A Supervised Learning Approach to Water Quality Parameter Prediction and Fault Detection

Kathleen Joslyn
Department of Mathematics & Statistics
Portland State University
Portland, OR, USA
kjoslyn@pdx.edu

John Lipor
Department of Electrical & Computer Engineering
Portland State University
Portland, OR, USA
lipor@pdx.edu

***Abstract*—Water quality parameters such as dissolved oxygen and turbidity play a key role in policy decisions regarding the maintenance and use of the nation's major bodies of water. In particular, the United States Geological Survey (USGS) maintains a massive suite of sensors throughout the nation's waterways that are used to inform such decisions, with all data made available to the public. However, the corresponding measurements are regularly corrupted due to sensor faults, fouling, and decalibration, and hence USGS scientists are forced to spend costly time and resources manually examining data to look for anomalies. We present a method of automatically detecting such events using supervised machine learning. We first present an extensive study of which water quality parameters can be reliably predicted, using support vector machines and gradient boosting algorithms for regression. We then show that the trained predictors can be used to automatically detect sensor decalibration, providing a system that could be easily deployed by the USGS to reduce the resources needed to maintain data fidelity.**

***Index Terms*—supervised learning, support vector regression, gradient boosting, water quality**

I. Introduction

Persistent monitoring of water quality factors such as dissolved oxygen (DO) and turbidity is an important task in science and engineering, with stakeholders ranging from scientific research to municipalities where policy decisions must be made. In particular, the United States Geological Survey (USGS) maintains a host of sensors throughout the country, and scientists at the USGS go to great lengths to maintain the fidelity of this data and make it publicly available. However, maintaining this vast network is a costly process, in large part due to the fact that the phenomena of interest can only be measured using costly in-situ sensors that require regular inspection and maintenance.

While significant advances to sensor technology have ushered in the age of "big data," numerous sources of error impact the resulting measurements, including those from sensor failure, decalibration, and bio-fouling. To combat these issues and maintain high-fidelity data, the USGS manually examines records for such corruptions, sending engineers to perform maintenance when a fault is detected. This is a costly process for data quality engineers, and an automated system to detect sensor faults would be highly valuable to both the USGS and other organizations collecting water quality data on a massive scale.

In this work, we show that the wealth of validated data provided by the USGS through its online portal that can be used to train supervised machine learning algorithms, which can then be used to detect anomalies in future data resulting from the sources listed above. Predicting water quality factors using supervised learning is an increasingly common task, but existing studies either focus on predicting a single parameter (e.g., dissolved oxygen (DO)) or rely on very small datasets with results that may not be indicative of general underlying trends. As mentioned, the USGS provides historical data for all its sensor sites, dating as far back as thirty years in some cases. In particular, we focus on the sensor suite near the Morrison Bridge on the Willamette River downtown Portland, OR.

Our contributions are as follows. We present an extensive study of water quality factor prediction, examining which of the measured factors can be reliably predicted by the others using two popular algorithms for supervised regression—support vector regression (SVR) [1] and gradient boosting (via the XGBoost implementation [2]). Our results show that six of the nine factors considered can be predicted with an R^2 value exceeding 0.9. We then show that decalibration in the form of multiplicative scaling can be reliably detected by using the trained regressor.

II. Problem Formulation & Related Work

In this work, we consider a total of nine parameters that are used when studying water quality: dissolved oxygen (DO), pH balance, chlorophyll, temperature, specific conductivity, turbidity, cyanobacteria, nitrate, and fluorescent dissolved oxygen matter (fDOM). A summary of these factors, including their units of measurement and a brief statistical summary, are given in Table I.

