Machine learning-based NO2 estimation from seagoing ships using TROPOMI/S5P satellite data

Kurchaba, S.

Chapter 3

Sensitivity analysis for the detection of NO_2 plumes from seagoing ships using TROPOMI data

Based on: Kurchaba, S., Sokolovsky, A., van Vliet, J., Verbeek, F.J., Veenman, C.J., 2024. Sensitivity analysis for the detection of NO_2 plumes from seagoing ships using TROPOMI data. Remote Sensing of Environment 304, 114041. doi:10.1016/j.rse.2024.114041.

Abstract The marine shipping industry is among the strong emitters of nitrogen oxides (NO_x) – a substance harmful to ecology and human health. Monitoring of emissions from shipping is a significant societal task. Currently, the only technical possibility to observe NO_2 emission from seagoing ships on a global scale is using TROPOMI data. A range of studies reported that NO_2 plumes from some individual ships can be visually distinguished on selected TROPOMI images. However, all these studies applied subjectively established pre-determined thresholds to the minimal speed/length of the ship – variables that to a large extent define the emission potential of a ship. In this Chapter, we investigate the sensitivity limits for ship plume detection as a function of their speed and length using TROPOMI data. For this, we train a classification model to distinguish TROPOMI image patches with a ship, from the image patches, where there are no ships. This way, we exploit ground truth ship location data to potentially exceed human visual distinguishability. To test for regional differences, we study four regions: the Mediterranean Sea, Biscay Bay, Arabian Sea, and Bengal Bay. For the Mediterranean and the Arabian Sea, we estimate the sensitivity limit to lie around a minimum speed of 10 knots and a minimum length of 150 meters. For the Biscay Bay – around 8 knots and 100 meters. We further show that when focusing the analysis on the biggest emitters (junctions of several ships in the area), the detectability can be improved up to above 0.8 ROC-AUC. Finally, we show that increasing the size of the dataset, beyond the dataset used in this study, yields further improvements in the detectability of smaller/slower ships. The rate of improvement in both experiments is dependent on the region studied.

3.1 Introduction

As it was mentioned in the Introduction of this thesis, the TROPOMI/S5P is the first satellite-based instrument that gives the possibility to visually detect NO_2 plumes from some individual seagoing ships [41]. This is due to significantly higher than its predecessor spatial resolution of the instrument. Such an improvement in the quality of the remote-sensing-based atmospheric monitoring allows to consider the TROPOMI/S5P instrument as a potential solution for the task of global and continuous monitoring of the emissions produced by seagoing ships [90]. However, in order to fully understand the potential of the TROPOMI for a given task, the first step is to estimate the limitations in terms of the sensitivity of the detection system for NO_2 plumes from seagoing ships using TROPOMI data.

To tackle the problem, we prepare image patches – small, regular-sized sections of the TROPOMI measurement (image). We use the created image patches to train a machine-learning classification model. The task of the model is to distinguish image patches with at least one ship from the image patches where there are no ships. The labels of the model were created using AIS ship location data, and, therefore, are independent of the distinctivity of ship plumes by a human. This way, we formulate the research questions of the study as follows:

- **RQ1**: What is the minimum speed and length of a seagoing ship so that the NO_2 plume from it can be detected with the detection system using TROPOMI data?
- **RQ2**: To what extent can the detectability of NO_2 plumes be improved if only the biggest emitters are taken into account? With the biggest emitters, we mean the biggest ships operating at the highest speeds, or several smaller or slower ships operating in proximity to each other.
- **RQ3**: Is there a potential for improvement of detectability of NO_2 plumes from the slow/small ships if more data were used to train the used classification model?

We conduct this study on four regions of interest: Mediterranean Sea, Biscay Bay, Arabian Sea, and Bengal Bay (the coordinate scope see in Table 3.1 and Figure 3.1). The study areas are directed towards the Europe – Middle East – Asia trade route, with selected areas representing low background pollution and common occurrence of clear skies.

The rest of the Chapter is organized as follows: In Section 3.2, we explain how the data was pre-processed in order to obtain datasets used for machine learning

models. In Section 3.3, we introduce the experimental setup for each stage of the study and present the obtained results. We discuss the obtained results in Section 3.4 and conclude in Section 3.5.

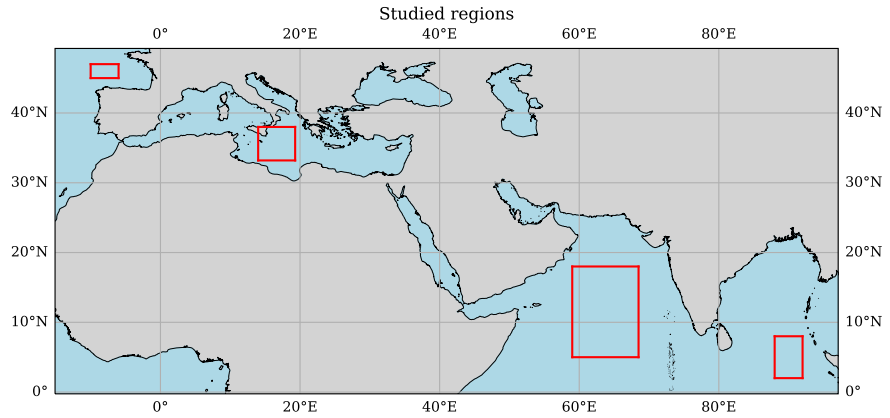


Figure 3.1: Four studied regions (from left to right): Biscay Bay, Mediterranean Sea, Arabian Sea, Bengal Bay.

Region	Longitude [deg]	Latitude [deg]	Studied period
Mediterranean	(14, 19.3)	(33.2, 38)	(31-03-20; 28-02-23)
Biscay Bay	(-10, -6)	(45, 47)	(01-04-20; 28-02-23)
Arabian Sea	(59, 68.5)	(5, 18)	(31-03-20; 30-11-22)
Bengal Bay	(88, 92)	(2, 8)	(03-06-20; 31-12-22)

Table 3.1: Geographical coordinates and analyzed periods defining the study scope for each region.

3.2 Dataset

The supervised learning task that is addressed in this study is to distinguish image patches with a ship plume on them. In this Section, we describe the process of the preparation of the dataset for the given supervised learning task. We first describe the undertaken steps of data pre-processing. We then introduce the features used for the model training and define the target variable.

Region	Ship image	No ship image
Mediterranean	16%	18%
Biscay Bay	48%	52%
Arabian Sea	49%	52%
Bengal Bay	54%	54%

Table 3.2: Percentage of data from the original dataset lost when a *qa value* of .75 is applied for filtering.

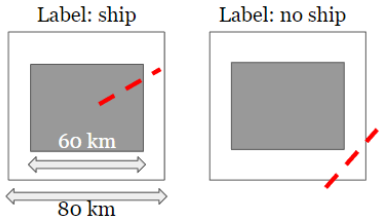


Figure 3.2: An illustration of the set-up used for counting the number of ships per image patch. White square – image patch. Grey square – a central part of the image patch. Red dashed lines – an example of ship trajectory starting from 2 hours before until the moment of the satellite overpass. Only ships, whose trajectories cross the central part of the image patch are considered to be present in the area covered by a patch.

