Machine learning techniques for real-time UV-Vis spectral analysis to monitor dissolved nutrients in surface water

Fei, Chuhong, Cao, Xiang, Zang, Devin, Hu, Cindy, Wu, Claudia, et al.

Chuhong Fei, Xiang Cao, Devin Zang, Cindy Hu, Claudia Wu, Eric Morris, Juan Tao, Ting Liu, George Lampropoulos, "Machine learning techniques for real-time UV-Vis spectral analysis to monitor dissolved nutrients in surface water," Proc. SPIE 11703, AI and Optical Data Sciences II, 117031D (5 March 2021); doi: 10.1117/12.2577050

SPIE.

Event: SPIE OPTO, 2021, Online Only

Machine Learning Techniques for Real-time UV-Vis Spectral Analysis to Monitor Dissolved Nutrients in Surface Water

Chuhong Fei, Xiang Cao, Devin Zang, Cindy Hu, Claudia Wu, Eric Morris,
Juan Tao, Ting Liu, George Lampropoulos

A.U.G. Signals Ltd, Toronto, Ontario, Canada

ABSTRACT

Ultraviolet-visible (UV-Vis) spectroscopy is a well-established technique for real-time analyzing contaminants in finished drinking water and wastewater. However, it has struggled in surface water because surface water such as river water has more complex chemical compositions than drinking water and lower concentrations of nutrient contaminants such as nitrate. Previous spectrophotometric analysis using absorbance peak at UV region to estimate nitrate in drinking water performs poorly in surface water because of interference from suspended particles and dissolved organic carbon which absorb light along similar wavelengths. To overcome these challenges, the paper develops a machine learning approach to utilize the entire spectral wavelengths for accurate estimation of low concentration of dissolved nutrients from surface water background. The spectral training data used in this research are obtained by analyzing water samples collected from the US-Canada bi-nationally regulated Detroit River during agricultural seasons using A.U.G. Signals' dual channel spectrophotometer system. Confirmatory concentrations of dissolved nitrate in these samples are validated by laboratory analysis. Several commonly used supervised learning techniques including linear regression, support vector machine (SVM), and deep learning using convolutional neural network (CNN) and long short-term memory (LSTM) network are studied and compared in this work. The results conclude that the SVM with linear kernel, CNN with linear activation function, and LSTM network are the best regression models, which are able to achieve a cross validation root-mean-squared-error (RMSE) less than 0.17 ppm. The results demonstrate effectiveness of the machine learning approach and feasibility of real-time UV-Vis spectral analysis to monitor dissolved nutrient levels in the surface watersheds.

Keywords: Machine learning, Deep learning, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Support Vector Machine (SVM), UV-Vis spectroscopy, Surface water nutrient monitoring

1. INTRODUCTION

Eutrophication, or the increase of undesirable nutrients in water bodies, is a global environmental problem.[1] Recent increases in eutrophic conditions in the Great Lakes have triggered multiple beach closures, lost tourism revenue, and a major water intake shutdown in Toledo, Ohio in 2014.[2] The rapidly growing concerns to the water quality in Great Lakes and surrounding watershed call for a more robust monitoring program, which can provide sufficient data related to water quality changes and respond to pollution in real time.[3,4] However, current monitoring practices, including total phosphorus, nitrogen, sulfur and organic carbon, rely on discrete sampling events and laboratory analysis. This approach is labor intensive, expensive, and can only account for limited information over short time period. Current field-applicable nutrient analyzers test only one parameter per set of instrumentation and are not cost-effective or practical for continuous evaluation of multiple nutrient species. There is an urgent need for inexpensive, low-maintenance analyzers to monitor real-time nutrient trending data in the field.