As stated in the introduction, we utilize the SVR (via the `scikit-learn` Python library [3]) and XGBoost algorithms. Both implementations are freely available and easily accessible to any data scientist with knowledge of Python. More importantly, these algorithms are shown to perform well across a large variety of datasets [4], and both are computationally efficient compared to popular methods from deep learning. We begin by forming a training dataset $\{X_i, Y_i\}_{i=1}^{N_{train}}$, where $Y_i \in \mathbb{R}$ is the (ith example of) the factor of interest to be predicted (e.g., DO) and $X_i \in \mathbb{R}^8$ is the feature vector

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

corresponding to the other eight factors considered in this study. Our goal is to learn a function $f : \mathbb{R}^8 \to \mathbb{R}$ such that $f(X_i) =: \hat{Y}_i \approx Y_i$ for all (X_i, Y_i) pairs in the training set. However, we also wish for f to generalize to data not in the training set (i.e, *test* data), and hence setting the regularization parameter in each algorithm is an important procedure that prevents overfitting the training data. We consider all nine factors as regression targets, i.e., we solve nine distinct regression problems. Further details on how the algorithms were trained are given in Section III.

A. Related Work

Existing studies on predicting water quality factors such as those considered here tend to focus either on predicting a single parameter with two or more algorithms or by predicting a small number of parameters with a single algorithm. These also tend to rely on much smaller datasets than that considered here; the previous studies [5]–[10] considered a range of examples from 132-2063. In contrast, we study a total of 52,563 examples collected over a period of three years, allowing for more accurate predictions and a more thorough understanding of the intrinsic relationships between parameters.

In [6], the authors report a high prediction for DO using SVR, linear genetic programming, and two types of artificial neural networks (multilinear perceptron and radial based function). They demonstrate that various regression algorithms can be used to gain a deeper understanding of the physical relationships between parameters, using a dataset of 2063 examples. In [5], the authors use support vector machines to classify DO values into three distinct bins (high, medium, and low levels) but only considered a dataset of 147 total examples. The authors of [7] employ the SVR, general regression neural network, back propagation neural network (BP-NN), and multilinear regression algorithms, but only consider predicting DO using a dataset of 240 total examples.

In [8], faecal coliform is predicted using a genetic programming (GP) algorithm as well as a variation of the SVM algorithm, LS-SVM. Finally, [10] predicts both DO and water temperature with SVR, using a genetic algorithm to guide the parameter tuning.

The end goal of this study is to bring to life a use for machine learning in water quality monitoring and to show that machine learning has an important role to play in this area of the sciences. Through a combination of accessible algorithms and a thorough investigation of a large dataset, we show that supervised learning has the potential to significantly streamline the data validation process at the USGS, saving valuable resources for the organization and improving the fidelity of their data.

III. Water Quality Factor Prediction

In this section, we describe our methodology for performing supervised regression and anomaly detection using SVR and XGBoost. As with all real-world datasets, the data we consider contains missing entries and a number of outliers, which can inhibit the performance of the supervised regression algorithms. Initial runs of the algorithms utilized all 14 parameters measured at the Morrison Bridge site. This resulted in poor prediction accuracy due to lack or inconsistency of data with five of the parameters. These parameters include discharge (tide filtered and not), gauge height, sensor depth, and mean water velocity. Even with the relative consistency of data in the final nine parameters, there were spots in data collection process that resulted in little to no data points collected. These missing data points were filled with the mean values of each individual water quality parameter.

Factor	Mean	Maximum	Minimum
dissolved oxygen (mg/L)	11.17	14.7	3.50
pH balance (std. unit)	7.33	8.80	6.80
chlorophyll (ug/L)	1.95	41.30	0.20
temperature (deg. C)	13.57	25.10	2.10
specific conductivity (uS/cm)	77.30	134	51
turbidity (FNU)	8.35	120	0.30
cyanobacteria (ug/L)	0.35	1.48	-0.05
nitrate (mg/L)	0.66	2.30	0.0
fDOM (ppd QSE)	6.89	25.60	-0.46

TABLE I
SUMMARY OF WATER QUALITY FACTORS STUDIED.

