This project has received funding from the European's Union Horizon 2020 research innovation program under grant agreement No. 769355

D4.4 Predictive Algorithms v2

Deliverable No.	D4.4	**Due Date**	30-Apr-2020
Type	Other	**Dissemination Level**	Public
Version	1.0	**Status**	Release 1
Description	This deliverable document contains the development of new prediction and forecasting algorithms in the ports where the need is greatest: to allocate energy more efficiently and to forecast the infrastructure needs and trend of environmental pollution in the ports. For different sectors within ports, there are different requirements and thus various predictive models/algorithms must be constructed to include different conditions. The proposed algorithms will be applied using available data measurements to test their performance and forecasting quality. Associated task: T4.5.		
Work Package	WP4		

Authors

Name	Partner	e-mail
Ignacio Lacalle	P01 UPV	iglaub@upv.es
Darío Alandes	P01 UPV	daalco1@teleco.upv.es
Benjamín Molina	P01 UPV	benmomo@upvnet.upv.es
Sergio Vivó Sánchez	P02 PRODEVELOP	svivo@prodevelop.es
Miguel Montesinos Lajara	P02 PRODEVELOP	mmontesinos@prodevelop.es
Dejan Štepec	P03 XLAB	dejan.stepec@xlab.si
Tomaž Martinčič	P03 XLAB	tomaz.martincic@xlab.si
Flavio Fuart	P03 XLAB	flavio.fuart@xlab.si
Damjan Murn	P03 XLAB	damjan.murn@xlab.si
Matija Cankar	P03 XLAB	matija.cankar@xlab.si
Gilda De Marco	P04 INSIEL	gilda.demarco@insiel.it
Charles Garnier	P05 CATIE	c.garnier@catie.fr
Zhe Li	P05 CATIE	z.li@catie.fr
Cynthia Perier	P05 CATIE	c.perier@catie.fr
Erwan Simon	P05 CATIE	e.simon@catie.fr
Grigoris Dimitriadis	P10 ThPA SA	gdimitriadis@thpa.gr
Eirini Tserga	P10 ThPA SA	etserga@thpa.gr
Dimitris Spyrou	P11 PPA SA	dspyrou@olp.gr
Kontogiorgi Chryssanthi	P11 PPA SA	kontogiorgich@olp.gr
Georgios Dioletis	P11 PPA SA	dioletisg@olp.gr
Athanasios Chaldeakis	P11 PPA SA	ahaldek@gmail.com
Thibault Guillon	P13 GPMB	t-guillon@bordeaux-port.fr
Michel Le-Van-Kiem	P13 GPMB	m-le-van-kiem@bordeaux-port.fr
Fabrice Klein	P13 GPMB	f-klein@bordeaux-port.fr

History

Date	Version	Change
10-Mar-2020	0.1	Table of contents and task assignments.
15-Mar-2020	0.2	Section 1 and annex 1 drafts.
20-Mar-2020	0.3	Writing sections 2, 3, 4.
10-Apr-2020	0.4	Integrating road traffic sections with contributions. from multiple partners.
25-Apr-2020	0.5	Integrating section 6 and annexes. Finalisation of the document.
05-May-2020	0.6	Version for internal review.
22-May-2020	1.0	Official release.

Key Data

Keywords	Predictive algorithms, data sources, port operations, machine learning, software, maritime traffic, road traffic, photovoltaic energy
Lead Editor	Dejan Štepec, P03 XLAB
Internal Reviewer(s)	Eirini Tserga P10 ThPA SA Luka Traven, P08 MEDRI Innovation Review, Joao Pita Costa, P03 XLAB

Abstract

AI (Artificial Intelligence) is becoming one of the main factors in a successful digital transformation of the ports. Larger ports are increasingly becoming aware of the value that is present, in a daily collected operational data. The ability to create operational insights from vast amounts of data that is collected in the ports will be one of the main advantages of future ports, in terms of energy efficiency, hinterland multimodal transport needs and better forecast of harmful actions.

This deliverable presents the second version of the task of predictive algorithms (T4.5) in WP4 due in M24. The results of the tasks that were identified in detail in D4.3 are presented, along with the methodology that was used to tackle the proposed tasks. The tasks were identified based on the existing documentation about requirements and use cases, as well as based on the review of the state-of-the-art in literature, existing trends and examples from the maritime industry, AI expertise and available internal and external data.

- **Prediction of vessel call data from FAL (Convention of Facilitation of Maritime Traffic) forms and other sources:** In this task, internal data about vessel calls are utilized to predict vessel calls and their durations (i.e. turnaround time). General statistical analysis and visualizations are also performed. Vessels call data is available in every port as is obtained from FAL forms, which are legally required, thus making this task generally applicable to every port at a low cost.
- **Use of AIS (Automatic Identification System) data:** AIS data is widely used in the maritime domain and is becoming extremely useful for data analytics tasks, especially because of its quantity. In this task, we visualize and analyse the data around the ports, provide port congestion indicators out of AIS data and to some extent ETA (Estimated Time of Arrival) prediction for the incoming ships, as well the capability to detect different events in the port area.
- **Use of satellite imagery:** Obtaining operational insights from remote sensing imagery presents an emerging field, offering the ports increased situational awareness by giving them the ability to monitor their port from Space and compare it in a global perspective to understand their unique differentiators in the global market. A novel approach is presented, that uniquely utilizes AIS data and satellite imagery data, to perform ship detection across operational satellite imagery, of medium spatial resolution.
- **Analysis and prediction of road traffic conditions with connection to port operations:** In this task hinterland multimodal transport requirements in the Port of Monfalcone, Port of Piraeus and Port of Thessaloniki are addressed. A common task of short-term traffic volume prediction has been identified and the results presented. Predictions are correlated with port operations, to provide estimates on the impact that congestions have on them. Different internal and external road and maritime traffic data sources were used to develop predictive models.
- **Prediction of renewable energy production:** In this task, ports are provided with the ability to estimate the potential of renewable energy production for different time resolutions. The task is focused on the Port of Bordeaux use case, but the developed methods are general and applicable for any port. Different open data sources are used about the weather and measured photovoltaic power and different predictive models presented.

Statement of originality

This document contains material, which is the copyright of certain PIXEL consortium parties, and may not be reproduced or copied without permission. This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

The information contained in this document is the proprietary confidential information of the PIXEL consortium (including the Commission Services) and may not be disclosed except in accordance with the consortium agreement.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the project consortium as a whole nor a certain party of the consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, and accepts no liability for loss or damage suffered by any person using this information.

The information in this document is subject to change without notice.

The content of this report reflects only the authors' view. INEA (Innovation and Networks Executive Agency) is not responsible for any use that may be made of the information it contains.

Table of contents

List of tables

List of figures

List of acronyms

Acronym	Explanation
AI	Artificial Intelligence
AIS	Automatic Identification System
API	Application Programming Interface
ASPM	Azienda Speciale per il Porto di Monfalcone
CATIE	CATIE Centre Aquitain des Technologies de l'Information et Electroniques
CFAR	Constant False Alarm Rate
COLREG	Convention on the International Regulations for Preventing Collisions at Sea
CSV	Comma Separated Values
DEBS	ACM International Conference on Distributed and Event-based Systems
DMA	Danish Maritime Authority
DOTA	Large-scale Dataset for Object DeTection in Aerial Images
DPO	Data Protection Officer
EAA	National Observatory of Athens
EDA	Exploratory Data Analysis
EO	Earth Observation
ESA	European Space Agency
ETA	Estimated Time of Arrival
ETD	Estimated Time of Departure
FAL	Convention of Facilitation of Maritime Traffic
FVG	Friuli Venezia Giulia
GCI	Gate Congestion Index
GDPR	General Data Protection Regulation
GPMB	Grand port maritime de Bordeaux
GPS	Global Positioning System
HMI	Human-Machine Interface
HRSC	High-Resolution Ship Collections
IMO	International Maritime Organization
INEA	Innovation and Networks Executive Agency
INSIEL	Insiel SpA
JSON	JavaScript Object Notation
KNN	k-nearest neighbours algorithm
LDA	Linear discriminant analysis
LSTM	Long Short-Term Memory network
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	machine learning
MMSI	Maritime Mobile Service Identity
NMEA	National Marine Electronics Association
PAS	Port Activity Scenario
PCA	principal component analysis
PIXSAT	PIXEL Satellite Dataset
PPA	Piraeus Port Authority S.A.
PVGIS	Photovoltaic Geographical Information System
PVI	Photovoltaic Installation
R-CNN	Region-based Convolutional Neural Networks
REST	Representational state transfer
RFID	Radio-frequency identification
RMSE	Root Mean Square Error
ROI	Region of Interest

Acronym	Explanation
S-AIS	Satellite-based AIS
SARIMAX	Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors
SDAG	Stazioni doganali autoportuali di Gorizia - autoporto
SILI	Sistema Informativo Logistico Integrato
SQL	Structured Query Language
TCP	Transmission Control Protocol
UNCTAD	United Nations Conference on Trade and Development
UPV	Universitat Politècnica de València
VHF	Very High Frequency
VHR	Very-high-resolution
XLAB	XLAB d.o.o.

1. About this document

This document presents the work performed and the results obtained with the development of predictive algorithms in PIXEL. The main achievement is the implementation of predictive models that apply to small and medium-sized ports and can be addressed with the data that is captured internally by ports, or accessible as open data to provide cost-efficient solutions, that are at the same time scalable to be used also in larger ports. One of the main drivers behind the definitions of the tasks and their scope was the data that can be obtained by ports as stakeholders and can be used in the project. Following task definitions, requirements and state-of-the-art presented in the previous report, this document focuses on a detailed overview of data sources, methods and obtained results.

1.1. Deliverable context

Keywords	Lead Editor
Objectives	**Objective 5 Develop predictive algorithms** This deliverable provides a selection of predictive algorithms that will be developed. Also, the initial results are presented for each section. Main results of objective 5 are being achieved through the development of predictive algorithms that have the potential of significantly increasing the efficiency in one of the following areas: **1. Energy demand:** achieved through prediction of renewable energy prediction (section 6). **2. Hinterland multimodal transport needs:** achieved as explained in section 2 (predicting vessel calls duration, type and quantity of cargo), section 3 (predicting ETA and port events), section 4 (ship traffic analytics through remote observation) and section 5 (traffic peaks, congestion). **3. The anticipation of environmentally harmful actions:** section 3 proposes the implementation of a Port Congestion Index that will be used to anticipate peaks in vessel emissions in ports, ETA (Estimated Time of Arrival) estimation could be used to trigger slow steaming and schedule port arrivals. Detection of additional events in the port area can help to prevent accidents involving harmful substances. This can be supported using remote sensing solutions proposed in section 4, which will help to increase situational awareness of the ports.
Exploitable results	The main exploitable result arising from this deliverable is the predictive algorithms and, to some extent, statistical analysis.
Work plan	This deliverable is the result of work performed from M7 to M24 in Task 4.5 - Predictive algorithms. Results (software) are being used as input to delivering Operational Tools in WP6, execution of pilots in WP7 and technical impact assessment in WP8.
Milestones	This deliverable, together with D4.3, is being used as verification of the achievement of MS5 "Predictive models/algorithms established" (M24).
Deliverables	This deliverable is a final report about the implementation of predictive algorithms defined in D4.3 and it is used as input for D6.4 (integration in PIXEL Operational tools), D7.2 (integration activities) and D8.3 (technical evaluation).
Risks	**WT5#6 Technical activities are not completed on time, are not aligned with the main objective, are not accurate or present a lack of consistency.** This deliverable shows that technical activities related to T4.5 have been executed in a timely fashion as defined in "Plan and future work" sections of D4.3. **WT5#9 Some processes cannot be modelled as they depend on too many factors or they are overmuch random.** Initial work shows promising results, thus minimising the risk of unfeasible models.

1.2. The rationale behind the structure

This is the final report about results obtained in T4.5 of the PIXEL project. Except for the introduction and conclusion, each section describes a specific subtask related to PIXEL data analytics. Those sections are organised in an introductory part summarising the problem statement, followed by the description of prediction and analytics developed for PIXEL use cases and scenarios, and, finally, the presentation of obtained results.

Appendix 1 lists data sources that have been used for data analytics in PIXEL. For each data source, a list of fields has been provided: Dataset name, Data Source, Description, Usage in PIXEL, Algorithms, Sharing of results, License and terms of use, Comments, DPO (Data Protection Officer) assessment.

Finally, Appendix 2, provides links source code and reports that are an integral part of this deliverable.

1.3. Version-specific notes

This is the final deliverable in a series of two.

Relation of the work performed in Task 4.5 - Predictive algorithms to overall PIXEL objectives, use cases and requirements are provided in sections ***2.1. Relation with PIXEL objectives and use cases*** and ***2.2. Relation to requirements*** of the first version of this report ***(D4.3 Predictive Algorithms v1)***. Those sections are based on the analysis of the following documents: ***Grant Agreement***, ***D3.4 Use cases and scenarios manual v2*** and ***D3.2 PIXEL Requirements Analysis.***

The reader is thus referred to D4.3 and the above-mentioned documents for a detailed elaboration of this topic.

This deliverable provides a full report about applied analytics and achieved results.

2. Predicting vessel calls data from FAL forms and other sources

Ports represent a rich source of data that can be utilized for optimizing port operations, affecting not only port processes but the whole transport chain. Vessel call data represents the most general information, available in a standardized form in each port. The forms that need to be submitted, were standardized in a Convention of Facilitation of Maritime Traffic (FAL)[1]. These documents are provided to the ports and public authorities through "single window" solutions, as a mandatory requirement[2], opening the potential for exploiting such data for gathering business insights that could add value and competitive advantage to the ports and maritime chain as a whole. Most of the ports have Port Community Systems in place, that store all the relevant FAL forms data, which is usually available also as historical data, particularly useful for performing data analytics and predictive modelling, addressed in this chapter.

2.1. Predictions and analytics for PIXEL

The three main results of the vessel call data analysis are:

- General statistical analysis and visualization of the vessel call data.
- Vessel turnaround time / ETD (Estimated Time of Departure) prediction.
- Analysis of the vessel patterns and cargo seasonality for long-term vessel call prediction

Statistical analysis and visualizations or EDA (Exploratory Data Analysis) of the vessel call data provide a deeper understanding of port operations and external factors that have an impact on them. Different long-term trends were analysed, such as the number of vessels through years, amount and types of cargo and average turnaround times. Turnaround time is one of the most important pieces of information that was extracted from the data. According to UNCTAD (United Nations Conference on Trade and Development)[3], it is one of the main indicators of port efficiency and trade competitiveness. A lot of effort to understand and explore all the factors that influence it was made. Some of the results of the vessel call EDA were not only useful in providing insight in port operations and performance but also the required understanding of the predictive algorithms built.

Ability to reliably predict turnaround time and at the same time also expected time of departure is an important step for the optimization of port operations and vessel scheduling. An ETD predictive algorithm was developed, using state-of-the-art gradient boosting machine learning algorithms, by utilizing vast amounts of vessel call data, as well as external environmental data. ETD predictive algorithm was evaluated on historical data, using cross-validation method and on live operational data from VIGIEsip system in Port of Bordeaux (GPMB). Different error metrics were calculated, and a comparison was also made on live operational data, which results were compared with baseline predictions from the Port of Bordeaux. The methodology and evaluation of the ETD predictive algorithms are presented separately.

The third main result is the analysis of vessel and cargo frequency and seasonality, through different visualizations and corresponding metrics, for separate cargo types and ships. Seasonal influences on cargo traffic were investigated, providing an understanding of cargo types and amounts that are transported at different times of the year. By looking at historical data, predictions can be extracted on similar seasonal variations for the future.

[1] http://www.imo.org/en/OurWork/Facilitation/ConventionsCodesGuidelines/Pages/Default.aspx

[2] http://www.imo.org/en/MediaCentre/PressBriefings/Pages/06-electronic-information-exchange-.aspx

[3] https://unctad.org/en/PublicationsLibrary/rmt2019_en.pdf

2.2. Results

In this section, the results for all the above-mentioned problems are presented, based on **Port of Bordeaux - Bassens terminal (GPMB)** historical vessel call data. GPMB was selected due to availability of all the needed data, especially large quantities of historical data, as well as live operational data, with baseline results available for comparison with the currently used approach in the port. The data used represents standardized data, applicable to any of the ports, with the same approach, as described in the following sections.

2.2.1. Exploratory data analysis of vessel calls data

Visualizations are one of the most important tools for understanding data and plays the most important role in exploratory data analysis (EDA), an important piece of our machine learning workflow. EDA can also be viewed as a stand-alone result, as it provides insights into port operations.

