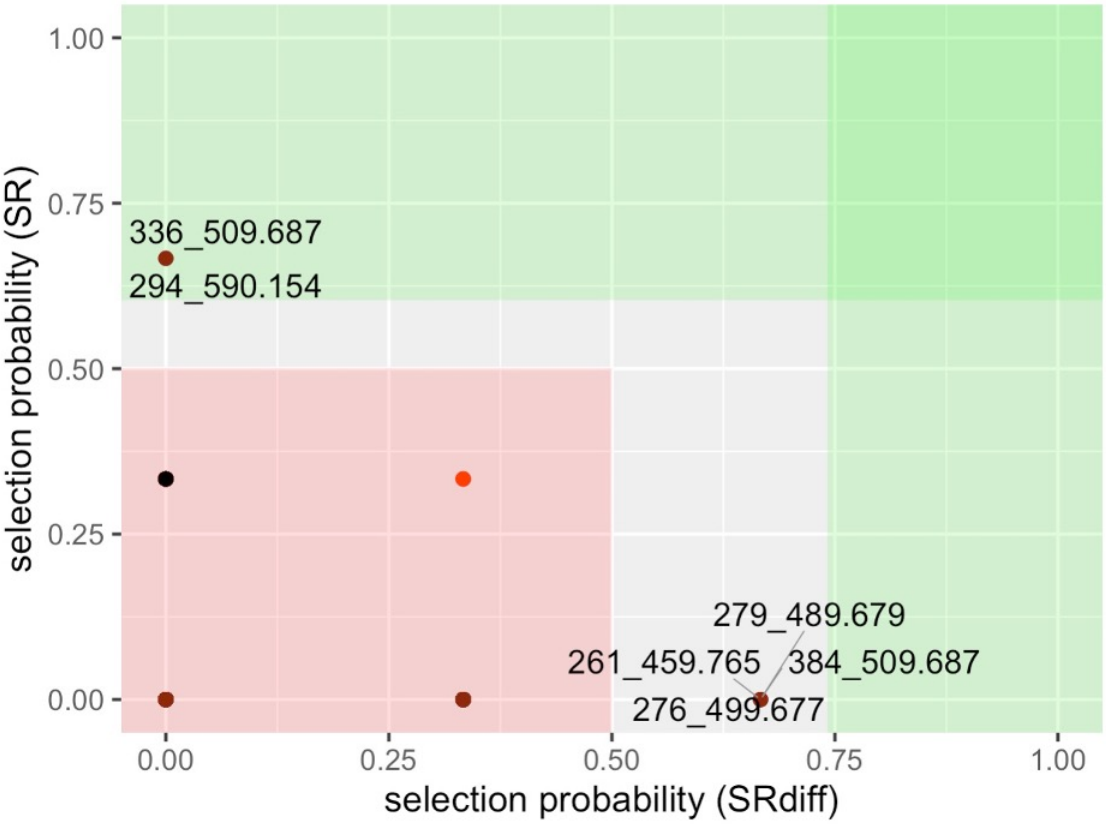

Figure 4: SR analysis on FEEM data with 12 environments

generalizable predictors.

However, no high scoring wavelength pairs selected does not imply there are not such generalizable and predictive pairs within the total 6990 pairs. As discussed in section 3.2.1, there is a total of 273 observations over all 12 environments and each environment only has 10 to 12 observations. Alternatively, to provide more observations within each environment, a grouping of treatment types was carried out in the next section.

4.2 SR analysis with combined 6 environments

The grouping of treatment types has almost twice the number of observations in each environment as shown in table 2. To decrease the total number of predictor variables to be considered, only wavelength pairs around geosmin-associated spectra regions were included in models (Preprint)[9].

