Supervised segmentation of NO2 plumes from individual ships using TROPOMI satellite data

Solomiia Kurchaba  
s.kurchaba@liacs.leidenuniv.nl  
Leiden University  
Leiden, The Netherlands

Jasper van Vliet  
jasper.van.vliet@ilent.nl  
Human Environment and Transport Inspectorate (ILT)  
Utrecht, The Netherlands

Fons J. Verbeek  
f.j.verbeek@liacs.leidenuniv.nl  
Leiden University  
Leiden, The Netherlands

Jacqueline J. Meulman  
jmeulman@math.leidenuniv.nl  
Leiden University  
Leiden, The Netherlands

Cor J. Veenman  
c.j.veenman@liacs.leidenuniv.nl  
Leiden University  
Leiden, The Netherlands

KEYWORDS  
TROPOMI/S5P satellite, NO2, maritime shipping, supervised learning, remote sensing application, ship plume segmentation

1 INTRODUCTION

The international shipping sector is one of the strongest sources of anthropogenic emission of NOx – a substance that has a negative impact both on ecology and human health. The contribution of the shipping industry is estimated to vary from 15% to 35% worldwide [17, 27], which leads to approximately 60,000 premature deaths annually [16]. While over the last 20 years the pollution produced by power plants, the industry sector, and cars has been constantly decreasing, the impact of maritime transport continues to grow [8]. This causes a big societal pressure, resulting in regulations [26] that put restrictions on emission levels that can be produced by individual ships.

All methods currently used for ship emission monitoring such as in-situ [3, 34], on-board [1], and airborne platform-based [44] have several disadvantages: they all require close proximity to a ship, they are costly, and they do not allow for monitoring on a global scale. A potential solution to the problem is the application of remote sensing instruments [39]. Studies [23, 29] show that NO2 traces of some individual ship plumes can be visually distinguished on images from the TROPOspheric Monitoring Instrument on board the Copernicus Sentinel 5 Precursor (TROPOMI/S5P) satellite [45] that was launched in 2018.

An important next step required for the development of an automated global-scale monitoring system is an automated interpretable method for the evaluation of NOx emission produced by individual ships. Among the main challenges for this development are low temporal sample rate and spatial resolution resulting in the extremely low signal-to-noise ratio. In addition, there is a high risk of interference of the ship plume with other NOx sources and a high frequency of occurrence of plume-like objects that cannot be associated with any ship. Finally, the ground truth for this task is not available. In this paper, we present a methodology that allows addressing the above-mentioned challenges. Using machine learning and taking advantage of spatial characteristics of a ship plume, the developed method allows automatic segmentation of NO2 plumes from the background, simultaneously assigning detected signal to a ship-emitter, circumventing the listed limitations.

We use NO2 retrievals from the TROPOMI/S5P as this is the only available remote sensing instrument that performs NO2 measurements with the resolution high enough1 to distinguish plumes from individual ships. To increase the number of potentially distinguishable plumes, we enhance the contrast between the ship plumes and the background. The used enhancement technique allows for a differentiation between the ship plumes and random co-occurring

1The ground pixel resolution of the TROPOMI/S5P instrument equals to 3.5 × 5.5 km² at nadir.
concentration peaks in the ships’ neighborhood. The application of the image enhancement technique also allows for an improvement of the low signal-to-noise ratio. Then, for each analyzed ship, we perform an automatic generation of a Region of Interest (ROI) that we call a *ship sector*. The purpose of the *ship sector* is to focus the area of analysis to the region where the ship plume is expected to be located. Subsequently, we normalize the *ship sector* and divide it into sub-regions. This way, we distinguish the plume of interest from all the other NO$_2$ plumes or land-origin outflows that potentially might be located within the *ship sector*. Based on the *ship sector* division, we create a set of spatial features that characterize the location of the NO$_2$ plume within the *ship sector*. Due to the absence of other sources of ground truth, each pixel of the *ship sectors* we manually label as a “plume” or “not a plume”. Trained on the manually labeled data, a machine learning model will enable us to automatically segment plumes in unseen images. We study five robust machine learning models of increasing complexity and compare their performance with the threshold-based methods used in previous studies. To validate the developed pipeline, we compare the estimated based on the segmentation results amount of NO$_2$ to the theoretically derived ship emission proxy [23].

The rest of this paper is organized as follows: In Section 2, we start with an overview of the related literature. We then provide a description of the data sources used in this study and introduce the developed methodology (Section 3). In Section 4, the reader can find the results of the study, which are followed by the conclusions in Section 5 and discussion in Section 6.

2 RELATED WORK

For more than a decade scientists were trying to use the available satellite data in order to quantify the NO$_2$ emission from the shipping industry as a whole. For instance, using the measurements from the Global Ozone Monitoring Experiment (GOME) [11] instrument onboard the second European Remote Sensing satellite (ERS-2), the authors estimated NO$_2$ emission level above the shipping lane between Sri Lanka and Indonesia [5]. With the images from the Scanning Imaging Absorption spectroMeter for Atmospheric Cartography (SCIAMACHY) [9] onboard the ENVironmental SatellitE (Envisat) mission, traces from the shipping industry over the Red Sea were quantified [37]. Finally, data from the Ozone Monitoring Instrument (OMI) [31] aboard the NASA Aura spacecraft was used to visualize a ship’s NO$_x$ emission inventory for the Baltic Sea [46]. However, due to the significantly lower resolution capabilities of the above-mentioned predecessors of TROPOMI/SCIAMACHY, OMI: 3.5 × 5.5 km$^2$, these studies were based on multi-month data averaging, which does not give the possibility to quantify the emission from individual ships.

There are many studies demonstrating the capability of TROPOMI NO$_2$ measurements to pinpoint the emission from urban (e.g. [4, 22, 32]) or industrial sources, such as mining industry [25], or a gas pipeline [43], as well as showing the possibility of the usage of TROPOMI data to quantify the positive effects of COVID-19 lockdown (e.g. [18]). Nonetheless, all studied emission sources are stationary (therefore, can be observed over an extended period of time) and emit much higher quantities of NO$_x$ than an individual ship. Thus, all the above-mentioned problems are arguably less complex than the one discussed in this paper.