11 years of historical port call data between the beginning of 2008 and the end of 2019 were collected. Data contains 6055 port calls which consist of 1905 unique vessels, unloading 55 and loading 46 types of unique cargo types. Average turnaround time of port calls in historical data is 53 hours, with 60 hours standard deviation. Recorded turnaround times are in a range between 0 and 965. Both values are examples of outliers, which are removed beforehand, using multipliers of standard deviation from the median turnaround time.

Different external factors that influence port operations were analysed. For example, weather data, water height (tide levels), holidays and congestions. The weather has an impact on turnaround time, but only for some types of cargo. So, combinations of factors must be considered, such as wind speed and cargo type or precipitations and cargo type. Because of changing water height in the estuary due to tides, vessels are unable to enter or exit the port when water is not high enough, at low tides. This kind of data represents a specific port data that needs to be considered along with standard vessel call data. At holidays there is reduced manpower available, so port operations are slowed down which results in reduced efficiency and increased turnaround times.

From the beginning of the available data, the yearly number of vessel arrivals is decreasing (Figure 1), but vessels carry more cargo (Figure 2). In 2008, vessels on average transported 4480 tons of cargo per arrival (sum of unloaded and loaded cargo tonnage), in 2018 they transported 5430 tons of cargo. The peak was in 2013 when the average vessel transported 6000 tons of cargo per arrival.

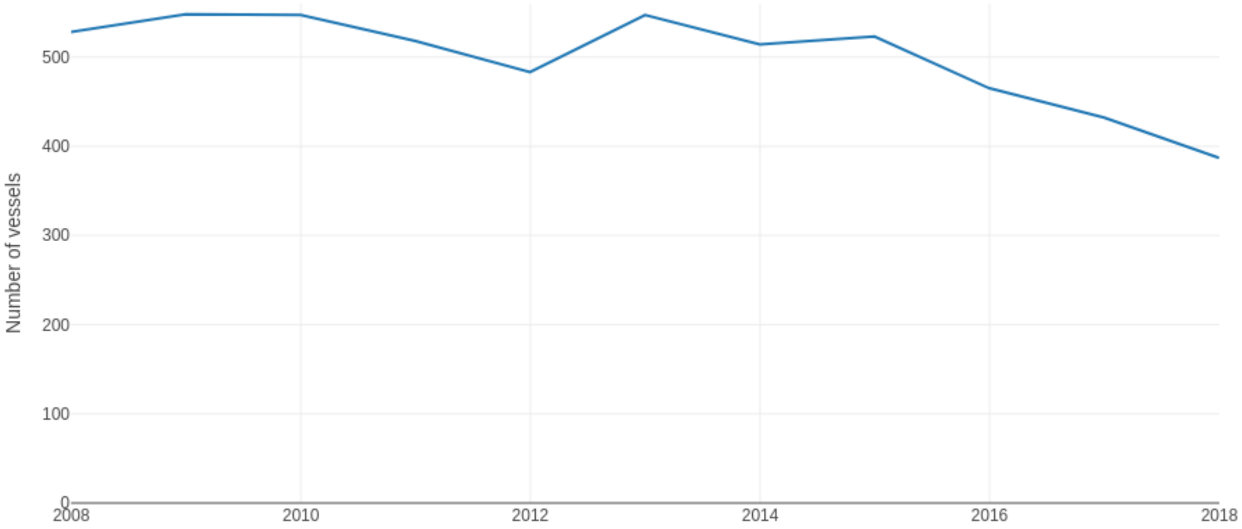


Figure 1. The number of vessels arriving at GPMB from 2008 to 2018.

In all historical data, 28,852,852 tons of cargo have been processed. The yearly average of loaded cargo is 1.43 Mt, and 1.18 Mt of unloaded cargo. Amount of processed cargo was increasing until 2013 and had seen a slight decrease since then (Figure 2).

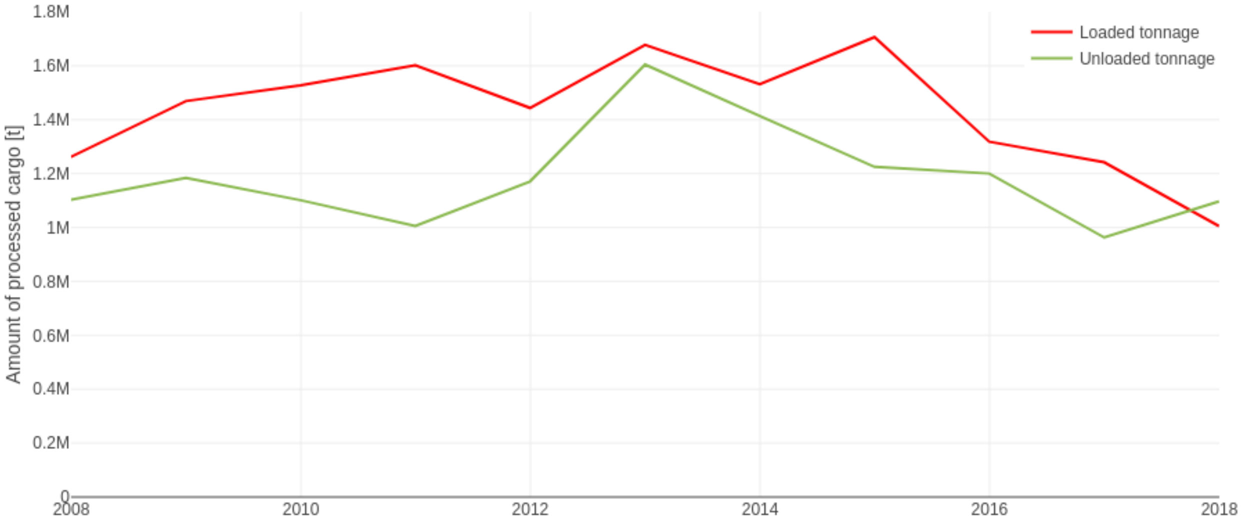


Figure 2. Amount of loaded and unloaded cargo at GPMB from 2008 to 2018.

Turnaround time distribution has clusters that occur due to tides. Most of the vessels enter or exit the port if there is high tide. More details about the correlation between turnaround time and water height will follow below.

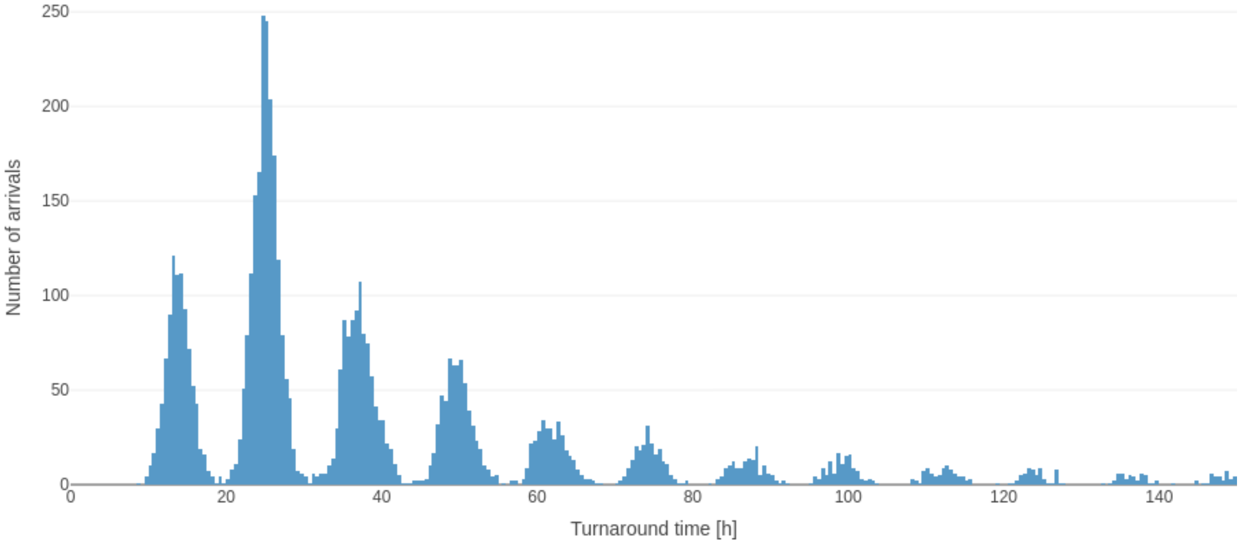


Figure 3. The number of arrivals according to the turnaround time.[4]

Most vessel arrivals occur in the middle of the week, on Wednesdays and Thursdays. Most departures happen on Fridays. There is a very small amount of departures on Sundays, which means that vessels that arrive over weekends, usually stay longer (Figure 4).

[4] Clusters are clearly visible and are formed due to water height.

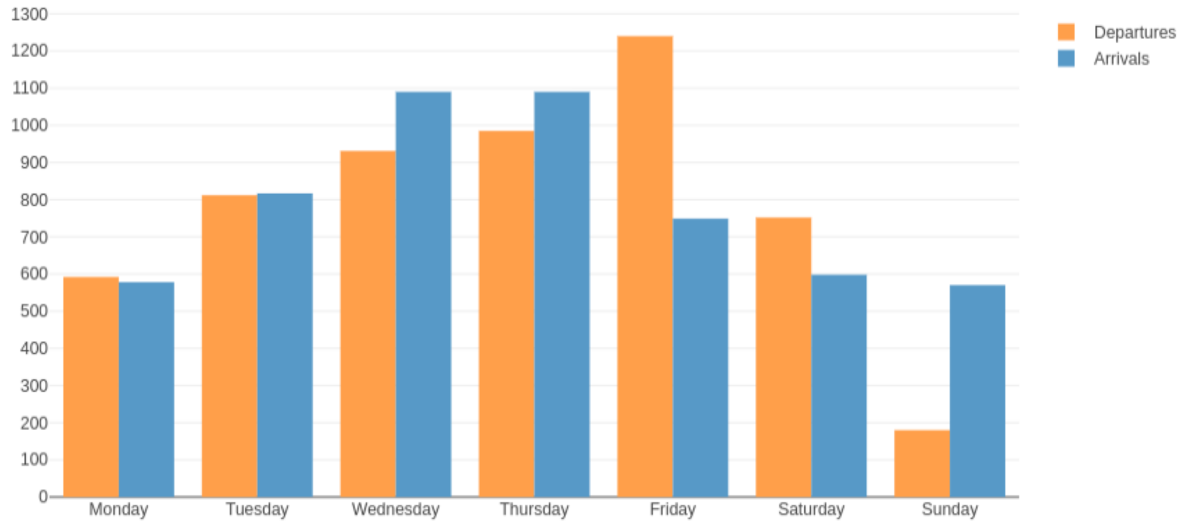


Figure 4. The number of arrivals and departures given day of the week.

Figure 5 presents how the day of the arrival influences turnaround time. There is reduced manpower in the port over the weekends or holidays, which reflects in longer turnaround times for vessels arriving on or just before those days.

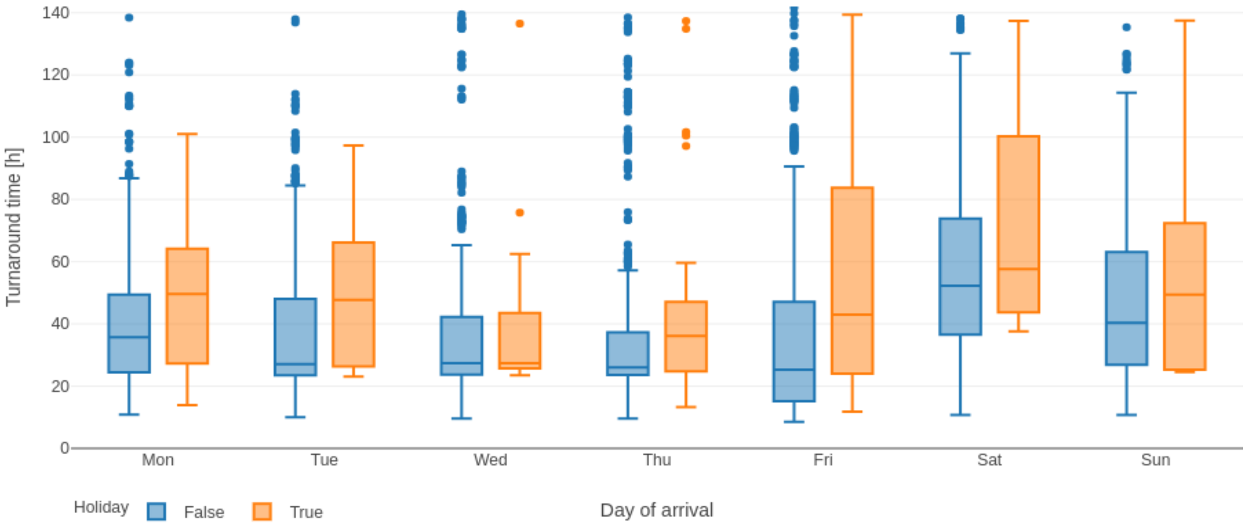


Figure 5. Influence of the arrival day and the presence of the holiday on the turnaround time.

There were 84 different cargo types processed in the port. They unloaded 55 and loaded 46 cargo types. The number of loading and unloading all cargo types that were processed more than 20 times are presented as stacked histograms in Figure 6.

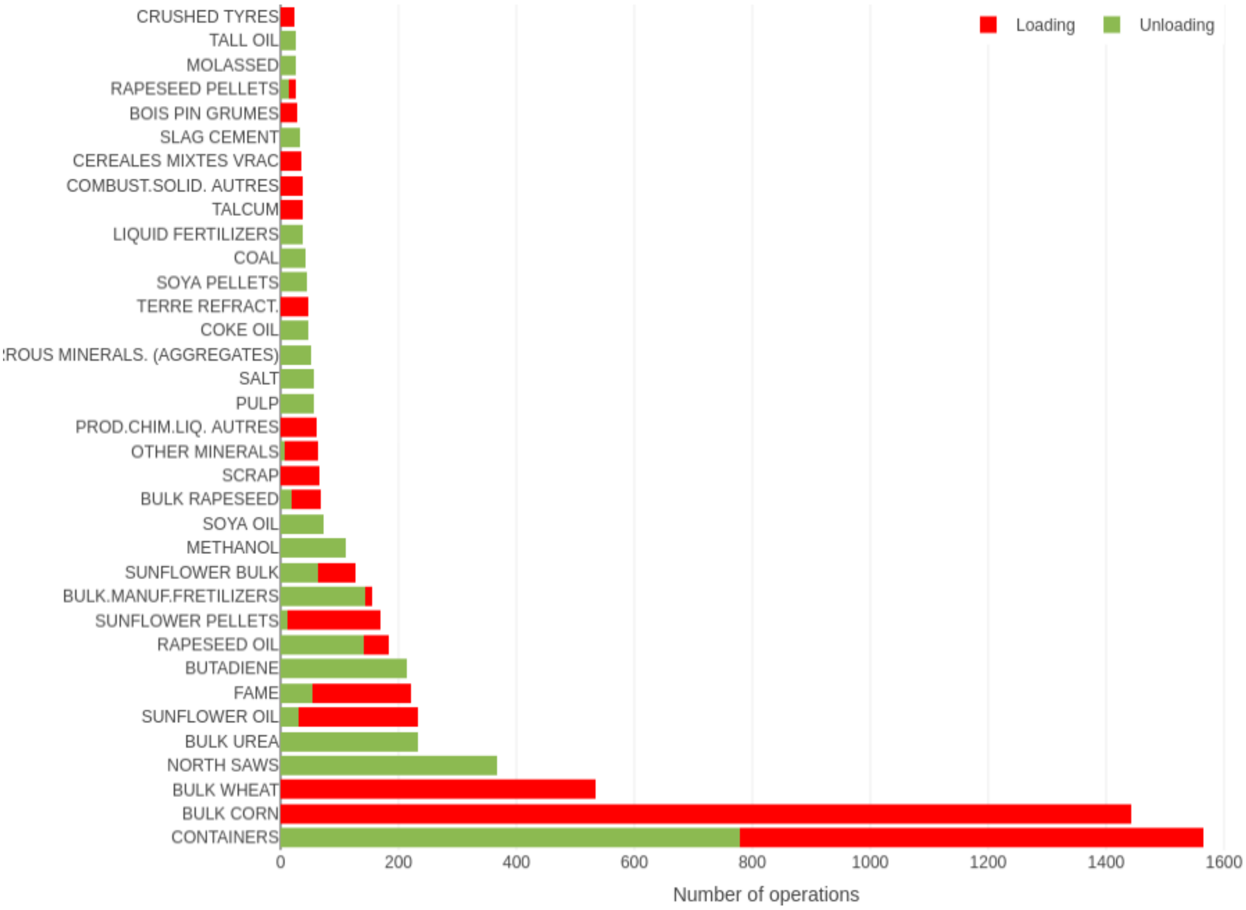


Figure 6. The most frequent cargo types and the number of operations.

