FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION FOR HIGHER PROFESSIONAL EDUCATION NATIONAL RESEARCH UNIVERSITY

«HIGHER SCHOOL OF ECONOMICS»

Faculty of Computer Science

Zhanar Orynbassar

Name, Surname

Картирование отходов из пластика с помощью данных дистанционного зондирования Земли

Russian title

Mapping Plastic Waste with Earth Remote Sensing Data

English title

Report for Research Seminar

Field of study 01.04.02 «Applied Mathematics and Informatics»

Program: Data Science

Student Name:

Zhanar Orynbassar

Supervisor Name:

Ramon Antonio Rodriges Zalipynis

INTRODUCTION

Marine debris is a growing environmental problem that has negative impacts on marine ecosystems and biodiversity. Remote sensing technologies offer a promising approach for detecting and monitoring marine debris, but there is a lack of standardized datasets and benchmarks for evaluating detection algorithms. Overall, the MARIDA dataset provides a valuable resource for researchers and practitioners working on the development of marine debris detection algorithms using remote sensing data. In this project, a gradient boosting algorithm was applied as a semantic segmentation algorithm, and achieved 92% of recall, 78% of f1 for marine debris classification task.

EXPLORATORY DATA ANALYSIS

Before choosing the dataset and main method, let's dive into the marine debris detection field. There are several papers for comparison.

Paper	Main focus	Result
Biermann et al.	The detection of patches of marine plastic debris using spectral signatures and classification of materials in the floating debris automatically. And the finding of the most efficient indices at detecting the macroplastic from the data.	For detecting the macroplastic from the data at subpixel level, most efficient ways are through Floating Debris Index (FDI) and Normalised Difference Vegetation Index (NDVI) showed that the naive Bayes classification algorithm can be detect plastic debris with accuracy of 86%.
Kikaki et al.	The observation of intertemporal satellite observations over the Bay islands and Gulf of Honduras and examined the possibility of detecting patches of floating plastic debris using spectral signals as well, in addition to investigating sources, trajectories and accumulation sites of the plastic debris.	The plastic debris can be detected using spectral signals as well and sources of plastic waste are major rivers of Honduras and Guatemala such as Motagua, Ulua, Cangrejal, Tinto and Aguan. The study also finds that the trajectories of plastic follow the direction of prevailing currents and there is no evidence of accumulation point of plastic waste.
Topouzelis et al.	Study the artificial plastic targets of large (10m x 10m) and small (5m x 10m) size and compiles a reference dataset by periodically observing them to examine the spectral characterization of the plastic debris in near real-world environment and to calibrate and validate plastic litter detection algorithms.	To tailor the spectral signature of plastic to the real-world environment, the authors use matched filtering and spectral unmixing. The authors' methodology of spectral characterization of plastic debris can detect pixels with the signature abundance fraction of 25% under suitable weather conditions.
Themistocleous et al.	Comparison of the predictive power of various indices derived from multispectral signals from Sentinel-2 satellite images.	The Plastic Index (PI) and the Reversed Normalized Difference Vegetation Index (RNDVI) are the best at identifying plastic in the marine environment. This result is verified by spectral signatures and images collected by Unmanned Aerial Vehicles.
Basu et al.	Comparison of the predictive power of 2 supervised learning (support vector regression (SVR) and semi-supervised fuzzy c-means (SFCM))and 2 unsupervised learning (K-means and fuzzy c-means (FCM)) classification techniques on Limassol, Cyprus and Mytilene, Greece.	The result is that the Support Vector Regression-based classification technique is the most accurate with figures 96.9–98.4%.
Kikaki et al. MARIDA	To propose a benchmark dataset and evaluation metrics for detecting marine debris using Sentinel-2 satellite imagery. The MARIDA dataset consists of high-resolution Sentinel-2 images with annotated marine debris objects and non-debris regions, along with metadata and ground truth labels. The paper presents the evaluation metrics used for assessing the performance of different machine learning models	The results of the benchmark show the effectiveness of deep learning approaches for marine debris detection, with the best performing model achieving an F1-score of 0.87. Overall, the paper provides a valuable resource for researchers and practitioners working on marine debris detection using remote sensing data.

After comparison, MARIDA was chosen as the dataset. The MARIDA paper presents a benchmark dataset for the detection of marine debris from Sentinel-2 remote sensing data. MARIDA provides 3399 Marine Debris pixels, labelled in different S2 tiles across various countries, different seasons, years and sea state conditions. Thus, MARIDA is an important geodata source for evaluating existing detection methods and developing new techniques based on available S2 data.