In [23] it was reported that the NO$_2$ plumes produced by individual ships can be visually distinguished in TROPOMI data. However, since the NO$_2$ traces of most of the ships in the area are not sufficiently stronger than the background concentrations, only plumes of the largest ships were addressed in that study. In addition, the presented approach requires multiple manual steps, which makes it impossible to apply the method on a global scale. In [29], we introduced the first attempt of a fully automated pipeline for the estimation of NO$_2$ emission from individual ships. In the study, we showed that pre-processing of the TROPOMI signal allows a visual distinction of a greater amount of ship plumes. The plume-background separation itself, however, was based on a locally optimized threshold, established individually for each of the analyzed ships. Due to the fact that the threshold was established on the basis of the only variable (NO$_2$ concentration), it was not flexible enough for a good quality of separation of the ship plume from the background. Finally, the one-feature-based method of thresholding does not allow for differentiation of the plume produced by the analyzed ship from all the other NO$_2$ plumes that might be located in the ship’s proximity.

Machine learning has proven to be an efficient technique for solving problems in geosciences and remote sensing [30, 33]. Recently emerging studies show the efficiency of applying machine learning models to the TROPOMI data as well. For instance, in [22], the authors have built a deep convolution neural network to classify images into those that contain an NO$_2$ plume from those that do not. The developed model, however, does not differentiate the sources of the detected plumes. Moreover, the study does not provide the attempts of segmentation of the detected plumes from the background, thus there is no possibility to quantify the amount of NO$_2$ that was emitted by a given source. Several studies reported a successful application of various multivariate machine learning models for the estimation of surface-level emissions. For example, in [13] the authors created a multivariate machine learning model for the estimation of NO$_2$ emission over Germany, while in [47] the O$_3$ concentrations were estimated for California. The areas of analysis of the above-mentioned studies, however, are restricted to over-land territories. In [28] was reported the first attempt to extend near-surface concentration predictions to an ocean region. The authors acknowledged that the performance of emission estimations over the ocean is significantly more challenging than the equivalent task for the over-land areas, mainly, due to the absence of in-situ measurements possibilities. In addition, the sources of the detected emission levels have not been studied in the above-mentioned paper.

In this study, we present a pipeline that for the first time allows the application of multivariate machine learning models for the estimation of NO$_2$ emission from seagoing ships. The developed method uses TROPOMI satellite data, AIS data on ship positions, as well as ECMWF wind data for segmentation of NO$_2$ plumes from individual ships, allowing differentiation between the plume produced by the ship of interest from all the other NO$_2$ plumes in the ship neighborhood. In the following Section, the used data sources and the developed methodology is presented in detail.
3 MATERIALS AND METHODS

3.1 Data sources

In this paper, we perform the integration of several data sources: TROPOMI satellite measurements, wind data, information about the ships’ positions, and information about the properties of analyzed ships. In this Section, the reader can find a general overview of the data sources that will be utilized throughout the study. We also provide here the criteria applied for data selection, which set the scope of this study.

3.1.1 TROPOMI data. TROPOspheric Monitoring Instrument (TROPOMI) [45] is a UV–Vis–NIR–SWIR (UV–visible–near-infrared–short-wave infrared) spectrometer on board the Copernicus Sentinel 5 Precursor (S5P) satellite. The satellite was launched in October 2017 and entered its operational phase in May 2018. It is a sun-synchronous satellite with a local equatorial overpass time at 13:30. In 24 hours, the satellite performs approximately 14 orbits and with these covers the full globe.

The TROPOMI spectrometer measures spectra of several trace gases including nitrogen dioxide (NO$_2$). Since the NO$_2$ gas is the most notable product of photochemical reactions of NOx emitted by ships, it can be utilized for the ships’ emission monitoring. The maximal ground pixel resolution of the TROPOMI instrument is equal to 3.5 × 5.5 km$^2$ at nadir. Due to the projection of the satellite images, the real size of the pixel will vary, depending on the true distance between the satellite and the part of the surface of the earth being imaged. To generate images of regular size, we regridded$^2$ the original TROPOMI data into a regular-size grid of size of 0.045° × 0.045°, which for the pixel in the middle of the analyzed area translates to approximately 4.2 × 5 km$^2$.

The following filtering criteria were applied to TROPOMI data: $qa\_value > 0.5$, $cloud\_fraction < 0.5$. Previous studies [23, 29] showed that such a selection of filtering values can yield considerably good results for a given task. The detailed description of the variables can be found in [40].

In this study, 68 days from the period between 1 April 2019 and 31 December 2019 of TROPOMI measurements were analyzed. The analyzed days were mostly sunny – the distribution of the variable $cloud\_fraction$ for the scope of this study is provided in Figure 1. The studied data product: tropospheric vertical column of nitrogen dioxide [19]. Data version: 1.3.0. For the analysis, an area in the Mediterranean Sea, restricted by the Northern coasts of Libya and Egypt from the south and South coast of Crete from the north$^3$ was chosen. This particular region was selected because of the presence of a busy shipping lane connecting Europe and Asia, the high frequency of occurrence of sunny days, and relatively low levels of NO$_2$ background concentrations, which are favorable conditions for the analysis. An outline of the area studied is presented in Figure 2.

3.1.2 Wind data. In this study, we use wind data from the European Center for Medium range Weather Forecasts (ECMWF). The wind fields (wind speed and wind direction) were taken from the

ECMWF operational model analyses at a spatial resolution of 0.25°$^4$, the temporal resolution of 6 hours and altitude of 10 meters. It was shown in [23] that wind data at 10 meters altitude is an optimal choice for a task of ship-plume matching. Starting from the product version upgrade from 1.2.2 to 1.3.0 that took place on March 27, 2019, the ECMWF 10 meters wind data for coinciding time is available as a support product in the TROPOMI/S5P data file.

3.1.3 Ship-related data. Since 2002 all commercial sea-going vessels are obliged to carry on board an AIS transponder [35] which transmits information about position, speed, heading (direction), and unique identifier (MMSI) and type of each ship. Thus, another source of data used in this study is data from Automatic Identification System (AIS) transponders. At the moment there is no open access AIS data available for the region and time period as described in Section 3.1.1, as well as the quality required for this study. The data, however, can be accessed through several commercial providers. For the scope of this study, the AIS data as well as information about the dimensions of the ships were provided by the Netherlands Human Environment and Transport Inspectorate (ILT). This is the Dutch national designated authority for shipping inspections, has access to commercial databases for the AIS data set used in this study, and is participating in this research.