Most of the exported cargo through the port is cereals (MAIS VRAC). Production is heavily affected by climate change[5]. Trend of the amount of exported cereals is declining and has almost halved since 2008 (Figure 7).

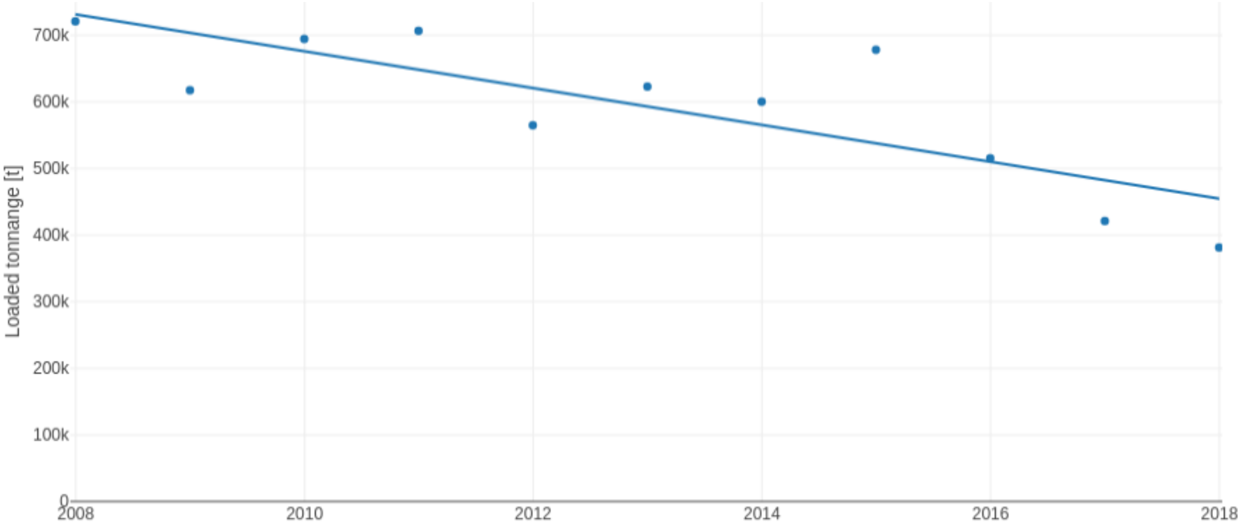


Figure 7. Decreasing amount of loaded cereals (MAIS VRAC) through the years.[6]

[5] http://www.occitanie.developpement-durable.gouv.fr/IMG/pdf/Etude_MEDCIE_GSO_cle1a7936.pdf (page 73)

[6] Scatter plot (dots) represent actual values, while the line represents the trend.

As intuitively expected, different types of cargo have different turnaround time distributions for unloading and loading cargo types. Processing capacity for unloading and loading cargo types in tonnage per hour, are presented in Figure 8 and Figure 9, respectively. Bulk cargo types have much larger turnaround times with large deviation. On the other hand, liquid cargo types have short turnaround times with little deviation.

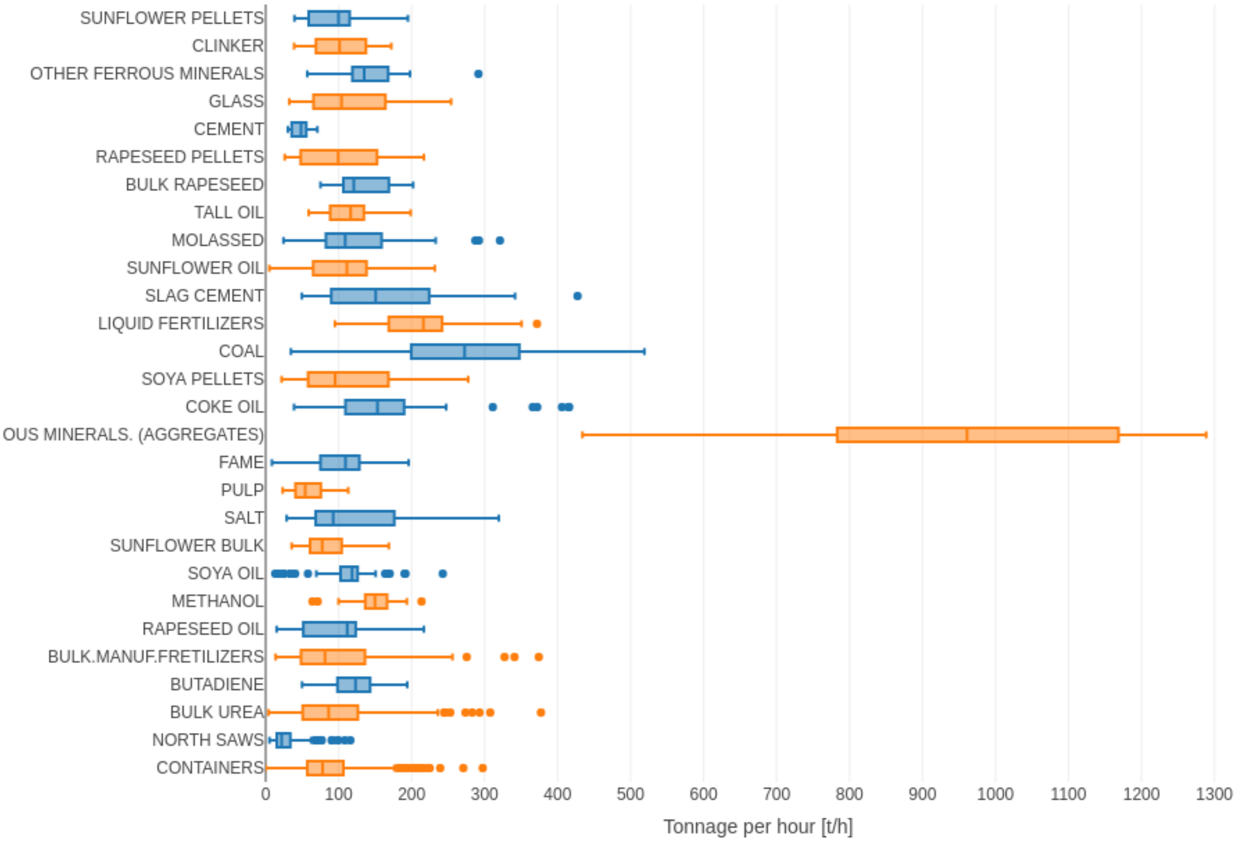


Figure 8. Cargo processing in tonnage per hour for different unloading cargo types.

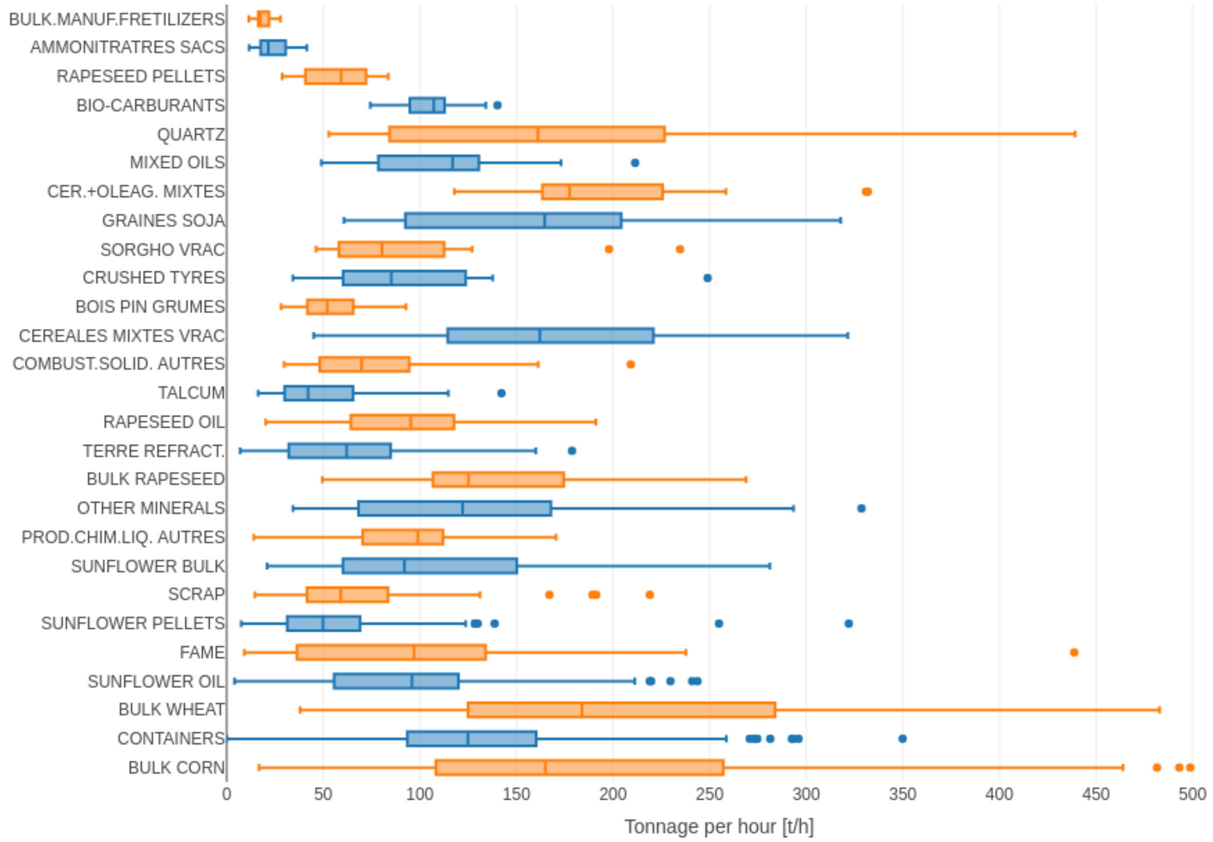


Figure 9. Cargo processing in tonnage per hour for different loading cargo types.

The "Port of Bordeaux - Bassens terminal is located about 90 km in the Gironde estuary. Water in the estuary is not deep enough to allow large ships to sail at low tide. There are multiple water height sensors along the Gironde estuary. The value from those sensors represents relative water height against the mean daily low, through multiple years for each sensor. As seen in Figure 10, most vessels depart when tidal levels are above 3 meters.

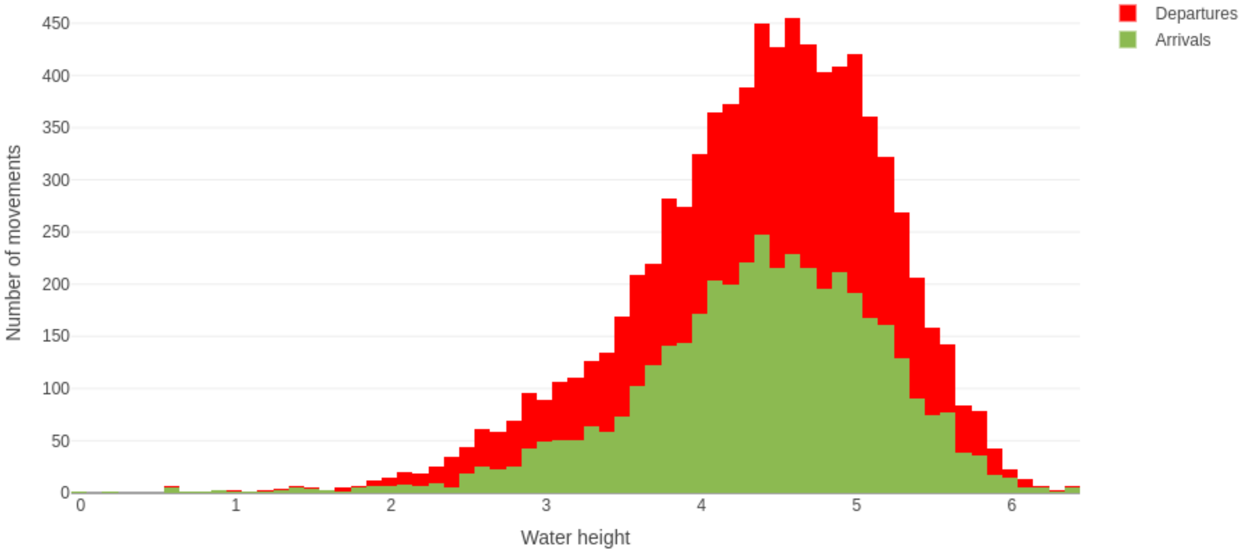


Figure 10. Tidal levels related to vessel departures and arrivals.

Another visualization that clearly shows ports dependence on water height is presented in Figure 11. The blue line represents the height of the water, green and red markers represent vessels arrivals and departures. They are clustering at higher water levels.

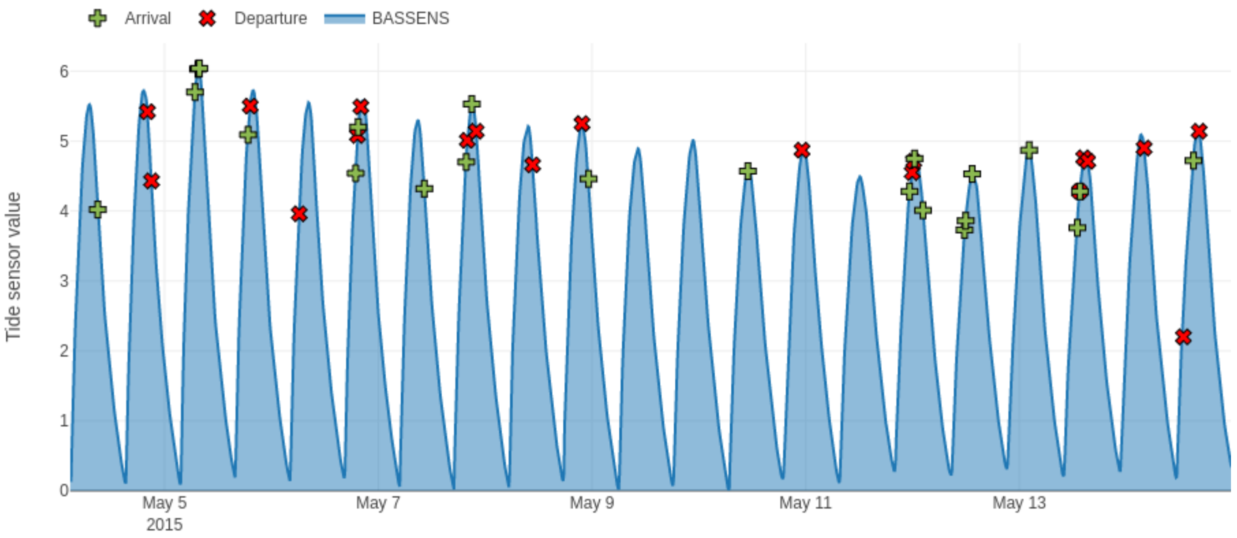


Figure 11. Tidal levels related to vessel departures and arrivals for 1 week.

Weather data, especially wind and precipitation data, that may influence the turnaround time, was examined. Hourly aggregated weather data was combined with historical port call data. As presented in Figure 12, the level of precipitation seems to influence turnaround times. Certain dry bulk cargo types experience noticeably longer turnaround times, in comparison with containers or liquid cargoes.

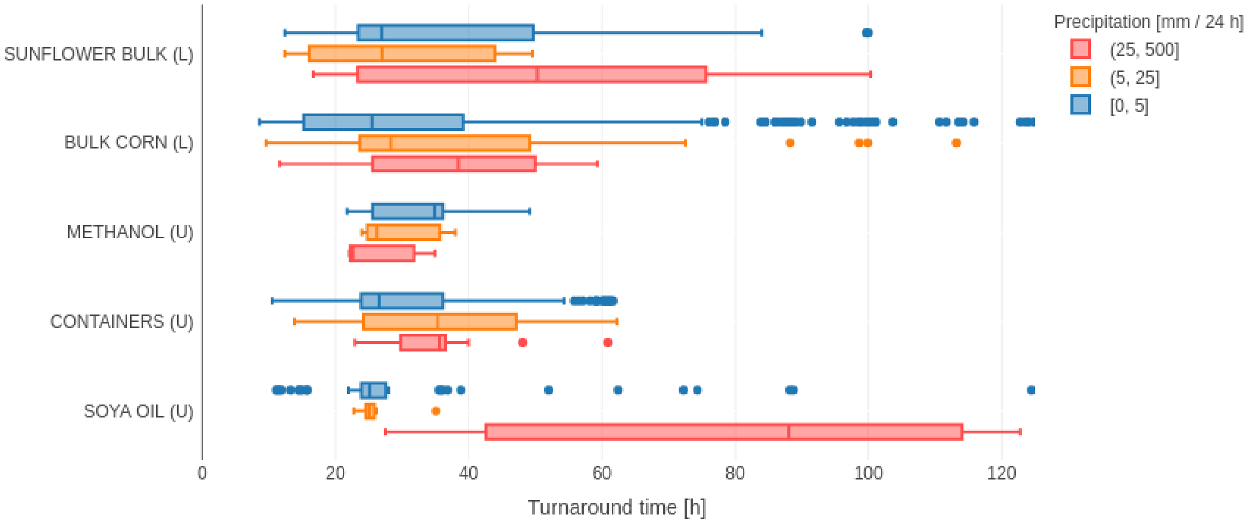


Figure 12. Influence of precipitation levels on turnaround time for loaded and unloaded cargo types.

