Using Machine Learning to estimate the Impact of Ports and Cruise ship traffic on urban air quality: the case of Barcelona

Alexandre Fabregat[a,*], Lluís Vázquez[a], Anton Vernet[a]

[a]*Deparment of Mechanical Engineering, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007, Tarragona (Spain)*

Abstract

Maritime activity is known to increase pollutant concentration levels in neighboring cities. In major touristic destinations, the singular need of cruise liners to keep supplying energy to on-board services and amenities while docked, has raised concerns about this industry contribution to pollutant emissions. To estimate the impact of port activities and that exclusively due to cruises, classical approaches would rely on atmospheric dispersion models. Although these tools retain the underlying physics, lack of details on background flow state and emission inventories limits their predictive capabilities. Using historical data on pollutant concentration, meteorology and traffic intensity at specific locations across the city of Barcelona, it was found that predictions of local pollutant concentration by the present Machine Learning tool are more accurate than those provided by the CALIOPE-Urban-v1.0 in our test cases. Estimated air quality impact due to cruise ships is shown to be limited in comparison to overall Port effects.

Keywords: Urban air pollution, Cruise ships, Generalized Boosted Regression Models, Machine Learning

*Corresponding author

Email addresses: alexandre.fabregat@urv.cat (Alexandre Fabregat), lluis.vazquez@urv.cat (Lluís Vázquez), anton.vernet@urv.cat (Anton Vernet)

Highlights

- Machine Learning is proposed as an alternative to classical dispersion models.
- Working dataset build from pollutant concentration, weather and traffic intensity.
- ML local predictions found to be more accurate than those from standard approaches.
- Main features explaining pollutant concentration variability are identified.
- Cruise ship activity impact on air quality of Barcelona metro area is quantified.

1. Introduction

Transportation represents a significant fraction of the total emissions of several chemical compounds that directly impact human health including Carbon Monoxide, Nitrogen and Sulfur Oxides, particulate matter, tropospheric Ozone and Volatile Organic Compounds (VOC) to mention the most relevant [1, 2]. Maritime traffic contribution to global emissions, both freight and passengers, has continuously grown over the last decades [3, 4, 5, 6]. Despite notable efforts to reduce its impact by, for instance, limiting the sulphur content in marine fuel [7], globally, maritime transport is still characterized by consuming low quality fuel in comparison to road and air traffic [8]. At port, shipping activities have been found to contribute to increased levels of air pollutants in neighboring urban areas [9, 10].

Cruise shipping has raised special concerns because of its particular mode of operation characterized by the continuous supply of fossil fuel energy to the various on-board services and amenities during docking ('hotelling') [11]. The average energy consumption used in manoeuvring cruise liners has been estimated to be only 5% of that used at dock to provide light, heating, ventilation, air-conditioning, cold storage, cooking and other services estimated at 2.5 MW per average cruise ship [12]. On one hand, previous efforts to quantify the impact of ports on air quality have been directed to characterize the pollutant emissions of docked vessels (number and location of sources, duration and rate, composition...) and use this information to feed more or less complex atmospheric dispersion models to predict the air quality in neighboring areas [12, 13, 9, 14]. Other studies analyzed pollutant concentration measurements to statistically infer connections between overall port

activity and air quality levels [15, 16, 17]. Isolated impact of cruise ships and its peculiar regime of sustained emissions during hotelling on major touristic destinations and areas of ecological value has also been the subject of research [15, 18, 19, 14, 13].

In general, the transport of a chemical species i is governed by the convection-diffusion-reaction equation that can be written as:

$$\frac{\partial \rho Y_i}{\partial t} + \nabla \cdot (\rho \vec{u}\, Y_i) = -\nabla \cdot \vec{J_i} + R_i + S_i \tag{1}$$

where ρ is the background (mixture) fluid density, t is time, $\vec{u}$ is the mixture velocity field, Y_i is the local mass fraction of chemical species i, $\vec{J_i}$ is the mass diffusion flux, R_i accounts for any production/consumption of i due to chemical reactions and S_i represents any other source/sink of i.

Full solutions of Eq. (1) are usually unavailable for, at least, two reasons. On one hand, local atmospheric velocity field $\vec{u}$ responsible for advecting the different chemical species in the turbulent planetary boundary layer is typically not well specified. Secondly, rates of emission, consumption by chemical reaction and surface deposition in pollutant inventories are, in many cases, approximations of varying accuracy. Several orders of magnitude smaller than the convective transport counterpart, diffusive contribution is usually neglected.

Classical approaches have sought for $Y_i(\vec{x}, t)$ by numerically integrating some approximated form of Eq. (1) [20, 21, 22]. In Gaussian Plume Models, one of the most commonly used approximation [23], wind velocity and direction are assumed constant and turbulent transport is parametrized using horizontal and vertical standard deviations of the emission distribution

[1]. Other approaches to urban pollutant dispersion modelling involve the spatial and temporal discretization of some convenient form of the chemical species transport equation Eq. (1) over a computational domain covering the geographical region and time period of interest. Instantaneous velocity and concentration are approximated by some form of averaged/filtered fields resulting from modelling the turbulent contribution to the transport of momentum, mass and heat. Common methodologies involve embedded computational meshes with a coarse far field level where boundary conditions are imposed and progressively finer grids where detail is needed. Although this approach retains part of the physics that governs atmospheric dispersion, it also has important inherent sources of error due to the impossibility of reconstructing the full turbulent hydrodynamic field and the lack of detailed characterization of each pollutant emission and chemical transformation rates.

The approach presented here uses Supervised Machine Learning (ML) techniques to elucidate the relation between the local concentration of a given pollutant (the target or response) and several relevant variables for pollutant transport including meteorology and traffic intensity (the features or predictors). Use of Machine Learning algorithms to exploit the vast amounts of data retrieved from city-wide monitor networks has grown in recent years [24, 25, 26, 27]. While classical approaches retain the underlying physics of pollutant dispersion and are well suited for assessing the impact of different potential scenarios on local air quality (by, for example, considering different spatial locations and emission rates for individual pollutant sources), when robust and accurate estimations of the functionality between the features and the response exist, ML methodologies can accurately predict the

isolated effect of each predictor on pollutant concentration levels regardless of the availability of precise descriptions of the background flow state and pollutant inventories.

The model presented in this work is especially intended to estimate the overall impact of port activity and that exclusively due to cruise ships on the air quality of a major coastal touristic destination. The city of Barcelona has been chosen as a benchmark because it is one of the major travel destination with a total of ∼9 million tourists in 2018 [28, 29] and its Port impact on the metropolitan area has already been the subject of significant research efforts [30, 31, 32].

This work is organized as follows: Sec. 2 describes the methods and materials involved in obtaining the working dataset used to train and tune the air quality predictive model. Sec. 3 presents the main results and quantifies the contribution of the Port of Barcelona activity to increased pollutant concentration for several chemical species and locations across the metropolitan area. Sec. 4 discusses the hypothesis, assumptions and potential future improvements to the current methodology and Sec. 5 summarizes the major findings.

2. Methods

2.1. Overview

In order to analyze the impact of the port activity and cruise ship traffic, we present and discuss an implementation of Supervised Machine Learning techniques for prediction of air quality levels in urban environments. The

specific goal is to obtain, for each pollutant species and selected spatial location, accurate estimations of the function f that captures the dependence between the local pollutant concentration — or *response* Y — and a set of n variables — or *predictors* X_j — this response depends upon, i.e.,

$$Y = f(X_j) + \varepsilon = \widehat{Y} + \varepsilon, \qquad j = 1 \dots n, \tag{2}$$

where $\widehat{Y}$ is the predicted concentration and ε is the error associated to unexplained variability in Y.

By inspection of Eq. (1), predictors must include information on both local atmospheric state and source emission rates of each pollutant. Precise estimations of pollutant concentration will be available as long as data on X_j and Y allow to elucidate the function f that explains most of the variability in Y and minimizes ε, this is, the difference between observed and predicted values.

2.2. Dataset description

The working dataset used in this work and described in this section contains pollutant concentration measurements, weather measurements and road, air and maritime traffic data.

2.2.1. Pollutant concentration

The response dataset is comprised of hourly concentrations measurements of Nitric Oxide (NO), Nitrogen Dioxide (NO_2), Total Nitrogen Oxides (NO_x), Sulfur Dioxide (SO_2), Ozone (O_3), Particulate Matter 10 µm or less in diameter (PM_{10}) and Carbon Monoxide (CO). The data is retrieved from $M = 8$ pollutant stations belonging to the Air Pollution Monitoring and Forecasting Network (XVPCA) of the Catalan Government [33] that are distributed

across the metropolitan area of Barcelona as shown in Fig. 1 (dark red markers). Height and type of station is shown in Tab. 1. All pollutant concentration is in $\mu g\,m^{-3}$ except CO that is in $mg\,m^{-3}$. Note that the relation between concentration and mass fraction of species i can be expressed as $c_i = Y_i\rho/M_i$ where M_i is the molar weight of i.