3.2.1 Data preprocessing

The first step of data preparation is regridding[1]. This is done so that for each region we have pixels with the same spatial coverage. The regridded pixel size for each region is approximately equal to 4×5 km^2. Following the set-up used in the previous studies [63, 64], for the regridding, we only use pixels with cloud coverage below 0.5, wind speed lower than 10 m/s, and *qa value* above 0.5 [93]. This level of *qa value* filtering was shown to be sufficient for the identification of NO_2 plumes from individual ships and is a trade-off between a high standard of data quality, and an attempt to preserve as many data points as possible. In Table 3.2, the reader can find an assessment of the data loss in case *qa value* filtering was set to the level of 0.75 – the level suggested in the TROPOMI manual [31].

As a next step, we split the studied area into non-overlapping patches of equal size 80×80 km^2. The selected size of the image corresponds to a distance that the fastest ships in the dataset will cover in 2 hours. The observation period of 2 hours

[1]The regridding is performed using the Python package HARP v.1.13.

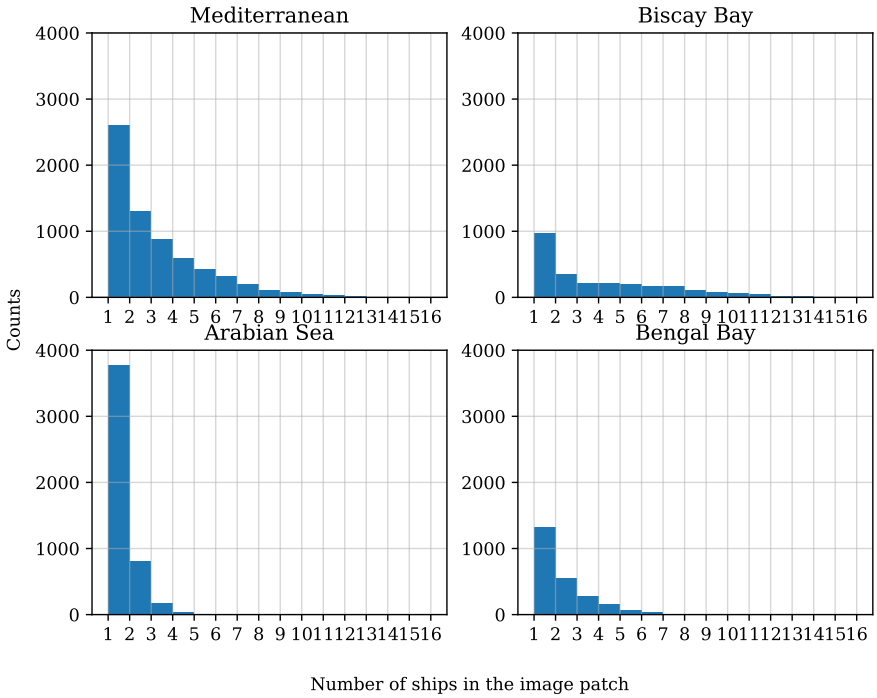


Figure 3.3: Distribution of the number of ships per image patch for the studied regions.

Region	Ship image	No ship image
Mediterranean	6652	9693
Biscay Bay	2641	2812
Arabian Sea	4804	24594
Bengal Bay	2444	6848

Table 3.3: Class-wise distribution of image patches for each studied region. The rate of imbalance depends on the traffic density in the region.

was motivated by the fact that due to the physical dispersion and limited lifetime of NO_2 within plumes, the detectability of ship plumes will fall sharply after 2 hours [107]. For each image patch, we calculate how many ships were in the central area of the patch within 2 hours before the overpass of the satellite. The central area of the patch is defined as 60×60 km^2 square. We do not take into account ships that do not pass through the central area of the image patch, as the probability that their plume will be located within the image patch is very low. An example is presented in Figure 3.2. The resulting distribution of the number of ships per image patch for each studied region can be found in Figure 3.3. Please note the regional differences in the distribution of ships among patches. The Arab Sea typically has a high number of patches with a single ship. The Biscay Bay, in comparison to other regions, has the highest number of patches with a high number of ships on it. These patterns illustrate the difference in shipping density among the studied regions.

3.2.2 Feature engineering

To study the sensitivity of the TROPOMI instrument with respect to the detection of NO_2 plumes from seagoing ships, we prepare a dataset for supervised machine learning. The NO_2 trace gas variable of our interest is *Tropospheric Slant Column Density – SCD trop* [31]. As mentioned in Chapter 2, the SCD variable is suitable for satellite sensitivity study [41] as its derivation is not based on airmass factor – a variable estimated based on, among others, historical NO_2 concentration within a certain area.

The objective is to distinguish image patches that cover the area where there are no ships, from image patches covering the area with at least one ship on it. Since this is a binary problem, the value of the output label is 1, if there is at least one ship that is faster than 6 kt, which is approximately 11.1 km/h and longer than 90 m in the area covered by an image patch. The output label is 0, if there is no ship in the area,

or the ship is shorter than 90 m or slower than 6 kt. The values of 90 m and 6 kt are sufficiently low to be well below detectable limits as will also follow from this study. Table 3.3 shows the resulting distribution of classes for studied regions. Examples of image patches without (label 0) and with at least one ship on it (label 1) are presented in Figure 3.4. We can see that not all image patches with a ship actually contain a visually distinguishable plume. This is because the NO_2 plumes produced by some ships are below the sensitivity limit of the TROPOMI instrument, or we are not able to distinguish it visually.

We address the classification problem with a multivariate classifier. Therefore, we represent the TROPOMI image patches in terms of a set of features - a statistical representation of the image patch. More specifically, for the regridded pixels of each image patch, we calculate the following statistics: *min(SCD), mean(SCD), median(SCD), max(SCD), std(SCD)*, where SCD stands for NO_2 slant column density. To give information about the level of plume dispersion, we add wind-related variables *zonal wind velocity (wind zon), meridional wind velocity (wind med)*, which represent the speed of the wind from the west to east and from south to north respectively. Finally, we add features *sensor zenith angle, solar zenith angle* and *solar azimuth angle* to represent the viewing geometry of the satellite. Values for wind information and satellite geometry are the average values of the pixels within the image patch. The resulting feature set is presented in Table 3.4. In Figure 3.5, the reader can find histograms of the dataset features for the studied regions. Clearly, the features related to the properties of ships cannot be included in the feature space, because the presence of a ship has to be established. Moreover, we deliberately do not include any features in the feature set related to the geographic locations of a given patch. This is because shipping lanes may bias the model. The dataset used in this study as well as the code used for generating the presented in this study results are available publicly as a reproducibility capsule [60]. Prior to the application of a machine learning model, all features were standardized using a median-interquartile range scaling[2] – a scaling technique that allows to reduce a negative impact of the outliers in the dataset [32].

3.3 Experiments and results

In this Section, we describe the experiments and show the results obtained. We start with the introduction of the classification model – we present model selection and hyperparameter optimization results. For the selected model, we provide the explain-

[2]RobustScaler implemented in scikit-learn v.1.2.2.