DATA PREPARATION

First step is to prepare labels and annotations. In this analysis, I have 14 labels as: 'Marine Debris', 'Dense Sargassum', 'Sparse Sargassum', 'Natural Organic Material', 'Ship', 'Clouds', 'Marine Water', 'Sediment-Laden Water', 'Foam', 'Turbid Water', 'Shallow Water', 'Waves', 'Cloud Shadows', 'Wakes', 'Mixed Water'.

To do the classification task, I have to rely on spatial signatures and calculated values, they are 'nm440',' nm490', 'nm560', 'nm665', 'nm705', 'nm740', 'nm783', 'nm842', 'nm865', 'nm1600', 'nm2200', 'NDVI', 'FDI', 'NDWI'. Each band is numbered as: 1:nm440 2:nm490 3:nm560 4:nm665 5:nm705 6:nm740 7:nm783 8:nm842 9:nm865 10:nm1600 11:nm2200. In which, except for NDVI, FDI and NDWI, all spatial signatures were directly obtained from the ACOLITE atmospheric corrector. Let's see the formulas of those three indices:

NDVI = (band8 - band4)/(band8 + band4)

NDWI = (band3 - band8)/(band3 + band8)

FDI = band8 - band6 + 10 * (band10 - band6) * (l_nir - l_redge) / (l_swir - l_redge) (non signature values are adjusted by individual)

As the dataset consists of Annotated polygons. The annotation procedure resulted in a vector dataset of the digitized polygons, in shapefile format. The dataset was converted into a raster structure, which was finally cropped into non-overlapping 256x256 pixel-sized patches. After the cropping, each patch was available for extra visual inspection.

DATA PROCESSING AND TRAINING

Two datasets were extracted in the form of numpy arrays, dataset.h5 only consists of 11 spatial signatures, dataset_si.h5 consists of spatial signatures and calculated indices. After obtaining those two datasets, split datasets into train, validation and test patches, then trained two gradient boosting models.

Gradient boosting algorithm works by combining multiple weak models or decision trees in a sequential manner, where each new model is built to correct the errors made by the previous ones. In other words, the algorithm starts with a single decision tree, and then iteratively builds additional decision trees that focus on correcting the errors of the previous tree.
The "gradient" part of the name comes from the fact that the algorithm uses gradient descent optimization to minimize the errors between the predicted and actual outputs. Gradient boosting classifier has become a popular algorithm for classification tasks because of its ability to handle complex non-linear relationships between variables and its high accuracy in predicting outcomes.
After training the model, we achieve a confusion matrix which shows the results of the model, as shown below.

```
root - INFO - Number of Input features: 14
root - INFO - Train: 429412
root - INFO - Test: 194863
root - INFO - Started training
root - INFO - Training finished after 37.37190508842468 seconds
root - INFO - Classifier is saved at: /Users/zhanarorynbassar/Desktop/Thesis/Gradient Boosting/semantic_segmentation/gradient_boosting/gb_classifier.joblib
root - INFO - Confusion Matrix:
                        Clouds  Dense Sargassum   Foam  Marine Debris  Marine Water  Natural Organic Material  Sediment-Laden Water  Shallow Water    Ship  Sparse Sargassum  Turbid Water       Sum  Recall
Clouds                 27766.0              0.0    0.0           16.0        4808.0                       0.0                   0.0            8.0   143.0              84.0          18.0   32843.0    0.85
Dense Sargassum            0.0            707.0    0.0            0.0           0.0                       1.0                   0.0            0.0     0.0              52.0           0.0     760.0    0.93
Foam                       8.0              0.0  118.0            3.0          21.0                       0.0                  20.0            1.0   216.0               0.0           0.0     387.0     0.3
Marine Debris              2.0              0.0    0.0          352.0          18.0                       7.0                   0.0            0.0     2.0               0.0           0.0     381.0    0.92
Marine Water            3844.0              0.0   67.0           77.0       25361.0                       1.0                   0.0          292.0   102.0               3.0         872.0   30619.0    0.83
Natural Organic Material   0.0              0.0    1.0           19.0           5.0                      22.0                   0.0            0.0     2.0               0.0           0.0      49.0    0.45
Sediment-Laden Water       0.0              0.0    1.0            4.0          20.0                       0.0               93002.0            0.0    10.0               0.0           0.0   93037.0     1.0
Shallow Water             10.0              0.0    1.0            0.0        1394.0                       4.0                   0.0          721.0     0.0               0.0         376.0    2506.0    0.29
Ship                     128.0              0.0   45.0           51.0          95.0                       4.0                   0.0            0.0   840.0               1.0          10.0    1174.0    0.72
Sparse Sargassum           0.0             34.0    0.0            5.0          15.0                      28.0                   0.0            0.0     0.0             799.0           0.0     881.0    0.91
Turbid Water              10.0              0.0   34.0            0.0        1638.0                       0.0                 551.0         8104.0     0.0               0.0       21889.0   32226.0    0.68
Sum                    31768.0            741.0  267.0          527.0       33375.0                      67.0               93573.0         9126.0  1315.0             939.0       23165.0       mPA:    0.72
IoU                        0.75             0.89   0.22           0.63          0.66                       0.23                  0.99           0.07    0.51               0.78           0.65      mIoU:    0.58
Precision                  0.87             0.95   0.44           0.67          0.76                       0.33                  0.99           0.08    0.64               0.85           0.94        OA:    0.88
F1-score                   0.86             0.94   0.36           0.78          0.79                       0.38                   1.0           0.12    0.67               0.88           0.79  F1-macro:    0.69
```

