Monitoring Water Contaminants in Coastal Areas Through ML Algorithms Leveraging Atmospherically Corrected Sentinel-2 Data

Razzano, Francesca; Mauro, Francesco; Di Stasio, Pietro; Meoni, Gabriele; Esposito, Marco; Schirinzi, Gilda; Ullo, Silvia Liberata

DOI
10.1109/IGARSS53475.2024.10641456

Publication date
2024

Document Version
Final published version

Published in
IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium, Proceedings

MONITORING WATER CONTAMINANTS IN COASTAL AREAS THROUGH ML ALGORITHMS LEVERAGING ATMOSPHERICALLY CORRECTED SENTINEL-2 DATA

Francesca Razzano[a,b,1], *Francesco Mauro*[a], *Pietro Di Stasio*[a], *Gabriele Meoni*[c], *Marco Esposito*[d], *Gilda Schirinzi*[b], *and Silvia Liberata Ullo*[a]

[a] University of Sannio, Engineering Department, Benevento, Italy
[b] University of Parthenope, Engineering Department, Naples, Italy
[c] Delft University of Technology, Space Engineering Department, Delft, Netherlands
[d] Cosine, Engineering Department, Sassenheim, Netherlands

ABSTRACT

Monitoring water contaminants is of paramount importance, ensuring public health and environmental well-being. Turbidity, a key parameter, poses a significant problem, affecting water quality. Its accurate assessment is crucial for safeguarding ecosystems and human consumption, demanding meticulous attention and action. For this, our study pioneers a novel approach to monitor the Turbidity contaminant, integrating CatBoost Machine Learning (ML) with high-resolution data from Sentinel-2 Level-2A. Traditional methods are labor-intensive while CatBoost offers an efficient solution, excelling in predictive accuracy. Leveraging atmospherically corrected Sentinel-2 data through the Google Earth Engine (GEE), our study contributes to scalable and precise Turbidity monitoring. A specific tabular dataset derived from Hong Kong contaminants monitoring stations enriches our study, providing region-specific insights. Results showcase the viability of this integrated approach, laying the foundation for adopting advanced techniques in global water quality management.

Index Terms— Remote Sensing, Artificial Intelligence, Water contaminants monitoring, Machine Learning

1. INTRODUCTION

Numerous studies are highlighting the importance of monitoring contaminants in water basins, coasts, or reservoirs. Paramount in this context is water security and access, explicitly acknowledged in the Millennium Development Goals, with a particular emphasis on the promotion of water resource sustainability. In this regard in our study, the aim is to monitor the Turbidity parameter, addressing significant negative impacts on human health, as highlighted by the U.S. Environmental Protection Agency (EPA). Elevated Turbidity or suspended solids levels can adversely impact aquatic health reducing light penetration. Turbidity acts as a shield for disease-causing pathogens, and it may lead to waterborne disease, which can cause intestinal sickness. Turbidity monitoring is essential, but traditional in-situ measurements, while precise, face challenges in creating regional-scale water quality databases due to their labor-intensive nature and time constraints. Implementing Remote Sensing (RS) techniques allows for a more expansive and efficient assessment of water quality on a larger scale, overcoming the limitations associated with traditional techniques. There are some interesting investigations useful in this end [1, 2] and, in addition, the evolution of (Machine Learning) ML models and their integration into RS have introduced a novel dimension to water quality metrics [3, 4].

Our study aims to pioneer a novel approach for monitoring Turbidity contaminants using a combination of CatBoost ML and high-resolution data from Sentinel-2 Level-2A, atmospherically corrected Sentinel-2 data. Our research endeavors to refine the existing methodology outlined in the state of the art by employing the CatBoost Machine Learning technique for regression. Motivated by a seminal study [5], we seek to elevate current practices, specifically addressing the use of Level-1C images in the Hong Kong article, which represent a distinct Sentinel-2 product. These images, while not atmospherically corrected initially, undergo several pre-processing steps that require other specific settings. Our innovative solution involves the integration of Sentinel-2 Level-2A images downloaded from GEE and incorporating atmospheric correction to optimize analysis phases and enhance the overall proposed methodology. The implementation of a highly effective ML technique, coupled with an increased number of training samples, significantly contributes to more comprehensive network training. As a result, our case study demonstrates superior metrics when compared to the samples reported in the Hong Kong article, which utilizes an Artificial Neural Network (ANN) with specific settings. This underlines the efficacy of our approach in addressing and surpassing the limitations identified in the prior study.