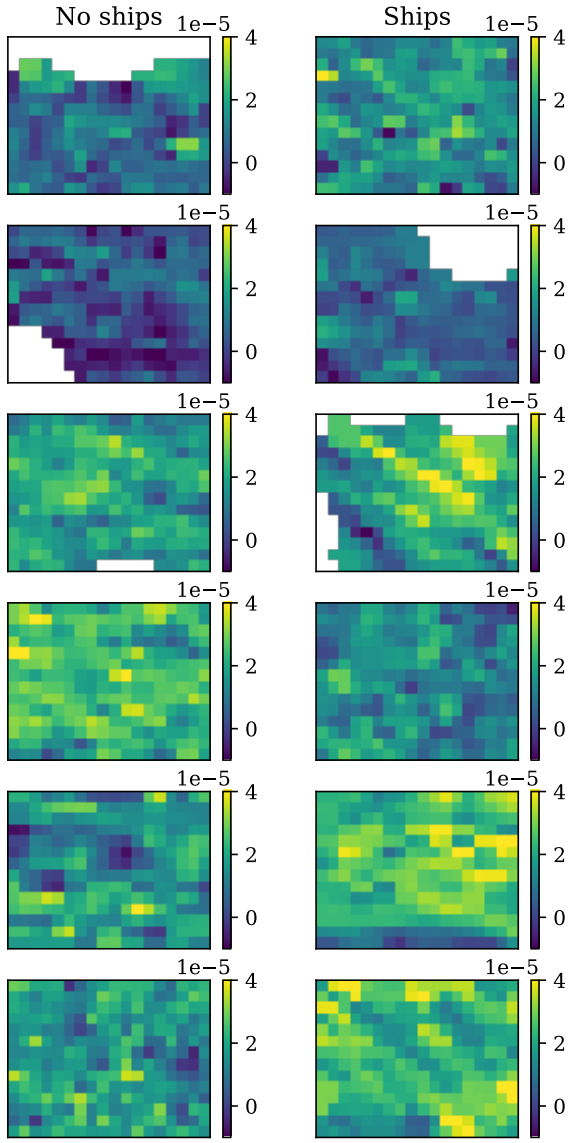


Figure 3.4: Examples of image patches without a ship and with at least one ship on it. The presented image patches were randomly sampled from the dataset of the region Biscay Bay.

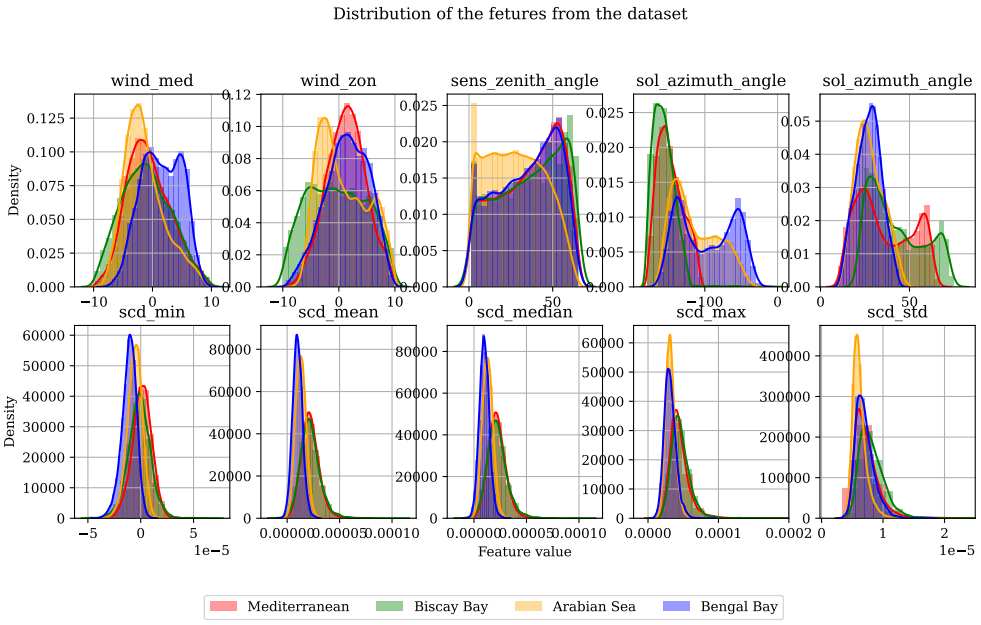


Figure 3.5: Histograms of the variables from the dataset.

Feature type	Feature name
NO_2 slant column density	min(SCD)
	mean(SCD)
	median(SCD)
	max(SCD)
	std(SCD)
Wind information	zonal wind velocity
	meridional wind velocity
Satellite geometry	sensor zenith angle
	solar zenith angle
	solar azimuth angle

Table 3.4: List of features used for classification model.

ability analysis. Next, in the consecutive subsections, we explain and provide the results of the experiments addressing the three research questions of this study.

3.3.1 Classification model

Experimental setup

As a first step, we compared the performance of several multivariate classifiers and selected the one that is going to be used in the remaining part of the Chapter for the sensitivity analysis. We studied four machine learning classifiers of increasing complexity: Logistic regression, Support Vector Machine (SVM) with the radial basis function (rbf) kernel, Random Forest[3], and Extreme Gradient Boosting[4] (XGBoost) [22]. All selected models are robust to noise and can be efficient even given the relatively small size of datasets. To make sure that we exploit the maximum potential of a given machine learning model, we optimized the hyperparameters of each studied model. The hyperparameters were optimized using a random search[5] technique with the objective metrics - *average precision*. The used search space of the hyperparameters for each of the models studied as well as the results of the hyperparameters optimization can be found in the original paper [61]. To be able to simultaneously perform the hyperparameter optimization and evaluation of the model performance, we use 5-fold nested cross-validation [96, 18] (for the explanation of the concept and visual example see Section 2.3). To maintain the same percentage of samples of a certain label in the training and test set, the cross-validation was based on *stratified K-fold* splits [47, 42].

Results

The classification results are presented in Table 3.5. Comparing the performances between different classifiers, we can see that the XGBoost classifier yielded the best results for most of the regions – we used this classifier for the remaining experiments of this study. Comparing the results between regions, we start with ROC-AUC. The highest achievable score of ROC-AUC is equal to 1. While the ROC-AUC score that will be obtained in case of random guessing is 0.5. The ROC-AUC score is calculated based on the ROC curve. For the XGBoost classifier, it is presented in the right-hand side plot of Figure 3.6. The scores for Biscay Bay and the Mediterranean Sea are higher than for the Arabian Sea and Bengal Bay. One of the reasons for this difference

[3] All above-mentioned models are implemented in Python scikit-learn v.1.2.2.
[4] XGBoost v. 1.7.0
[5] Implemented in Python scikit-learn v.1.2.2.

Region	Model	Average Precision	ROC-AUC
Mediterranean	XGBoost	**0.636 ± 0.013**	**0.712 ± 0.011**
	Random Forest	0.629 ± 0.018	0.706 ± 0.016
	SVM (rbf)	0.615 ± 0.015	0.694 ± 0.013
	Logistic	0.448 ± 0.008	0.546 ± 0.009
Biscay Bay	XGBoost	**0.704 ± 0.021**	**0.713 ± 0.015**
	Random Forest	0.620 ± 0.025	0.652 ± 0.022
	SVM (rbf)	0.573 ± 0.020	0.589 ± 0.014
	Logistic	0.523 ± 0.013	0.541 ± 0.018
Arabian Sea	XGBoost	0.226 ± 0.007	0.610 ± 0.008
	Random Forest	**0.229 ± 0.006**	**0.618 ± 0.006**
	SVM (rbf)	0.195 ± 0.004	0.545 ± 0.007
	Logistic	0.169 ± 0.003	0.498± 0.008
Bengal Bay	XGBoost	**0.379 ± 0.017**	**0.601 ± 0.01**
	Random Forest	0.364 ± 0.016	0.601 ± 0.010
	SVM (rbf)	0.346 ± 0.006	0.560 ± 0.016
	Logistic	0.289 ± 0.015	0.542 ± 0.016

Table 3.5: Results of the optimization of the classification models' hyperparameter. The reported results were obtained on the hold-out test sets based on nested 5-fold cross-validation [96, 18]. The bold font indicates the performance of the best model for a given region.