Ultraviolet-visible (UV-Vis) spectroscopy is a well-established technique for real-time analyzing contaminants in finished drinking water and wastewater.[5] Spectrophotometers have been used to measure nitrate-nitrogen (NO3-N) in wastewater, drinking water, and brackish tidal water[6] but have struggled in rivers. River waters have

Author correspondence: A.U.G. Signals Ltd, 103-73 Richmond St W, Toronto, Ontario, Canada M5H4E8
E-mails: {fei,tony,yi,dhu,claudia,eric.morris,juan,tliu,lampro}@augsignals.com, Telephone: +1(416)9234425

AI and Optical Data Sciences II, edited by Bahram Jalali, Ken-ichi Kitayama,
Proc. of SPIE Vol. 11703, 117031D · © 2021 SPIE · CCC
code: 0277-786X/21/$21 · doi: 10.1117/12.2577050

more complex chemical compositions than drinking water and lower concentrations of NO3-N relative to industrial applications, greenhouses growing operations, and wastewater treatment plants.[7] Previous spectroscopy analysis using the absorbance peak at UV region to estimate nitrate in drinking water performs poorly in surface water because of interference from suspended particles and dissolved organic carbon which also absorb light along similar wavelengths. To overcome these challenges, the paper develops a machine learning approach to utilize the entire spectral wavelengths for accurate estimation of low concentration of dissolved nutrients from surface water background. Machine learning techniques can be employed to take advantage of high dimension of spectral features in order to estimate the nutrients in surface water.

Machine learning techniques have been previously used with UV-Vis spectroscopy for applications such as concentration prediction of organic acids,[8] wastewater quality monitoring,[9] and dissolved organic matter studies.[10] Recent advances in parallelized implementation of machine learning algorithms have expanded their applications in spectroscopy analysis. Here we provide a comparison study among many established multivariate supervised machine learning models in order to enhance precision of nutrient estimation and streamline adoption of UV-Vis spectroscopy to surface water monitoring.

The paper is organized in the following sections. Section 2 will describe data collection procedures and training data structure. Section 3 will elaborate machine learning regression algorithms that are considered and their cross validation results. Conclusion will be discussed in Section 4.

2. TRAINING DATA COLLECTION

Surface water samples were collected on a regular basis at Windsor, Ontario Canada side of the US-Canada bi-nationally regulated Detroit River to generate a UV-Vis spectral dataset and develop algorithms to predict nitrate concentration. Samples were field filtered through 0.45µm nitrocellulose filters and collected in pre-washed polyethylene bottles. Water samples were collected in duplicate, with one sample used for UV-Vis spectral absorption analysis and the other submitted for laboratory analysis to confirm nitrate concentration. spectrophotometer sampling was completed with a Filtrax Sample Filtration System (Hach, Inc) which provides filtration of 0.15 µm to remove suspended particles in the river water. UV-Vis spectral absorption was measured for discreet samples using an A.U.G. Signals' TRITON® spectrophotometer. The spectrophotometer is a dual-channel instrument that records light in the UV-Vis region from 200 nm to 850 nm as light passes through a single optical flow cell containing the water sample. For each water sample, the spectrophotometer is recalibrated using deionized water and nine (9) spectral absorption results with respect to deionized water background were collected. The corrected spectral absorption data were communicated to a desktop computer via the Intelligent Water Surveillance software from A.U.G. Signals Ltd. Fig. 1a depicts 9 spectral samples collected from the water sample taken from the Detroit river on May 21, 2018.