[1]Corresponding author. *Email addresses*: francesca.razzano002@studenti.uniparthenope.it (FR), f.mauro@studenti.unisannio.it (FM), p.distasio@studenti.unisannio.it (PD), G.Meoni@tudelft.nl (GM), m.esposito@cosine.nl (ME), gilda.schirinzi@uniparthenope.it (GS), ullo@unisannio.it (SLU)

Our work discusses key aspects, such as the use of the atmospherically corrected Sentinel-2 data via GEE platform, the description of the Area of Interest (AOI), the creation of a tabular dataset derived from the Hong monitoring stations [5], and the application of the CatBoost ML method developed through Python language. The final sections present the obtained results and the conclusions.

2. GEE ENVIRONMENT AND DATA SOURCES

In the following sections, the significance of the GEE platform is addressed with the chosen dataset for the specific scenario.

2.1. GEE environment

The GEE platform is a powerful environment, that has been meticulously designed for the visualization of petabyte-scale geographic data and to facilitate advanced scientific research.

It offers an extensive collection of data characterized by diverse bands, projections, and resolutions, along with pivotal tools for enabling swift multisensor analysis. Its inherent capabilities empower users to employ both fundamental and sophisticated techniques, including ML-based algorithms and models, for the comprehensive analysis of Earth Observation (EO) data.

2.2. Sentinel-2 Level-2A

From GEE Sentinel-2 Level-2A images have been retrieved and a dedicated developed, as described in Section 4.1. In RS applications within GEE, the utilization of Sentinel-2 Level-2A products holds particular prominence, since they are suitable for various applications, including land cover mapping, vegetation monitoring, and change detection. For completeness, it is emphasized that although 13 spectral bands are made available by the advanced MultiSpectral Instrument (MSI) sensor on board Sentinel-2, Band 10 is not present in the Sentinel-2 Level-2A products.

3. AREA OF INTEREST

The AOI of our study has been identified with the Hong Kong region motivated by the initial study described in [5]. The coastal waters of Hong Kong are physically and chemically complex because of a mix of anthropogenic activities and fluctuating hydrographic conditions. Based on the hydrodynamic properties and pollution status, the Environmental Protection Department (EPD) of the Hong Kong government separated Hong Kong waterways into 10 water management zones. Every month, EPD uses a specialized marine monitoring vessel outfitted with a Differential Global Positioning System (DGPS) and an advanced Conductivity, Temperature, and Depth (CTD) profiler to analyze water quality and collect samples at 76 monitoring sites in coastal areas and open sea, as shown in the map of Fig. 1, created to identify the distribution of the monitoring stations. A specific database collects all water quality data and has been made available online. This database has been

Fig. 1. Map of the 76 monitoring stations in coastal areas and open sea of Hong Kong region - the Area of Interest (AOI)

instrumental in our study to determine the Ground Truth (GT) for the Turbidity parameter that we want to monitor.

4. DATASET CREATION AND METHODOLOGY

In our case study, we decided to employ a CatBoost as an ML model for the regression analysis, to predict the values of contaminants. This decision is grounded in a comprehensive examination of the State of the Art (SOTA), which indicates that CatBoost, along with many ML techniques, demonstrates superior performance compared to DL models [6, 7] in the contest of interest. For our regression model, a set of independent variables as input to the model is needed. In this case, the spectral bands of the Sentinel-2 Level-2A have been selected to this end. These variables are believed to have an impact on the output variables we aim to predict, namely the target or dependent variables representing the values of the contaminants. Independent variables can be taken of various types, such as numeric, categorical, or others. CatBoost is especially well-suited for handling different types of variables without the need for specific conversion, thereby simplifying the data preparation process.