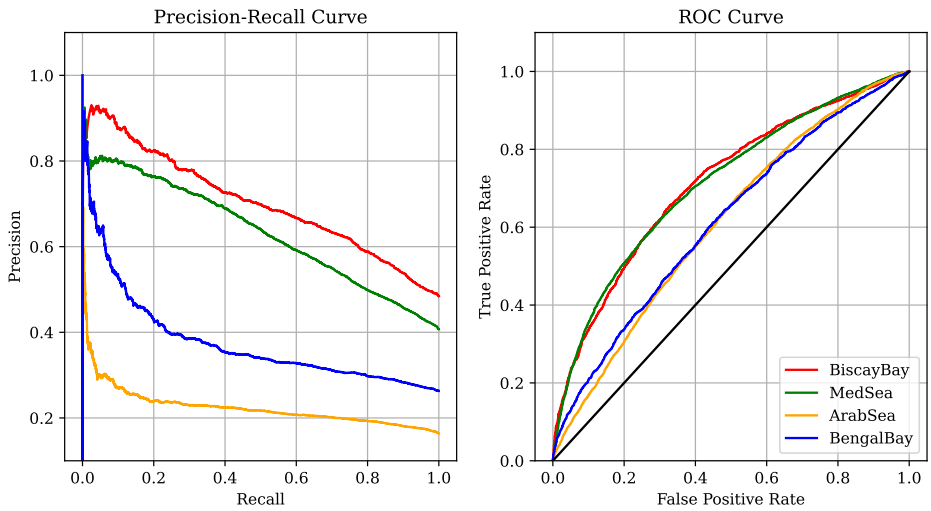


Figure 3.6: Precision-recall and ROC curves for the studied regions. The black line in the right panel – performance of a random guess classifier.

might be that the regions Biscay Bay and the Mediterranean Sea have a higher overall number of ships per image patch (and, therefore, a higher percentage of potentially well-recognizable plumes) than the two remaining regions, c.f. Figure 3.3. Next, we compare the scores of average precision. Also in the case of this metric, a perfect classifier would have a score of 1.0, while a random guess classifier would have an average precision score equal to the ratio of positive samples in the dataset. The average precision score is calculated based on a precision-recall curve, which is presented in Figure 3.6, left-hand-side plot. Due to the different rates of class imbalance of datasets from different regions, the average precision scores from the Table are difficult to compare directly. However, analyzing the precision recall-curves, we can conclude the following: the performance of the classifiers on Biscay Bay and Mediterranean Sea regions are very close to each other and the difference between the obtained average precision scores is mainly caused by a slightly different class imbalance. The lower average-precision scores for the regions Bengal Bay and Arabian Sea are also to a big extent a result of the fact that those datasets contain fewer image patches with a ship than two other regions. However, in the case of Bengal Bay, for the lower rates of recall, we can observe quite high values of precision. This signalizes the fact that there is a set of images that the model can quite confidently correctly recognize. This is not the case for the Arabian Sea, which implies better performance of the classification model on the Bengal Bay region in comparison to the Arabian Sea. For all regions, it is important to underline that the reported performances of the models were negatively affected by the presence of ships whose size and speed are known to be too small or slow to be detected by the TROPOMI instrument, which is a cause of the topic of this research, that is the study of the detection limits.

Explainability analysis

As a next step, we would like to understand which of the used features are the strongest indicators of the presence of a ship in the area for the XGBoost model. For this, we perform the explainability analysis using the SHapley Additive exPlanations (SHAP) [70] summary plots (see Figure 3.7). The plots indicate the strength of the impact of a value of a certain model feature on the model outcome (positive or negative) for individual samples from the test set. The red and blue colors show the effects of a certain feature's high and low values respectively.

We can see that for the Mediterranean Sea, and Biscay Bay, the feature having the strongest impact on the decision of the model the most is *scd std*, representing the standard deviation of stratospheric column density within the image patch. In the

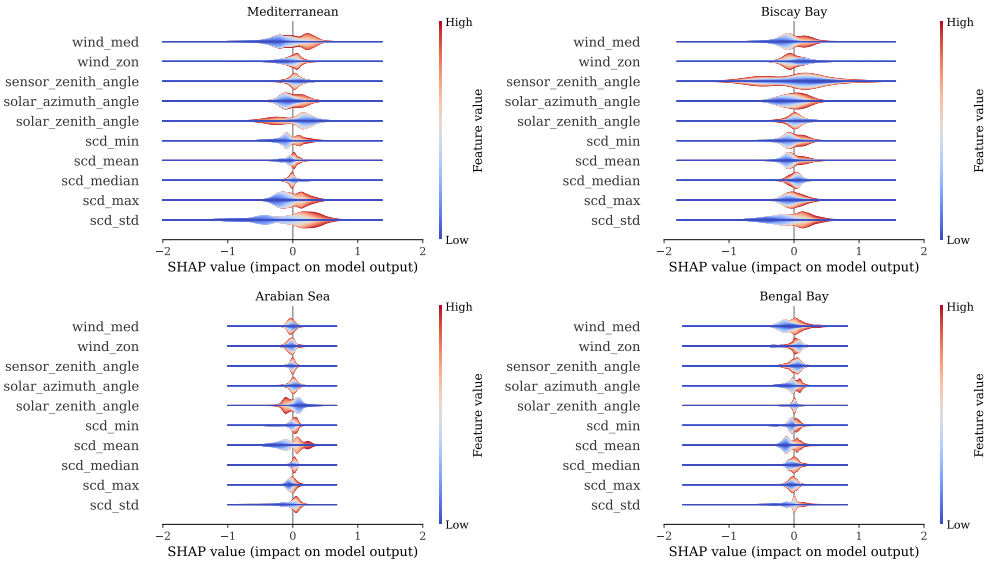


Figure 3.7: SHAP violin plots on concatenated test sets for each studied region.

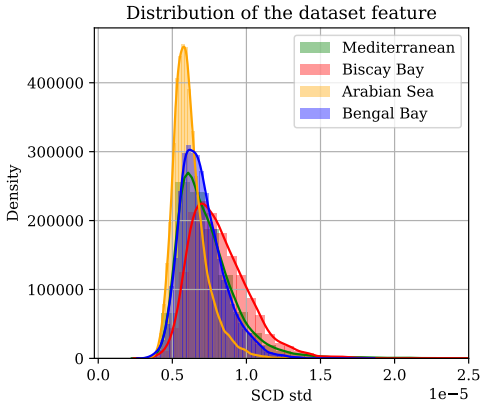


Figure 3.8: Distribution of the variable *scd std* for four studied regions. For the Arabian Sea, the distribution is noticeably more narrow than for other regions.

case of the Mediterranean Sea, *scd max* and *solar zenith angle* also play significant roles. Interestingly, the direction of the meridional wind also has a strong influence on the model's decision in the Mediterranean Sea. From the plot, we see that the negative meridional wind corresponds to strong negative model responses, potentially due to land outflow from Europe affecting ship plume visibility. In the Arabian Sea and Bengal Bay regions, the strongest impact on the model response is attributed to the values of the feature *scd mean*. Notably, for the Arabian Sea, high values of *scd std* do not necessarily indicate the presence of a plume, possibly because as we can see from Figure 3.8, standard deviations of NO_2 concentrations in this region are typically lower compared to others. Low values of *scd std*, however, are used by the model as a strong suggestion of the absence of a plume in the image patch. Finally, one can notice that for Biscay Bay, the feature *sensor zenith angle* is of great importance. However, since we do not see a clear split into high/low values for positive/negative model outcomes, the influence of the feature on the model response will depend on the values of other features [40, 47]. From this experiment, we can conclude that the same machine learning models applied to different studied regions not only yield different quality of results but are also driven by different sets of features.