With the aim of reduction of the number of images where the ship plume cannot be visually detected, in our study, we only focus on ships with a speed that exceeds 14 kt. If two ships move in immediate proximity to each other, only the ship with the highest speed was taken into consideration. From the analysis were also excluded ships that are not involved in global trade, such as Yachts, Leisure Vessels, or Research Vessels. In Figure 3, the information about the dates used for this study as well as the number of ships per day studied is depicted. The differences in the number of ships per studied day can be caused by bad weather conditions on the measurement day.

---

$^2$For the data regridding HARP v.1.13 Python package was used.

$^3$lon: [19.5°; 29.5°]; lat: [31.5°; 34.2°].

$^4$For the analyzed area the spatial resolution of 0.25° × 0.25° translates to $\approx 23.4 \times 27.6$ km$^2$. 
3.2 Method

In this Subsection, we present the developed methodology. Taking advantage of the characteristics of the analyzed ship as well as wind conditions in the studied region, our approach allows the segmentation of NO$_2$ plume produced by the particular ship of interest distinguishing it from all the other concentration peaks in the surrounding area. The results produced by the proposed approach are easily interpretable and thus can be used as a reliable source of information by ship inspectors.
The method consists of the following steps: an AIS data-based interpolation of the ship tracks at the moment and just before the satellite overpass 3.2.1, definition 3.2.2 and enhancement 3.2.3 of a ship plume image, definition of a ship sector 3.2.4 that allows the further restriction of the analyzed area, normalization of the defined ship sector and split of the normalized sector into sub-regions 3.2.5 that, finally, give the possibility to retrieve the set of necessary features. These steps are described below.

3.2.1 Ship tracks. The first step is to estimate the tracks of the studied ships. Taking into account a lifetime of NO₂ equal to a few hours, we study the track of the ship starting from two hours before the moment of the satellite overpass. We distinguish two types of ship-track: the ship track, obtained based on resampling and interpolation of the AIS data, and the wind-shifted ship track. For calculation of the wind-shifted ship track, we assume that the plume emitted by a ship has moved in accordance with wind direction by a distance \( d = v \times |\Delta t| \), where \( v \) is the speed at a given location at 10 m above sea level from ECMWF for 12:00 UTC, and \( |\Delta t| \) is a time difference between the time of the satellite overpass and the time of a given AIS ship position. Both wind speed and wind direction are assumed to be constant for the whole time during which we study the plume. Such an assumption may create uncertainties in the expected position of the plume of the ship. However, with the methodology presented further in this section, such uncertainties will not affect the correctness of the ship-plume allocation.

Summing up, the ship track provides us the information on the position of the ship from where the studied ship plume was emitted. The wind-shifted ship track indicates the expected position of the center line of the NO₂ plume after displacement driven by local wind conditions. Figures 3.2.1(a) and 3.2.1(b) give examples of the ship track and its corresponding wind-shifted ship track respectively.

3.2.2 Ship plume image. Utilizing the knowledge of a ship’s position summarized in its ship track and wind-shifted ship track, we are able to focus our attention on the area that lies within immediate proximity to the analyzed ship. For this, the concept of a ship plume image (see Figure 5(a)) is introduced. The area of the ship plume image is determined based on the wind-shifted ship track as follows: the average coordinate of the studied wind-shifted ship track defines the center \((\text{longitude}_{\text{centr}}, \text{latitude}_{\text{centr}})\) of the ship plume image, the borders of the image are defined as \(\text{longitude}_{\text{centr}} \pm 0.4^\circ\). This particular size of a ship plume image was determined in order to allow for optimal plume coverage for the most typical range of ship speeds (14 kt - 20 kt)\(^6\).

Given the size of the pixel grid, such an offset results in an image of a maximal dimension of 18 × 18 pixels.

3.2.3 Pre-processing of a ship plume image. To improve the quality of the satellite signal, in the data pre-processing step, on each of the analyzed ship plume images we apply spatial auto-correlation statistic local Moran’s I [2]. In [29] we showed that the application of this technique substantially improves separability between the ship plume and the background.

The local Moran’s I spatial auto-correlation statistic allows the enhancement of the intensity of high-value pixels located in a cluster while suppressing isolated concentration peaks randomly occurring in the background. We characterize a ship plume as a cluster of pixels adjacent to each other with a concentration higher than the background average. This way, calculating the spatial auto-correlation of a ship plume image, we combine image denoising with the enhancement of the relevant part of the image.

An example of a result of an enhanced ship plume image is provided in Figure 5(b).

Formally, the Moran’s I spatial auto-correlation statistic is defined as follows: for each pixel \( i \) of a ship plume image, the local Moran’s I is calculated as:

\[
I_i = \frac{(x_i - \mu)}{\sigma^2} \sum_{j=1, j \neq i}^{N} w_{ij}(x_j - \mu), 
\]

where \( x_i \) is the value of the respective pixel, \( N \) is the number of analyzed pixels of a ship plume image (in our case 18 × 18), \( \mu \) is a mean value of all \( N \) pixels, \( \sigma^2 \) their variance, and \( w_{ij} \) is the value of an element in a binary spatial contiguity weight matrix \( W \) at location \( j \) with regards to the analyzed pixel \( i \). The value of an element of the binary spatial contiguity matrix \( w_{ij} = 1 \) for pixels that are considered to be the neighbors of the analyzed pixel \( i \), and 0 otherwise. For the study, the queen spatial contiguity [24], which is the 3 x 3 8-connected neighborhood of the analyzed (central) pixel was applied. The value \( I_i \) becomes the value of the corresponding pixel of the resulting enhanced image.

From Equation 1 we can imply the weak side of using the Moran’s I statistic for ship plume enhancement – apart from the clusters of high values, the given method enhances clusters of low-value pixels. For the area in Mediterranean Sea, in horizontal direction 0.4° = 37.4 km, in vertical direction 0.4° = 44.2 km.

\(^6\)kt - knot, a unit of speed equal to a nautical mile per hour. 14 kt = 26 km/h. 20 kt = 37 km/h.
at the same time. The methodology proposed in this study, however, is designed in a way so that this negative impact is minimized.

| Parameter                        | Value  |
|----------------------------------|--------|
| Trace track duration             | 2 hours|
| Wind speed uncertainty           | 5 m/s  |
| Wind direction uncertainty       | 40°    |

Table 1: Parameters applied for ship sector definition.