RESULT
Overall, the MARIDA benchmark dataset and evaluation framework provide a valuable resource for evaluating object detection algorithms for marine debris detection using Sentinel-2 remote sensing data. By only using ndvi, fdi and ndwi parameters, the Gradient Boosting gives 92% of recall, 78% of f1.

DISCUSSION
There are challenges and limitations of the data wrangling process, such as the presence of clouds and shadows in the imagery, the lack of ground truth data for some areas, and the subjective nature of annotating marine debris objects.
The challenges associated with marine debris detection include the high variability in debris types and sizes and the difficulties in accurately identifying and distinguishing debris from other objects in the marine environment.
Additionally, the authors note that existing object detection algorithms have not been optimized for detecting marine debris and may struggle with the variability and complexity of the marine debris environment.

REFERENCES

Code link: https://github.com/Zhanarik/PS

[1] Kikaki, Aikaterini, Konstantinos Karantzalos, Caroline A. Power, and Dionysios E. Raitsos 2020. "Remotely Sensing the Source and Transport of Marine Plastic Debris in Bay Islands of Honduras (Caribbean Sea)" Remote Sensing 12, no. 11: 1727. https://doi.org/10.3390/rs12111727

[2] Biermann L, Clewley D, Martinez-Vicente V, Topouzelis K. Finding Plastic Patches in Coastal Waters using Optical Satellite Data. Sci Rep. 2020 Apr 23;10(1):5364. doi: 10.1038/s41598-020- 62298-z. Erratum in: Sci Rep. 2020 May 12;10(1):8069. PMID: 32327674; PMCID: PMC7181820.

[3] Lacerda, A.L.d.F., Rodrigues, L.d.S., van Sebille, E. et al. Plastics in sea surface waters around the Antarctic Peninsula. Sci Rep 9, 3977 (2019). https://doi.org/10.1038/s41598-019-40311-4

[4] Basu, Bidroha, Srikanta Sannigrahi, Arunima Sarkar Basu, and Francesco Pilla. 2021. "De velopment of Novel Classification Algorithms for Detection of Floating Plastic Debris in Coastal Waterbodies Using Multispectral Sentinel-2 Remote Sensing Imagery" Remote Sensing 13, no. 8: 1598. https://doi.org/10.3390/rs13081598

[5] Themistocleous, Kyriacos, Christiana Papoutsa, Silas Michaelides, and Diofantos Hadjimitsis. 2020. "Investigating Detection of Floating Plastic Litter from Space Using Sentinel-2 Imagery" Remote Sensing 12, no. 16: 2648. https://doi.org/10.3390/rs12162648

[6] Topouzelis, Konstantinos, Dimitris Papageorgiou, Alexandros Karagaitanakis, Apostolos Pa pakonstantinou, and Manuel Arias Ballesteros. 2020. "Remote Sensing of Sea Surface Artificial Floating Plastic Targets with Sentinel-2 and Unmanned Aerial Systems (Plastic Litter Project 2019)" Remote Sensing 12, no. 12: 2013. https://doi.org/10.3390/rs12122013

[7] Kikaki K, Kakogeorgiou I, Mikeli P, Raitsos DE, Karantzalos K (2022) MARIDA: A benchmark for Marine Debris detection from Sentinel-2 remote sensing data. PLoS ONE 17(1): e0262247. https://doi.org/10.1371/journal.pone.0262247