Factor	Mean	Maximum	Minimum
dissolved oxygen	0.9789	0.9835	0.9756
pH balance	0.8654	0.8746	0.8557
chlorophyll	0.8820	0.8972	0.8627
temperature	0.9843	0.9880	0.9818
specific conductivity	0.8784	0.8946	0.8687
turbidity	0.8145	0.8367	0.7932
cyanobacteria	0.8852	0.8880	0.8811
nitrate	0.9469	0.9493	0.9443
fDOM	0.9357	0.9434	0.9268

TABLE II
SVR PREDICTION ACCURACY RESULTS (R^2)

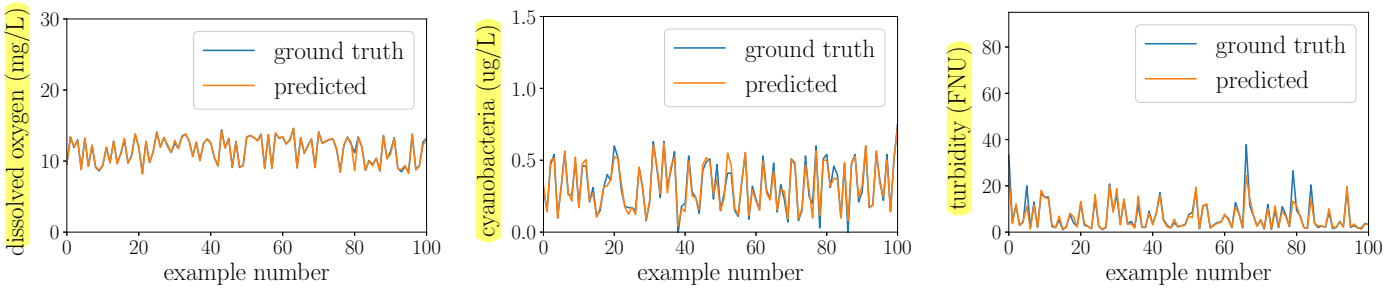
A. Supervised Regression

As described above, we considered 52,563 examples of all parameters. We then performed a 90% / 10% training/test data split, where the training data is used to learn the corresponding regression function (using SVR or XGBoost), and the test data is used to evaluate the accuracy of the resulting regression. In contrast to some studies mentioned in Section II, we selected our training and test sets *randomly*, i.e., we did not use the first 90% of the data to predict the final 10%. To quantify algorithm performance, we use the R^2 coefficient, defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{N_{test}} \left(\hat{Y}_i - Y_i\right)^2}{\sum_{i=1}^{N_{test}} (Y_i - \bar{Y})^2},$$

where $\bar{Y} = \sum_{i=1}^{N_{test}} Y_i$ is the sample mean of the test data, and a value of $R^2 = 1$ implies a perfect prediction. Since the training/test split was performed randomly, we trained each algorithm for 10 independent instances and report the mean, maximum, and minimum R^2 values.

The results of both algorithms are given in Tables II and III. The best performing parameters depended on the algorithm



(a) (b) (c)

Fig. 1. Subset of ground truth and predicted values for (a) dissolved oxygen using SVR ($R^2 = 0.9810$), (b) cyanobacteria using XGBoost ($R^2 = 0.9429$), and (c) turbidity using SVR ($R^2 = 0.8283$). The training and test sizes are $N_{train} = 47,306$ and $N_{test} = 5,257$.

Factor	Mean	Maximum	Minimum
dissolved oxygen	0.9446	0.9471	0.9414
pH balance	0.8164	0.8286	0.8042
chlorophyll	0.8794	0.8869	0.8673
temperature	0.9849	0.9857	0.9836
specific conductivity	0.8857	0.8896	0.8760
turbidity	0.9376	0.9493	0.9195
cyanobacteria	0.9493	0.9486	0.9409
nitrate	0.9160	0.9197	0.9128
fDOM	0.8831	0.8886	0.8769

TABLE III
XGBOOST PREDICTION ACCURACY RESULTS (R^2)

used. For SVR, temperature and dissolved oxygen performed well with an average prediction accuracy score of 0.9889 and 0.9789, respectively. For XGBoost, performance was similar for temperature, with a prediction accuracy of 0.9849. Dissolved oxygen and three other parameters performed above prediction accuracy of 0.9. Fig. 1 shows example predictions of (a) DO using SVR, (b) cyanobacteria using XGBoost, and (c) turbidity using SVR. The worst prediction accuracies were likewise different for both algorithms. For SVR, turbidity performed poorly in relation to the other water quality parameters. To better understand why turbidity performed so poorly in comparison to the other parameters, we studied a plot of the predicted values, shown in Fig. 1(c). The plot reveals that the predicted values track the actual values well except in the cases of large spikes in the data, which SVR has trouble predicting. For XGBoost, the poorest performance was pH balance, but the reason for this poor performance was not apparent from the plotted predictions.