4.1. Dataset creation

To derive the values at the different bands and therefore the independent variables of our regression problem, a dedicated code was developed to retrieve Sentinel-2 Level-2A images from GEE. Based on the method proposed in the previous paper by Mauro et al. [8], where an automated methodology for downloading satellite data from GEE was introduced, a modified version of the code has been adapted for the analysis at hand. Furthermore, all bands have been resampled to a spatial resolution of 10 meters, and the code has been entirely automated. This enables effortless extension of the dataset to other regions in the world. An array containing latitude and longitude coordinates for the points of interest is used, as well as an array of dates corresponding to each point. For each pair of geographic coordinates and dates, a pre-processing operation calculating a time window of three days before and after the in-situ date is performed and a square of size 0.2 km x 0.2 km is created around the central point specified by each pair of latitude and longitude coordinates. Satellite images from the Level-2A Sentinel-2 dataset ('COPERNICUS/S2_SR') within this geographical area and for the specified time window are filtered. The bands considered in this regard are: B1, B2, B3, B4, B5, B6, B7, B8, B8A, B9, B11, B12, at the different spatial resolution of 10, 20 and 60 meters. The final images are in total 660 and their peculiarity is to be all atmospherically corrected.

At this stage, to each image, along with its corresponding bands (representing the independent variables of the regression problem), is assigned the value of the dependent variable, constituted by the GT contaminant value extracted from the database of Hong Kong. Subsequently, the constructed dataset is in tabular form, obtained in this way through the calculation of the average over the small squares, and with this dataset, the specific ML algorithm is trained to learn the associated contaminant values. In Figure 2 is reported the workflow of our work. Dataset and code are available on a specific GitHub page that will be made publicly available after paper acceptance

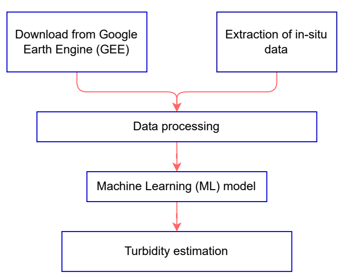

Fig. 2. Workflow of the proposed work

4.2. The ML algorithm: CatBoost

To monitor the Turbidity in coastal areas, we chose as ML algorithm, the CatBoost, pioneered by Yandex and extending upon the principles of decision trees and gradient boosting.

Numerous studies have affirmed the efficacy of CatBoost as an ML algorithm [6, 9, 10]. In our work, a Python library has been utilized to implement the CatBoost model for regression task and new code has been developed to tailor the model to our case study. Since it represents a high-level, versatile, and interpreted programming language known for its readability and simplicity.

4.2.1. Methodology applied

As previously described, the model was trained with input-output pairs, specifically the independent and dependent variables from the created dataset. In this way, the model endeavors to learn the relationship between the features extracted from satellite images and the levels of water contaminants. In the prediction phase, after the training process, the model is employed to forecast water contaminant levels on new data.

During this stage, only the independent variables are supplied in input. Furthermore, the data are systematically prepared, incorporating a suitable splitting between training, testing, and validation sets with percentages of 55%, 20% and 25% respectively. Lastly, the model performance is assessed using appropriate regression metrics, such as the Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) values. These metrics are typically employed for regression problems, and particularly RMSE and MAE are utilized in the referenced work [5] and this facilitates a comparative analysis. The selection of parameters for the CatBoost algorithm is reported in Table 1 where the 'depth' parameter manages the depth of the tree structure, while the 'Learning_rate' influences the size of adjustments applied to the tree model, determining the speed at which the model learns. The 'Iterations' parameter corresponds to the number of trees (rounds), emphasizing the total number of boosting iterations, and the 'l2_leaf_regularization' represents the L2 regularization coefficient to prevent overfitting. In summary, these parameters were tuned to control learning speed and overfitting. Lastly, the chosen Loss function has been the RMSE, a common choice for regression tasks.