### 3.2.4 Ship sector

A plume produced by a ship at a given moment will be displaced, over time, in the direction of the wind in the analyzed area. Having the wind information available, we would like to restrict the analysis to the part of the ship plume image, where the probability to find the plume of the ship is the highest. We perform the area restriction by defining an ROI of an analyzed wind-shifted ship visualization and has no influence on further modeling.

As a starting point of the ship sector’s definition, we use the wind-shifted ship track, calculated as described in 3.2.1. We then calculate the extreme wind-shifted tracks by adding the margin of wind-related uncertainty to either side of the wind-shifted ship track. The extreme wind-shifted tracks delineate the borders of the ship sector, showing the extreme possible positions of the plume, taking the wind measurement uncertainty into account. The wind uncertainty is assumed due to the limited spatial and temporal resolution of wind data [23], based on reported in several studies ([6, 42]) measurement bias, as well as assumptions of constant wind mentioned in Section 3.2.1. Illustrations of the extreme wind-shifted tracks and the resulting ship sector are shown in Figures 5(e) and 5(f), respectively. As a result of the delineation of the ship sector area, the plume should always lie within the ship sector boundaries. Only pixels lying within the ship sector are taken into consideration in further analysis. Parameters related to the ship sector can be found in Table 1.

#### 3.2.5 Feature set construction

In order to obtain a multivariate description of the ship sector pixels, we encode the spatial information into a set of generic features. First, we perform a ship sector normalization to make spatial information in the sector comparable into a set of generic features. First, we perform a ship sector standardization of the orientation and the scale of the original normalized sector by splitting the ship sector’s image can be characterized in terms of two values: a level and a sub-sector. An illustration of the normalized sector divided into a set of levels and sub-sectors is presented in Figure 7.

### 3.3 Experiment design

Here, we describe the experimental setup used in this study: first, we describe the dataset used for the training of the multivariate models, then we explain the models used for the benchmarking, and provide a list of used multivariate classifiers. In addition, in this Section, the reader can find the description of the methods used for hyperparameters’ optimization and measures utilized for performance evaluation.

#### 3.3.1 Dataset composition

Following the steps provided in the previous subsections, we created 754 images and cropped them to an area of the ship sector. The ship sector images were enhanced by Moran’s l operator and manually labeled so that they can be used for training machine learning models. Not all ship sector images contained a visually identifiable NO2 plume. Moreover, due to the dispersion and chemical transformation of a ship plume, some parts of the plume will always be under the detection limit of the satellite and therefore, indistinguishable. Thus, labeling errors are possible. To minimize the chance of mistakes the labeler was
supported with several representations of the area of interest: the original not enhanced NO\textsubscript{2} tropospheric vertical columns for the area of a ship plume image, the enhanced with Moran’s \( I \) area of a ship plume image, and NO\textsubscript{2} tropospheric vertical columns for the full studied area in Mediterranean Sea with the positions of the neighboring ships. The descriptive statistics of the resulting dataset are provided in Figure 8. In Table 2, the information on the data distribution within the two classes of the dataset is shown. All mentioned numbers correspond to the full dataset before the training/test set division.

|                      | No plume | Plume |
|----------------------|----------|-------|
| Number of pixels     | 68646    | 6980  |
| Number of images     | 208      | 535   |

Table 2: A number of measurement points per class in the dataset.

3.3.2 Multivariate models. To exploit the potential of multivariate modeling we used several classifiers of increasing complexity: Logistic Regression, Support Vector Machines with linear kernel...
The above-mentioned models are multivariate, and thus are able to enhance and low-value enhanced NO\textsubscript{2}. A supervised learning model will be able to differentiate between the negative impact of this side effect, apart from the Moran’s I statistic. The enhancement of clusters of low-value pixels. To mitigate the result in the creation of additional high-value pixels resulting from the experiment are highly robust. Therefore, the potential mistakes in human labeling, if present in reasonable amounts, should still allow for models’ proper training.

The first feature of the model is enhanced by Moran’s I values of the pixels that were translated into a one-dimensional feature vector. As mentioned in 3.2.3, the application of Moran’s I may result in the creation of additional high-value pixels resulting from the enhancement of clusters of low-value pixels. To mitigate the negative impact of this side effect, apart from the Moran’s I, the feature set was composed of the corresponding value of NO\textsubscript{2}. This way, a supervised learning model will be able to differentiate between high and low-value enhanced NO\textsubscript{2} clusters. Other features used by the model are Wind Speed, Wind Direction\textsuperscript{10}, Ship Speed, and Ship Length. Finally, the position of an analyzed pixel within the normalized sector in terms of levels and sub-sectors was translated into the feature vectors using one-hot encoding. The resulting feature set was composed of 17 features in total. For the full feature list, see Figure 13. The used binary label indicates whether the given pixel is a part of the ship plume or not.

3.3.3 Benchmarks. To quantify the performance improvement gained by the usage of multivariate supervised models, we performed ship plume segmentation by applying a thresholding method on a single selected feature. First, we applied a thresholding method on the tropospheric vertical column of NO\textsubscript{2} TROPOMI product regressed in accordance to the description in Section 3.1.1. No image enhancement technique was applied. This simplest way of plume-background separation was used, among the others, in [37] for the quantification of NO\textsubscript{2} emission from the international shipping sector. In [23], separation of pixels related to NO\textsubscript{2} plumes from individual ships was also performed based on solely TROPOMI NO\textsubscript{2} data. In this paper, we will refer to this benchmarking method NO\textsubscript{2} threshold. Visualization of the input data for this thresholding technique can be found in Figure 9(a).

As a second benchmarking method, following the suggestion made in [29] we performed a ship plume segmentation based on images enhanced with Moran’s I statistic. The satellite image enhancement allows effective separation of a greater amount of NO\textsubscript{2} plumes. However, as mentioned in Section 3.2.3, the application of Moran’s I statistic may result in the enhancement of low-value clusters that are not a part of a plume. Visualization of the input data for this benchmarking technique is presented in Figure 9(b). In the rest of the article, we call this method Moran’s I threshold.

To overcome the problem of enhancement of low-value clusters by Moran’s I, we propose to assign zero value to all pixels of the image with intensity lower than the median of the given ship sector picture, and afterward apply the Moran’s I enhancement. This is the third benchmarking method used in this study. We call it Moran’s I on high NO\textsubscript{2}. Visualization of the results of the application of Moran’s I only on high NO\textsubscript{2} values can be found in Figure 9(c). As presented in Figure 9, for all three benchmarking methods only pixels that lie within the ship sector area were taken into account for segmentation.