While SVR had more parameters that exceeded the $R^2 = 0.9$ prediction accuracy threshold, the algorithm itself took much longer to process the large dataset than XGBoost. For each run of the SVR algorithm, the training time ranged from two to four minutes on a personal computer. In contrast, the XGBoost algorithm averaged ten to thirty seconds per run. However, while the gradient boosting algorithm proved to be fast, SVR had more parameters that received prediction accuracy scores above 0.9.

The above results are encouraging in that they indicate strong correlations among the measured water quality factors. However, the question remains as to whether the trained regression algorithms can be used for the purpose of anomaly detection. In what follows, we show that this is indeed the case. Moreover, our empirical results suggest that fault detection is possible even when the initial R^2 score is not particularly high.

B. Fault Detection

The ultimate aim of this study is to produce a system that would make detecting sensor faults a more streamlined process. Intuitively, if the prediction of a parameter is within a reasonable range (e.g., close to those achieved in the previous section), then it is likely that the sensor itself is performing well. If, however, the prediction accuracy suddenly begins to perform poorly, that is a strong indication that there is a problem with calibration, sensor failure, or bio-fouling. To show that the regression algorithm can detect a sudden change in data, a scalar value was applied to the testing data. This is a way to mimic a sensor failure due to short-term weather events such as storms. A sudden drop in prediction accuracy could be utilized as a way to warn engineers of a problem with the sensors. This could create a way to detect sensor failures in a time sensitive way, potentially saving time and money for the USGS.

We now show that the regression functions learned in the previous section can be used to reliably detect sensor faults in the form of a multiplicative scaling. Formally, we train a regressor on N_{train} *successive* samples, taking the remaining examples as the test data, in a manner similar to time-series prediction. Note however, that no temporal aspects have been considered when training the regression algorithms, and doing so is an important topic of future work. We then choose a point in time t samples into the test set and apply a multiplicative scaling, so that

$$\tilde{Y}_i = \begin{cases} Y_i & i < t \\ \alpha Y_i & i \geq t \end{cases}.$$

We use the updated values $\tilde{Y}_i$ as a test set. This scaling model is motivated by previous studies in sensor calibration [11]. The key question is whether the resulting prediction score changes significantly after applying such a scaling.

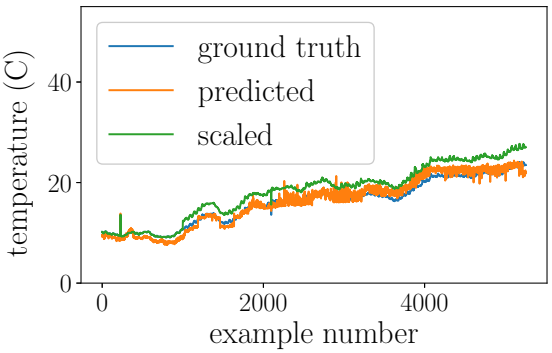


Fig. 2. Subset of ground truth and predicted values for (a) dissolved oxygen using SVR ($R^2 = 0.9810$), (b) cyanobacteria using XGBoost ($R^2 = 0.9429$), and (c) turbidity using SVR ($R^2 = 0.8283$). The training and test sizes are $N_{train} = 47,306$ and $N_{test} = 5,257$.

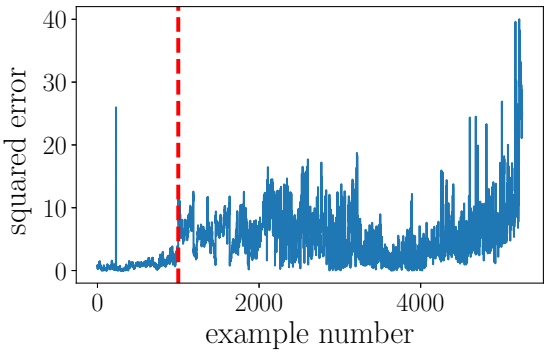


Fig. 3. Subset of ground truth and predicted values for (a) dissolved oxygen using SVR ($R^2 = 0.9810$), (b) cyanobacteria using XGBoost ($R^2 = 0.9429$), and (c) turbidity using SVR ($R^2 = 0.8283$). The training and test sizes are $N_{train} = 47,306$ and $N_{test} = 5,257$.