2.2.2. ETD predictive algorithm

EDA, presented in the previous section presents an important part in a machine learning workflow (Figure 13), that was used to develop the ETD prediction system. The same data from GPMB was used, representing operational data, with some data preparation needed for stable model training and increased generalization performance. All vessels without any cargo to load and unload (i.e. empty ships) were initially removed. Ships, whose turnaround time was less than 1 hour and those of which turnaround time exceeded two times the standard deviation from the median, of the corresponding cargo type (i.e. outliers, corresponding to erroneous inputs in VIGIEsip), were also removed, as well as vessel calls, with combinations of loading and unloading cargo types that occurred in less than 5 occurrences. This filtering procedure reduced the number of port calls to 5492, with 1768 unique vessels and 34 unloading and 32 loading cargo types, respectively.

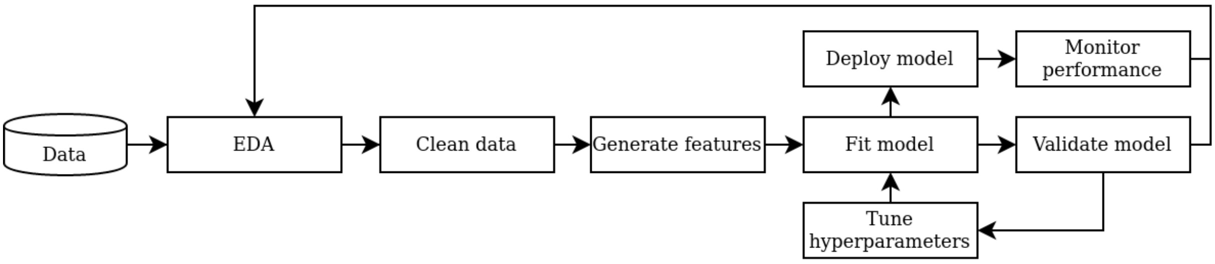


Figure 13. Machine learning workflow used for developed ETD prediction system.

Data cleaning and constructing meaningful features, that hold predictive value, represents a major part of the machine learning workflow. Data was investigated in exploratory data analysis, where the influence of different parameters on turnaround time was analysed and already presented in Section 2.2.1. Besides identifying predictive features, encoding them appropriately is also particularly important. Timestamp data (e.g. arrival time) needs to be transformed into multiple categorical features (e.g. day of the week, hour). Most of the machine learning models also require the data to be normalized and categorical features transformed into numerical ones (e.g. one-hot-encoding).

For predictive modelling, a recently presented, gradient boosting-based method was utilized, CatBoost[7], which offers the most convenient use in operational environments with heterogeneous, structured data. The CatBoost method uses decision trees as base predictors and thus omits the need for data normalization as the pre-processing step. CatBoost also handles categorical features during training, as opposed to pre-processing time, thus omitting the need for special transformations. Decision trees also have an increased level of explainability of the results, in comparison with other black-box methods (e.g. neural networks). CatBoost also outperforms other state-of-the-art gradient boosted decision methods (e.g. XGBoost[8]) in terms of predictive accuracy and speed, out-of-the-box, without the need for extensive hyperparameter fine-tuning. Nevertheless, the importance of hyperparameters on data was analysed, using Grid Search parameter tuning procedure, provided with the CatBoost implementation and present the results of this analysis.

To evaluate the proposed method, a combination of historical and live operational data from the port was used. Turnaround time is computed as a difference between arrival and departure times. To effectively use the available data, the method using the cross-validation procedure with a 1-year left-out strategy on historical data was evaluated. In this way, the proposed method was evaluated on all the 11 years of data. The method on the splits was evaluated, using the MAE (Mean Absolute Error), RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error). With such a combination of evaluation metrics, we ensure explainability of the performance (MAE, MAPE), as well as to evaluate the influence of the large errors (RMSE). The best performing model on historical data was also evaluated on live operational data of 100 vessels from the port. In this case, the comparison was also made against the expertly provided predictions from the port, based on their current manual approach.

2.2.2.1. Experiment results

Date of vessels arrival, loading and unloading berth numbers, cargo types and tonnage were the most important attributes that were used from FAL forms for the predictive algorithm. Date of arrival was encoded as a day of the week, which is a categorical feature. Hours of arrival were segmented into 6 bins (therefore bin size is 4 hours) and encoded as a categorical feature. Loading and unloading tonnage is already a numerical value, so there is no need to transform them. Cargo types and berth numbers were used as categorical features. From external data sources holidays, weather, and tidal data were used. Holidays were encoded as Boolean features, indicating the presence of holiday in +-3 days from the vessel arrival. The weather was encoded as an aggregation of precipitations and wind speed in the first 24 hours after vessel arrival. Tidal data was encoded as current water height, time to next and since previous low and high

[7] https://catboost.ai/

[8] https://xgboost.ai/

tide. Categorical features were encoded with a one-hot encoding method, except for the CatBoost model. CatBoost library can handle categorical features out of the box. Some models required normalisation of the data.

A cross-validation procedure and evaluation metrics were used to first evaluate predictive performance on large-scale historical data. Results, for the most frequent unloading and loading cargo types, are reported in Table 1 and Table 2, respectively. Results are also compared to the linear regression model, to demonstrate the superiority of the used CatBoost method, over baseline machine learning models. Note that careful feature normalization and categorical feature transformation was performed for linear regression.

It is noticed that MAE is around 15 hours for unloading and 11 hours for loading operations or expressed with MAPE - around 30% error. The comparison is made against ground truth turnaround time, computed out of exact arrival and departure times. The error on predicted turnaround time reduces greatly for certain cargo types (e.g. liquid cargo, containers), even below 10% and it is evident that the error correlates with specific cargo types and findings presented in the EDA section. Overall error is similar for both, unloading and loading operations and CatBoost significantly outperforms baseline linear regression method.

Table 1. Evaluation with historical data for 10 most frequent unloading (U) cargo types.

Unloading cargo type	MAE		RMSE	MAPE
	CatBoost	Linear R.	CatBoost	CatBoost
BUTADIENE	2.51	4.46	4.33	13.87
METHANOL	2.69	6.18	4.25	8.51
SOYA OIL	7.16	8.47	13.28	22.49
CONTAINERS	7.23	8.42	9.79	22.81
RAPESEED OIL	11.12	11.68	17.84	22.81
SALT	16.67	19.3	25.4	41.15
BULK. MANUF. FRETILIZERS	17.49	18.76	27.87	34.84
BULK UREA	23.19	26.7	32.03	37.07
NORTH SAWS	24.87	26.83	33.31	45.84
SUNFLOWER BULK	83.03	82.37	105.47	41.53
Top 10 cargo types (U)	14.62	16.33	27.46	28.43
All cargo types (U)	15.75	17.55	28.22	30.46

Table 2. Evaluation with historical data for 10 most frequent loading (L) cargo types.

Loading cargo type	MAE		RMSE	MAPE
	CatBoost	Linear R.	CatBoost	CatBoost
CONTAINERS	7.24	8.41	9.79	22.79
BULK CORN	8.65	10.51	13.89	29.64
LIQ. CHEM. PROD. OTHER	9.42	11.48	15.58	32.78
BULK WHEAT	10.04	11.53	14.98	31.34
OTHER MINERALS	12.12	13.55	17.53	38.25
SUNFLOWER BULK	13.34	14.62	21.41	38.88
SUNFLOWER OIL	13.64	13.54	18.58	29.27
SCRAP	22.60	26.50	32.22	34.15

Loading cargo type	MAE		RMSE	MAPE
	CatBoost	Linear R.	CatBoost	CatBoost
SUNFLOWER PELLETS	23.06	24.47	33.02	35.75
FAME	23.61	28.83	39.77	42.34
Top 10 cargo types (L)	10.64	12.37	17.93	29.53
All cargo types (L)	11.57	12.74	19.99	30.81

The deployed system on 2 months of live operational data was additionally evaluated. The data consisted out of 93 port calls, with at least 3 arrivals for each cargo type, to have reliable statistics, comparable with historical data. The best performing CatBoost model was used, from the offline evaluation with historical data. Results, presented in Table 3 show, that the performance (MAE) is consistent with the performance on offline data and consistently better from the current simplistic model used by the port, by a large margin.

Table 3. Evaluation with live operational data for cargo types.[9]

Operation	Cargo type	PIXEL MAE	Port MAE
Unloading	LIQUID FERTILIZERS	2.03	18.66
	METHANOL	2.34	9.49
	RAPESEED OIL	2.58	32.16
	BUTADIENE	3.39	16.73
	SOYA OIL	6.84	24.74
	TALL-OIL	7.32	25.49
	BULK UREA	16.93	57.29
	NORTH SAWS	25.37	55.48
	SUNFLOWER BULK	89.77	188.93
	Combined	**16.52**	**45.81**
Loading	CORN BULK	8.35	19.41
	CRUSHED TYRES	10.75	9.75
	BULK WHEAT	10.78	17.72
	SUNFLOWER OIL	12.68	14.00
	SCRAP	13.56	56.88
	Combined	**9.97**	**22.75**

Other features that may influence turnaround time were evaluated, but it did not improve the predictive performance of the model. Tidal levels, as presented in the EDA section, have an influence on arrivals but encoding water height and the time since the last high and low water did not improve the results of the model. Similarly, the congestion in the port should influence the turnaround time. Multiple features, that encoded congestion (e.g. number of the vessels in the port, number of vessels with the same (U/L) cargo, the average turnaround time for the last N ships that visited the port in the last M days) were experimented with, but no improvement in the predictive performance was noticed. Weather data was also investigated, especially wind and precipitation data, that should influence on the turnaround time. Hourly aggregated weather data was obtained and combined with historical port call data. The level of precipitation influences

[9] The evaluation has been performed for cargo type with at least 3 arrivals in an evaluation period of 2 months. Results are reported for unloading and loading operation, as well as overall results for separate operations.

the turnaround time (as presented in the EDA section), but it did not improve the predictive performance. Note that these conclusions relate to GPMB data, which is a relatively small port, but it is doubtful, that such features should be useful in general. He most important features of the best performing model are presented in Table 4.

Table 4. Features used in the best performing model and their importance.[10]

Feature	Importance
cargo type (U)	17.40
cargo tonnage (U)	16.15
day of entry	12.71
berth (U)	11.13
cargo type (L)	8.54
hour of the entry (round 4)	8.21
berth (L)	8.03
fiscal cargo type (L)	6.70
fiscal cargo type (U)	5.56
cargo tonnage (L)	4.72
holiday on entry	0.31
holiday in 2 days	0.20
holiday 1 day ago	0.18
holiday in 1 day	0.16

2.2.3. Analysis of vessel calls seasonality

Historical vessel call data can be used for traffic trends analysis, which can be used to predict future traffic growths or declines, as well as for the seasonality analysis of the specific cargoes. Vessel calls are usually announced well in advance and are thus not beneficial to predict the actual vessel calls, but rather to give the ports the analytic capabilities that can provide insights into vessel and cargo movements and future trends that can be used for strategic planning.

From 11 years of historical data from the Port of Bordeaux, seasonality of the cargo can be analysed. When different cargo types and their volume throughout the years are analysed, an increase in volume is observed, in specific months. By understanding seasonal demands, the increased demand for equipment and space for specific cargo can be foreseen. For example, maize is harvested between late summer and early to mid-autumn, which results in a higher volume of exported maize cargo, transported through GPMB, presented in Figure 14.

[10] Feature importance values are normalized so that the sum of importance of all features is equal to 100.

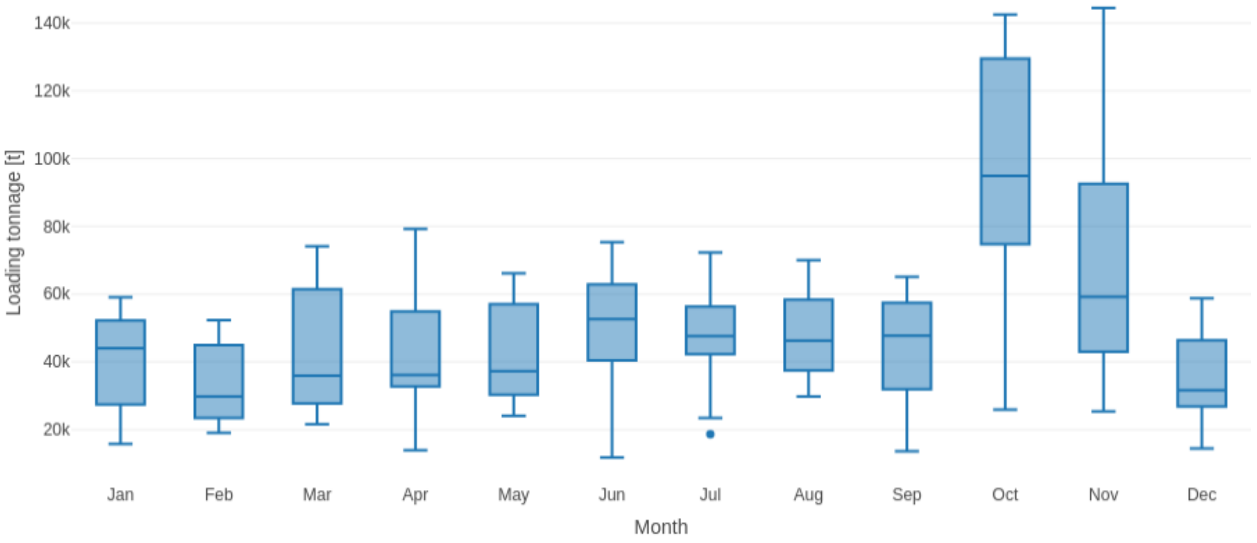


Figure 14. The amount of exported maize cargo through GPMB.

Fertilizer is mostly imported in two cycles. First one is during late winter and spring and the second one during autumn, which corresponds with the farming needs and is presented in Figure 15.

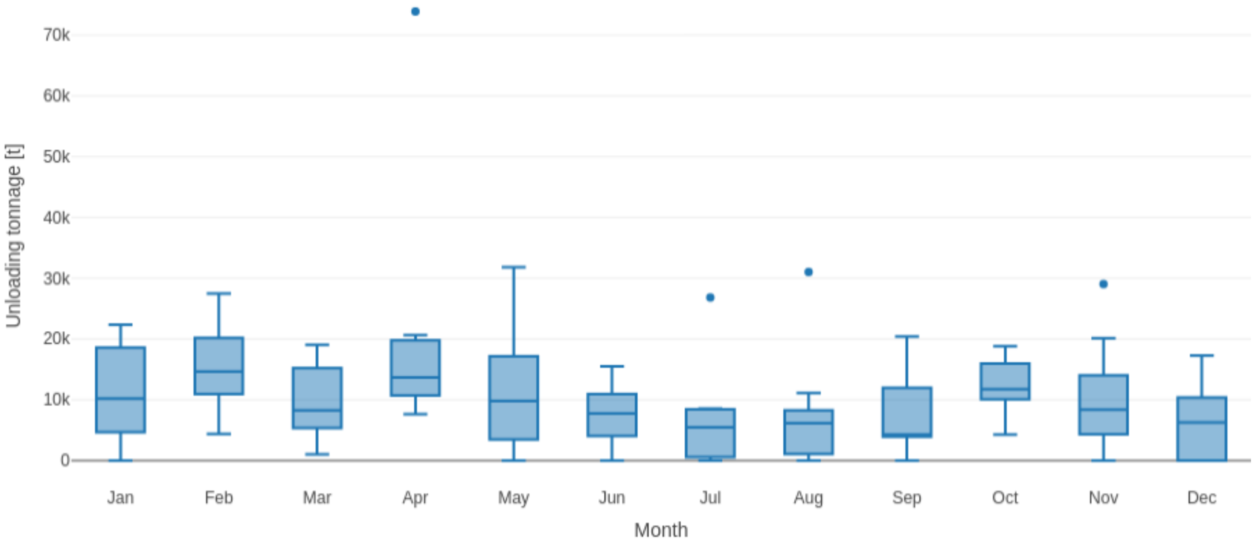


Figure 15. The amount of imported fertilizer through GPMB.