2.2.2. Weather

The weather dataset includes measurements on wind velocity ($km\,h^{-1}$) and direction (°), relative humidity (%), atmospheric pressure (hPa), precipitation (mm), temperature (°C) and solar irradiance ($W\,m^{-2}$) with a 30-minutes sampling rate from $N = 11$ different meteorological stations across the metropolitan area of Barcelona and surroundings (the location of the closest stations to the metropolitan area are shown in Fig. 1 in blue). These data have been obtained from the network of automatic weather stations (XEMA) of the Meteorological Service of Catalonia [34].

This predictor set is extended with daily measurements of cloud cover from a single station located at the Observatori Fabra station (see Fig. 1) obtained from the European Climate Assessment and Dataset (ECA&D) [35].

2.2.3. Traffic

Pollutant emissions from combustion of maritime fossil fuels are incorporated into the model by parsing traffic log files of the Port of Barcelona [36] (magenta marker in Fig. 1). The available data includes arrival and departure time stamps for each vessel and ship type (used to identify cruise liners) and size (length and beam). This data has been used to generate two predictors for the hourly number and median size of total vessels and cruise

liners.

Predictors for air traffic activity consist of hourly number of arrival and departures from the Airport of Barcelona [37] (green marker in Fig. 1).

Hourly predictors for road traffic intensity expressed as an index ranging from 1 (fluid) to 6 (total congestion) have been obtained from the extensive network of traffic measurement points across the city [38] (orange markers in Fig. 1).

Distributions of the number of total vessels and cruise ships simultaneously docked at the Port of Barcelona over the time span considered in this work ranging from October 2017 to March 2020 (both included) as well as the distribution of flight operations at the Airport of Barcelona are shown in Appendix B.

2.3. Preprocessing

Weather and pollutant concentration measurements were joined by their respective observation time stamps into a synchronized single dataset. Daily measurements of cloud cover from a single weather station (Barcelona—Observatori Fabra) were assumed constant over the corresponding 24 hour period and over the entire metropolitan area. The working dataset was completed by merging the three traffic data collections by their time stamps.

Estimations for road traffic intensity σ and every weather predictor θ on each pollutant station $i = \{1 \dots M\}$ were obtained by weighted interpolation of each observation. The weighting factor was set to the 4-th power of the distance d_{ij} between the pollutant station i and each of the $j = \{1 \dots N\}$ meteorological stations to ensure that closest observations contributes the most to the interpolated weather predictor. Analogously, the distance d_{ik}

between the pollutant station i and each of the $k = \{1 \dots Q\}$ road transit measurement points were used in the spatial interpolation of traffic intensity. Specifically,

$$\theta_i = \sum_{j=1}^{N} \hat{\theta}_j \eta_{ij}, \qquad \sigma_i = \sum_{k=1}^{Q} \hat{\sigma}_k \eta_{ik}, \tag{3}$$

where η_{ij} is the distance weight from each pollutant station i to each meteorological station j and η_{ik} is the distance weight from each pollutant station i to each to each traffic measurement location k:

$$\eta_{ij} = \frac{s_{ij}}{\sum_{j=1}^{N} s_{ij}}, \qquad \eta_{ik} = \frac{s_{ik}}{\sum_{k=1}^{Q} s_{ik}}, \tag{4}$$

where

$$s_{ij} = \frac{1}{d_{ij}^4}, \qquad s_{ik} = \frac{1}{d_{ik}^4} \tag{5}$$

The wind direction predictor is expressed in terms of its alignment with respect to the direction connecting each pollutant station i and the two major transportation infrastructures likely to affect the air quality in the metropolitan area, the Port and the Airport of Barcelona.

Given each pollutant station bearing angle with respect to the Port α_{iP} and the Airport α_{iA} and β_i being the direction the wind blows to, alignment with respect to the Port and Airport can be computed as:

$$a_{iP} = \left| \frac{|\alpha_{iP} - \beta_i|}{180} - 1 \right|, \qquad a_{iA} = \left| \frac{|\alpha_{iA} - \beta_i|}{180} - 1 \right|, \tag{6}$$

where α_{iP}, α_{iA} and β_i are expressed in degrees with the origin pointing to the North and increasing in the clockwise (east) direction. The values of a_{iP} and a_{iA} are constrained in the range $0 \leq a_i \leq 1$ where 0 indicates wind blowing against the i-th station as seen from the corresponding infrastructure and 1 indicates complete alignment.

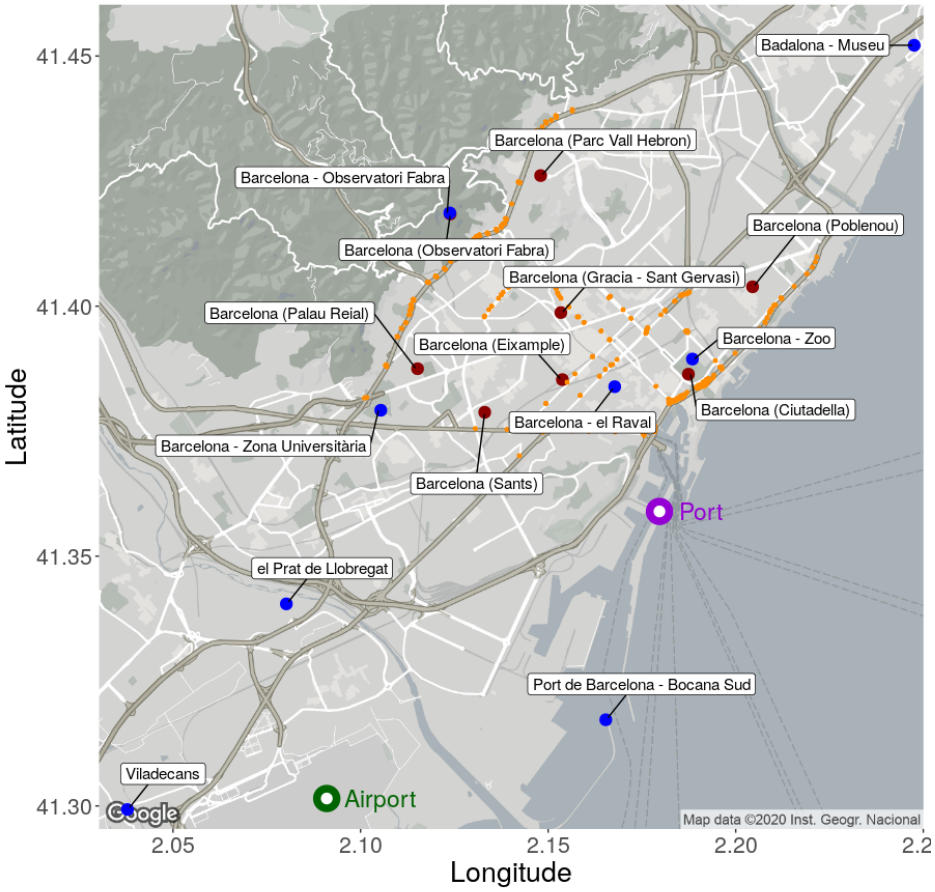

Figure 1: Map of Barcelona showing the location of pollutant (dark red) and meteorological (blue) measurement stations and traffic intensity measurements (orange). Cloud cover data comes from a single station (Observatori Fabra). Representative locations of the Port and Airport of Barcelona are shown in magenta and green respectively.

2.4. Working dataset

The working dataset is only comprised of complete observations, this is, all predictors for a given hourly time stamps are known. The temporal span of the working dataset ranges from October 2017 to March 2020 both included. The list of the 25 predictors used in this work including a description and the variable name is shown in Tab. 2. The data completeness for each pollutant and station defined as the ratio between the actual number of observations and the theoretical number of hourly observations over the 30 months period is shown in Tab. 1. Note that while nitrogen oxides concentration data exists for all stations, other pollutants are not available for all measurement points.

Station name	Data completeness (%)							Height (m)	Type
	CO	NO	NO_2	NO_x	O_3	PM_{10}	SO_2		
Ciutadella	—	80	80	80	79	—	—	7	BU
Eixample	79	80	80	80	79	79	80	26	TU
Gràcia-Sant Gervasi	80	79	79	79	80	68	80	57	TU
Observatori Fabra	—	63	63	63	61	43	—	415	BSU
Palau Reial	72	74	74	74	71	67	74	81	BU
Parc Vall Hebron	75	78	78	78	78	75	79	136	BU
Poblenou	—	64	64	64	—	62	—	3	BU
Sants	—	79	79	79	—	—	—	35	BU

Table 1: Pollutant station data completeness (%) for each pollutant, height and type ('BU' = Background Urban, 'BSU' = Background Suburban, 'TU' = Traffic urban).

To illustrate the data contained in the working dataset, Fig. 2 shows the temporal evolution of some variables (including the response for NO_x) for the Barcelona—Eixample station over March 2019.