The goal is to see if the regression algorithm is able to pick up on the change in testing data. The prediction score for the unscaled data is compared to the prediction score for the scaled data to determine if the change in accuracy is detected. We considered the case of predicting temperature using XGBoost, since temperature was among the top factors in terms of performance and the low computational cost of XGBoost makes it a strong candidate for running in real time. Scalar values of 1.15, 1.5, 3, and 10 were considered; however, all values above 1.15 were trivial to detect, and hence we discuss only the case of $\alpha = 1.15$. The results for this test are promising. The prediction of the regular data when ran at a training and testing split of 90% / 10% received a prediction score of 0.9449. Keeping this as a comparative value, a scalar of $\alpha = 1.15$ was applied after the first 1000 samples of the testing data to simulate a fault in the sensor. The resulting prediction score was 0.8248, which indicates a performance reduction greater than 10%. This actively shows the anomaly in the scaled data and reflects the ability of the regression algorithm to detect scaled changes in the data.

While the reduction in R^2 value is encouraging, the users of such a fault detection method may wish to detect anomalies in real time, rather than averaged over a window. Fig. 3 shows the squared error between the prediction and the scaled data as a function of the example number (i.e., time). The dashed red line indicates the point where the scaling factor was applied. The figure clearly demonstrates a visible increase in squared error after the time at which the linear scaling was applied. Hence, we conclude that supervised regression has strong potential for use in fault detection for water quality measurement systems.

IV. Conclusion

In this work, we have shown that two supervised regression algorithms (support vector regression and gradient boosting) can be used to predict a variety of water quality factors with a high degree of accuracy. We then demonstrated how such a trained regressor could be used to detect sensor faults in the form of a linear scaling, which could occur in practice as a result of weather events. While the results here are promising, there is still a rich opportunity for research in this area. First, would be interesting to include the temporal correlations in the data when predicting future values. As a second line of work, it may be more realistic to apply a scaling factor that grows linearly over time, simulating a gradual sensor drift.

References

[1] D. Basak, S. Pal, and D. C. Patranabis, "Support vector regression," *Neural Information Processing-Letters and Reviews*, vol. 11, no. 10, pp. 203–224, 2007.

[2] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[4] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems," *J. Mach. Learn. Res*, vol. 15, no. 1, pp. 3133–3181, 2014.

[5] S. Malek, M. Mosleh, and S. M. Syed, "Dissolved oxygen prediction using support vector machine," *Int. Comput. Inf. Sci. Eng. World Acad. Sci. Eng. Technol*, vol. 8, no. 1, pp. 46–50, 2014.

[6] E. Olyaie, H. Z. Abyaneh, and A. D. Mehr, "A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in delaware river," *Geoscience Frontiers*, vol. 8, no. 3, pp. 517–527, 2017.

[7] X. Ji, X. Shang, R. A. Dahlgren, and M. Zhang, "Prediction of dissolved oxygen concentration in hypoxic river systems using support vector machine: a case study of wen-rui tang river, china," *Environmental Science and Pollution Research*, vol. 24, no. 19, pp. 16 062–16 076, 2017.

[8] M. S. Jadhav, K. C. Khare, and A. S. Warke, "Water quality prediction of gangapur reservoir (india) using ls-svm and genetic programming," *Lakes & Reservoirs: Research & Management*, vol. 20, no. 4, pp. 275–284, 2015.

[9] Y. Park, K. H. Cho, J. Park, S. M. Cha, and J. H. Kim, "Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, korea," *Science of the Total Environment*, vol. 502, pp. 31–41, 2015.

[10] 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, no. 3-4, pp. 458–465, 2013.

[11] J. Lipor and L. Balzano, "Robust blind calibration via total least squares," *simulation*, vol. 2, p. 8, 2014.