Wood from the north is also imported a lot. Most of it is during the spring until June and the second peak is during the autumn until December, presented in Figure 16.

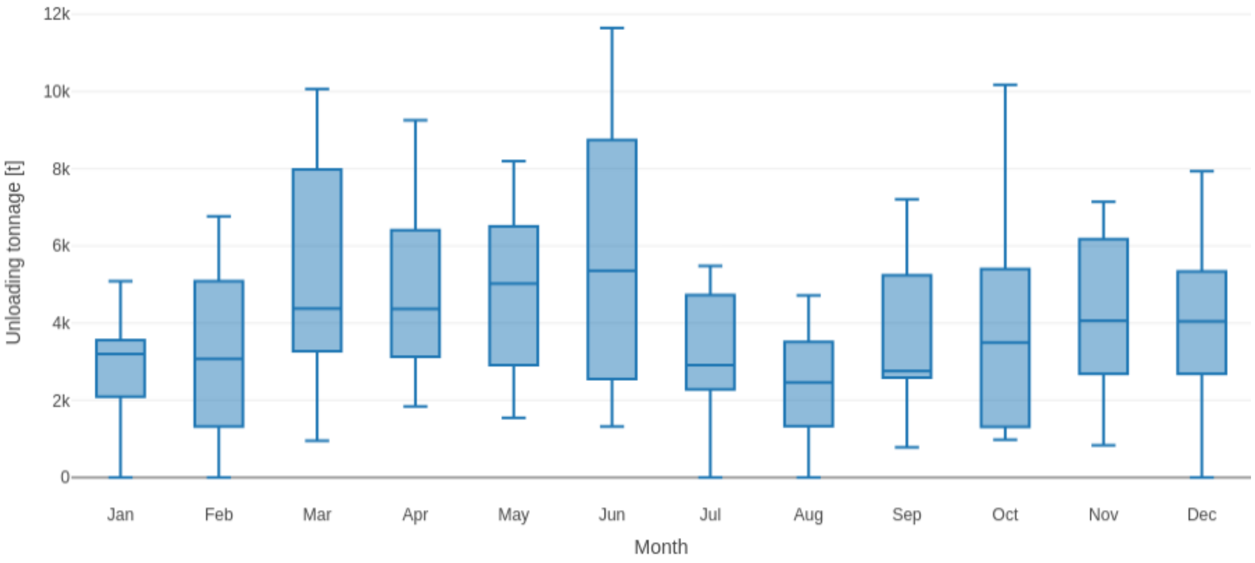


Figure 16. The amount of wood imported through GPMB.

Rapeseed is exported with an obvious peak in July and is presented in Figure 17.

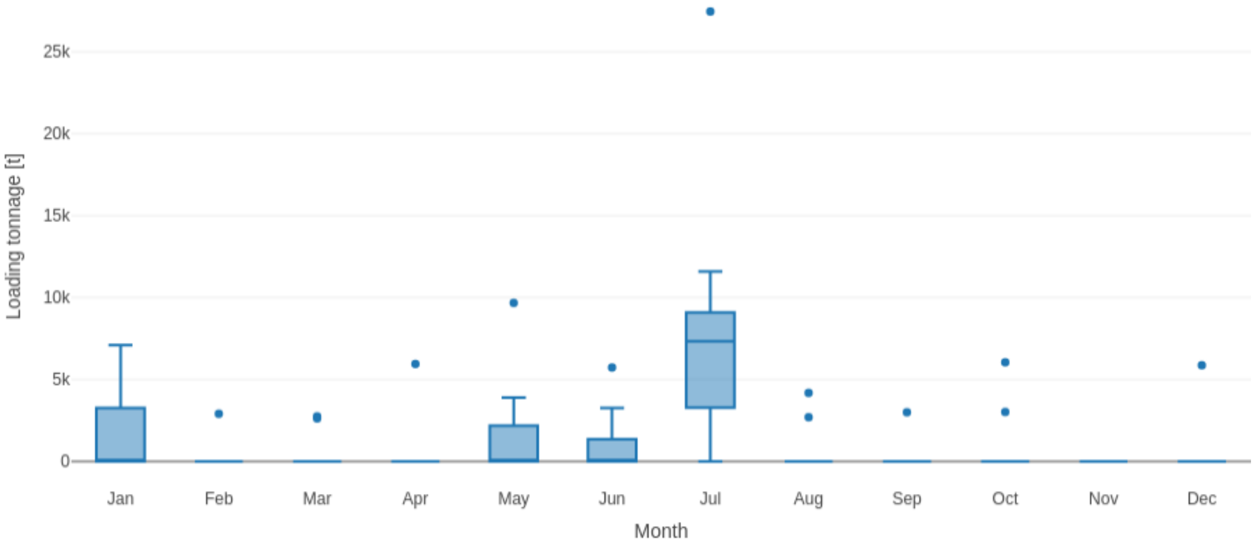


Figure 17. The amount of exported rapeseed exported through GPMB.

Sunflower seeds import peak is at harvesting time in autumn, presented in Figure 18. Such analysis was performed, for all the cargo types and an interactive tool will be integrated into the PIXEL platform.

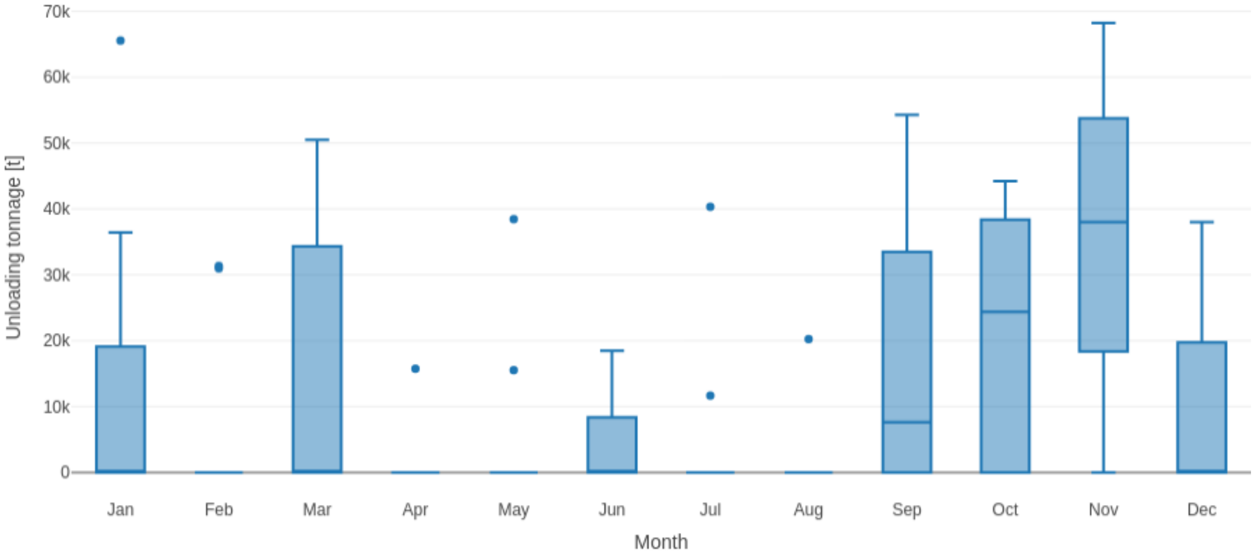


Figure 18. The tonnage of imported sunflower seeds through GPMB.

Vessels that transport certain types of cargo, are arriving regularly on the same days of the week (Figure 19 and Figure 20). Most obvious peaks have vessels transporting containers, maize, or wood. These peaks are in the middle of the week (around Thursday). Vessels transporting some other types of cargo have arrivals distributed evenly over the week. Vessels transporting *bulk urea* have decreased the number of arrivals in the middle of the week, while butadiene has decreased arrivals over the weekend. This kind of data and analysis can be used to better schedule the port resources, far in advance, offering the capability for strategic long-term planning.

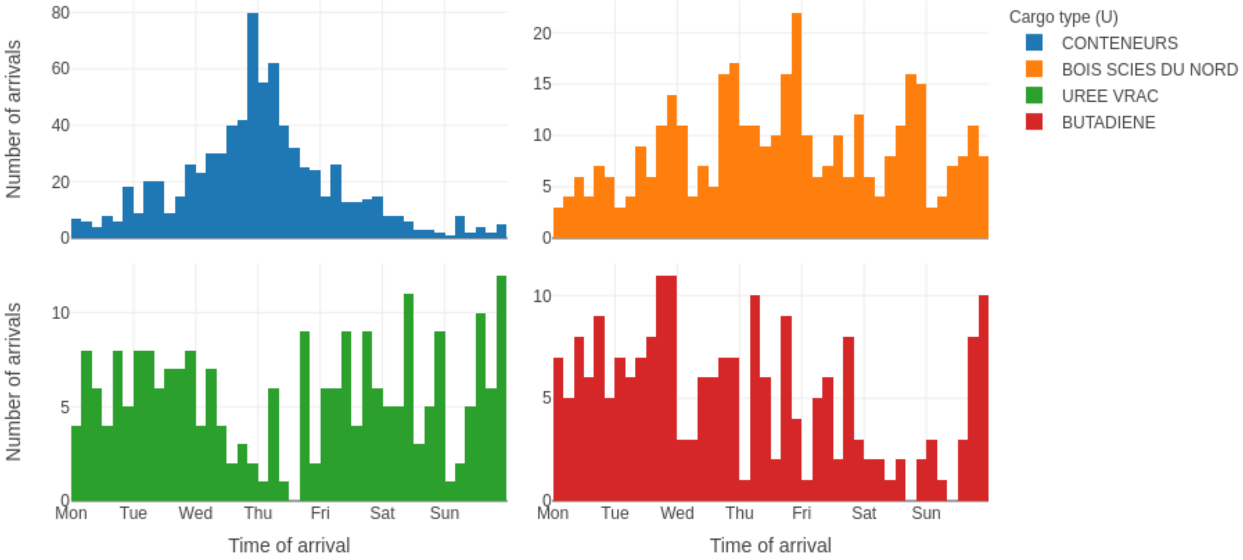


Figure 19. Unloading cargo arrival times distribution over weeks.

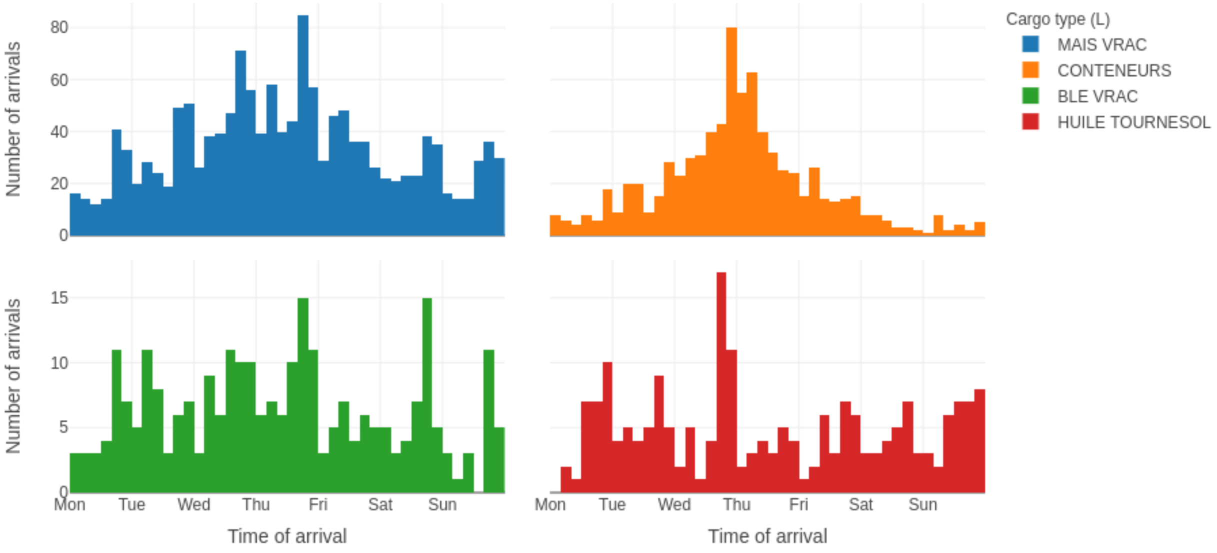


Figure 20. Loading cargo arrival times distribution over weeks.

Most of the vessels arriving in GPMB are not regular. The median number of same vessel arrivals is one (1). Nevertheless, some of the vessels are coming regularly. Some of them even have similar numbers of days between two consecutive arrivals. Distributions of days between two consecutive arrivals for some frequently arriving vessels with low variation in days between arrivals are presented in Figure 21.

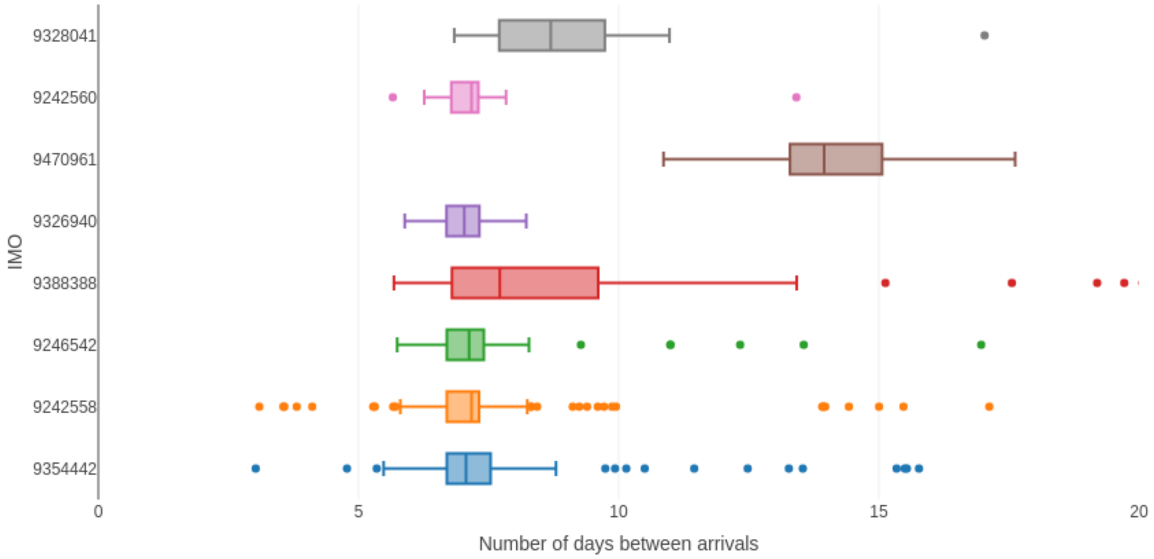


Figure 21. Days between two consecutive arrivals from the same vessel.

Some of them have a median number of days between arrivals around 7. This means they are arriving every week on the same day, even on a similar time of the day. Histogram of arrivals for 4 frequent vessels is presented in Figure 22.

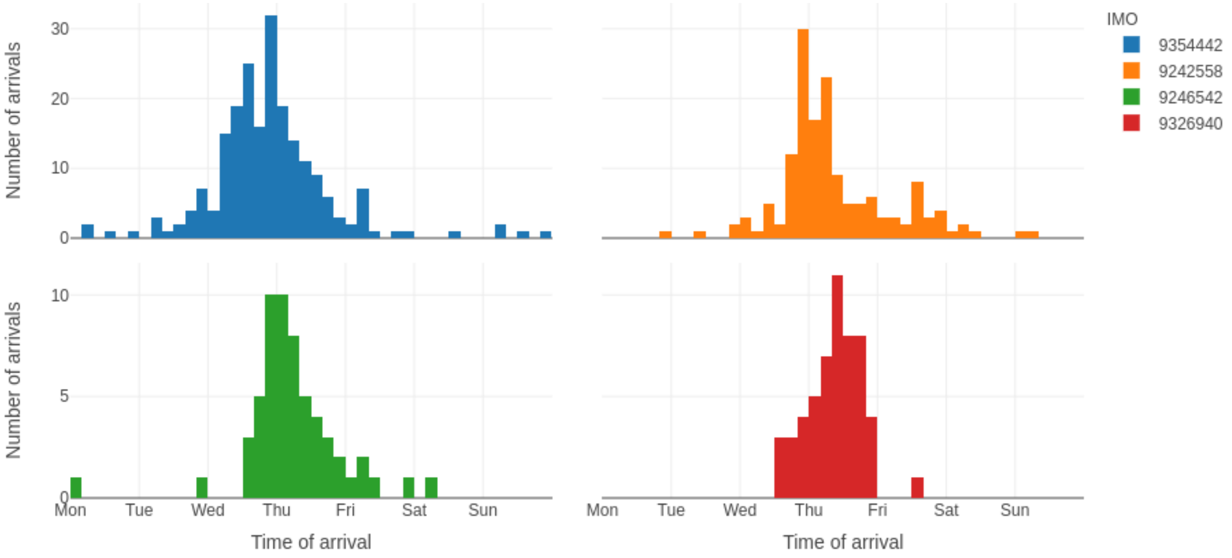


Figure 22. Vessels arrival time distribution over the weekdays.