Predictor description	Variable
Hourly wind alignment between Port and pollutant station	port_align
Hourly wind velocity	wind_velocity
Daily cloud cover	cloud_cover
Hourly wind alignment between Airport and pollutant station	airport_align
Hourly precipitation	precipitation
Hourly temperature	temperature
Hourly atmospheric pressure	pressure
Hourly solar irradiance	irradiance
Hourly relative humidity	rel_humidity
Hourly traffic intensity	road_traffic
Hourly total number of vessels	number_vessels
Hourly number of cruise ships	number_cruises
Median length of hourly total vessels	vessel_length
Median length of hourly cruise ships	cruise_length
Day is Monday	monday
Day is Tuesday	tuesday
Day is Wednesday	wednesday
Day is Thursday	thursday
Day is Friday	friday
Day is Saturday	saturday
Day is Sunday	sunday
Day number of the year	day_number
Year number	year
Hour of the day	hour
Hourly number of flights into/from the Airport of Barcelona	air_traffic

Table 2: List of predictors (or features) and corresponding variable names.

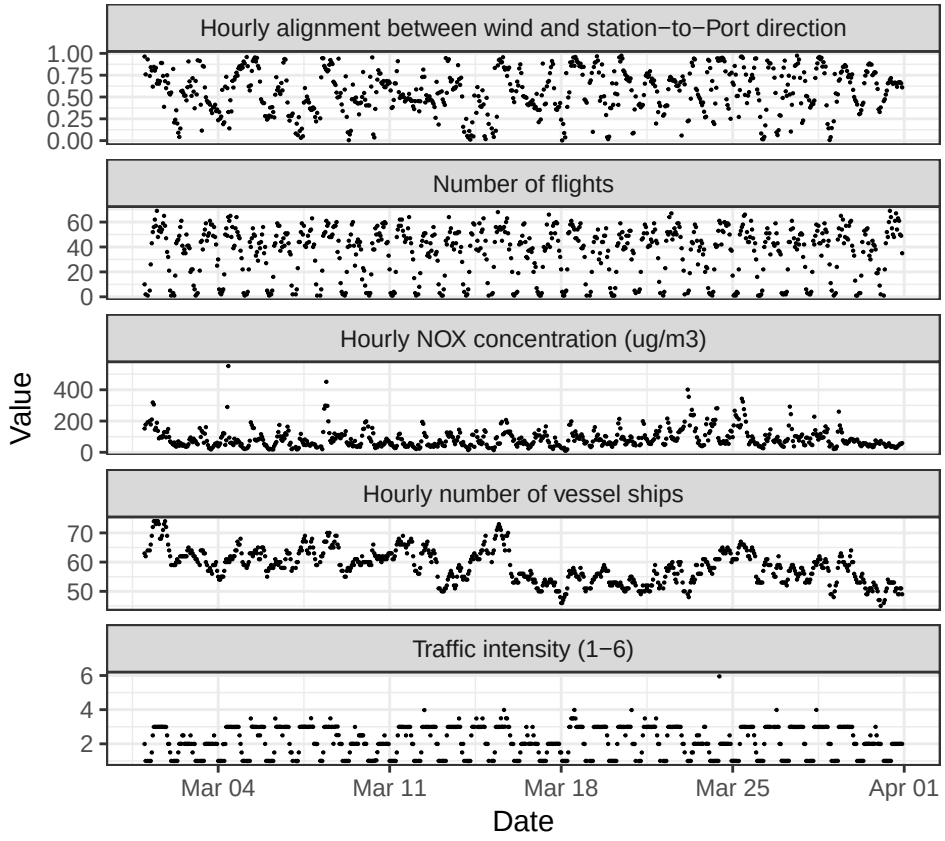

Figure 2: Temporal evolution of hourly wind alignment with respect to the Barcelona—Eixample station, number of flights, hourly concentration of NO_x, Hourly number of total vessels and traffic intensity over March 2019.

2.5. ML procedure

The R package *mlr 2.17.1* [39] has been used for data pre-processing, resampling, hyperparameter tuning and post-processing. Regarding the pre-processing, all continuous features have been normalized by subtracting the average value and dividing by the standard deviation for each pollutant species and measurement station. Categorical variables have been converted

to continuous features using a *1-of-1* method.

In this work, six different learning algorithms (or learners) widely used in ML applications have been tested, namely Gradient Boosting Machine or GBM (*gbm*), Support Vector Machines or SVM (*ksvm*), Random Forest (*h2o.radnomForest*), Multivariate Adaptive Regression Splines (*earth*), Feed-Forward Multilayer Artificial Neural Network (*h2o.deeplearning*) and Feed-Forward Neural Networks for Multinomial Log-linear models (*nnet*) where the name of the integrated regression model in the R package *mlr 2.17.1* [39] appears in parenthesis. A 5-fold cross validations resampling strategy has been used to assess the performance of each learner. Thus, the working data set described above has been repeatedly split into a training and a test set. After optimizing each learner hyperparameters, the training set has been used to derive the corresponding model which performance has then been estimated using the testing set. The benchmark results in terms of mean squared error for each pollutant across all stations for each learner are shown in Appendix C. Results clearly indicate that, for this specific problem and working dataset, the Gradient Boosting Machine (GBM) outperforms the rest of the algorithms considered in this work. Optimal GBM hyperparameters were found to be around 5000 trees and an interaction depth of 25 for most of pollutants and measurement stations. All results in Sec. 3 have been obtained with the GBM methodology.

All computations have been performed on a Dell AIO 7770 workstation with a Intel(R) Core(TM) i9-9900 CPU @ 3.10GHz. Learner training and testing, hyperparameter tuning and resampling typical CPU time is 45 minutes for each station and pollutant. Single point prediction takes less than 1

ms.

3. Results

3.1. Validation

To assess the predictive performance of the ML-based tool presented in Sec. 2, predicted concentration levels are compared with those reported by Benavides *et al.* [40] who used the CALIOPE-Urban v1.0 *physics-based* model to predict NO_2 concentration at three of the pollutant measurement stations listed in Tab. 1. CALIOPE-Urban v1.0 is the result of coupling CALIOPE [41] — an operational mesoscale air quality forecast system based on the HERMES (emissions) [42], WRF (meteorology) [43] and CMAQ (chemistry) [44] models — with the urban roadway dispersion model R-LINE [45].

The results shown in Tab. 3 indicate that the *ML-based* model outperforms the CALIOPE-Urban v1.0 platform in every metric for all available stations. Definitions of Geometric Mean Bias (GeoMean), correlation coefficient (R), Mean Bias (MB) Root Mean Square Error (RMSE) and Fraction of predictions within a factor of 2 in observations (FAC2) can be found in the Appendix A.

3.2. Feature importance

The relative importance of each predictor has been determined using the entropy-based *information gain* [39] that returns an *importance weight* for each feature that maximizes the amount of captured variability in the response. The four most important features for each station are shown in

Metric	Eixample		Palau Reial		Grcia-Sant Gervasi	
	Benavides *et al.* [40]	Current	Benavides *et al.* [40]	Current	Benavides *et al.* [40]	Current
GeoMean	0.83	0.97	1.10	0.92	1.07	0.96
R	0.55	0.81	0.57	0.78	0.52	0.82
MB	8.57	-0.04	1.23	0.07	6.00	-0.10
RMSE	26.70	14.16	21.57	13.12	25.11	14.74
FAC2	0.86	0.97	0.73	0.85	0.79	0.94

Table 3: Comparison of statistics between Benavides *et. al* and the current study.

Fig. 3 in descending order from top to bottom rows. Results indicate that the day of the year is the the most important feature for almost all background urban (in black) and the only background suburban (in red) stations. Although this predictor also explains most of the variability in traffic urban stations (in blue), this leading role is shared with the year number in the case of CO, the traffic intensity for NO_x and the temperature for PM_{10}. The hour of the day is the second most important feature for NO_x and O_3 at most stations, with some exceptions for which this position is occupied by traffic intensity, wind speed and temperature. In the case of CO, the second most relevant predictor at the traffic urban stations is the traffic intensity while the year number dominates in background urban sites. This latter feature is also the second most important variable in the case of SO_2. This suggests that, despite the relatively small number of years in the working dataset, the concentrations for these two species exhibit significant changes over the 2017-2020 period. In contrast to the other species, the PM_{10} concentration at most locations has been found to be controlled in second place by the temperature.

While the top two most important features were dominated by temporally-related predictors, namely the hour of the day, the day of the year and the year number, the third and fourth positions, with a much smaller impact on the concentration variability as discussed later, exhibit notable variability across stations and pollutants. Thus, carbon Monoxide concentration eventually exhibits some dependence on the docked ship size. Similar behavior is exhibited by the sulphur dioxide. NO_x, mostly controlled by the day of the year and the hour of the day, has also been found to change with the intensity of the traffic, the wind velocity and the vessel size. Third and fourth features for Ozone include the wind velocity and alignment with respect to the Port and the temperature. Finally, the particulate matter, strongly dependent on the day of the year and the temperature, exhibits also some dependence on the traffic intensity, the hour of the day, the wind alignment with respect to the Port and the number of flights at the Airport.

The *importance gain* method for each station and pollutant are shown in Fig. 4. This metric has been normalized such that its sum across all predictors is 100. In general, the results indicate a consistent distribution of importance between stations and pollutants with most of the variability in concentration explained by the hour of the day, the day number of the year, the road traffic intensity and the temperature. On the other side, weekday predictors are found to be the least important features.