3.3.2 Sensitivity limits estimation

In this Subsection, we address the first research question: What is the minimum speed and length of a seagoing ship so that the NO_2 plume from it can be detected with the detection system based on TROPOMI data? With the detection system we mean a sequence of steps needed to automatically detect an NO_2 plume from a ship on a TROPOMI image patch. The first step of this sequence is a measurement performed by the TROPOMI sensor. The last step is the application of a trained machine-learning model on the set of unseen image patches with the aim of distinguishing patches covering the area with a ship. In [41], it was shown that the length and the speed of the ship are the main factors determining the emission potential of the ship. Following the considerations presented in [41], in order to decrease the level of problem complexity, we represent the length/speed of the studied ship in terms of one variable – the ship emission proxy E_s [41], as defined in Section 2.4. For the purpose of this study, we define the sensitivity limit of the detection system for NO_2 plumes from seagoing ships using TROPOMI data for a given region as the level of ship emission proxy E_s, starting from which the classification model can distinguish image patches without a ship from image patches with a ship.

Region	Average Precision	ROC-AUC
Mediterranean	0.538 ± 0.036	0.518 ± 0.038
Biscay Bay	0.539 ± 0.053	0.513 ± 0.067
Arabian Sea	0.563 ± 0.035	0.560 ± 0.031
Bengal Bay	0.564 ± 0.054	0.540 ± 0.060

Table 3.6: Model performance when only considering the one-ship patches with the emission proxy below 10% quantile.

Given the provided definition of the sensitivity limit, our initial investigation evaluates the classification model's performance using image patches with the lowest total emission proxy. For this, we first exclusively chose patches covering a single ship. Then, from the selected subset, we further narrowed our selection to those patches with an emission proxy falling below the 10% quantile of all one-ship patches. To ensure comparability of performance metrics between areas and samples with different ship proxy values, we took a sample with an equal number of patches with and without a ship covered by the patch. To make sure that all image patches with and without ships that satisfy the above-provided criteria are used for the model training and evaluation, we repeated the sampling procedure 5 times. Subsequently, we conducted a 5-fold cross-validation for each set of sampled data points. The averaged results over the five folds are presented in Table 3.6. The outcomes indicate that none of the regions allowed for distinguishing patches with a ship, as the ROC-AUC/Average precision values obtained were not significantly higher than 0.5. Consequently, we infer that the ships with the lowest emission proxies in our dataset fall below the sensitivity limit of the detection system for NO_2 plumes from seagoing ships using TROPOMI data.

In the next experiment, we checked what the emission proxy threshold for the ship plumes detectability is. Here, we again considered only image patches with one ship on it. We then gradually removed ships with the lowest emission proxy from the dataset, analyzing the changes in the model performance. The applied emission proxy thresholds were determined as a range of quantiles starting from 10% and gradually increasing by 10%, until it reaches 90%. If after reaching a certain level of threshold, the number of patches with a ship (label 1) went below 300, the experiment was terminated and the next thresholding levels were not tested[6]. The criterion of 300 patches was established based on the number of patches with a ship left after a 90%

[6]This way, the highest applied threshold for Biscay Bay was 70% and for Bengal Bay 80% quantile.

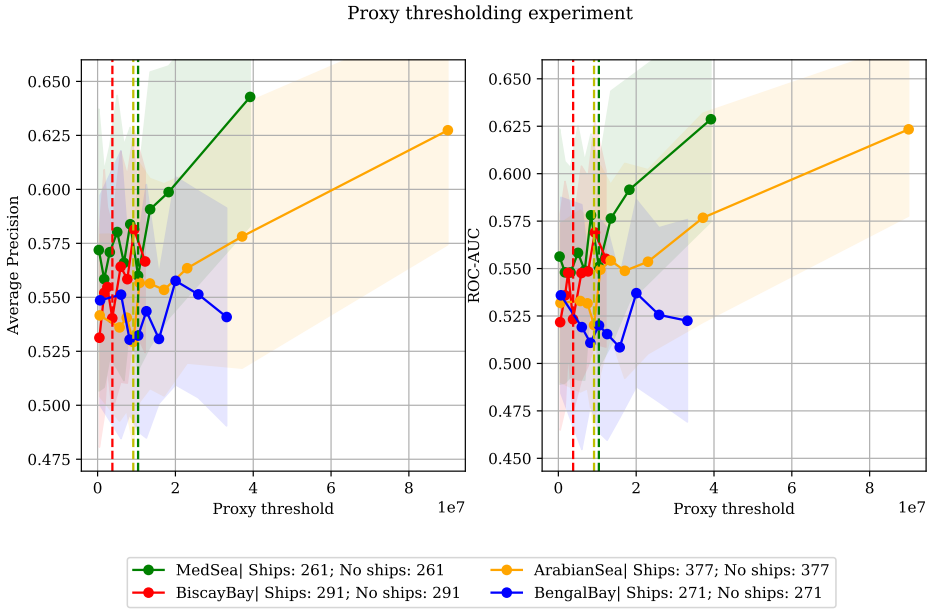


Figure 3.9: Step-wise removal of the patches (containing one ship) with the lowest emission proxy. Dashed lines indicate estimated levels of sensitivity limits for the Biscay Bay, Mediterranean, and Arabian Seas. To assure the comparability of the results, a similar size of training/test datasets was used at each threshold level.

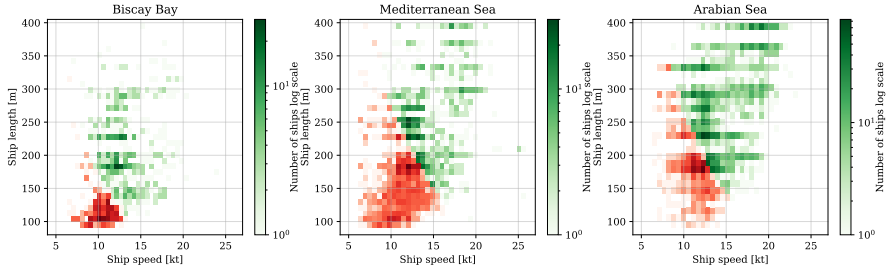


Figure 3.10: 2D histograms of speed and lengths for ships that are above (green) and below (red) the estimated sensitivity limits for the Biscay Bay, Mediterranean, and Arabian Seas.

threshold applied for the region with the highest number of one-ship patches available (Arabian Sea). Clearly, by removing the image patches with the proxy values below a certain threshold, we decreased the size of the dataset. To eliminate the potential effect of the dataset size on the model performance, throughout the experiment, we kept the dataset size constant. To achieve this, for each applied threshold, we sampled the number of data points equal to the number of data points available for the highest applied threshold. As in the previous experiment, we repeated the sampling procedure 5 times. For each set of sampled data points, we performed a 5-fold cross-validation.