3.3.4 Segmentation validation metrics. For the assessment of classification quality, we used a precision-recall curve, an average precision score (AP)\textsuperscript{11} that is defined as the area under the precision-recall curve (PR-AUC), a receiver operating curve (ROC), and, finally, area under ROC curve (ROC-AUC). Precision and Recall are respectively defined as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (4)
\]

\[
\text{Recall} = \frac{TP}{TP + FN}, \quad (5)
\]

where TP stands for true positive and corresponds to the pixels that were labeled as a “plume” and were correctly identified by

\[\text{Wind Direction feature vector was encoded into its sine and cosine components, in order to enable a continuous feature space for various wind directions.}\]

\[\text{All evaluation methods were implemented in Scikit-learn v. 0.24.2 package[36].}\]
3.3.5 Cross-validation and parameters optimization. For the model fine-tuning and model performance evaluation, nested cross-validation [12, 41] was used. In the inner loop, we performed a randomized grid-search [7] with 5-fold cross-validation to optimize the hyperparameters of the used models. The AP score was used as a target function for optimization. The performance of the best model identified during the inner loop of cross-validation is evaluated on the "hold-out" test set, generated during the outer loop of cross-validation. The above-mentioned procedure is repeated five times, generating five independent training and test sets. An illustration of the applied cross-validation scheme can be found in Figure 10. The search space of the hyperparameters for each of the analyzed multivariate models is provided in Appendix A. The optimal parameters selected for each model by the randomized grid search can be found in Appendix B.

3.3.6 NO$_2$ validation metrics. So far, we have been measuring models’ performance based on manually created labels. To evaluate
the uncertainty hidden in human labeling, the reference value is required. Due to the fact that there are no on-site emission measurements available at the scale of this analysis, it is, therefore, necessary to use a ship emission proxy to represent the reference value. Here, we propose to use a theoretically derived NOx emission proxy $E_x$ defined as follows:

$$E_x = L_s^2 \cdot U_s^3,$$

where $L_s$ is the length of the ship in m, and $U_s$ is its speed in m/s. The details of the derivation of the given measure can be found in [23], where the proxy was introduced. As it is noted in [23], the advantage of $E_x$ in comparison to other ship emission proxies (e.g. [20]) is that it can be calculated based on AIS data only, while other existing emission proxies require ship information that is not in the AIS data and is not available publicly.

The ship emission proxy is calculated for each ship of the test sets (see Figure 10). We compare the obtained values of emission proxy with the estimated on the basis of segmentation results amount of produced NOx. We estimate the amount of produced NOx by summing up NOx concentration within the pixels classified as a “plume” by each of the studied models. For the comparison between the emission proxy and the estimated amount of NOx, Pearson linear correlation was used.

## 4 RESULTS

In this section we present the results of our study. We begin with the presentation of the results of the plume segmentation model in Subsection 4.1. Appropriate segmentation quality is necessary for a correct estimation of NO2 produced by ships. In Subsection 4.2, we validate the concept presented in this paper. In the Subsection, we compare the obtained on the basis of segmentation model results of NO2 estimation with the theoretical ship emission proxy.

### 4.1 Plume segmentation

In Table 3, we report the results of the pixel classification based on a 5-fold cross-validation for all models and benchmarks studied. Figure 11 provides the corresponding precision-recall curves, obtained by averaging the scores over all cross-validation test sets. In Figure 13, we visualize the model coefficients for the linear models studied, as well as the impurity-based feature importance coefficients for the tree-based models (Random Forest and XGBoost). The obtained results can be summarized as follows:

(i) From Table 3, Figure 11, as well as Figure 12 we can conclude that nonlinear classifiers clearly outperform both linear classifiers and threshold-based univariate benchmarks. Both used measures: AP score and ROC-AUC resulted in a similar rank of the studied classifiers. With XGBoost, Random Forest, or RBF SVM models, a very high level of precision can be achieved. For the task of ship plume segmentation, our biggest interest lies in the correct segmentation of the most representative pixels of the ship plume. Thus, the obtained level of recall we consider as reasonably satisfactory. From Table 3, we can also see that the level of the standard deviation of AP scores for multivariate non-linear models is significantly lower than for linear or univariate models. This suggests that the results obtained with the nonlinear classifiers are more robust.

(ii) From Figure 13, we can see that Linear SVM, Logistic Regression, Random Forest, and XGBoost multivariate models utilize the spatial information provided by sub-sectors and levels. The complexity of the RBF SVM model does not allow the direct calculation of the importance of the utilized features. Even though due to the different nature of the models, the coefficients’ values depicted in Figure 13 cannot be compared directly, the relative differences between the models’ features go along with our intuition on where the plume produced by an analyzed ship should be located within a normalized sector. For instance, high negative coefficients for the features corresponding to a feature Sub-sector 2, tell us that if a high-value pixel occurs in the middle of the sector, it is most probably a part of the plume produced by the studied ship.

| Model                  | AP       | ROC-AUC  |
|------------------------|----------|----------|
| Linear SVM             | 0.609±0.063 | 0.935±0.009 |
| Logistic               | 0.610±0.064 | 0.936±0.010 |
| RBF SVM                | 0.742±0.031 | 0.951±0.008 |
| Random Forest          | 0.743±0.030 | 0.952±0.008 |
| XGBoost                | 0.745±0.030 | 0.953±0.007 |
| NO2 threshold          | 0.375±0.062 | 0.823±0.017 |
| Moran’s I threshold    | 0.493±0.063 | 0.912±0.011 |
| Moran’s I on high NO2  | 0.607±0.056 | 0.922±0.010 |

Table 3: Results on the test set with 5-fold cross-validation. Bold font indicates the best-obtained result. Under the dashed line: results obtained from univariate threshold-based methods that, in this study, we considered as benchmarks.
4.2 Validation with emission proxy