Notably though, the model reveals a few notable peculiarities. First, the results in the suburban station (Observatori Fabra) located at an altitude more than twice that of the second one in height, suggest that ozone concentration is mostly controlled by the day of the year. These results are in

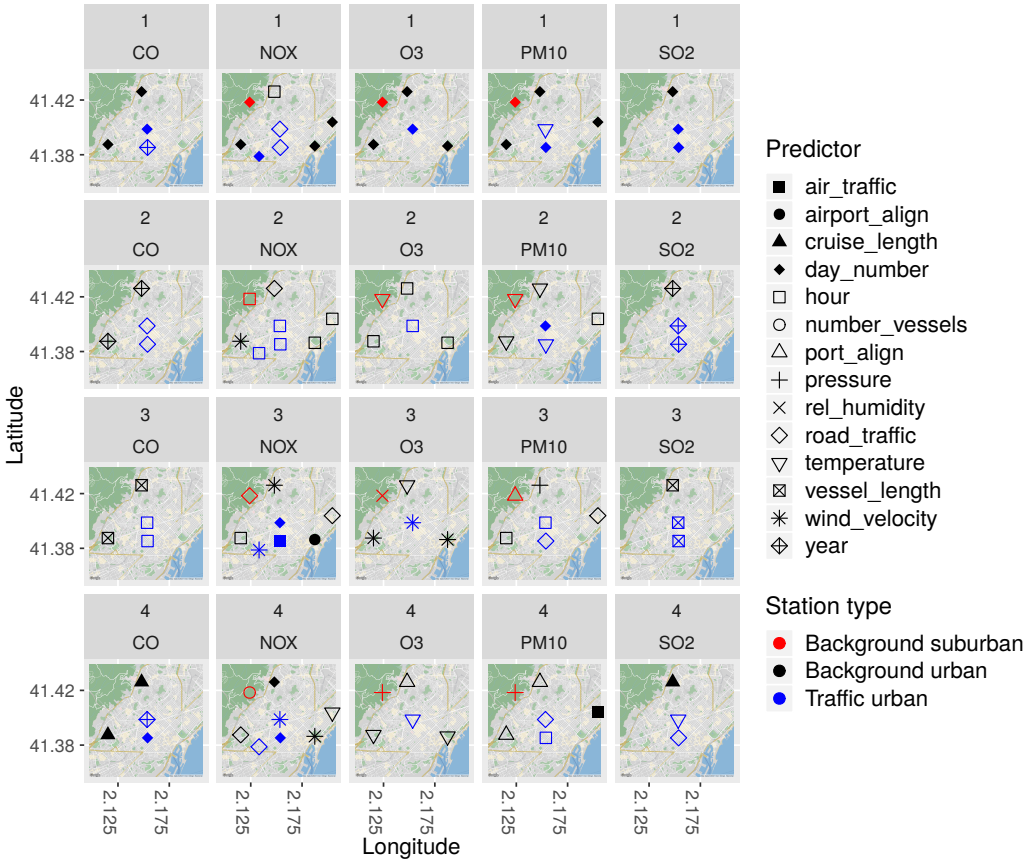

Figure 3: Map showing the top four most important predictors as measured by the entropy-based information gain for CO, NO_x, O_3, PM_{10} and SO_2. Colors indicate the type of pollutant station as shown in Fig. 1. Each symbol identifies the corresponding predictor as listed in Tab. 2.

agreement to the well-known seasonal behavior of tropospheric ozone with higher values in the summer months when atmospheric conditions favour O_3 formation.

As expected, traffic intensity scores high in importance in urban traffic stations (Eixample and Gràcia - Sant Gervasi) for nitrogen oxides and carbon monoxide. Contrarily, limited effects on these pollutant concentrations for this feature are found in background and suburban locations. The impact of traffic intensity on nitrogen oxides concentration at the Observatori Fabra station can be explained by the proximity of a major road traffic route (Ronda de Dalt). Wind velocity seems to affect also the nitrogen oxides level in background urban and suburban stations.

Wind velocity impact on concentration levels significantly varies across stations with Palau Reial and Sants exhibiting the largest effect on nitrogen oxide levels. Wind alignment with respect to the Airport and the Port are found to have similar importance with the latter exhibiting larger impact on PM_{10} concentration at the two only 'Traffic Urban' stations, namely, Eixample and Gràcia - Sant Gervasi.

As happens for the two predictors for the number of total vessels and cruise liners, the ship size of the latter type is found to be the least relevant. The number of airport movements, the pressure and irradiance predictors exhibit modest importance and large cross-station and cross-pollutant variability.

The overall least important features are those related to the day of the week. As shown, except Sunday — a holiday—, the workweek days do not show a great influence on pollutant concentration. Negligible importance has

also been found for hourly precipitation, cloud cover, relative humidity and notably, the hourly number of docked cruise ships.

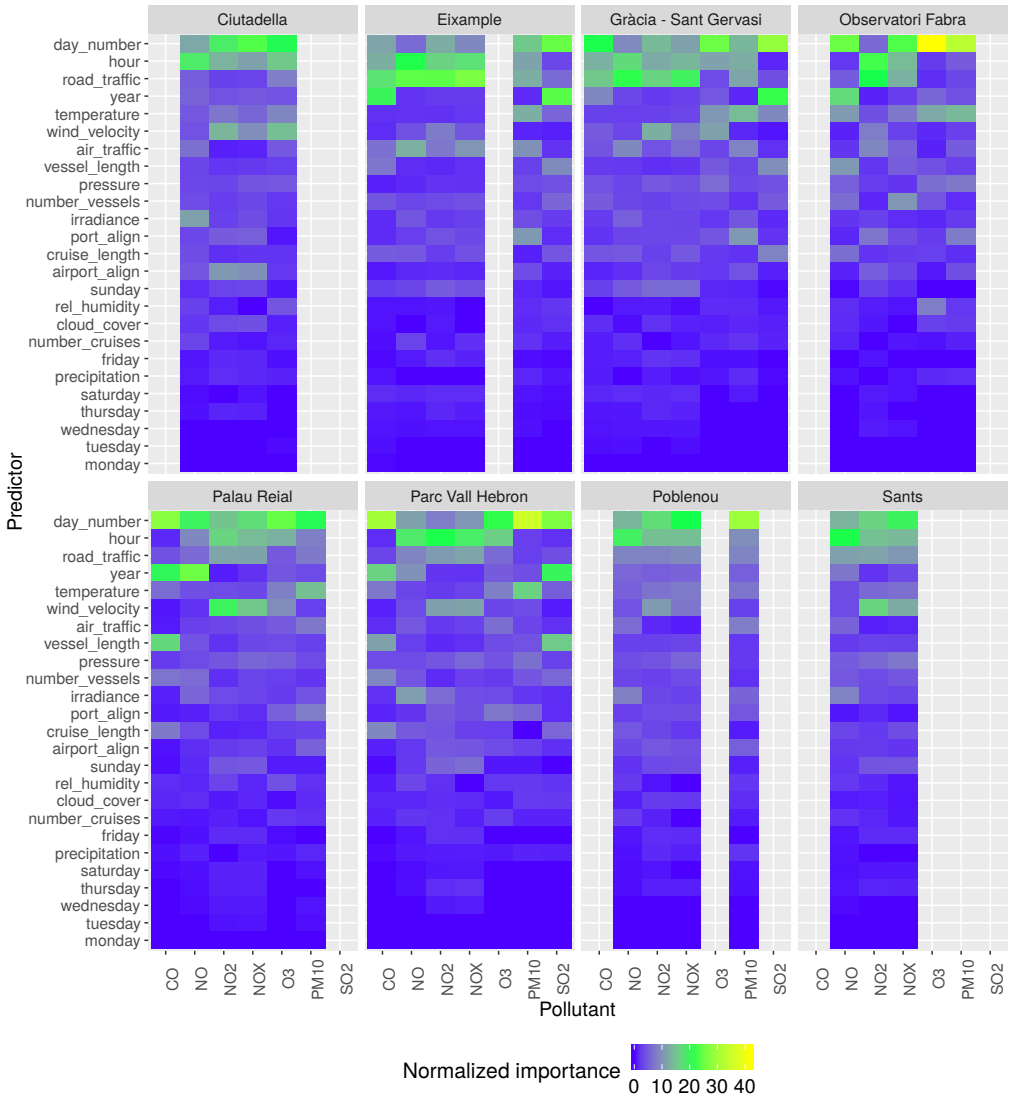

Figure 4: Normalized information gain attributable to each predictor in Tab. 2 as measured by the entropy-based information gain for each pollutant. Each panel corresponds to a different station (see Fig. 1).

3.3. Overall Port impact

Using the hourly number of total vessels as a proxy for the Port of Barcelona activity, the isolated impact of this predictor on the air quality at each pollutant station has been estimated by comparing the actual and the predicted concentration levels for several values of this predictor. Thus, the x-axis of Fig. 5 shows the difference between the total number of vessels used in the prediction and the actual value. The y-axis shows the average difference over all observations in the working dataset between the observed concentration and average prediction of the GBM learner (using a 3-fold re-sampling). For reference, mean and median concentration values for each station and pollutant in the working dataset are shown in Tab. 4.

As shown in Fig. 5, the excellent accuracy of the ML model results in negligible averaged difference between the actual and the predicted concentration ($y \approx 0$) when the number of vessels used in the prediction coincides with the actual one, i.e., $x = 0$.