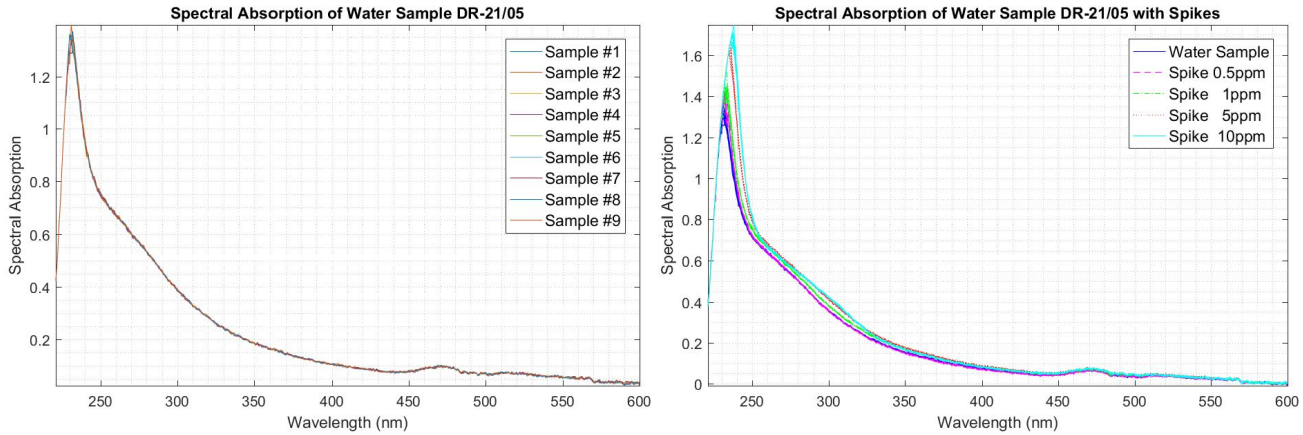
Each water sample is also spiked with nitrate standard solutions and analyzed using the same spectrophotometer to obtain more spectral data associated to the collected water sample. Spiking (or fortification) is a common procedure performed to test whether the response to a sample is the same as that expected from a calibration curve in analytical chemistry. By adding nitrate standard solution of known concentration to the collected water samples, we are able to get more spectral calibration data that define the underlying spectral response to nitrate in background river water sample. Fig. 1b illustrates the spectral absorption results collected from the spiked samples with various concentrations.

In total, there are 31 water samples that were collected from the Detroit river during the Summer season in 2018. From the 31 water samples, there are 1665 valid spectral samples successfully acquired from the spectrophotometer. The UV-Vis spectral region from 220 nm to 600 nm is selected for nitrate analysis so the feature dimension size is 380 since optical resolution of the spectrophotometer apparatus is 1 nm.

3. MACHINE LEARNING APPROACH

3.1 Supervised machine learning algorithms

Given the training data generated from the previous section, we evaluate different supervised learning algorithms in order to accurately estimate nitrate concentration in water samples. The following commonly used regression algorithms have been considered.



(a) Spectral absorption from a water sample (b) Spectral absorption from spiked samples

Figure 1: The spectral training data of water sample "DR-21/05" taken from Detroit River on May 21 2018 and the corresponding spiking solutions with 0.5 ppm, 1 ppm, 5 ppm, and 10 ppm nitrate concentration respectively.

3.1.1 linear regression

Linear regression considers a linear relationship of spectral absorption features to concentration estimation. It is the simplest and widely used regression method in practical applications. From the the Beer–Lambert law,[11] the optical absorption due to nitrate is approximately linear to its concentration, thus naturally linear regression is one of our candidate algorithms.

3.1.2 Partial least squares regression

Partial least squares (PLS) regression[12] is a statistical method that reduces the predictors to a smaller set of uncorrelated components and performs least squares regression on these components. PLS regression is especially useful when the predictors are highly collinear such as spectral measurements in our application.

3.1.3 Ensemble regression

Ensemble regression combines several base regression models and create a stronger model, to achieve better robustness and accuracy over a singular model. We use a popular bagging ensemble method, random forest,[13] which generates averaged prediction of randomized decision trees.

3.1.4 Gaussian process regression

Gaussian process regression (GPR) is a nonparametric Bayesian approach probabilistic regression method where the interpolated values are modeled by a Gaussian process governed by prior covariances, i.e. kernel. GPR has several benefits, working well on small datasets and having the ability to provide uncertainty measurements on the predictions but may lose efficiency in high dimensional feature spaces.

3.1.5 Support vector machine regression

Support vector machine (SVM) regression[14] is considered one of the most robust prediction methods, which can efficiently perform a non-linear regression using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. Two popular kernels: the radial basis function (RBF) and linear kernel are considered in our regression analysis.