Table 2: Grouped treatment types and corresponding number of observations

	Treatment types	**Number of observations**
1	Raw	22
2	UV only	23
3	H2O2 High	23
4	H2O2 Low	23
5	O3 High	23
6	O3 Low	22

The region with excitation wavelength greater than 400 *nm* and emission wavelength grater than 550 *nm* was analyzed firstly and the result of this SR analysis is shown in figure 5. The wavelength pair with excitation 405 *nm* and emission 583.426

nm was of interest. From the previous work[10]–[14], this wavelength pair is between fluorophore regions associated with humic acids and microbial pigments.

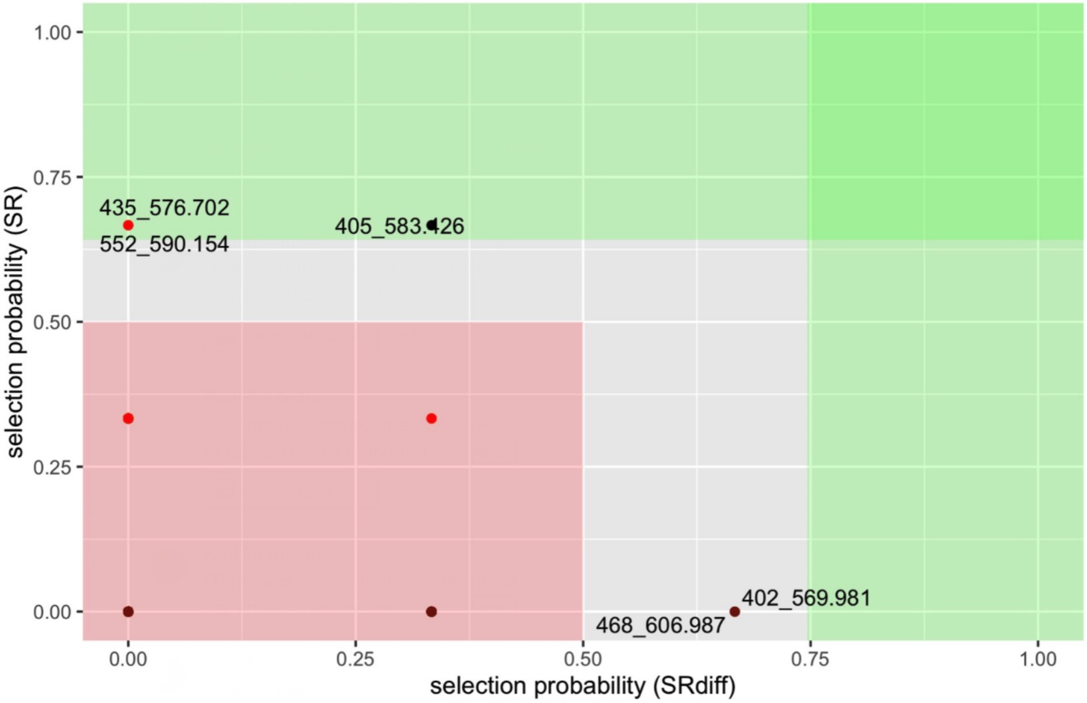

Figure 5: SR analysis with the combined 6 environments on the region with excitation wavelength greater than 400 *nm* and emission wavelength grater than 550 *nm*.

The region with excitation wavelength less than 250 *nm* and emission wavelength greater than 300 *nm* was also analyzed and the result of SR is shown in figure 6. Two wavelength pairs, $246 - 486.589$ and $246 - 483.021$, were selected were selected having the same excitation wavelength 246 *nm*. The pair with emission wavelength 486.589 *nm* scored higher for generalizability while the one with emission wavelength 483.021 *nm* was more predictive. Both have the same fluorophore characteristic that is associated with humic-like products.