In general, the model predicts that an increase in the activity in the Port of Barcelona as measured by the total vessel traffic leads to an increase in the local concentrations of carbon monoxide, nitrogen oxides and particulate matter. The overall trend indicates that the impact of the maritime traffic weakens as the distance from the Port increases. Thus, the impact of the number of vessels on the concentration levels in Ciutadella and Eixample stations is notably larger than that in Observatori Fabra and Palau Reial.

The largest impacts predicted by the model are found at the closest to the Port station (Eixample) where NO and NO_2 concentrations are expected to rise around 12% and 10% respectively in comparison to the mean value

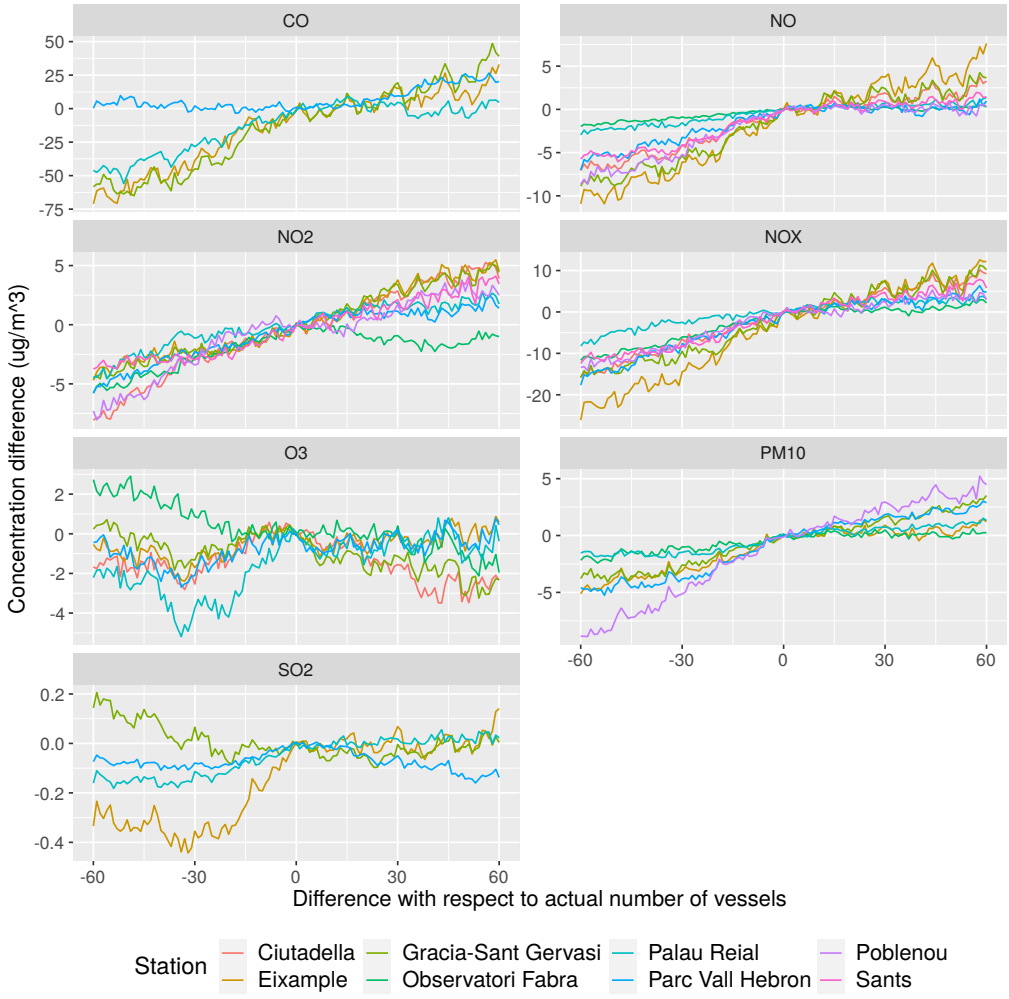

Figure 5: Predicted average concentration increase in $\mu g\,m^{-3}$ due to overall Port traffic for each pollutant measurement station (see Fig. 1). Each panel corresponds to a pollutant species.

in Tab. 4. In contrast, the largest increases in CO (9%) and PM_{10} (11%) concentrations are found at slightly further stations (Grcia - Sant Gervasi and Poblenou respectively). The signature of the Port activity on the ozone and sulphur dioxide concentration is clearly weaker suggesting lack of correlation

between the number of vessels and their concentration levels.

Station	Metric	CO	NO	NO_2	NO_x	O_3	PM_{10}	SO_2
Ciutadella	Mean	-	13.08	34.29	53.96	41.64	-	-
	Median	-	3	31	38	41	-	-
Eixample	Mean	410.0	32.47	52.16	101.65	36.51	26.61	1.79
	Median	300	19	49	79	35	24	1
Gràcia - Sant Gervasi	Mean	414.0	19.41	45.54	74.81	42.54	24.34	1.90
	Median	300	9	41	56	43	22	1
Observatori Fabra	Mean	-	2.75	11.71	16.17	82.85	16.57	-
	Median	-	1	9	11	82	14	-
Palau Reial	Mean	284.7	7.47	28.49	38.91	50.78	18.96	1.56
	Median	200	2	22	25	51	18	1
Parc Vall Hebron	Mean	294.0	8.17	29.15	41.36	54.14	21.42	1.42
	Median	300	3	23	27	53	19	1
Poblenou	Mean	-	15.07	38.01	60.84	-	28.11	-
	Median	-	5	34	43	-	24	-
Sants	Mean	-	9.03	32.65	46.24	-	-	-
	Median	-	3	27	32	-	-	-

Table 4: Mean and median concentration in $\mu g\,m^{-3}$ for every pollutant and station.

3.4. Cruise ship impact

Following an analogous approach, the hourly number of liners has been used as a proxy for cruise shipping activity in the Port of Barcelona. In Fig. 6, the x-axis corresponds to the average difference between the actual number of cruise ships and that used in the prediction while the y-axis shows the corresponding difference between the observed and average predicted concentration over all observations in the working dataset.

Of course, since cruise ships represent a relatively small fraction of the total vessels, its isolated effect on the air quality of Barcelona is expected

to be less pronounced. Predictions suggest that the largest impact of cruise ships on nitrogen oxide concentration levels occurs at the two stations closest to the Port (Ciutadella and Eixample) with 1.3 $\mu g\,m^{-3}$ of NO_x for each additional cruise liner in Port. Impact on nitrogen oxides levels is predicted to decay as distance between the station and the Port increases.

In contrast to this marked variability across stations for NO_x, predicted concentrations of PM_{10} are found to vary in a consistent way across stations with an average slope of 0.04 $\mu g\,m^{-3}$ for each additional cruise liner.

CO is predicted to be insensitive when changes in the number of docked cruise liners is modest ($\pm$ 3 ships). For variations above $\pm$ 5 cruise ships, CO concentration levels start exhibiting significant variations at the closest stations to the Port. Farther locations are predicted to remain constant independently of the number of liners.

Cruise ship activity seems to have a negligible impact on sulfur dioxide with maximum variations in concentration less than 2.5% change with respect to the across-station mean value of 1.67 $\mu g\,m^{-3}$ (obtained from the four station mean values in Tab. 4). Predicted impact on ozone concentrations is found to be negligible in most stations and negatively correlated in others, specially Observatori Fabra. The relatively large horizontal and vertical distance between the Port and this station suggest that, on overall, cruise ships do not affect the ozone levels across the metropolitan area.

The overall modest increase in pollutant concentration of the hourly number of cruise liners and this predictor relatively small *importance gain* suggest that, in comparison to the overall effects of the Port, the impact of the cruise activity on the air quality of the city is very limited.