Table 1. Parameters for CatBoost

Iterations	Learning_rate	depth	Loss function	L2_leaf_regularization
600	0.6	12	RMSE	1.0

5. RESULTS

From the distribution of the 660 Turbidity values extracted from the Hong Kong database for the selected period and area of interest the mean of the created dataset values is 3.6882 NTU (Nephelometric Turbidity Unit). To evaluate the performance of the GradientBoostingRegressor, XGBRegressor and CatBoostRegressor models, we focused on MSE, RMSE, and MAE values. The metrics on test set reported in Table 2 suggest that all models have learned the patterns effectively.

Table 2. Comparison of performance evaluation metrics on test set between GradientBoostingRegressor, XGBRegressor and CatBoostRegressor

Models	MSE	RMSE	MAE
GradientBoosting	0.0553251	0.2352128	0.2141934
XGBoost	0.0484808	0.2201837	0.2143259
CatBoost	**0.0453542**	**0.2129653**	**0.1945048**

Specifically, compared with others ML algorithms, CatBoost and XGBoost demonstrate strong performance across all metrics, with CatBoost showing slightly better performance in terms of MSE, RMSE, and MAE values. Besides the quantitative analysis, it may be beneficial to visually inspect the model

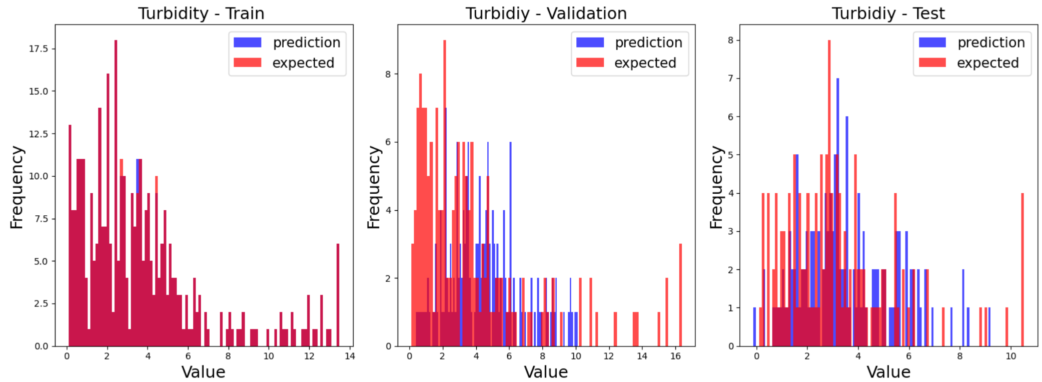

Fig. 3. Comparison of expected (red) and predicted (blue) values for Training, Validation, and Testing sets

predictions against the actual values for identifying specific patterns. As can be observed from the representation in Figure 3, the distributions of the expected and predicted data in the training phase are comparable and practically overlapping.

This indicates effective learning from the training set. If we examine the distributions on the validation and test sets, it can be verified that the model reasonably learns and predicts new data. This implies a promising starting point for enhancing the input dataset for the model, thereby enriching its predictive capability.

To further validate our study, we conducted a comparison for Turbidity contaminant, reported in Table 3, with the metrics used to evaluate the Artificial Neural Network (ANN) proposed by [5] and [11]. We trained the network on a common region, with our focus limited to a sub-area within the broader region they considered. Given that, since in the mentioned studies, the dataset was split into training and validation sets while in our study the splitting involves training, validation, and test sets, to ensure a comparable comparison, the Table reflects accuracies of RMSE and MAE exclusively originating from the validation sets while the Sample size parameter integrates both the phases of training and validation. Furthermore, considering the mean values of the authors, which are 4.80 NTU for [5] and 9.4 NTU for [11], our mean, when compared to the metrics we obtained, lead us to infer that we have achieved superior results in terms of both RMSE and MAE.