The results of the experiment are presented in Figure 3.9. We can see that for the lowest thresholds, for all four regions, the average performance quality did not change. This means that the removed ships were still below the sensitivity level of the detection system for NO_2 plumes from seagoing ships using TROPOMI data. From a certain threshold (indicated with dashed lines on the plot), however, the model performance started to increase. The level of the ship emission proxy threshold starting from which we observe the improvement of the performance of the model is the sensitivity limit of the detection system for NO_2 plumes from seagoing ships using TROPOMI data for a given region. For the Mediterranean and the Arabian Sea, the sensitivity limit in terms of ship emission proxy was established to be around $1 \times 10^7 m^5/s^3$. For the Biscay Bay, the sensitivity limit is lower and is around $3.8 \times 10^6 m^5/s^3$. To get the intuition around these numbers, we return to the values of speed and length of the ship. To achieve this, for the regions of the Biscay Bay, Arabian, and Mediterranean Seas, in Figure 3.10, we present 2D histograms of the speed and length of ships that are above (green color) and below (red color) the estimated sensitivity limits. From the histograms, we conclude that to distinguish NO_2 plumes, the minimum speed of the ship for the Arabian and Mediterranean Seas should range between 10 and 15

kt depending on the length of the ship. Ships that are slower than 10 kt or shorter than 150 m are below the sensitivity limit. For Biscay Bay, the limit lies around 8 kt and 100 m. For Bengal Bay, the sensitivity limit cannot be determined since the available amount of data did not allow us to raise the proxy threshold high enough to see the increase in the performance of the model. However, when comparing the curve dynamics of the Bengal Bay with other regions, the obtained pattern suggests that the sensitivity limit for this region is higher than for the Arabian and Mediterranean Seas.

3.3.3 On detection of biggest emitters

Our second research question is how the detectability of NO_2 plumes can be improved if only the biggest emitters are taken into account. Our aim here is to understand the potential of the detectability of NO_2 plumes when the total emission proxy is very high. The high emission proxy can result from a big ship operating at a high speed, or smaller or slower ships operating in proximity to each other. Therefore, in this experiment, we considered all image patches (without, with one, or with more than one ship on it). This way, in some of the image patches, there will be more than one ship with a high emission proxy present. As in the previous experiment, we gradually removed from the dataset the image patches with the lowest total emission proxy. Once again we studied how the removal of the low emitters affects the quality of classification. The thresholds used for the proxy filtering were determined as quantiles of the proxy values of the dataset of a given region. For the Mediterranean and Arabian Sea, the applied quantiles ranged from 0 to 90%. For the Biscay and Bengal Bay, due to the smaller sizes of the datasets, the applied quantiles ranged from 0 to 80%. In Figure 3.11, we present the results of the experiment. For each of the studied regions, we can observe an increase in the model performances. We can see that for the Mediterranean Sea, for the patches with the highest total emission proxy, the ROC-AUC score can exceed 0.8. For the regions Arabian Sea and Bengal Bay, the level of the results is noticeably lower. This pattern in the results is similar to what we observed in Subsection 3.3.1.

As a next step, we checked if the dependency between the applied proxy threshold and classification performance is impacted by a certain hyperparameter configuration of the XGBoost model. We would like to know to which extent we can improve the quality of classification for the image patches with the highest total emission proxy. For this, we studied two configurations of the dataset. In the first case, we applied the highest proxy threshold for the given region (the last data point from the corresponding

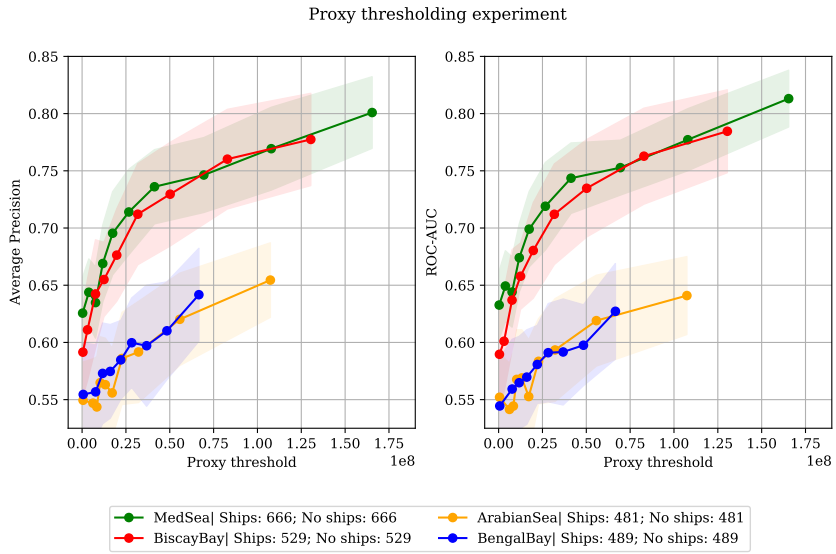
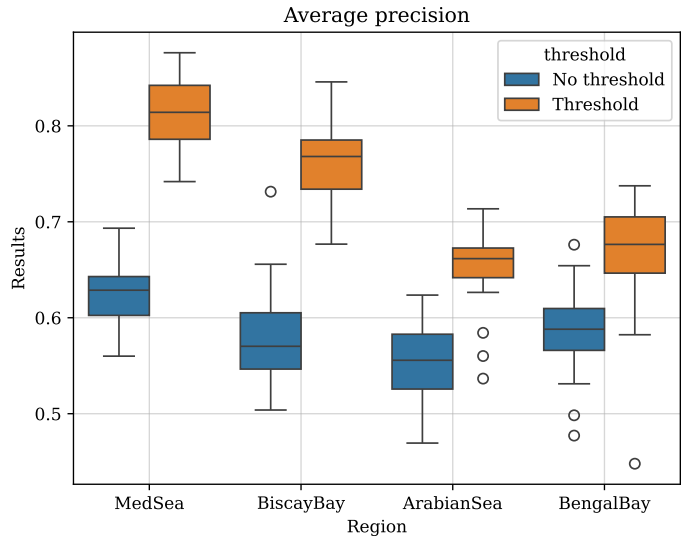


Figure 3.11: Illustration on how the step-wise removal of the image patches with the lowest total emission proxy from the dataset affects the performance of the classification model.

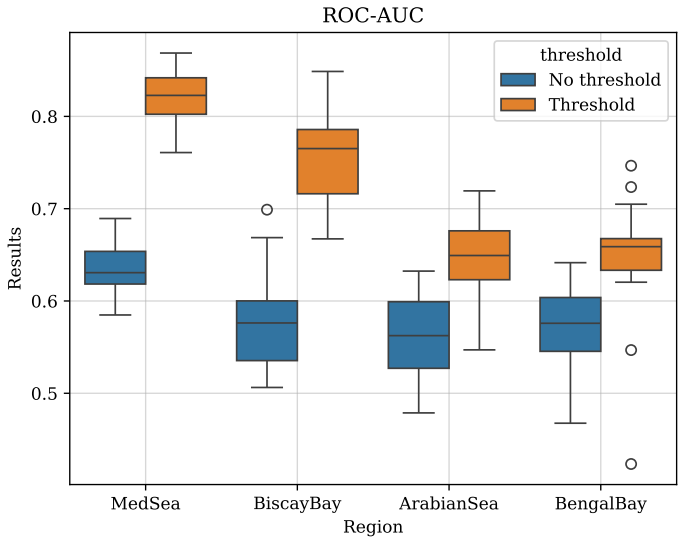
plots of Figure 3.11). In the second case, we did not apply any proxy threshold but kept the dataset size equal to the case when the proxy threshold was applied (the scenario corresponds to the first data point of the corresponding plots of Figure 3.11). For each of the datasets, we performed optimization of the hyperparameters of the classification model, in the same way as it is explained in 3.3.1. We then compared the performance of the models for both scenarios. The results are presented in Figure 3.12. For all studied regions, we can see that the quality of detecting NO_2 plumes from ships can be improved if only the image patches with the highest total emission proxy are considered. Based on this, we conclude that the dependencies shown in Figure 3.11 are not the results of a particular model configuration, but rather a property of data. However, we can see that the optimization of the hyperparameters of the model did not result in the improvement of the model performance.