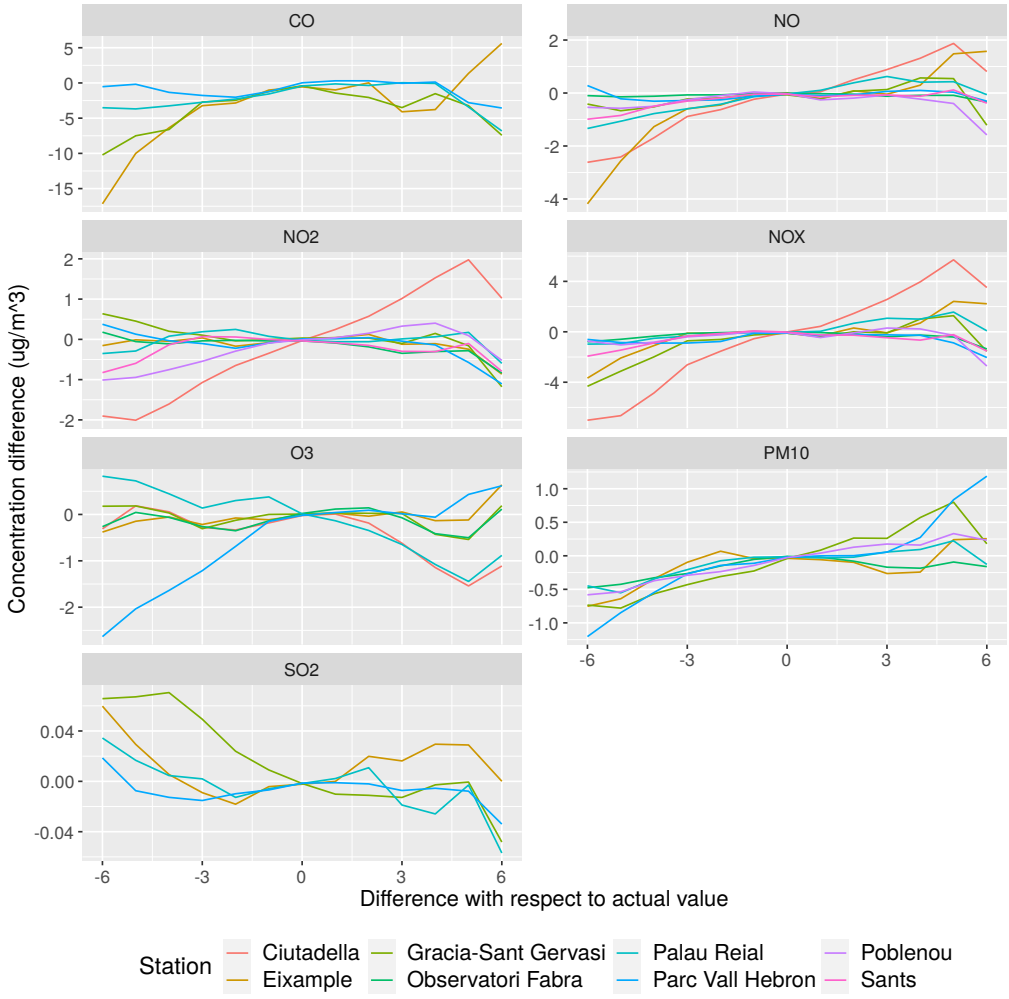

Figure 6: Predicted average concentration increase in $\mu g\,m^{-3}$ due to cruise ship traffic for each pollutant measurement station (see Fig. 1). Each panel corresponds to a pollutant species.

4. Discussion

The availability and ease of access to large amounts of data from uncountable sources of scientific and technological relevance has made it possible to

use Machine Learning techniques to improve our understanding of processes and phenomena until now only accessible through classical approaches.

This paradigm is illustrated in this work where, provided enough data on both the atmospheric state and several proxies on pollutant inventories, it is shown that ML-based predictors are capable of outperforming traditional pollutant dispersion models. While traditional, physics-based approaches suffer from uncertainties on local weather and pollutant inventories, ML approaches are generally criticized by their *black box* nature offering small insight on the inner workings of the physical processes under study.

Besides these general considerations, and despite the excellent predictive capabilities of the tool developed here, this model has two main limitations in comparison to traditional approaches. On one hand, the relatively small number of stations reporting concentration levels across the metropolitan area of Barcelona hinders the potential of the model to be used as an operational tool capable of predicting air quality at any location across the region of interest. In other words, the current model lacks of spatial features. Therefore, air quality predictions are restricted to the locations and heights of each pollutant station for which an independent model has been derived.

Pollutant measurement campaigns with fixed or mobile equipment might provide in the future high spatial resolution information that make possible to use measurement location and altitude as predictors. This key feature would enable the transition of the current ML model into a fully-operational tool capable of providing accurate, fast and cheap predictions of air quality.

Second, in this work pollutant emissions are accounted via some proxy variables including the number of flights, vessels and road traffic congestion.

Although these are prime contributions to urban pollution, industrial and agricultural emission may also play a significant role in worsening the air quality of neighboring cities. Overall model accuracy could be potentially improved provided better characterizations of pollutant emission inventories in addition to transportation. Of course, the performance of the ML tool presented here is expected to increase as the size and quality of the time series datasets used to train the learners increase.

5. Conclusions

Weather and pollutant concentration measurements collected by monitoring networks across the metropolitan area of Barcelona have been combined with road, air and maritime traffic intensity data to train a Machine Learning model to predict the air quality in the city. In terms of predictive capabilities of local NO_2 concentration levels at several station locations across Barcelona, the resulting tool has been found to have better performance than the CALIOPE Urban v1.0 platform.

Using Mean Squared and Absolute Errors as metrics for predictive accuracy, the benchmark of different classical learners suggested that, for this specific dataset and across all stations and pollutants, the Gradient Boosting Machine (GBM) has been found to be the best performer.

Importance analysis based on information gain indicates that the three top features explaining pollutant level variability are the time of the year, the daytime and the road traffic intensity. The least important predictors include the workweek day, the precipitation, relative humidity and, notably, the number of cruise ships in Port.

Using the difference between observed and predicted pollutant concentrations for several difference values with respect to the actual number of vessels, the ML tool has been used to estimate the contribution to worsened air quality of the overall Port activity and that exclusively due to cruise liners.

Results suggest that, close to the Port of Barcelona, the overall activity of this infrastructure leads to 12% and 10% increased mean concentration of NO and NO_2 respectively. Largest increases in CO and PM_{10} have been estimated to be 9% and 11%. In comparison, the Port impact on SO_2 and O_3 concentration levels is predicted to be significantly weaker.

Using the hourly number of docked cruise ships as a proxy to estimate the isolated effect of this industry, predictions suggest that the impact of liners on the air quality is very limited in comparison to the overall Port contribution. Predictions suggest that the largest impact of cruise ships is 1.3 $\mu g\,m^{-3}$ of NO_x for each additional docked cruise ship in Port.

Besides the better accuracy in predicting pollutant concentration, the numerical simulations carried out with the ML-based model presented here required only a personal workstation with typical run-times in the order of hours. This computational costs are well below those typically required by state-of-the-art dispersion models.

6. Acknowledgements

This study was financially supported by the Spanish Ministry of Economy, Industry and Competitiveness — Research National Agency (under projects DPI2016-75791-C2-1-P and RTI2018-100907-A-I00), by FEDER funds and

by the Generalitat de Catalunya — AGAUR (under project 2017 SGR 01234). Authors would like to thank the Government of Catalonia for granting access to atmospheric pollutant data from the XVPCA network and the Servei Meteorològic de Catalunya for weather measurements.

[1] J. Seinfeld, S. Pandis, Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, 1998.

[2] T. R. Walker, O. Adebambo, M. C. Del Aguila Feijoo, E. Elhaimer, T. Hossain, S. Johnston Edwards, C. E. Morrison, J. Romo, N. Sharma, S. Taylor, S. Zomorodi, Chapter 27 - Environmental Effects of Marine Transportation, in: C. Sheppard (Ed.), World Seas: an Environmental Evaluation (Second Edition), Academic Press, second edition edn., 505 – 530, 2019.

[3] S. Sorte, V. Rodrigues, C. Borrego, A. Monteiro, Impact of harbour activities on local air quality: A review, Environmental Pollution 257 (2019) 113542.

[4] M. Viana, P. Hammingh, A. Colette, X. Querol, B. Degraeuwe, de Vlieger I., J. van Aardenne, Impact of maritime transport emissions on coastal air quality in Europe, Atmospheric Environment 90 (2014) 96 – 105.

[5] V. A. Profillidis, G. N. Botzoris, Chapter 2 - Evolution and Trends of Transport Demand, in: V. Profillidis, G. Botzoris (Eds.), Modeling of Transport Demand, Elsevier, 47 – 87, 2019.

[6] M. Zhao, Y. Zhang, W. Ma, Q. Fu, X. Yang, C. Li, B. Zhou, Q. Yu, L. Chen, Characteristics and ship traffic source identification of air pollutants in China's largest port, Atmospheric Environment 64 (2013) 277286, doi:\bibinfo{doi}{10.1016/j.atmosenv.2012.10.007}.

[7] C. Schembari, F. Cavalli, E. Cuccia, J. Hjorth, G. Calzolai, N. Perez, J. Pey, P. Prati, F. Raes, Impact of a European directive on ship emissions on air quality in Mediterranean harbours, Atmospheric Environment 61 (2012) 661669.

[8] D. Lack, J. Corbett, Black carbon from ships: a review of the effects of ship speed, fuel quality and exhaust gas scrubbing, Atmospheric Chemistry & Physics Discussions 12 (2011) 3509–3554.

[9] L. Fileni, E. Mancinelli, M. Morichetti, G. Passerini, U. Rizza, S. Virgili, Air Pollution in Ancone Harbour, Italy, in: Maritime Transport 2019, 199–208, 2019.

[10] J. Isakson, T. Persson, E. Lindgren, Identification and assessment of ship emissions and their effects in the harbour of Göteborg, Sweden, Atmospheric Environment 35 (2001) 3659–3666, doi:\bibinfo{doi}{10.1016/S1352-2310(00)00528-8}.

[11] E. Merico, A. Donateo, A. Gambaro, D. Cesari, E. Gregoris, E. Barbaro, A. Dinoi, G. Giovanelli, S. Masieri, D. Contini, Influence of in-port ships emissions to gaseous atmospheric pollutants and to particulate matter of different sizes in a Mediterranean harbour in Italy, Atmospheric Environment 139.