3. Use of AIS data

The Automatic Identification System (AIS) was proposed and mandated by the IMO (International Maritime Organization) and its main intention is to prevent collisions on the sea. It provides additional information, however, it does not replace existing solutions on board, such as radar and other means that are regulated by COLREG (Convention on the International Regulations for Preventing Collisions at Sea). Since December 31st, 2004, all vessels exceeding 300GT are obligated to have an AIS transceiver installed and operational. Navigational data, information about the ship and voyage related data, is transmitted via VHF (Very High Frequency) radio between ships and shore stations. The range is limited to the VHF range, which is about 10-20 nautical miles but S-AIS (Satellite-based AIS) is available, which can track ships on the open sea. Kinematic information (ship location, speed, course, heading, etc.) and some static information, like MMSI (Maritime Mobile Service Identity), ship type, ship size, etc, are provided every couple of seconds when a ship is underway and every couple of minutes when the ship is anchored or moored. This data is available in almost real-time, while historical data is also available. Besides collision avoidance, AIS data is used for many other applications in the maritime domain, such as fishing fleet monitoring, maritime security, search and rescue, accident investigation, fleet and cargo tracking and many others. AIS data is utilized in a novel way for data analytics and predictive modelling in port areas, demonstrating its applicability well beyond its initial purpose.

3.1. Predictions and analytics for PIXEL

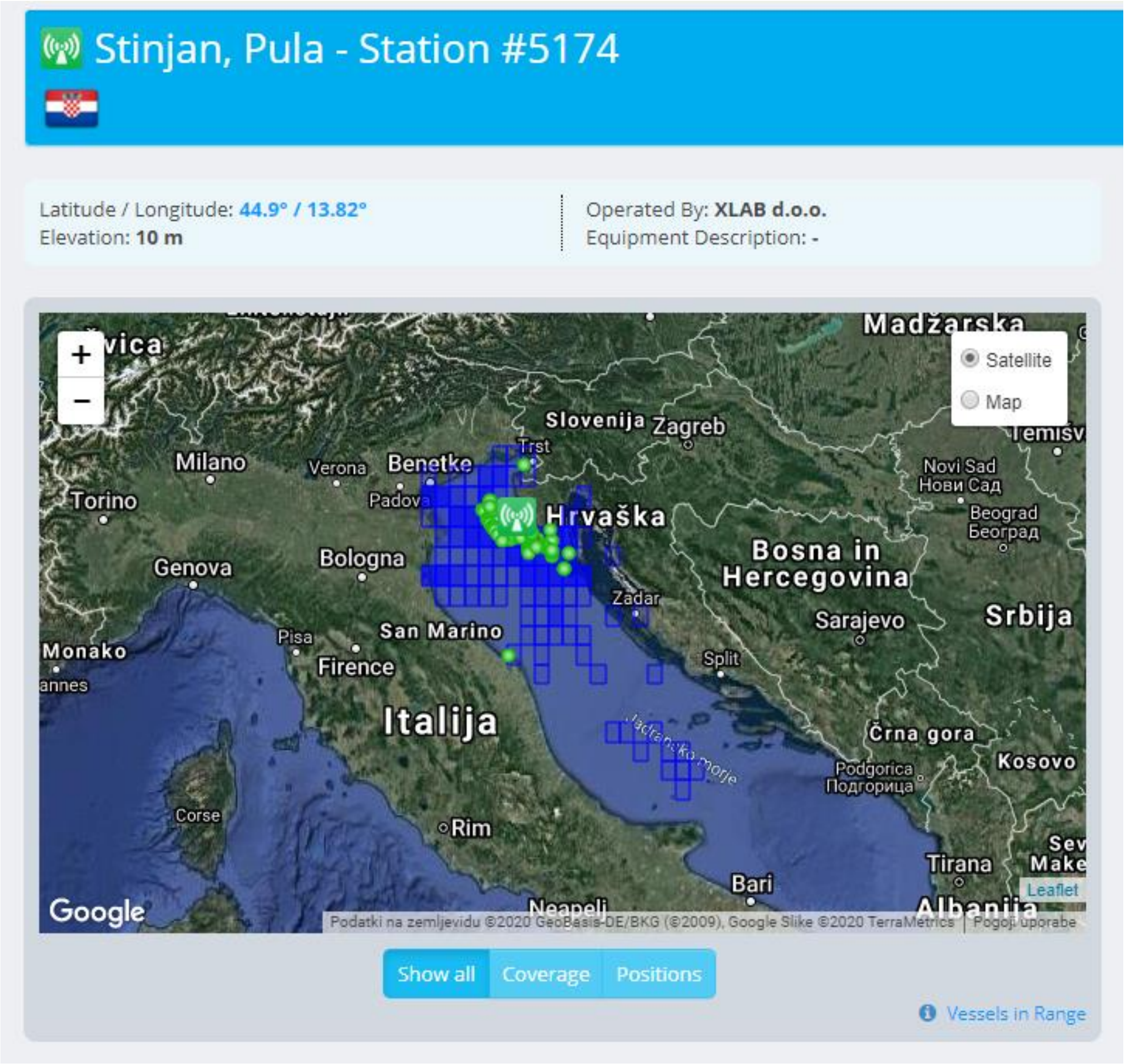
AIS data represents a rich source of data about maritime traffic and offers tremendous potential for data analytic solutions and predictive modelling, which can help at optimizing logistic chains and reducing environmental impacts. The focus in PIXEL project was to investigate AIS data for the following tasks:

- Visualization and analysis of AIS data around the ports
- Port congestions indicators
- ETA prediction from AIS data and other sources

AIS data is a geospatial data and different visualizations offer a unique view on maritime traffic, especially when visualization represents spatial and temporal view over the AIS contained fields, as well as derived AIS based metrics and products. AIS data comes in vast quantities, spatially and temporarily, thus visualization represents the most significant tool to summarize and present the main insights. Real-time availability of AIS data, as well as the possibility of having large quantities of historical data, represents a great potential to compute various port business and environmental metrics in a selected region of interest. A variety of such metrics was computed, to detect different patterns, which can be used by the ports to numerically express their business and environmental metrics and express them over time, with the actual operational AIS data. AIS data also offers a potential for predicting vessel arrival times more accurately, which can supplement official FAL forms data, as well as ETA data captured in AIS messages, which is often erroneous and not accurate. Errors were also noticed in other AIS fields (e.g. navigational status, positional data, speed), thus a special procedure was developed, to validate and to the most possible extent also correct the erroneous data, with the help of data analysis and predictive modelling.

In D4.3, different AIS data sources were presented, that were investigated for the use in the PIXEL project. Most of the AIS data that was used for the analysis, were collected from AISHub[11], which is a platform for sharing AIS data. To gain access to data from AISHub, data was shared, from the AIS receiver that is installed in Pula, Croatia. AIS data to other AIS providers, such as MarineTraffic (Figure 23) and others were also shared, which provides access to some of their data features', free of charge.

[11] http://www.aishub.net/



Stinjan, Pula - Station #5174

Latitude / Longitude: 44.9° / 13.82°
Elevation: 10 m

Operated By: XLAB d.o.o.
Equipment Description: -

Figure 23. PIXEL AIS antenna coverage, as calculated by MarineTraffic.

By connecting the AIS receiver to AISHub, it provided access to the whole network of amateur stations around the world. One of the problems in data from AISHub is weak coverage in some parts of the world. Coverage for the EU region is presented in Figure 24.

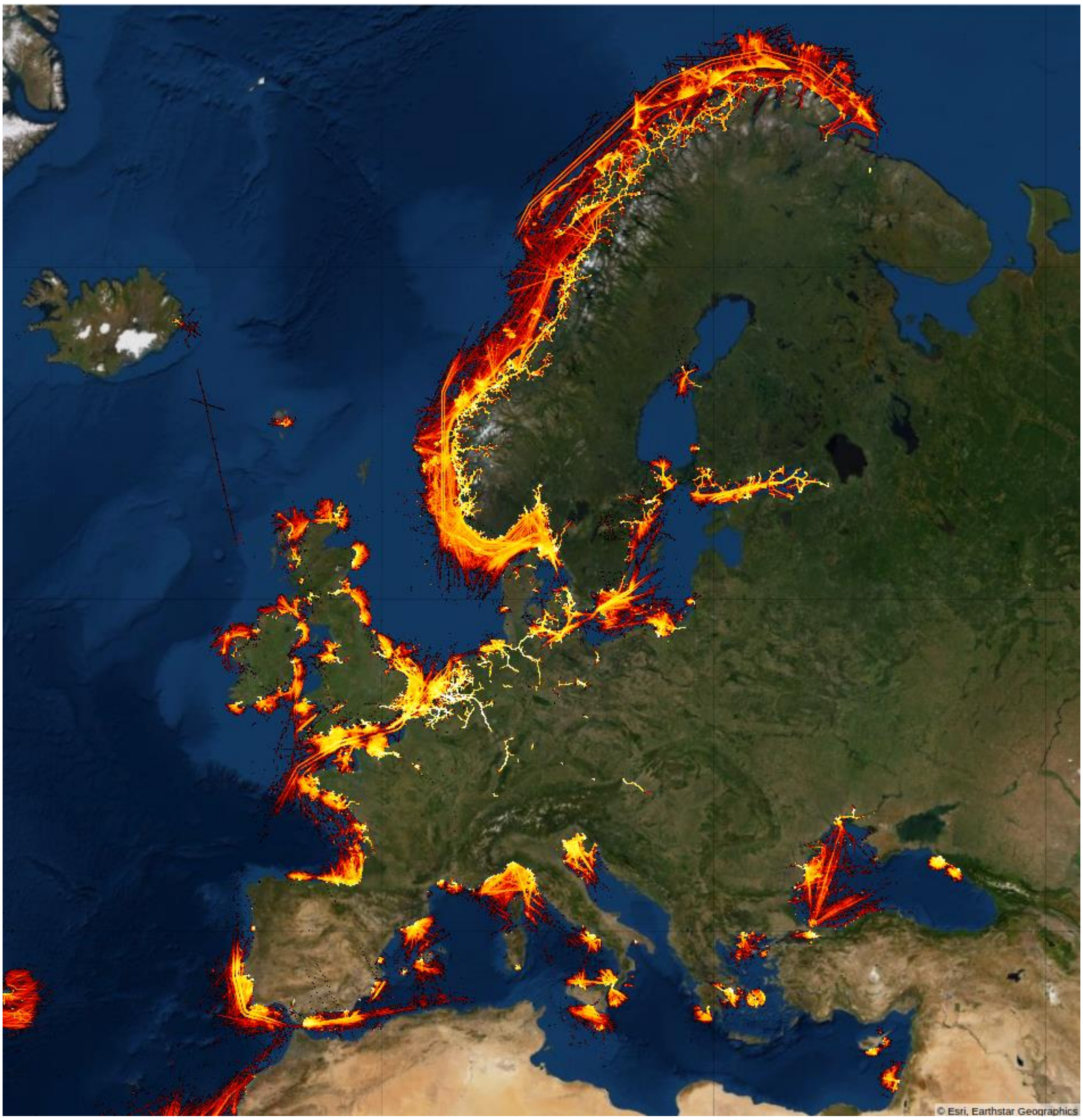


Figure 24. AISHub coverage over Europe.

AIS data for the port areas of 3 out of 4 partners – GPMB (Grand port maritime de Bordeaux), PPA (Piraeus Port Authority S.A.) and ASPM (Azienda Speciale per il Porto di Monfalcone) was collected. Among them, GPMB is not fully covered, as the antenna is only at the entrance of the Gironda estuary. For the Port of Thessaloniki, there is no AIS coverage and a separate antenna would need to be installed.

3.1.1. Visualization and analysis of AIS data around the ports

It was noticed that AIS data comes with a lot of errors in the reported fields, which prevents us from using this data directly for analytics and predictive modelling and this presented the first problem that needed to be addressed for a successful implementation of our aforementioned objectives. Errors can appear in manually entered data, such as vessel's draught, destination, ETA or even vessel's size, as well as in automatically collected data, such as location, speed, and navigational status. Examples of reporting incorrect navigational status are presented in Figure 25 and Figure 26.

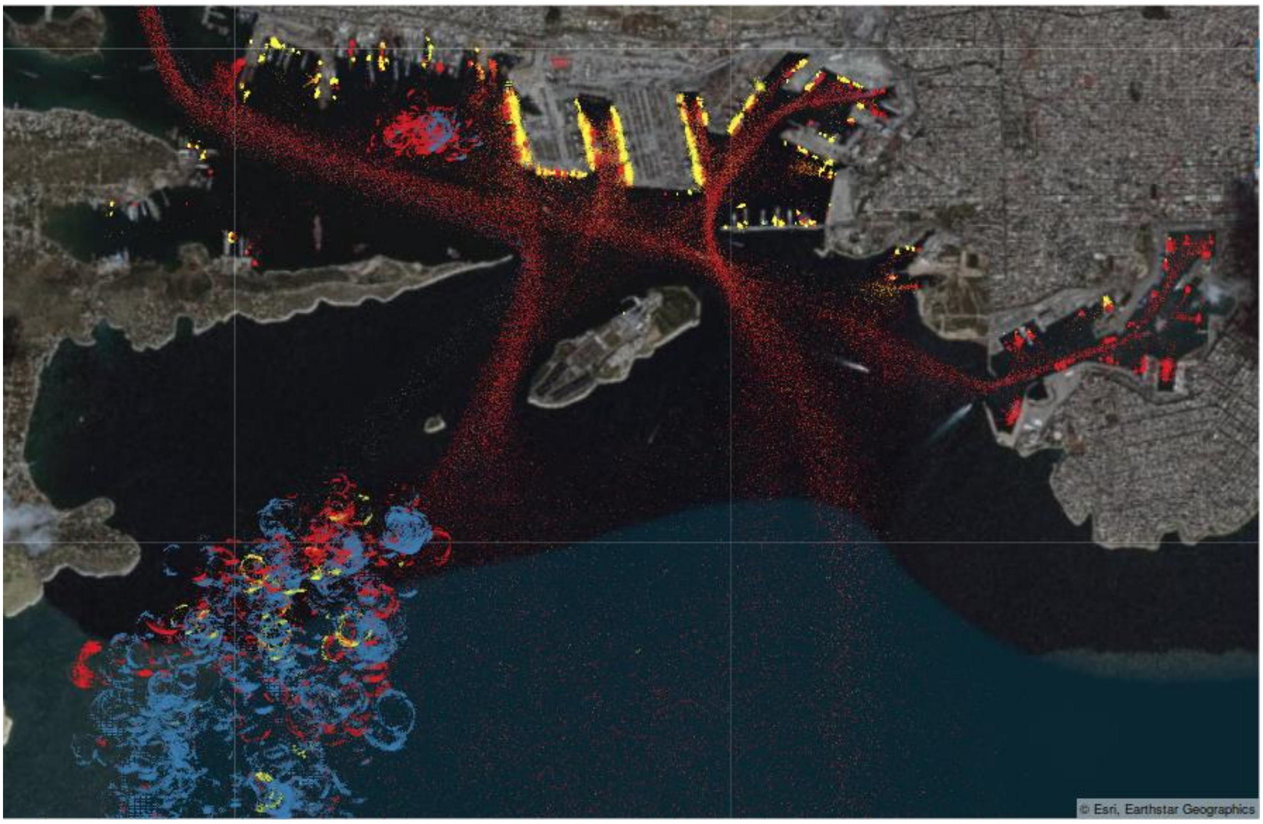


Figure 25. Navigational statuses in PPA.[12]

[12] Red dots: AIS navigational status 0 (under way using engine); blue dots - navigational status 1 (at anchor); yellow dots - navigational status 5 (moored).