3.1.6 Multilayer perceptron regression

A multilayer perceptron (MLP) is a class of feedforward artificial neural network to mimic the working of neuron in human brain for learning of non-linear regression relationship. To distinguish from deep neural network with multiple hidden layers, a multilayer perceptron network only has 1 or 2 hidden layers. In our implementation, we consider a multilayer perceptron network with one hidden layer of size 20.

3.1.7 Convolutional neural network

Convolutional neural network (CNN)[15] is a class of deep neural networks, most commonly applied to analyzing visual imagery. Convolutional neural networks are variants of multilayer perceptrons, designed to exploit the strong spatially local correlation present in natural images. UV-Vis spectrophotometer acquires spectra data using a detection array such as charge coupled device (CCD) or complementary metal oxide semiconductor (CMOS) detector, thus having the same strong spatially local feature correlation as 2D images. In a convolutional neural network, the hidden layers are built from three main types of layers: convolutional layer, pooling layer, and dense/fully-connected layer. Two activation functions are considered in convolutional neural network algorithms: the linear activation function and the rectified linear activation function (ReLU). In our application, we choose a CNN architecture of 5 layers in the sequence of CONV→POOL→CONV→POOL→DENSE.

3.1.8 Long short-term memory network

Long short-term memory (LSTM)[16] is a class of recurrent neural networks which can capture temporal relation in the data. It is learned to recognize input and store it in a long-term state, and extract it when needed. Although training data are not temporal type of data that have time series relation, their correlation along spectral wavelength pixels present similar short-term memory behavior, thus also worth investigation by applying LSTM in spatial data.

3.2 Leave-one-group-out cross validation

Cross validation is a technique for assessing how the results of a statistical analysis will generalize to an independent data set, which is powerful preventative measure against overfitting or selection bias. The above mentioned commonly used regression algorithms are evaluated using cross validation on the nitrate training data collected from the Detroit river water samples.

The objective of developing a nitrate estimator on water sample spectral features is to apply to next water samples such that nitrate nutrient still can be accurately estimated from history water samples even when the background river water is changing over time. Thus, cross validation of the regression algorithm should apply to water samples instead of spectral samples. With this objective in mind, we use the leave-one-group-out cross validation method. Leave-one-group-out is a cross-validation scheme which holds out the training data samples according to a group labelling. In our case, each river water sample can be regarded as a experiment and the entire training data are built on multiple experiments. The Leave-one-group-out method is then based on different experiments (i.e., groups) and each training set is thus constituted by all the training data except the one group related to a specific water sample. This water sample group information such as "DR-21/05" can be used to encode and group the corresponding spectral sample folds for cross validation. This leave-one-group-out scheme is different from k-fold or holdout methods where the original training data samples are randomly partitioned into sub-samples. Thus k-fold or holdout methods only validate the regression models under spectral reading noise among spectral samples, but the leave-one-group-out scheme validates them under background water variations among water samples.

4. RESULTS AND CONCLUSION

The supervised machine learning algorithms described in Section 3.1 are implemented using Python Deep Learning Library TensorFlow or MATLAB Statistics and Machine Learning Toolbox. The popular root-mean-squared-error (RMSE) between the measured concentration and the predicted concentration is selected as the performance metric to evaluate these machine learning algorithms using the leave-one-group-out cross validation scheme.

Table 1 shows the corresponding RMSE metric values of various regression algorithms. From the results, we can see that random forest, and SVM with RBF kernel algorithms perform poorly with their cross validation RMSEs greater than 1 ppm. Linear regression, partial least squares, SVM with linear kernel, multilayer perceptron, convolutional neural network and long short-term memory network have good performance with RMSEs less than 0.3 ppm. In particular, the SVM with linear kernel, convolutional neural network with linear activation function, and LSTM are the best, which can achieve RMSE less than 0.17 ppm. The best prediction results of RSME 0.1637 ppm achieved by the convolutional neural network model with linear activation function are illustrated in Fig. 2 with comparison to ground truth measurements.

Table 1: Cross validation results on commonly used regression algorithms.