Figure 14 provides the correlation plots of NO\textsubscript{2} values estimated for a given ship on a given day based on the segmentation results of a given model and the theoretically derived NO\textsubscript{x} ship emission proxy \( E_s \). Table 4 gives information on the achieved level of Pearson correlation and the number of plumes that were segmented by a certain model. Here, our baseline result is the level of Pearson correlation and the number of plumes that were identified by Manual Labeling. We can see that majority of the models detected more plumes than the labeler. However, in all cases apart from XGBoost, the higher number of segmented plumes caused the decrement of the correlation score. The XGBoost model, on the other hand, was able to detect more plumes than the manual labeler, while achieving the highest correlation score. Such a result allows us to form a hypothesis that the developed machine learning-based methodology is able to segment plumes better than a human labeler. An example of a case where the XGBoost classifier identifies a plume better than the human labeler can be found in Figure 15. More experiments are, however, required in order to make final conclusions.
Figure 13: Coefficients of the features in the decision function of the linear models and the impurity-based feature importance values for tree-based models.

| Segmentation Method       | Pearson Correlation | Number of detected plumes |
|---------------------------|---------------------|---------------------------|
| XGBoost                   | 0.834               | 371                       |
| Manual Labeling           | 0.781               | 334                       |
| Random Forest             | 0.775               | 436                       |
| NO2                       | 0.774               | 334                       |
| Logistic                  | 0.766               | 452                       |
| Linear SVM                | 0.765               | 452                       |
| RBF SVM                   | 0.757               | 447                       |
| Moran’s I on high NO2     | 0.733               | 422                       |
| Moran’s I                 | 0.681               | 448                       |

Table 4: Results on the comparison between the estimated amount of NO2 and theoretically derived NOx ship emission proxy. Sorted in accordance with the achieved level of Pearson correlation. Italic font indicates baseline results.

The highest contrast between the scores of the performance metrics and the correlation with the emission proxy can be noted for the NO2 threshold benchmark model. This is due to the fact that the ship plumes composed out of one pixel in our dataset were not labeled as plumes. The substantially high correlation with the emission proxy suggests that the single-pixel plumes were, nevertheless, identified by the method correctly. An illustration of such an example is provided in Figure 16.

5 CONCLUSIONS

In this study, we presented a new supervised-learning-based method for the automatic evaluation of emission plumes produced by individual ships using satellite data. The experiments were performed using NO2 measurements from the TROPOMI/S5P satellite. We started with the enhancement of the satellite data in order to increase the contrast between the ship plume and the background. The applied image pre-processing technique enhances the intensity of high-value pixels located in a cluster (plume) and suppresses random concentration peaks in the background. We then automatically assigned a ship sector to each analyzed ship, which excludes from the analysis parts of the image where the plume of the studied ship cannot be located based on wind conditions and speed/direction of the ship.

As a next step, we presented a feature construction method consisting of the normalization of the ship sector and its division into smaller sub-regions. Each sub-region has a different probability to contain a plume produced by the ship of interest. This way, we differentiate the plume produced by the ship of interest from all the other plumes potentially located within the ship sector. The set of newly created spatial ship sector-based features allows us to perform ship plume segmentation using multivariate machine learning models. The application of the multivariate models gives the possibility to support the ship plume segmentation process with a set of additional one-dimensional features such as ship characteristics and speed.

We integrated several data sources into a multivariate dataset. We manually labeled the data, so that the problem of individual ship-plume segmentation can be addressed with supervised learning.
Figure 14: Pearson correlations between estimated (based on classification results) values of NO$_2$ emitted by each ship on a given day and a theoretical ship emission proxy. Black lines indicate a fitted linear trend. Grey lines show 30% deviations from the fitted linear trend.

We trained a set of robust linear and nonlinear multivariate classifiers and compared their performance with the segmentation results of thresholding-based univariate benchmarks. All studied non-linear classifiers showed superior results in comparison to both linear models and univariate benchmarks. With the XGBoost model, we were able to achieve more than a 20% increase in the segmentation average precision in comparison to the best benchmark univariate model.

We validated the proposed methodology using an independent measure, i.e. a theoretically derived NO$_x$ ship emission proxy that we use as a reference value. For the comparison, we estimated
the amount of NO$_2$ produced by each of the analyzed ships and calculated the Pearson correlation of the obtained results with the ship emission proxy. We compared the obtained correlations and the number of plumes segmented by each of the studied models with the results obtained from manual segmentation. We showed that with the XGBoost model we are able to segment more plumes while achieving a 6.8% higher correlation with the emission proxy than when the plumes were segmented manually. That might suggest that the proposed method is able to find plumes that are hardly or not detectable by the human eye.
The presented approach opens new perspectives for the application of remote sensing in the domain of ship emission monitoring. However, there are several points on the generalization of results, the methodology, and the TROPOMI detection limit we would like to address here.

Firstly, we would like to discuss the possibility of the application of the proposed methodology to other regions. In this study, we presented a general approach that allows for the application of machine learning models for more efficient, automated segmentation of plumes from individual ships using TROPOMI data. All steps of feature preparation can be performed on the data from any region of the globe. Nevertheless, the machine learning models will have to be retrained on the region-specific datasets.

Secondly, not all regions will be equally suitable for the performance of ship emission monitoring with remote sensing. In particular, at the moment there is no scientific evidence that under the thick layer of land-based emission outflow it will still be possible to differentiate plumes produced by ships. Therefore, areas that lie in close proximity to big cities, ports, or industrial objects are currently challenging to analyze.

The next point is the validation approaches used in this study. For the training of the machine learning model, we used human labels. Human labeling is the basis of all machine learning methods and this study pioneers ship plume segmentation with more efficient supervised learning based on human labeling. However, the dispersion and chemical transformation of a ship plume, as well as its non-rigid structure mean that there are always some parts of this plume that are at or beyond the visible detection limit of the combination of the TROPOMI instrument and the retrieval algorithm. This can cause errors in labeling as is demonstrated in Figure 15. Such mistakes if present in reasonable amounts should not affect the performance of the model, but, if the number of labeling errors is too high, the machine learning model will not be able to learn properly, and thus, the resulting performance will be very poor. The fact that non-linear models were able to easily outperform thresholding-based benchmarks suggests that the models were able to use the provided labels for training and thus, the labeling error rate is low. Nevertheless, an independent measure of the method evaluation is needed. Since the interest of our study centers around seagoing ships, the in-situ measurements cannot be considered as a potential way of method validation. The option of on-board measurement of fuel samples, cannot be performed at the scale of the study. Therefore, a theoretical measure of ship emission potential which is ship emission proxy turns out to be the only available option of a reference value for the results of this study.