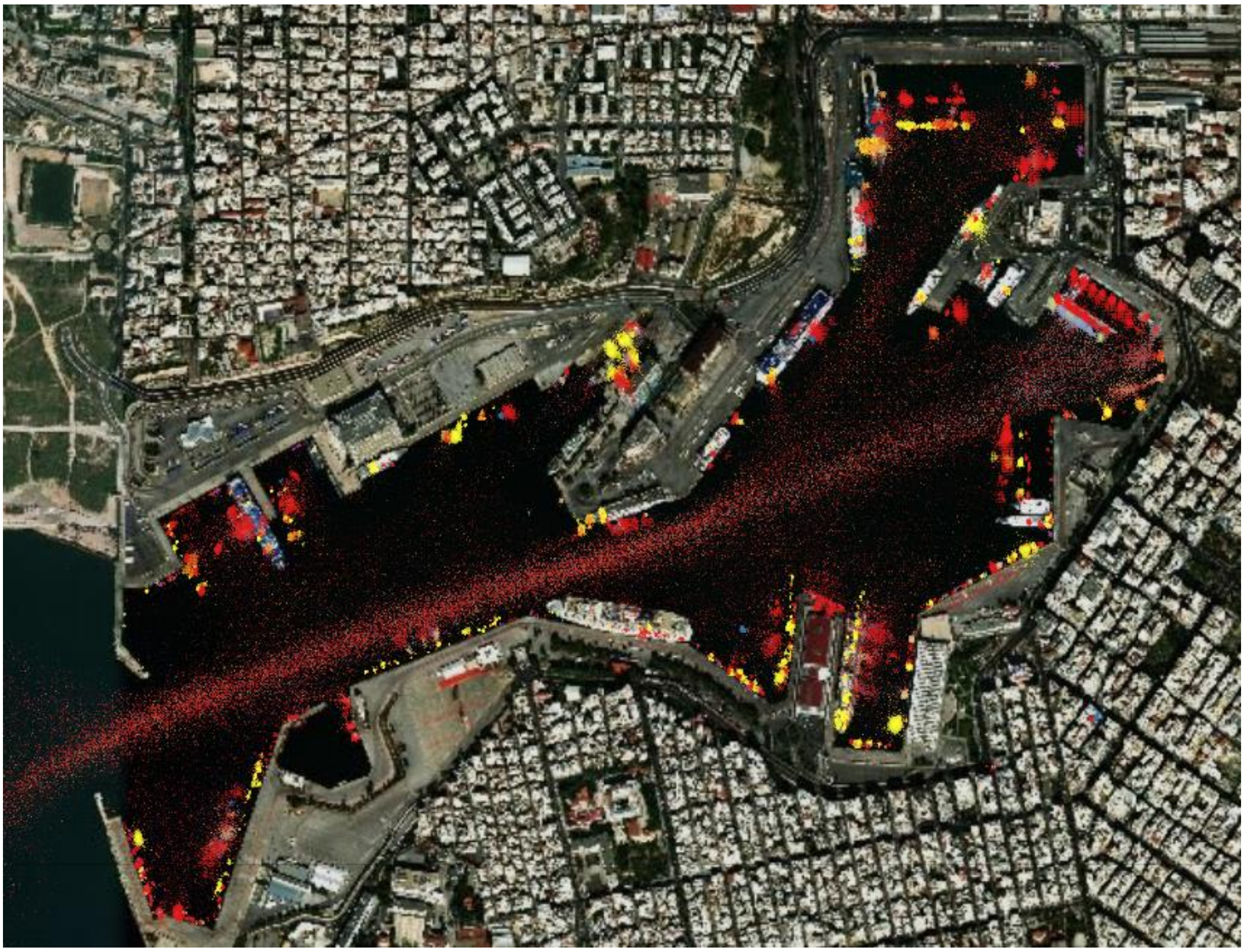


Figure 26. Navigational statuses in PPA passenger terminal.[13]

Location data that is reported automatically can also contain noise; an example is presented in Figure 27. The vessel that is moored, appears, as it is jumping multiple hundred meters in different directions.

[13] Red dots: AIS navigational status 0 (under way using engine); blue dots - navigational status 1 (at anchor); yellow dots - navigational status 5 (moored).

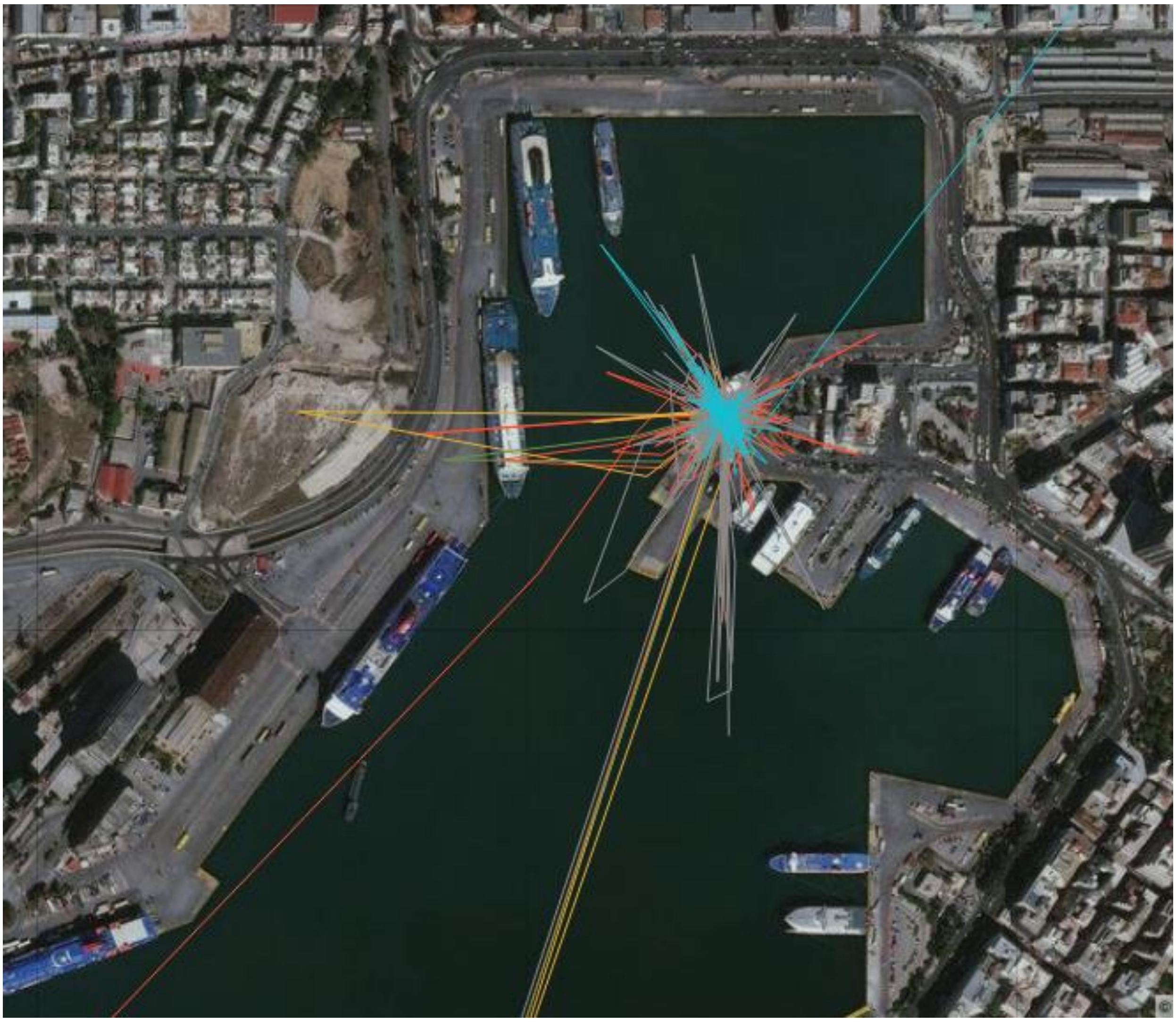


Figure 27. Errors in AIS data.[14]

Different methods for determining navigational status were developed, each with its pros and cons, for a subset of navigational statuses, that are relevant for the analysis. Whether the vessel (e.g. cargo vessel, tanker, passenger ship) is moving, is anchored, or moored at the terminal, is vital information. Example of originally reported navigational statuses in AIS messages is presented in Figure 28, while the improved results are presented in Figure 29.

[14] Quick and large changes of location of moored vessels, even over ground. Each line colour presents the vessel's journey (from arrival to departure from the passenger terminal in PPA).

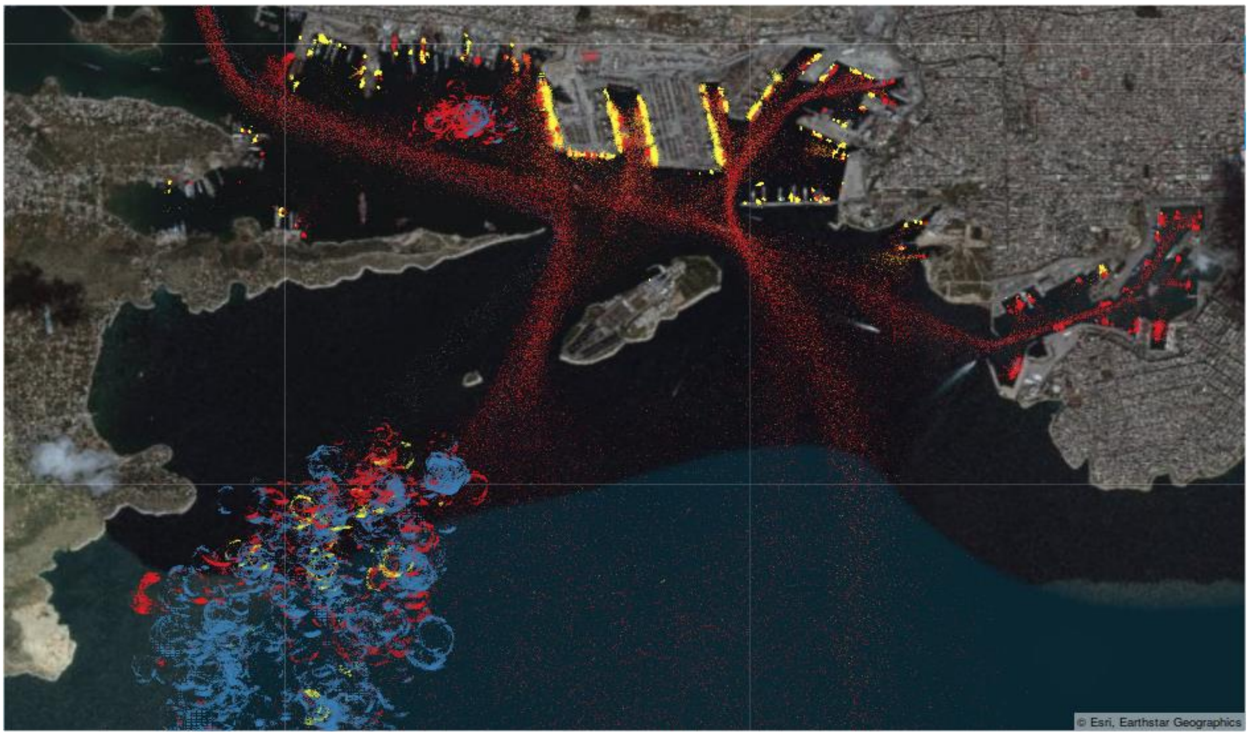


Figure 28. Reported AIS navigational statuses with errors, as reported in original data.

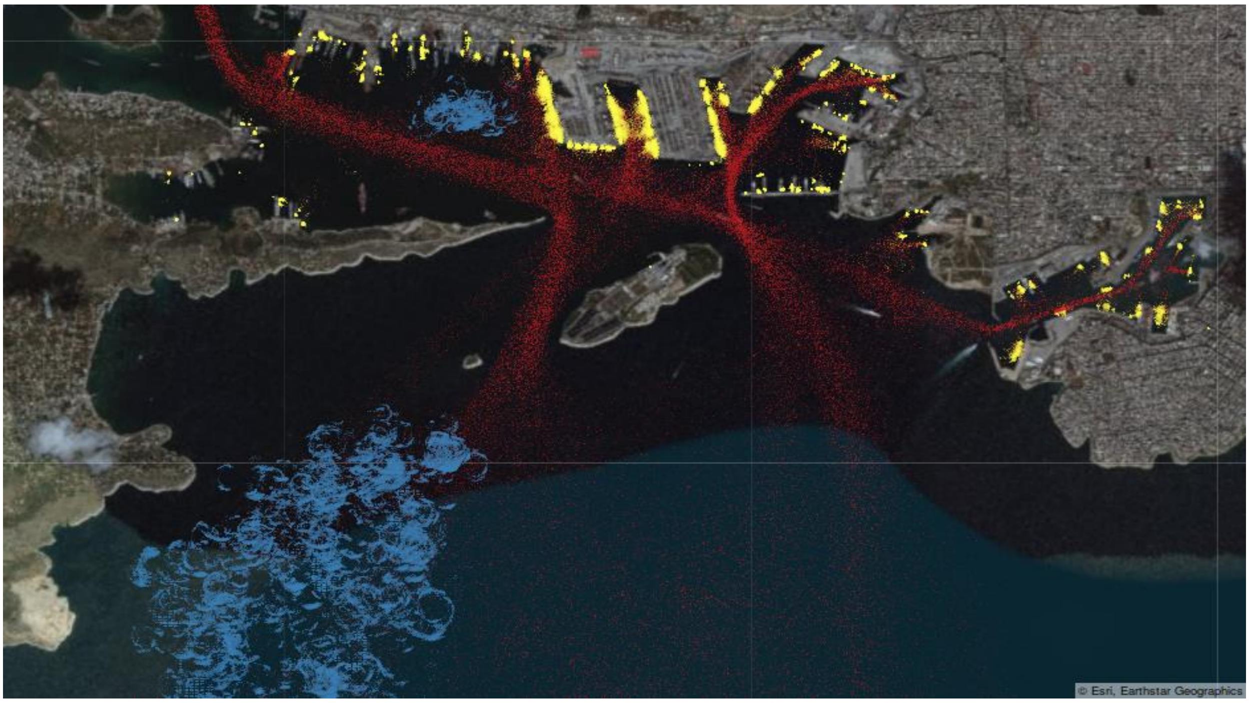


Figure 29. Reported AIS navigational statuses corrected with our proposed method.

The first method that determines navigational status is based on the vessel's speed and its location. If speed is under the manually determined threshold for moving vessels (e.g. 0.5 kn) and it is located in the area of anchoring or terminals (Figure 30), its navigational status is anchored or moored, respectively. A good thing about this method is that navigational status can be determined relatively quickly. The downside is that polygons of specific areas (terminal and anchorage areas) must be drawn manually. The problem is also that the vessel can anchor at the border of the drawn polygon and thus affect the accuracy of such an approach, as it will present cases with a false change of its navigational status, changes such as the vessel drifting in or out of the area, due to winds or water flows.