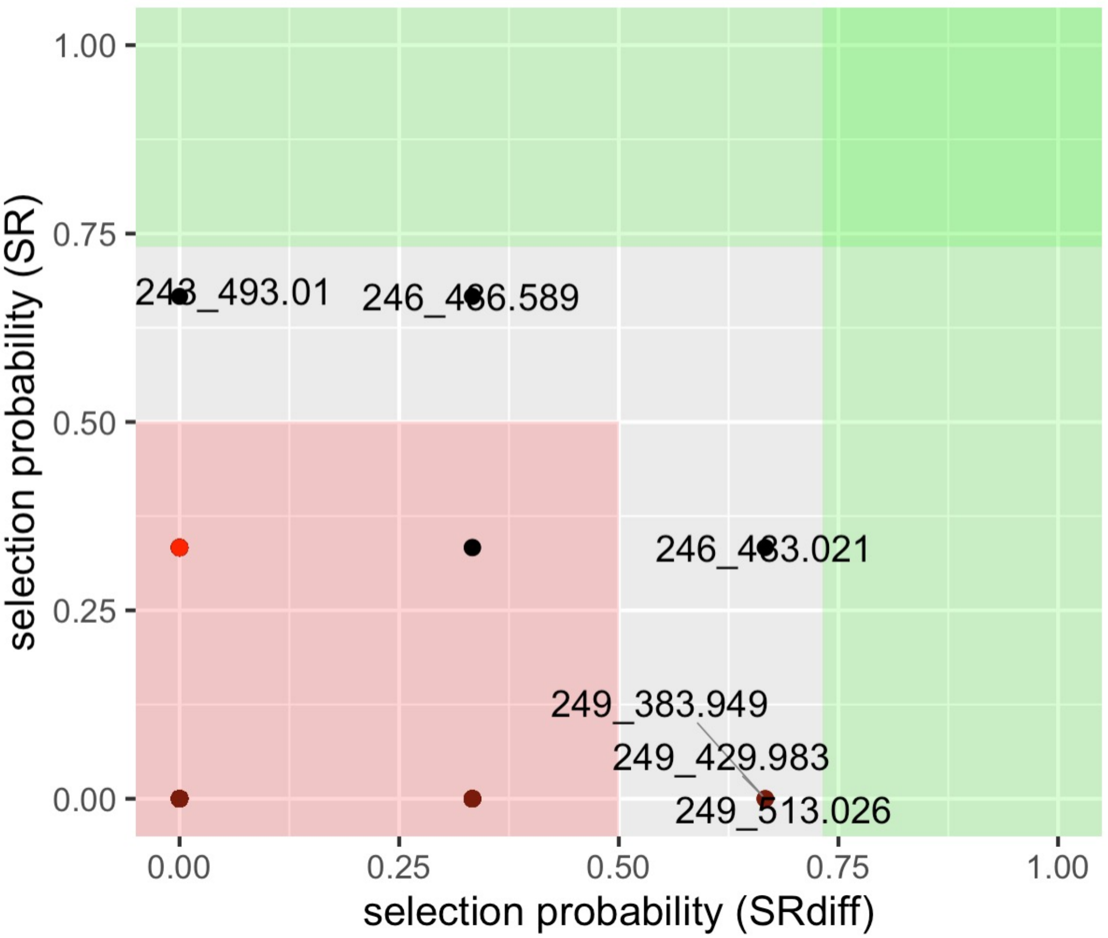

Figure 6: SR analysis with the combined 6 environments on the region with ex < 250 nm and em > 300 nm.

5 Conclusion

During this project, geosmin levels of the water from JD Kline waterworks in Nova Scotia, Canada were characterized by fluorescence excitation-emission matrix (FEEM) wavelengths pairs via stabilized regression (SR) analysis for variable selection. Based on the results from SR analysis, humic-like substances, fulvic acids, and potentially, microbial pigments are related to the geosmin levels. These wavelength pairs could be used as stable predictors to predict the geosmin concentration of source water from an unknown environment setting. However, more FEEM samples are needed to verify the accuracy before being put into industrial use. With more FEEM samples, SR analysis could be done with more environment types, which could yield better stability scores and prediction accuracy for more peak fluorescent regions. Overall, this project provides a general idea to find the stable variables for predicting geosmin levels and a simple way to improve the assessment of the geosmin outbreaks in surface water. Also, other compounds like 2-methyl isoborneol and β-cyclocitral should be studied by this method, so that a complete statistical model could be built for the prediction of taste-and-odor water qualities.