Regression Algorithm	**CV RMSE (ppm)**	**Remarks**
Linear regression	0.2825	Good
Partial least squares Regression	0.2822	Good
Random forest regression	1.0553	Poor
Gaussian process regression (GPR)	0.7890	Fair
SVM regression (Linear kernel)	0.1659	Excellent
SVM regression (RBF kernel)	1.1346	Poor
Shallow neural network Regression	0.2509	Good
CNN Regression (Linear activation function)	**0.1637**	Excellent
CNN regression (ReLU activation function)	0.2049	Good
Long short-term memory (LSTM) regression	0.1690	Excellent

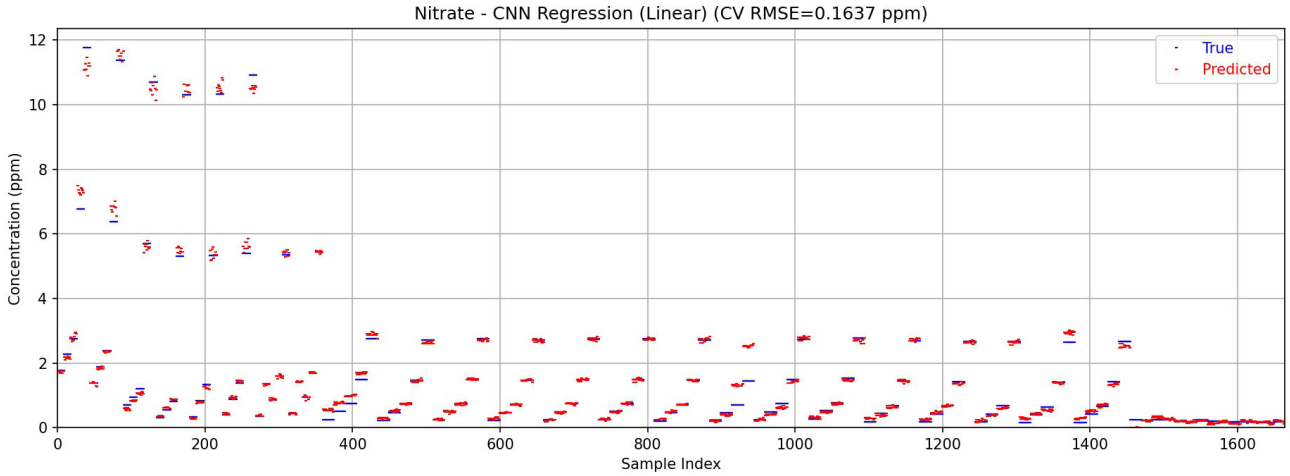


Figure 2: Prediction results of the CNN regression algorithm with linear activation function with comparison to ground truth measurements

From the regression results, we can also observe that linear regression, SVM with linear kernel, or CNN with linear activation function is generally better than their non-linear counterparts. This is consistent to the approximately linear relationship of optical absorption to nitrate concentration, which under ideal situation is governed by the Beer–Lambert law.

The results demonstrate effectiveness of the machine learning approach and feasibility of real-time UV-Vis spectra analysis to monitor dissolved nutrient levels in the surface watersheds. The two best regression candidates, SVM with linear kernel and CNN with linear activation function, will be implemented in A.U.G. Signals' online nitrate analyzer prototype using a UV-Vis spectrophotometer system to continuously monitor dissolved nutrient levels in the surface watersheds in real time.

Further improvement of nitrate regression models is possible by collecting more training data to contain seasonal river samples. Investigation of possible interfering contaminants such as dissolved organic carbon (DOC) and their spectral behaviour may also further increase nitrate regression accuracy in surface water samples.

ACKNOWLEDGMENTS

The authors would like to thank the project partner, Prof. Scott Mundle's research team from the Great Lakes Institute for Environmental Research at the University of Windsor, for collecting and providing the raw data

used in the paper. The authors also want to thank the funding support from Advancing Water Technologies (AWT) program by the Ontario Water Consortium (OWC) and Engage program by the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

[1] Chislock, M. F., Doster, E., Zitomer, R. A., and Wilson, A. E., "Eutrophication: Causes, consequences, and controls in aquatic ecosystems," *Nature Education Knowledge* **4**, 10 (2013).