[12] H. Saxe, T. Larsen, Air pollution from ships in three Danish ports, Atmospheric Environment 38 (2004) 4057–4067.

[13] K. Poplawski, E. Setton, B. McEwen, D. Hrebenyk, M. Graham, C. Keller, Impact of cruise ship emissions in Victoria, BC, Canada, Atmospheric Environment - Atmos Environ 45 (2011) 824–833.

[14] F. Murena, M. Luigia, F. Quaranta, D. Toscano, Impact on air quality of cruise ship emissions in Naples, Italy, Atmospheric Environment 187.

[15] S. Eckhardt, O. Hermansen, H. Grythe, M. Fiebig, M. Cassiani, A. Baecklund, A. Stohl, The influence of cruise ship emissions on air pollution in Svalbard — A harbinger of a more polluted Arctic?, Atmospheric Chemistry & Physics Discussions 13 (2013) 3071–3093.

[16] A. Donateo, E. Gregoris, A. Gambaro, E. Merico, R. Giua, A. Nocioni, D. Contini, Contribution of harbour activities and ship traffic to PM2.5, particle number concentrations and PAHs in a port city of the Mediterranean Sea (Italy), Environmental science and pollution research international 21.

[17] D. Contini, A. Gambaro, F. Belosi, S. Pieri, W. Cairns, A. Donateo, E. Zanotto, M. Citron, The direct influence of ship traffic on atmospheric PM2.5, PM10 and PAH in Venice, Journal of Environmental Management 92 (2011) 2119–2129.

[18] A. Maragkogianni, S. Papaefthimiou, Evaluating the social cost of cruise ships air emissions in major ports of Greece, Transportation Research Part D: Transport and Environment 36 (2015) 10 – 17.

[19] J. Celic, S. Valcic, M. Bistrović, Air pollution from cruise ships, Proceedings Elmar - International Symposium Electronics in Marine (2014) 75–78.

[20] L. T. Fan, Y. Horie, H. J. Paulus, Review of atmospheric dispersion and urban air pollution models, CRC Critical Reviews in Environmental Control 2 (1-4) (1972) 431–457.

[21] A. Leelossy, F. Molnár, F. Izsák, A. Havasi, I. Lagzi, R. Mészáros, Dispersion modeling of air pollutants in the atmosphere: a review, Central European Journal of Geosciences 6 (2014) 257–278, doi:\bibinfo{doi}{10.2478/s13533-012-0188-6}.

[22] M. Lateb, R. N. Meroney, M. Yataghene, H. Fellouah, F. Saleh, M. C. Boufadel, On the use of numerical modelling for near-field pollutant dispersion in urban environments — A review, Environmental Pollution 208 (2016) 271 – 283, ISSN 0269-7491, doi:\bibinfo{doi}{https://doi.org/10.1016/j.envpol.2015.07.039}, URL http://www.sciencedirect.com/science/article/pii/S0269749115003723, special Issue: Urban Health and Wellbeing.

[23] C. Hood, I. MacKenzie, J. Stocker, K. Johnson, D. Carruthers, M. Vieno, R. Doherty, Air quality simulations for London using a coupled regional-to-local modelling system, Atmospheric Chemistry and Physics 18 (2018) 11221–11245, doi:\bibinfo{doi}{10.5194/acp-18-11221-2018}.

[24] H. Relvas, A. Miranda, An urban air quality modeling system to support

559 decision-making: design and implementation, Air Quality, Atmosphere
560 & Health 11, doi:\bibinfo{doi}{10.1007/s11869-018-0587-z}.

561 [25] J. Song, K. Han, Deep-MAPS: Machine Learning based Mobile Air Pol-
562 lution Sensing, 2019.

563 [26] Y. Zheng, F. Liu, H.-P. Hsieh, U-Air: When urban air quality in-
564 ference meets big data, 1436–1444, doi:\bibinfo{doi}{10.1145/2487575.
565 2488188}, 2013.

566 [27] J. Kleine Deters, R. Zalakeviciute, M. Gonzalez, Y. Rybarczyk, Model-
567 ing PM2.5 Urban Pollution Using Machine Learning and Selected Mete-
568 orological Parameters, Journal of Electrical and Computer Engineering
569 2017 (2017) 1–14, doi:\bibinfo{doi}{10.1155/2017/5106045}.

570 [28] Observatori del Turisme a Barcelona, Barcelona tourism activity re-
571 port - 2018, `https://www.observatoriturisme.barcelona/sites/`
572 `default/files/IAOTB18.pdf`, 2018.

573 [29] Eurostat Table [mar_mg_aa_pwhd], Top 20 ports-gross weight of goods
574 handled in each port, by direction, `https://appsso.eurostat.ec.`
575 `europa.eu/nui/show.do?dataset=mar\_mg\_aa\_pwhd\&lang=en`,
576 2020.

577 [30] N. Perez, J. Pey, C. Reche, J. Cortés, A. Alastuey, X. Querol, Impact of
578 harbour emissions on ambient PM10 and PM2.5 in Barcelona (Spain):
579 Evidences of secondary aerosol formation within the urban area, The
580 Science of the total environment 571 (2016) 237–250.

[31] J. Perdiguero, A. Sanz, Cruise activity and pollution: The case of Barcelona, Transportation Research Part D: Transport and Environment 78 (2020) 102181, ISSN 1361-9209.

[32] G. Rodrguez, E. Martin-Alcalde, J. Murcia-González, S. Saurí, Evaluating air emission inventories and indicators from cruise vessels at ports, WMU Journal of Maritime Affairs 16, doi:\bibinfo{doi}{10.1007/s13437-016-0122-8}.

[33] Departament de Territori i Sostenibilitat, Dades d'immissió dels punts de mesurament de la Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica, `https://analisi.transparenciacatalunya.cat/Medi-Ambient/Dades-d-immissi-dels-punts-de-mesurament-de-la-Xar/uy6k-2s8r`, portal de Dades Obertes de la Generalitat, 2020.

[34] Servei Meteorològic de Catalunya, Xarxa d'Estacions Meteorològiques Automàtiques (XEMA), `https://www.meteo.cat/observacions/llistat-xema`, 2020.

[35] A. Tank, J. Wijngaard, G. Können, R. Böhm, G. Demarée, G. AA, M. Mileta, S. Pashiardis, L. Hejkrlik, C. Kern-Hansen, R. Heino, P. Bessemoulin, G. üller Westermeier, M. Tzanakou, S. Szalai, T. Pálsdóttir, D. Fitzgerald, S. Rubin, M. Capaldo, P. Petrovic, Daily surface air temperature and precipitation dataset 19011999 for European Climate Assessment (ECA), International Journal of Climatology 22 (2002) 1441 – 1453, doi:\bibinfo{doi}{10.1002/joc.773}.

[36] Barcelona Port Authority, Open Data of the Port of Barcelona, URL http://opendata.portdebarcelona.cat, 2020.

[37] Nicolas Bruno, Hourly flight movements at the Airport of Barcelona from 2015-2020, Private Communication between the authors and EUROCONTROL, 2020.

[38] Barcelona's City Hall Open Data Service, Traffic state information by sections of the city of Barcelona, URL https://opendata-ajuntament.barcelona.cat/data/en/dataset/trams, 2020.

[39] B. B., M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, Z. M. Jones, mlr: Machine Learning in R, Journal of Machine Learning Research 17 (170) (2016) 1–5, URL http://jmlr.org/papers/v17/15-066.html.

[40] J. Benavides, M. Snyder, M. Guevara, A. Soret, C. Pérez García-Pando, F. Amato, X. Querol, O. Jorba, CALIOPE-Urban v1.0: Coupling R-LINE with a mesoscale air quality modelling system for urban air quality forecasts over Barcelona city (Spain), Geoscientific Model Development Discussions (2019) 1–35doi:\bibinfo{doi}{10.5194/gmd-2019-48}.

[41] J. Baldasano, M. Pay, O. Jorba, S. Gassó, P. Jimenez-Guerrero, An annual assessment of air quality with the CALIOPE modeling system over Spain, The Science of the total environment 409 (2011) 2163–78, doi:\bibinfo{doi}{10.1016/j.scitotenv.2011.01.041}.

[42] M. Guevara, F. Martínez, G. Arévalo, S. Gassó, J. Baldasano, An improved system for modelling Spanish emissions: HERMESv2.0, Atmo-

spheric Environment 81 (2013) 209 – 221, doi:\bibinfo{doi}{10.1016/j.atmosenv.2013.08.053}.

[43] W. Skamarock, J. Klemp, A time-split nonhydrostatic atmospheric model for research and NWP applications, J.Comput. Phys. 135.

[44] D. Byun, K. Schere, Review of the Governing Equations, Computational Algorithms, and Other Components of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System, Applied Mechanics Reviews 59 (2006) 51–77.

[45] M. G. Snyder, A. Venkatram, D. K. Heist, S. G. Perry, W. B. Petersen, V. Isakov, RLINE: A line source dispersion model for near-surface releases, Atmospheric Environment 77 (2013) 748 – 756.