3.3.4 Potential improvements in small ship detectability

In this Subsection, we address the third research question of the study. Namely, we investigate whether there is a potential for improvement of detectability of NO_2 plumes from the slow/small ships if more data would be used for the training of the



(a)



(b)

Figure 3.12: Comparison of the performance of the model when all ship images are in the dataset and when only images with the proxy above the predetermined proxy threshold are used.

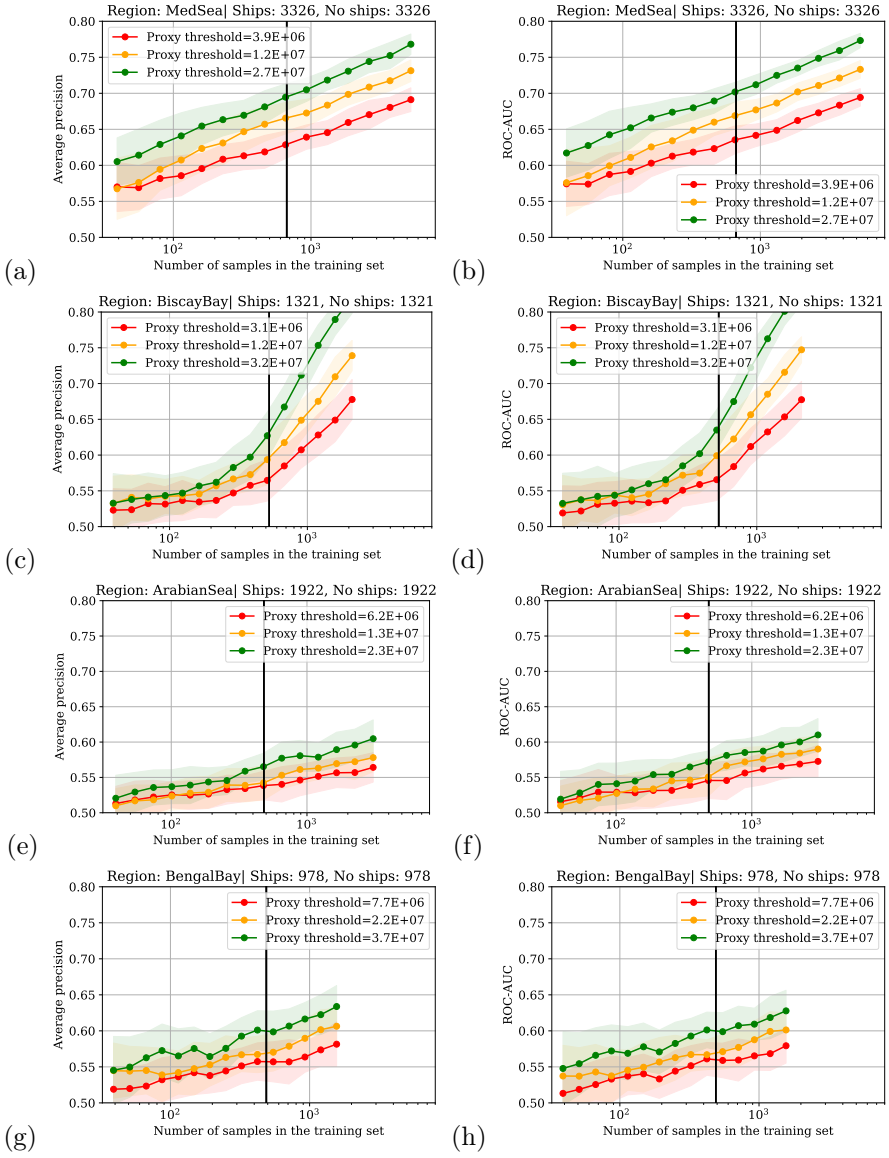


Figure 3.13: Learning curves for different levels of the applied thresholds. The black line indicates the dataset size that was used for the experiments reported in Figures 3.11, 3.12.

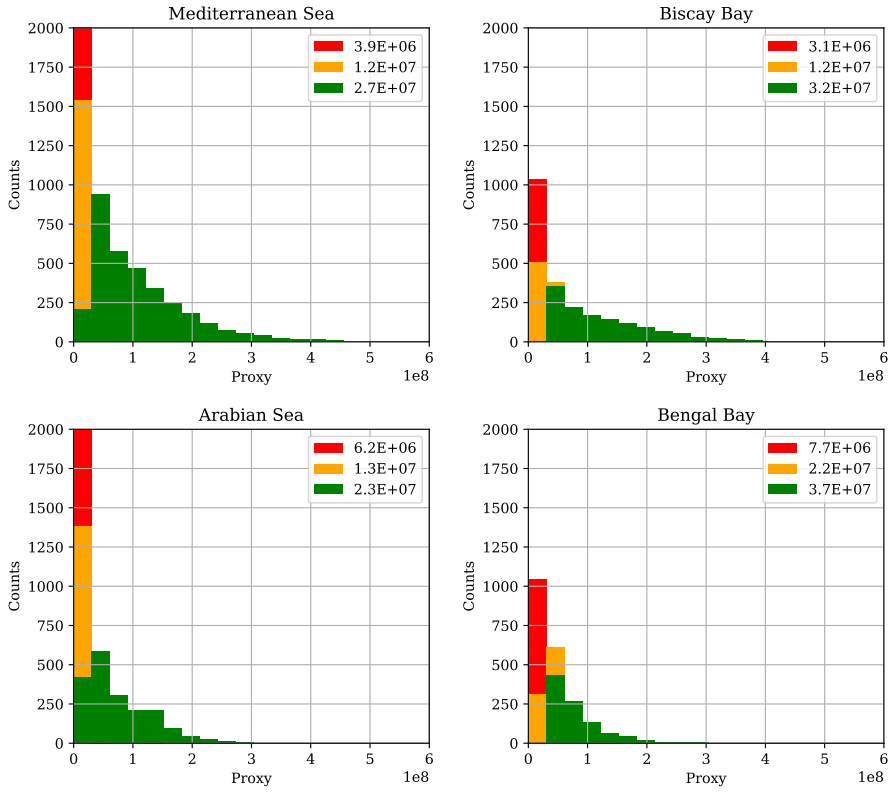


Figure 3.14: Change of the ship proxy distribution after applying thresholds as in Figure 3.13.

classification model. For each region, we selected three proxy thresholding levels and studied the change in the model performance with the growth of the size of the dataset used for the model training. We focus here on the low thresholds. The used thresholds were set as 10%, 30%, and 50% quantiles of the proxy value for the Mediterranean Sea and Biscay Bay. For the Arab Sea and Bengal Bay, the applied thresholds were 10%, 40%, and 60% due to the fact that the model performances on the lowest quantiles were indistinguishable. Similarly to the previous experiment, the maximum size of the dataset was defined by the number of data points in the dataset with the proxy value higher than the highest among the three applied thresholds.