[2] Steffen, M. M., Davis, T. W., McKay, R. M. L., Bullerjahn, G. S., Krausfeldt, L. E., Stough, J. M., Neitzey†, M. L., Gilbert, N. E., Boyer, G. L., Johengen, T. H., Gossiaux, D. C., Burtner, A. M., Palladino, D., Rowe, M. D., Dick, G. J., Meyer, K. A., Levy, S., Boone, B. E., Stumpf, R. P., Wynne, T. T., Zimba, P. V., Gutierrez, D., , and Wilhelm, S. W., "Ecophysiological examination of the Lake Erie microcystis bloom in 2014: Linkages between biology and the water supply shutdown of Toledo, OH," *Environmental Science & Technology* **51**, 6745–6755 (2017).

[3] "The 2012 great lakes water quality agreement. annex 4–nutrients." https://binational.net/annexes/a4/ (2012). Accessed: 2021-02-01.

[4] Byappanahalli, M. N., Nevers, M. B., Shively, D. A., Spoljaric, A., and Otto, C., "Real-time water quality monitoring at a great lakes national park," *Journal of Environmental Quality* **47**, 1086–1093 (2018).

[5] Rieger, L., Langergraber, G., Thomann, M., Fleischmann, N., and Siegrist, H., "Spectral in-situ analysis of NO2, NO3, COD, DOC and TSS in the effluent of a WWTP," *Water Science & Technology* **50**, 143–152 (2004).

[6] Etheridge, J. R., Birgand, F., Osborne, J. A., Osburn, C. L., II, M. R. B., and Irving, J., "Using in situ ultraviolet-visual spectroscopy to measure nitrogen, carbon, phosphorus, and suspended solids concentrations at a high frequency in a brackish tidal marsh," *Limnology and Oceanography, Methods* **12**, 10–22 (January 2014).

[7] Maguire, T. J., Wellen, C., Stammler, K. L., and Mundle, S. O. C., "Increased nutrient concentrations in Lake Erie tributaries influenced by greenhouse agriculture," *Science of The Total Environment* **633**, 433–440 (August 2018).

[8] Wolf, C., Gaida, D., Stuhlsatz, A., Ludwig, T., McLoone, S., and Bongards, M., "Predicting organic acid concentration from UV/vis spectrometry measurements - a comparison of machine learning techniques," *Transactions of the Institute of Measurement and Control* **35**, 5–15 (2013).

[9] Carré, E., Pérot, J., Jauzein, V., Lin, L., and Lopez-Ferber, M., "Estimation of water quality by UV/Vis spectrometry in the framework of treated wastewater reuse," *Water Science & Technology* **76**, 633–641 (2017).

[10] Li, P. and Hur, J., "Utilization of UV-Vis spectroscopy and related data analyses for dissolved organic matter (DOM) studies: A review," *Critical Reviews in Environmental Science and Technology* **47**, 131–154 (2017).

[11] Ingle, J. D. J. and Crouch, S. R., [*Spectrochemical Analysis*], Prentice Hall, New Jersey (1988).

[12] Cheng, J. and Sun, D., "Partial least squares regression (PLSR) applied to NIR and HSI spectral data modeling to predict chemical properties of fish muscle," *Food Engineering Reviews* **9**, 36–49 (2017).

[13] Breiman, L., "Random forests," *Machine Learning* **45**, 5–32 (2001).

[14] Chang, C.-C. and Lin, C.-J., "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology* **2**, 27:1–27:27 (2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[15] LeCun, Y., Bengio, Y., and Hinton, G., "Deep learning," *Nature* **521**, 436–444 (2015).

[16] Hochreiter, S. and Schmidhuber, J., "Long short-term memory," *Neural Computation* **9**(8), 1735–1780 (1997).