Appendix A. Metrics

The Geometric Mean Bias (GeoMean), the Fraction of Model Results Within a Factor of 2 of Observations (FAC2), the Correlation Coefficient (R), the Mean Bias (MB) and the Root Mean Square Error (RMSE) are

defined as follows:

$$\text{GeoMean} = \exp\left(\overline{\ln \widehat{Y}_i} - \overline{\ln Y_i}\right) \tag{A.1}$$

$$\text{FAC2} = \sum_{i}^{n} \left[0.5 \leq \frac{\widehat{Y}_i}{Y_i} \leq 2.0\right] / n \tag{A.2}$$

$$R = \frac{\overline{\left(\widehat{Y}_i - \overline{\widehat{Y}_i}\right)\left(Y_i - \overline{Y_i}\right)}}{\sigma\left(Y_i\right)\sigma\left(\widehat{Y}_i\right)} \tag{A.3}$$

$$\text{MB} = \overline{Y_i - \widehat{Y}_i} \tag{A.4}$$

$$\text{RMSE} = \left(\overline{\left(Y_i - \widehat{Y}_i\right)^2}\right)^{1/2} \tag{A.5}$$

where Y_i and $\widehat{Y}_i$ are predicted and observed concentration data points. Average and standard deviation are respectively represented by the symbols $\overline{(\cdots)}$ and $\sigma(\cdots)$ and $[\ldots]$ are the Iverson brackets.

Appendix B. Air and maritime traffic

The distribution of number of simultaneously docked total vessels and cruise ships at the Port of Barcelona over the October 2017 to March 2020 period (both included) are shown in Fig. B.7 and B.8 respectively.

Analogously, the distribution of number of flight opeartions at the Airport of Barcelona over the October 2017 to March 2020 period (both included) is shown in Fig. B.9.

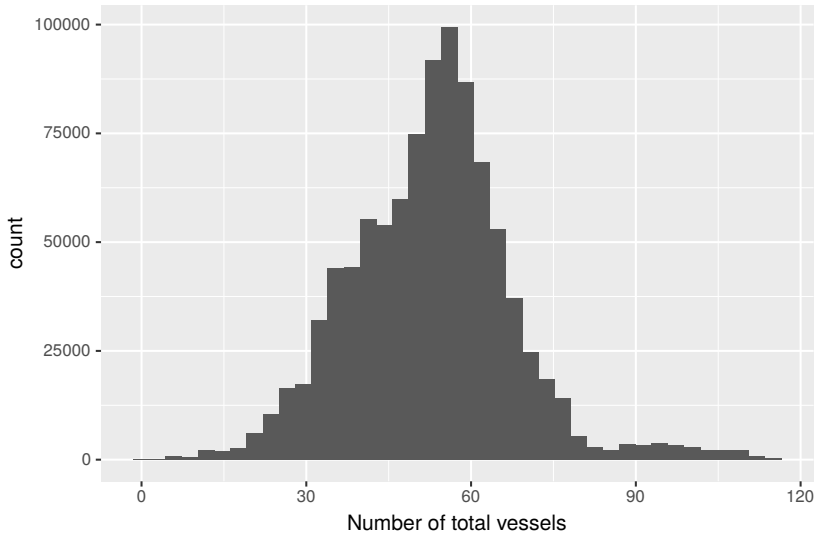

Figure B.7: Distribution of simultaneously docked total vessels at the Port of Barcelona from October 2017 to March 2020 both included.

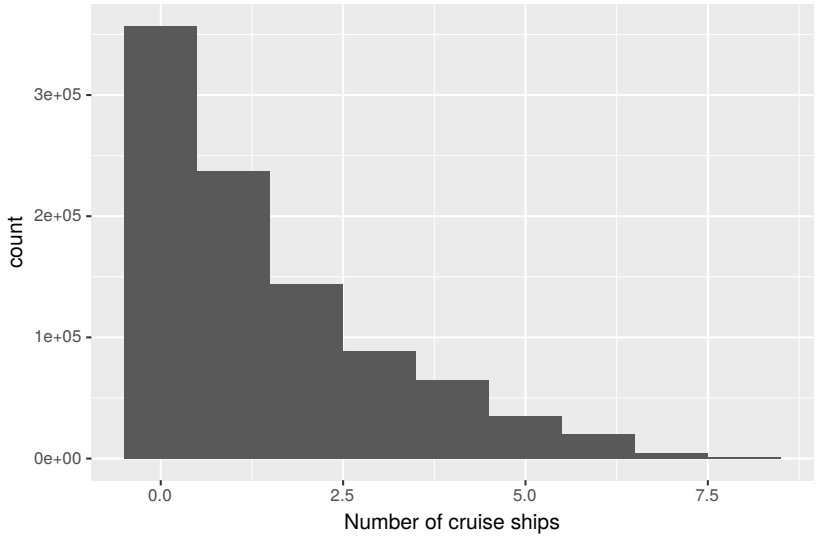

Figure B.8: Distribution of simultaneously docked cruise ships at the Port of Barcelona from October 2017 to March 2020 both included.

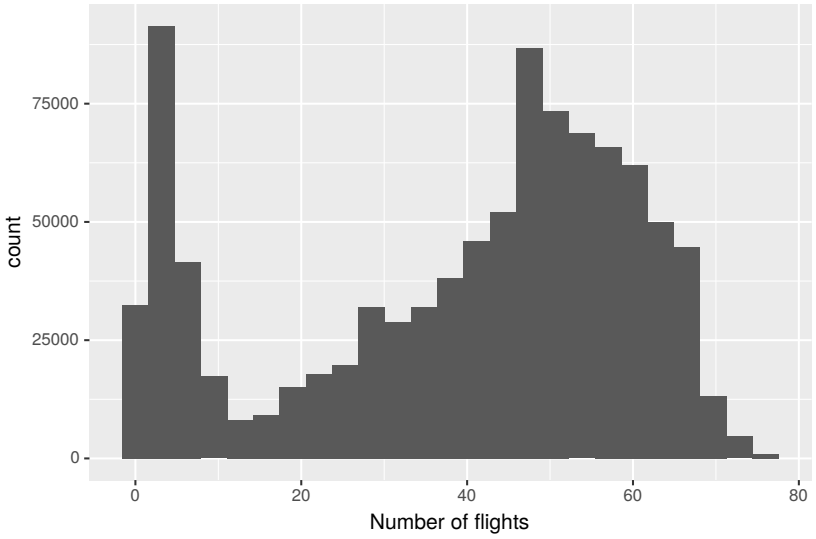

Figure B.9: Distribution of simultaneously docked cruise ships at the Port of Barcelona from October 2017 to March 2020 both included.

Appendix C. Learner benchmark

The comparison of performance for each optimized learner is shown in Fig. C.10 using the mean squared error for each pollutant averaged across stations.

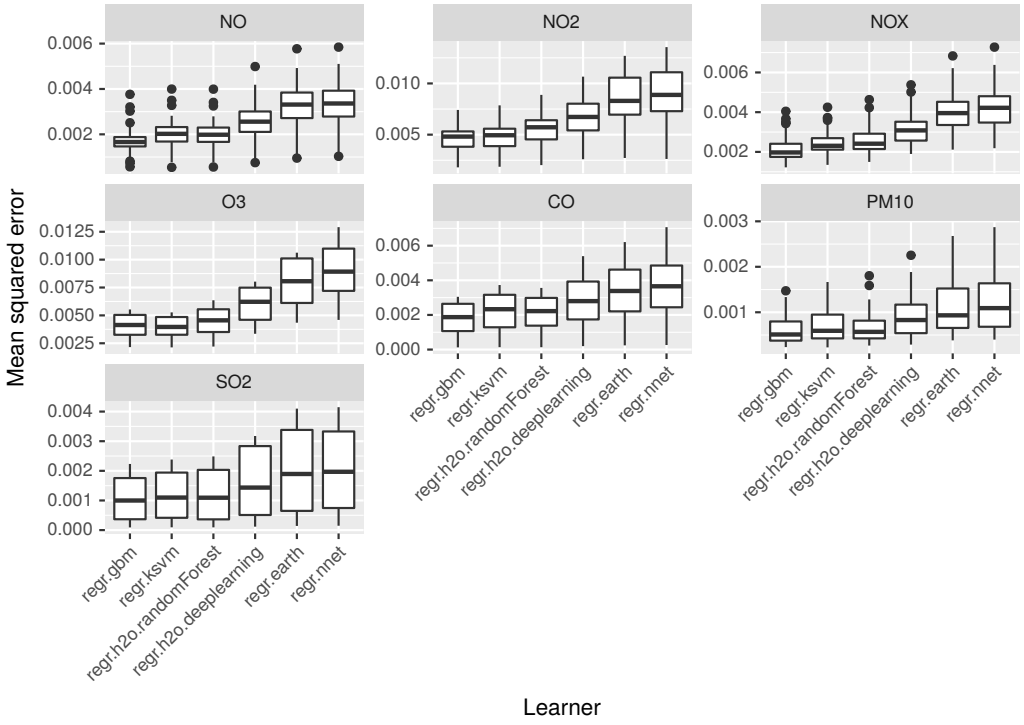

Figure C.10: Mean squared error for each pollutant across all stations for each learner using a 5-fold cross-validation resampling.