The usage of the ship emission proxy, however, has its limitations. Namely, the used ship emission proxy does not take into account many factors that influence the expected level of emission for a given ship. Nonetheless, the used proxy allows us to rank the emission potential of the analyzed ships properly.

As a following, we would like to discuss the fact that even though only fast ships were taken into consideration in this study, the number of ships for which the plume was possible to distinguish is higher than the number of ships for which the plume was invisible for the labeler. This study focuses on observing emission sources at the edge of the detection limits of the TROPOMI instrument.
We would like to express gratitude to Dr. Robert Hoek (ILT) for his valuable comments and suggestions on improvements of the presentation of this article.

ACKNOWLEDGMENTS

We would like to express gratitude to Dr. Robert Hock (ILT) for his valuable comments and suggestions on improvements of the presentation of this article.

This work is funded by the Netherlands Human Environment and Transport Inspectorate, the Dutch Ministry of Infrastructure and Water Management, and the SCIPPER project, which receives funding from the European Union’s Horizon 2020 research and innovation program under grant agreement Nr.814893.

REFERENCES

[1] Harshit Agrawal, Quentin GJ Malloy, William A Welch, J Wayne Miller, and David R Cocker III. 2008. In-use gaseous and particulate matter emissions from a modern ocean going container vessel. Atmospheric Environment 42, 21 (2008), 5504–5510. https://doi.org/10.1016/j.atmosenv.2008.02.053

[2] Luc Anselin. 1995. Local indicators of spatial association—LISA. Geographical analysis 27, 2 (1995), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

[3] Jörg Beecken, Johan Mellqvist, Kent Salo, Johan Ekhollm, and J-P Jalkanen. 2014. Airborne emission measurements of SO2, NOx, and particles from individual ships using a sniffer technique. Atmospheric Measurement Techniques 7, 7 (2014), 1957–1968. https://doi.org/10.5194/amt-7-1957-2014

[4] Steffen Beirle, Christian Borger, Steffen Dörner, Ang Li, Zhaokun Hu, Fei Liu, Yang Wang, and Thomas Wagner. 2019. Pinpointing nitrogen oxide emissions from space. Science advances 5, 11 (2019), eaax9800. https://doi.org/10.1126/sciadv.aax9800

[5] S Beirle, U Platt, R Von Glasow, M Wenig, and T Wagner. 2004. Estimate of nitrogen oxide emissions from shipping by satellite remote sensing. Geophysical Research Letters 31, 18 (2004), 18202. https://doi.org/10.1029/2004GL020312

[6] Maria Belmonte Rivas and Ad Stoffelen. 2019. Characterizing ERA-Interim and ERA5 surface wind biases using ASCAT. Ocean Science 15, 3 (2019), 831–852. https://doi.org/10.5194/os-15-831-2019

[7] Yves Bengtsson, Jan F. M. L. Elberink, and David H. M. O. Miller. 2011. The proposed methodology also opens new research directions. https://doi.org/10.5194/amt-7-1957-2014

[8] K Folkert Boersma, Geert CM Vinken, and Jean Tournadre. 2015. Ships going slow in reducing their NOx emissions: changes in 2005–2012 ship exhaust inferred from satellite measurements over Europe. Environmental Research Letters 10, 7 (2015), 074007. https://doi.org/10.1088/1748-9362/10/7/074007

[9] Henrich Bovensmann, JP Burrows, M Buchwitz, Johnathan Ferekich, Sreekish Noël, VN Rozanov, KV Chance, and A P H Goede. 1999. SCIAMACHY: Mission objectives and measurement modes. Journal of the atmospheric sciences 56, 2 (1999), 127–150. https://doi.org/10.1175/1520-0469(1999)056<0127:SGSMMO>2.0.CO;2

[10] Leo Breiman. 2001. Random Forests. Machine learning 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[11] John P Burrows, Mark Weber, Michael Buchwitz, Vladimir Rozanov, Annette Ladstätter-Weißennayer, Andreas Richter, Rüdiger Delbek, Ricarda Hooghen, Klaus Bramstedt, Kai-Uwe Eichmann, et al. 1999. The global ozone monitoring experiment (GOME). Mission concept and first scientific results. Journal of the Atmospheric Sciences 56, 2 (1999), 151–175. https://doi.org/10.1175/1520-0469(1999)056<0151:FGOEGM>2.0.CO;2

[12] Luc Anselin. 1995. Local indicators of spatial association—LISA. Geographical Analysis 27, 2 (1995), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

[13] Ka Lok Chan, Ehsan Khorsandi, Song Liu, Frank Baier, and Pieter Valks. 2021. Automated detection of NOx emissions from individual ships with the TROPOMI/S5P satellite sensor. Environmental Research Letters 16, 10 (2021), 105033. https://doi.org/10.1088/1748-9362/ac2d45

[14] Mohammad Jærvi Pagliari, et al. 2018. Gridded emissions of air pollutants for the period 1974–2017. Atmospheric Measurement Techniques 11 (2018), 721–733. https://doi.org/10.5194/amt-11-721-2018

[15] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. Journal of machine learning Research 15, 12 (2014), 1–27. https://doi.org/10.5194/os-15-831-2019

[16] Gavin C Cawley and Nicola LC Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research 11 (2010), 2079–2107. https://doi.org/10.1515/joma.2013.560

[17] Klaus Bramstedt, Kai-Uwe Eichmann, et al. 1999. The global ozone monitoring experiment (GOME). Mission concept and first scientific results. Journal of the Atmospheric Sciences 56, 2 (1999), 151–175. https://doi.org/10.1175/1520-0469(1999)056<0151:FGOEGM>2.0.CO;2

[18] Tianq Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 785–794. https://doi.org/10.1145/2939672.2939785

[19] James J Corbett, James J Winebrake, Erin H Green, Prasad Kasibhatla, Veronika Eyring, and Axel Lauer. 2007. Mortality from ship emissions: a global assessment. Environmental science & technology 41, 24 (2007), 8512–8518. https://doi.org/10.1021/es070166e

[20] Monica Criska, Diego Guizzardi, Marilena Muntean, Edwin Schaaf, Frank Den- tener, John A Van Aardenne, Suvi Monni, Ulrike Doering, Jos Olijier, Va- lero Pagliari, et al. 2018. Gridded emissions of air pollutants for the period 1974–2017 within EDGAR v4. 3.2. Earth Syst. Sci. Data 10, 4 (2018), 1987–2013. https://doi.org/10.5194/essd-2018-31

[21] Jieying Ding, Ronald J van der A, JH Eske, Mas Miyu, Trisvesiyni Stavrovakut, JBRGM Van Geffen, and JP Vekklin. 2020. NOx emissions reduction and rebound in China due to the COVID-19 crisis. Geophysical Research Letters 47, 19 (2020), e2020GL091912. https://doi.org/10.1029/2020GL091912

[22] Ken Eske, Jos van Geffen, Folkert Boersma, Kai-Uwe Eichmann, Arnoud Apit- lerio, Mattia Federnaga, Maarten Smeep, Jep Peppijn Vekklin, and Diego Loyola. 2022. Sentinel-5 precursor/TROPOMI Level 2 Product User Manual Nitrogen dioxide. Technical report SP-TKNMI-L2-0021-MA.