6 Reference

[1] N. C. for Biotechnology Information. “Pubchem compound summary for cid 29746, geosmin”. (2022), [Online]. Available: `https://pubchem.ncbi.nlm.nih.gov/compound/Geosmin.` (visited on 04/10/2022).

[2] F. Juttner and S. B. Watson, “Biochemical and ecological control of geosmin and 2-methylisoborneol in source waters”, *Applied and environmental microbiology*, vol. 73, no. 14, pp. 4395–4406, 2007.

[3] R. Srinivasan and G. A. Sorial, “Treatment of taste and odor causing compounds 2-methyl isoborneol and geosmin in drinking water: A critical review”, *Journal of Environmental Sciences*, vol. 23, no. 1, pp. 1–13, 2011.

[4] G. Izaguirre, C. J. Hwang, S. W. Krasner, and M. J. McGuire, “Geosmin and 2-methylisoborneol from cyanobacteria in three water supply systems”, *Applied and Environmental Microbiology*, vol. 43, no. 3, pp. 708–714, 1982.

[5] J. Yu, Y. Zhao, M. Yang, *et al.*, “Occurrence of odour-causing compounds in different source waters of china”, *Journal of Water Supply: Research and Technology—Aqua*, vol. 58, no. 8, pp. 587–594, 2009.

[6] Y. S. Vadasarukkai, G. A. Gagnon, D. R. Campbell, and S. C. Clark, “Assessment of hydraulic flocculation processes using cfd”, *Journal-American Water Works Association*, vol. 103, no. 11, pp. 66–80, 2011.

[7] E. Wright, H. Daurie, and G. A. Gagnon, “Development and validation of an spe-gc-ms/ms taste and odour method for analysis in surface water”, *International Journal of Environmental Analytical Chemistry*, vol. 94, no. 13, pp. 1302–1316, 2014.

[8] N. Pfister, E. G. Williams, J. Peters, R. Aebersold, and P. Bühlmann, "Stabilizing variable selection and regression", *The Annals of Applied Statistics*, vol. 15, no. 3, pp. 1220–1246, 2021.

[9] P. Bjorndahl, "The identification of novel geosmin-associated fluorescence regions to inform development of monitoring strategies for to outbreaks in drinking water supplies impacted by climate pressures", 2022.

[10] B. F. Trueman, S. A. MacIsaac, A. K. Stoddart, and G. A. Gagnon, "Prediction of disinfection by-product formation in drinking water via fluorescence spectroscopy", *Environmental Science: Water Research & Technology*, vol. 2, no. 2, pp. 383–389, 2016.

[11] Y. Park, S. A. MacIsaac, P. Kaur, M. Brophy, and G. A. Gagnon, "Monitoring the influence of wastewater effluent on a small drinking water system using eem fluorescence spectroscopy coupled with a parafac and pca statistical approach", *Environmental Science: Processes & Impacts*, vol. 23, no. 6, pp. 880–889, 2021.

[12] F. Choo, A. Zamyadi, K. Newton, *et al.*, "Performance evaluation of in situ fluorometers for real-time cyanobacterial monitoring", *H2Open Journal*, vol. 1, no. 1, pp. 26–46, 2018.

[13] D. Vione, C. Minero, and L. Carena, "Fluorophores in surface freshwaters: Importance, likely structures, and possible impacts of climate change", *Environmental Science: Processes & Impacts*, vol. 23, no. 10, pp. 1429–1442, 2021.

[14] L. Moberg, G. Robertsson, and B. Karlberg, "Spectrofluorimetric determination of chlorophylls and pheopigments using parallel factor analysis", *Talanta*, vol. 54, no. 1, pp. 161–170, 2001.