Table 3. Comparison of Metrics Modeling results of Turbidity water quality parameter

Study	Years of Interest	Image data	Sample size	RMSE	MAE
[11]	1999–2015	Landsat-5, 7, 8	120	3.10	2.61
[5]	2015–2021	Sentinel-2 L1C	352	1.95	1.61
Our study	**2019-2020**	**Sentinel-2 L2A**	**528**	**0.24**	**0.20**

6. CONCLUSIONS

In conclusion, the results affirm CatBoost's excellent performance in capturing intricate dataset relationships, making it a promising tool for advancing water contaminants monitoring through RS. This study serves as a foundation for future development, enriching global coastal data for adaptable ML models. The network's ability to predict coastal parameters, especially in challenging regions, positions it as a significant asset. Future efforts will extend monitoring to include chemical contaminants, highlighting the potential of ML for coastal ecosystem preservation and global well-being.

7. REFERENCES

[1] K. Toming, T. Kutser, A. Laas, M. Sepp, B. Paavel, and T. Nõges, "First experiences in mapping lake water quality parameters with Sentinel-2 MSI imagery," *Remote Sensing*, vol. 8, no. 8, pp. 640, 2016.

[2] M. Potes, G. Rodrigues, A.M. Penha, M.H. Novais, M.J. Costa, R. Salgado, and M.M. Morais, "Use of Sentinel 2–MSI for water quality monitoring at Alqueva reservoir, Portugal," *Proceedings of the International Association of Hydrological Sciences*, vol. 380, pp. 73–79, 2018.

[3] E. Sharaf El Din, Y. Zhang, and A. Suliman, "Mapping concentrations of surface water quality parameters using a novel remote sensing and artificial intelligence framework," *International Journal of Remote Sensing*, vol. 38, pp. 1023–1042, 02 2017.

[4] H. Guo, J.J. Huang, B. Chen, X. Guo, and V.P. Singh, "A machine learning-based strategy for estimating non-optically active water quality parameters using Sentinel-2 imagery," *International Journal of Remote Sensing*, vol. 42, no. 5, pp. 1841–1866, 2021.

[5] I. Kwong, F. Wong, and T. Fung, "Automatic Mapping and Monitoring of Marine Water Quality Parameters in Hong Kong Using Sentinel-2 Image Time-Series and Google Earth Engine Cloud Computing," *Frontiers in Marine Science*, vol. 9, pp. 871470, 05 2022.

[6] S. Berkani, I. Gryech, M. Ghogho, B. Guermah, and A. Kobbane, "Data Driven Forecasting Models for Urban Air Pollution: MoreAir Case Study," *IEEE Access*, vol. 11, pp. 133131–133142, 2023.

[7] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[8] F. Mauro, B. Rich, V.W. Muriga, F. Janku, A. Sebastianelli, and S.L. Ullo, "SEN2DWATER: A Novel Multispectral and Multitemporal Dataset and Deep Learning Benchmark for Water Resources Analysis," in *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 2023, pp. 297–300.

[9] R. Nagaraj, V. Arulvadivelan, K. Gouthamkumar, K. Dharshen, and L.S. Kumar, "Surface water mapping and volume estimation of Lake Victoria using Machine Learning Algorithms," in *2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)*, 2023, pp. 1–6.

[10] H. Mazumdar, M.P. Murphy, S. Bhatkande, H.P. Emerson, D.I. Kaplan, and H.A. Gohel, "Optimized machine learning model for predicting groundwater contamination," in *2022 IEEE MetroCon*. IEEE, 2022, pp. 1–3.

[11] S. Hafeez, M.S. Wong, H.C. Ho, M. Nazeer, J. Nichol, D. Abbas, S.and Tang, K.H. Lee, and L. Pun, "Comparison of Machine Learning Algorithms for Retrieval of Water Quality Indicators in Case-II Waters: A Case Study of Hong Kong," *Remote Sensing*, vol. 11, no. 6, 2019.