[23] Qianzhu Fan, Yan Zhang, Weichuan Ma, Huixun Ma, Junlan Feng, Qi Yu, Xin Yang, Simon KW Ng, Qingyan Fu, and Limin Chen. 2016. Spatial and seasonal dynamics of ship emissions over the Yangtze River Delta and East China Sea and their potential environmental influence. Environmental science & technology 50, 3 (2016), 1322–1329. https://doi.org/10.1021/acs.est.5b03965

[24] Rong-En Fan, Wei-Kai Chang, Chou-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. The journal of machine learning research 9 (2008), 1871–1874. https://doi.org/10.5194/os-15-175-2019

[25] Douglas P Finch, Paul I Palmer, and Tianran Zhang. 2022. Automated detection of atmospheric NO₂ plumes from satellite data: a tool to help infer anthropogenic combustion emissions. Atmospheric Measurement Techniques 15, 3 (2022), 721–733. https://doi.org/10.5194/amt-15-721-2022

[26] Aristeidis K Georgoulias, K Folkert Boersma, Kaiser van Vliet, Xinmei Zhang, Prodomos Zanis, Joz de Laat, et al. 2020. Detection of NO2 pollution plumes from individual ships with the TROPOMI/S5P satellite sensor. Environmental Research Letters 15, 12 (2020), 124037. https://doi.org/10.1088/1748-9362/aba445

[27] Arthur Getis and Jared Aldstadt. 2004. Constructing the spatial weights matrix using a local statistic. Geographical analysis 36, 4 (2004), 90–104. https://doi.org/10.5194/ga-36-90-2004
Appendices

Appendix A  HYPERPARAMETERS SETTINGS
Below the reader can find the hyperparameters’ search space that was used for the optimization of the multivariate model’s performance, along with the hyperparameters that were always used for the model training.

- **Linear SVM** (random_state=0)
  - C: (2.0e-2, 2.0e-1, 2.0, 2.0e1, 2.0e2)
- **Logistic** (solver='saga', l1_ratio=0.5, random_state=0)
  - penalty: (‘l1’, ‘l2’, ‘elasticnet’, ‘none’)
  - C: (0.0001, 0.001, 0.1, 1)
  - max_iter: (100, 120, 150)
- **RBF SVM** (kernel='rbf', gamma = 'scale', random_state=0)
  - C: (2.0e-2, 0.5e-1, 1.0e-1, 1.5e-1, 2.0e-1, 2.0)
- **Random Forest** (n_estimators=500, oob_score=True, random_state=0)
  - min_samples_leaf: [2; 36]
  - max_features: (‘sqrt’, 0.4, 0.5)
  - criterion: (‘gini’, ‘entropy’)
- **XGBoost** (objective='binary:logistic', eval_metric='aucpr', n_estimators=500, booster='gbtree', random_state=0)
  - gamma: [0.05; 0.5]
  - max_depth: (2, 3, 5, 6)
  - min_child_weight: (2, 4, 6, 8, 10, 12)
  - subsample: [0.6; 1.0]
  - colsample_bytree: [0.6; 1.0]
  - colsample_bylevel: [0.6; 1.0]
  - learning_rate: (0.001, 0.01, 0.1, 0.2, 0.3, 0.4)
  - reg_alpha: (0, 1.0e-5, 5.0e-4, 1.0e-3, 1.0e-2, 0.1, 1)

Appendix B  HYPERPARAMETERS SELECTED BY THE OPTIMIZATION PROCESS
Here we provide the set of the hyper-parameters that were identified as optimal through the performance of randomized grid search procedure.

- **Linear SVM** In every iteration of cross-validation procedure the optimal value of parameter C was: C=0.02.
- **Logistic regression** In every iteration of cross-validation procedure the optimal value of parameter C was: C=0.001.
- **SVM RBF** In every iteration of cross-validation procedure the optimal value of parameter C was: C=0.1.
- **Random Forest**
  - CV0: criterion = entropy, max_features = 0.4, min_samples_leaf = 18
  - CV1: criterion = entropy, max_features = sqrt, min_samples_leaf = 24
  - CV2: criterion = entropy, max_features = 0.4, min_samples_leaf = 24
  - CV3: criterion = entropy, max_features = sqrt, min_samples_leaf = 18
  - CV4: criterion = entropy, max_features = 0.4, min_samples_leaf = 18
- **XGBoost**
  - CV0: gamma = 0.2, max_depth = 6, min_child_weight = 10, subsample = 0.6, colsample_bytree = 0.89, colsample_bylevel = 0.79, learning_rate = 0.01, reg_alpha = 1e-05
  - CV1: gamma = 0.15, max_depth = 5, min_child_weight = 2, subsample = 0.6, colsample_bytree = 0.89, colsample_bylevel = 0.6, learning_rate = 0.01, reg_alpha = 0.01
  - CV2: gamma = 0.2, max_depth = 6, min_child_weight = 2, subsample = 0.6, colsample_bytree = 0.89, colsample_bylevel = 0.6, learning_rate = 0.01, reg_alpha = 0.0005
  - CV3: gamma = 0.2, max_depth = 6, min_child_weight = 2, subsample = 0.6, colsample_bytree = 0.89, colsample_bylevel = 0.6, learning_rate = 0.01, reg_alpha = 0.0005
  - CV4: gamma = 0.15, max_depth = 5, min_child_weight = 2, subsample = 0.6, colsample_bytree = 0.89, colsample_bylevel = 0.6, learning_rate = 0.01, reg_alpha = 0.01