The resulting learning curves for each of the studied regions are presented in Figure 3.13. We can see that for all studied regions, the results shown in Figure 3.11 can be improved by using more data for model training. We also observe that for the regions Biscay Bay and Mediterranean Sea, more data results in a more significant increase in performance, than for the Arabian Sea and Bengal Bay. To explain this, in Figure 3.14, we present the distribution of the variable ship emission *Proxy* for each consecutive threshold applied. The histograms show that for the Biscay Bay and the Mediterranean Sea, there are many more image patches with high values of total emission proxy than for the Arabian Sea and Bengal Bay. As a result, even after removing from the dataset the image patches with the lowest total emission proxy, for such regions as the Arabian and Bengal Bay, the models are still trained on significantly lower total emission proxies than the models for the Biscay Bay and the Mediterranean Sea.

3.4 Discussion

The main objective of this study was to investigate the sensitivity limits of a detection system for NO_2 plumes from seagoing ships using TROPOMI data, considering the speed and length of the ships that we expressed through the means of ship emission proxy. By the detection system, we mean a sequence of steps starting from the signal measurement by the sensor, followed by data retrieval, and finally the application of the developed methodology of automated detection of ship plumes. Each of these steps influences the numbers obtained in this study.

To be able to address the problem of sensitivity estimation, we build a methodology based on machine-learning classification models. This approach allowed us to effectively exploit the TROPOMI signal and contextual information while automatically separating the image patches into those, where the NO_2 plumes can and cannot

be detected. The choice of a multivariate model enabled us to take into account features important for satellite sensitivity, such as wind and satellite/solar viewing angles. Studying several machine learning classifiers of increasing complexity, we found that the XGBoost model yielded the best performance across most regions. This shows the importance of the application of complex machine-learning models for the effective identification of TROPOMI image patches with NO_2 plumes from ships with a relatively low number of features.

With the first research question (**RQ1**), we attempted to determine the minimum speed and length of seagoing ships for which the TROPOMI data-based detection system can detect NO_2 plumes. We first showed that while the smallest ships considered in our dataset are below the detection limit of the system, once reaching a certain level of ship speed/size, the signal becomes detectable. Second, for the Mediterranean Sea and the Arabian Sea, we estimated sensitivity limits of approximately $1 \times 10^7 m^5/s^3$. For Biscay Bay, the obtained limit lies around $3.8 \times 10^6 m^5/s^3$. Comparing the obtained numbers with the ship emission estimation provided in [41], we can see that our detection system allows us to correctly recognize some plumes with concentrations close to the background concentrations estimated for the Mediterranean Sea. The obtained values of emission proxy translate to the minimum detectable speed of 10 kt and minimum detectable length of 150 m for the Mediterranean and Arabian Seas and 8 kt and 100 m for Biscay Bay. Unfortunately, due to the insufficient amount of data, the sensitivity limits for the Bengal Bay region could not be determined.

With the second research question (**RQ2**), we examined the potential improvement in NO_2 plume detectability when considering only the biggest emitters. With our results, we numerically confirmed that restricting the analysis to faster/larger ships leads to enhanced detectability of NO_2 plumes. For the Mediterranean Sea region, the performance of the classification model can exceed 0.8 ROC-AUC and average precision scores. This finding suggests concentrating the focus on the larger emitters, could potentially increase the efficiency of the application and accuracy of ship emission monitoring using the TROPOMI instrument. Our analysis also revealed distinct differences in model performance quality between regions. Notably, the Mediterranean Sea and Biscay Bay consistently show better performance compared to the Arabian Sea and Bengal Bay. We can see that these variations could be attributed to variations in ship traffic density between the regions. Additional factors that potentially can influence the performances of the models are measurement conditions (e.g., number of cloudy days), differences in data quality between regions (c.f. Table 3.2), and different scales of temperature fluctuations or concentration of ozone in the background. The

last two factors affect the lifetime of NO_2. However, an in-depth understanding of this problem requires a separate study and we leave it as future work.

Our investigation into the third research question (**RQ3**), regarding the potential for improving NO_2 plume detectability from slow or small ships by utilizing more training data, again showed the variability of the results across the regions. For the Mediterranean Sea and Biscay Bay regions, an increase in data volume led to a notable enhancement in model performance. While, for the Arabian Sea and Bengal Bay, the impact of increased data, even though present, was less pronounced. One of the established reasons was the fact that for European regions we had a higher ratio of data points with a high value of emission proxy in the dataset than for the Bengal Bay and Arabian Sea. Nevertheless, the obtained results indicate that the accuracy of currently determined detection limits is perhaps constrained not by the methodology or the sensor, but by data availability.

Implications and future work

The insights gained from this study have important implications for satellite-based ship emission monitoring. By identifying sensitivity limits and optimal ship characteristics for detectability, our findings guide the scope of future studies on ship's NO_2 estimation using TROPOMI data and give an overview of the potential application of the TROPOMI instrument for ship emission monitoring. Moreover, the obtained results can be used as a benchmark sensitivity level for future satellite missions, such as, for instance, TANGO [67].

In future research, it would be valuable to explore factors beyond ship speed and length that influence detectability, such as temperature regimes, clouds, background ozone concentrations, effect of the sunglint or satellite viewing angle. Moreover, it would be valuable to perform an in-depth study explaining the observed multi-regional differences in ship plume detectability. Finally, studying different types of machine-learning architectures or including more data features in the used datasets can provide additional insights into understanding if the ship plume detectability limits can be lowered further by means of potential improvement information extraction from image patches. A possible candidate is Convolutional Neural Networks (CNN), as it was done in [38] for the detection of visually distinguishable ship NO_2 plumes. However, [63, 64] provide indications that CNN architecture might not be a suitable option for the detection of plumes that are poorly distinguishable on the TROPOMI data.

3.5 Conclusions

In this study, we investigated the sensitivity limits of the TROPOMI data-based detection system with respect to the detection of NO_2 plumes from individual seagoing ships. To the best of our knowledge, no previous research has examined this aspect, making our findings novel and significant in understanding the capabilities of the TROPOMI instrument. Our results are obtained through the analysis of four regions of interest (the Mediterranean Sea, Biscay Bay, Arabian Sea, and Bengal Bay) and can be summarized as follows:

1. We quantified the sensitivity limits of a detection system for NO_2 plumes from seagoing ships using TROPOMI data in terms of the length and speed of a ship beyond which the NO_2 plumes from individual ships cannot be distinguished anymore.

2. We also numerically showed that, as expected, the ships with higher emissions (through either greater length or speed) are more easily detected. We demonstrated such an effect by analyzing model performances with the removal from the dataset ships with the lowest emission proxy. This is agnostic to the model or studied region.

3. Then, we demonstrated that the detection of the NO_2 plumes from the ships with lower emission proxy can be improved, once more training data are added.

4. Finally, we obtained different levels of results between the studied regions. We showed that for different regions a machine learning model not only yields different levels of results but also uses different features as main indicators of the presence of a plume in an image patch. A discrepancy is noticeable when comparing the Arabian Sea and Bengal Bay to the Mediterranean Sea and Biscay Bay.

To sum up, our findings suggest that, while efficient monitoring of seagoing ships from the TROPOMI satellite is possible, the quality of ship plume detectability depends on many factors. We believe that our results provide guidelines for establishing the research scope for future studies on NO_2 ship plume detection as well as contribute to the successful application of satellite-based instruments for the monitoring of NO_2 emission from seagoing ships.