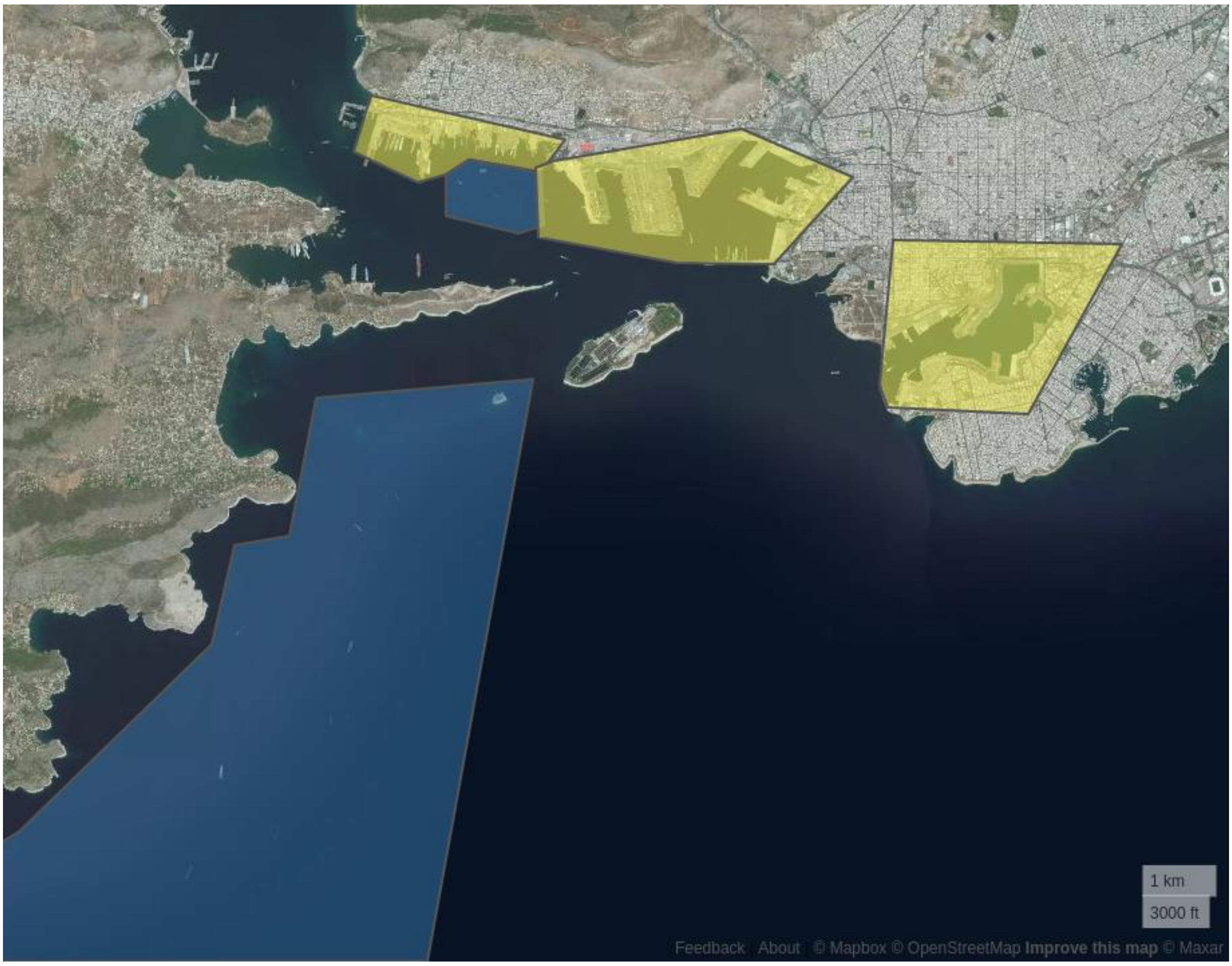

1 km
3000 ft
Feedback About © Mapbox © OpenStreetMap Improve this map © Maxar

Figure 30. Manually drawn areas of anchoring (blue) and mooring (yellow) over wider PPA ROI.

As the second approach, vessels speed and rotation were analysed. If a vessel is not moving but only rotating, then it is anchored. If the vessel is not moving and not rotating, it is moored at one of the terminals. The good thing about this method is that polygons do not have to be determined and thus it is easy to apply it globally. Vessel rotation is obtained from the heading attribute in AIS messages, which must be encoded with sinus and cosine to prevent overflow from 360° to 0°. Three days of a particular vessel's speed (SOG_SMOOTH - speed over ground smoothed with moving average) and heading (HEADING_SIN and HEADING_COS) are plot (Figure 31). For instance, a vessel is not moving from 11 AM September 30 until 4 PM October 1, as its speed is 0 and heading is not changing much. This means the vessel is moored. The speed is then raised until 7 PM, as the vessel is sailing and later the speed falls again under the threshold of moving vessels with changing heading information until 5 AM on October 3. This means that the vessel was anchored during this time. To automatically detect navigation status, a moving threshold to 0.5 kn was set. If the vessel's speed is under this threshold, then the vessel is not sailing. Then standard deviations of *heading_sin* and *heading_cos* over the period when the vessel is not sailing are calculated. If the value of any of them is more than the selected threshold (e.g. 0.5), then the vessel is anchored, else it is moored. The threshold of 0.5 was found to be a good boundary for distinguishing anchored and moored vessels.

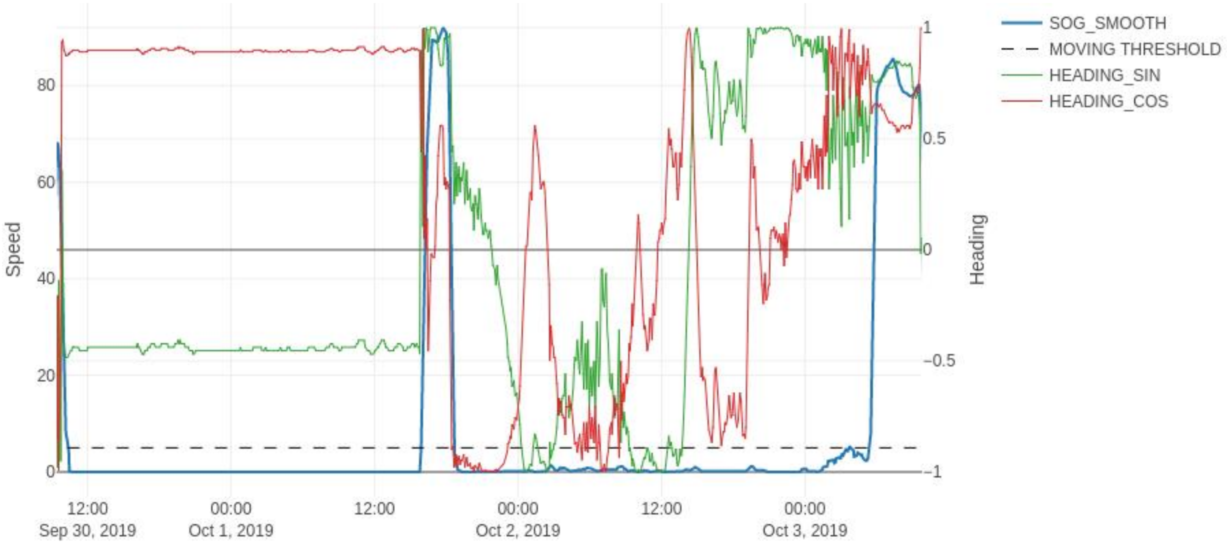


Figure 31. Variations in vessels' speed and heading when the vessel is moored, sailing and anchored.[15]

With such clean and valid AIS data, vast amounts of AIS data can be easily grouped into separate journeys or voyages. In most of the presented problems and results, one voyage is defined, as a vessel movement from the entrance into the region of interest (port area) and its exit. Different statistical values are calculated for each voyage. Vessel's waiting/anchored time, sailing time and speed, turnaround time present just some of them, which can now be computed much more accurately.

3.1.2. Use of AIS data for port analytics

Some of the key indicators of the port's performance are waiting and turnaround times, which constitute the so-called Port Congestion Indicators. A method that calculates such statistics automatically from AIS data was developed and as such, providing real-time data-driven metrics for measuring port efficiency, as well as certain environmental metrics. This means that no input from the ports is needed, thus the proposed methods are widely applicable to any port, with available AIS coverage. Ports can be compared with other ports or with the past performance of the same port. Weekly medians of waiting (anchored vessels) and cargo processing (moored) times are presented in Figure 32. The time needed to process cargo is more consistent, as it depends mostly on cargo capacities the port can process in a given time. On the other side, waiting time mostly depends on how many vessels are also waiting in front of the port.

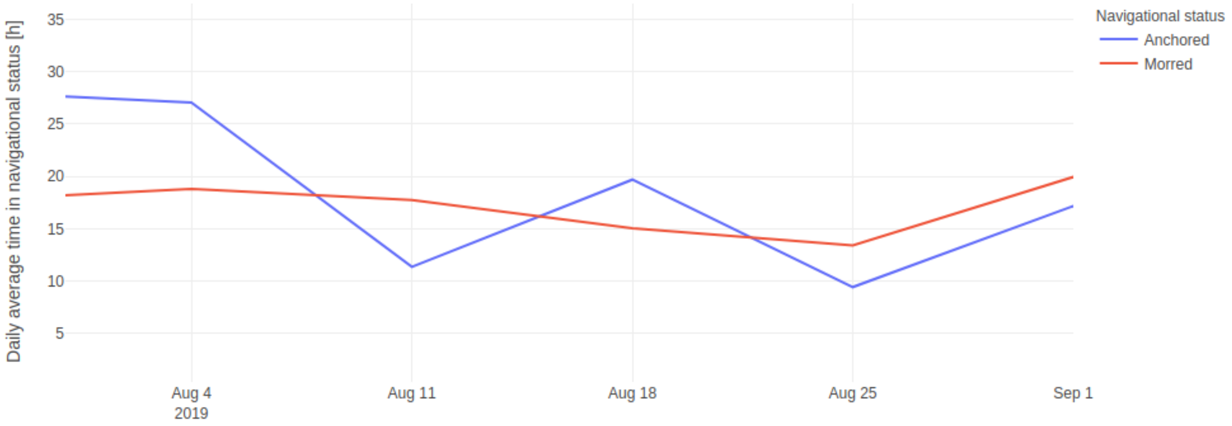


Figure 32. Weekly medians of PPA vessels' waiting and cargo processing times, August 2019.

As an example, Figure 33, shows movements for a high-speed passenger vessel. The vessel sails a regular passenger line, scheduled at the PPA passenger terminal at predefined hours. After removing outliers, the

[15] Provided in sine and cosine encoding

average time of departure is at 4:31 ±1h 47 min (std) and average time of arrival is 14:24 ±1 h 18 min (std). For departures between 3:00 and 7:00 and arrivals between 11:00 and 17:00, standard deviation gets significantly smaller. Then the average time of departure is 4:21 ±2 min (std) and the average time of arrival is 14:30 ±31 min (std).

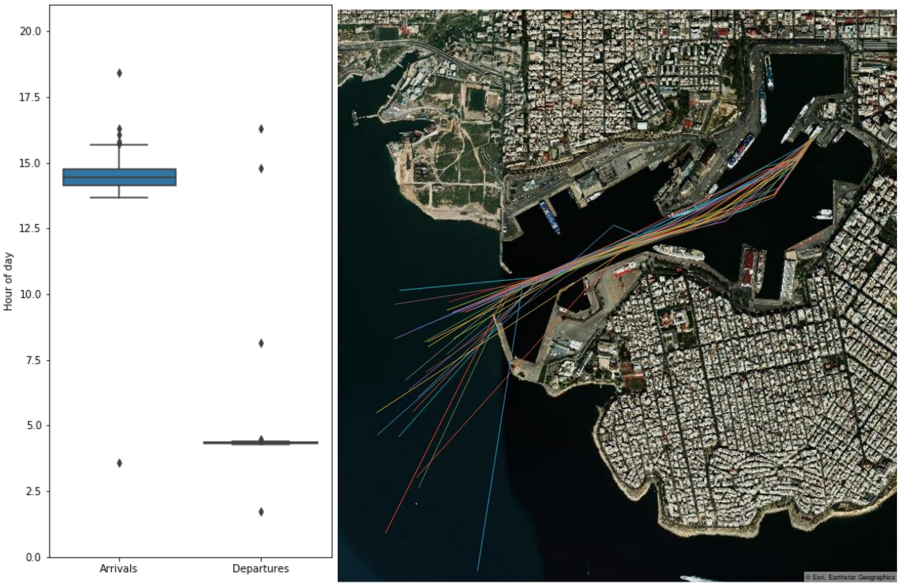


Figure 33. High-speed ship arrivals and departures from PPA passenger terminal in June 2019.[16]

Detailed extract of this vessel arrivals and departures are presented in Table 5. With a closer look at the data, it is noticed that the vessel stayed moored at the terminal for more than 36 hours, twice. This was due to extreme weather conditions[17] and a general strike[18].

Table 5. subsection of high-speed craft arrivals and departures from the passenger terminal in PPA.

Arrival time	Departure time	Time moving	Time moored	Average speed	Missed departure
…	…	…	…	…	…
2019-09-12 14:27:26	2019-09-13 4:25:06	0:11:27	0 days 13:46:13	7.70	
2019-09-13 15:41:34	2019-09-15 4:20:56	0:12:52	1 days 12:26:30	7.39	Yes
2019-09-15 16:17:09	2019-09-16 4:22:46	0:18:15	0 days 11:47:22	6.87	
2019-09-16 14:14:27	2019-09-17 4:22:04	0:14:56	0 days 13:52:41	7.84	
…	…	…	…	…	…

[16] Boxplots of arrival and departure times in the left figure, paths on the right. MMSI=239658000. Vessel missed departures two times.

[17] https://www.ferryhopper.com/en/blog/ferry-news/strong-winds-greek-ferries-latest

[18] https://news.gtp.gr/2019/09/17/ferry-strike-greece-announced-september-24/

Arrival time	Departure time	Time moving	Time moored	Average speed	Missed departure
2019-09-21 14:33:38	2019-09-22 4:21:26	0:16:41	0 days 13:31:07	6.64	
2019-09-22 14:34:46	2019-09-23 4:22:22	0:13:15	0 days 13:34:21	6.25	
2019-09-23 13:57:10	2019-09-25 4:21:26	0:14:35	1 days 14:09:41	7.44	Yes
2019-09-25 14:09:55	2019-09-26 4:22:08	0:15:16	0 days 13:56:57	7.19	
...	...	...	...	...	

Using clustering methods, anchoring and terminal areas from vessels' locations and navigational statuses can be obtained. Clusters were formed based on navigational status and distances between locations of messages. Drawing convex hulls around clusters yields areas that are used for anchoring or terminals in ROI (Region of Interest). Areas are obtained automatically without other data sources or input from the ports. The method can be easily used globally without large modifications.

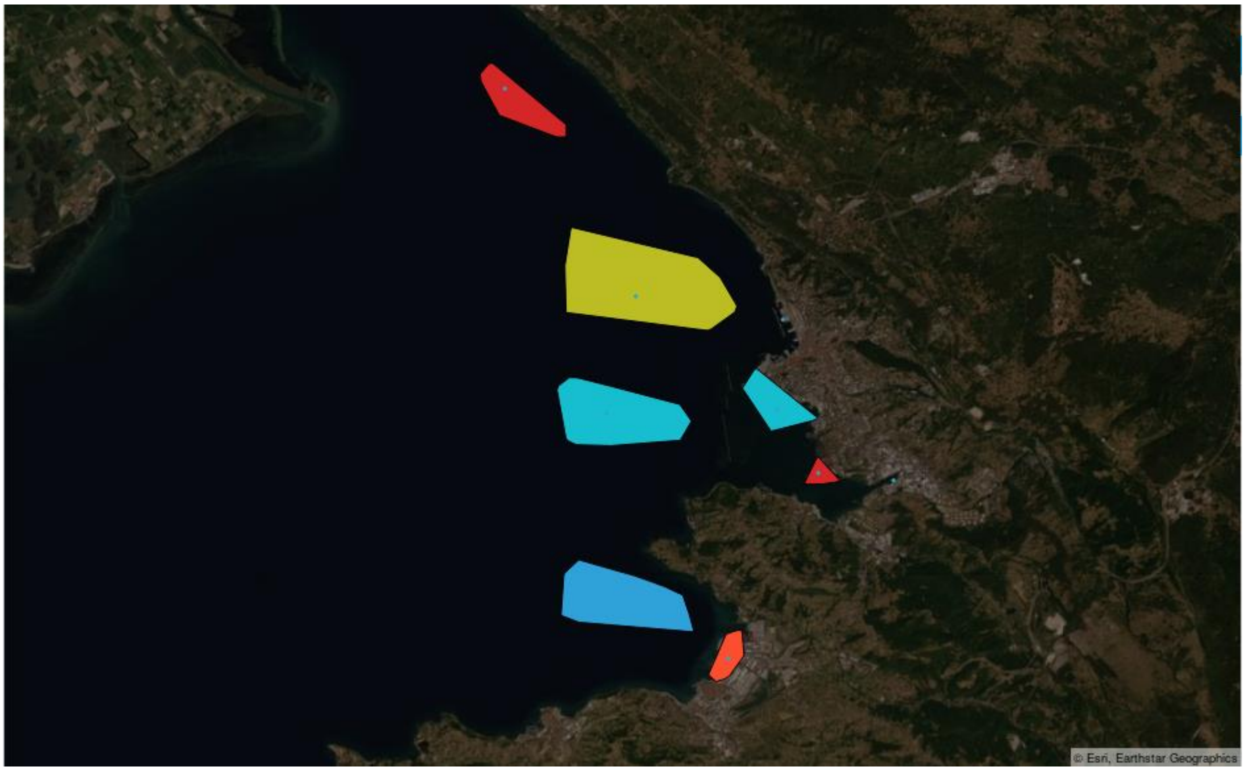


Figure 34. Convex hulls around anchoring and terminal areas of Koper, Trieste and Monfalcone.

Having identified areas of ports, one can analyse vessels' movements between them. It is possible to see regular connections between different ports and identify which anchoring areas belong to which port. Using methods from the field of graph theory could provide a lot of interesting insights in shipping lines.

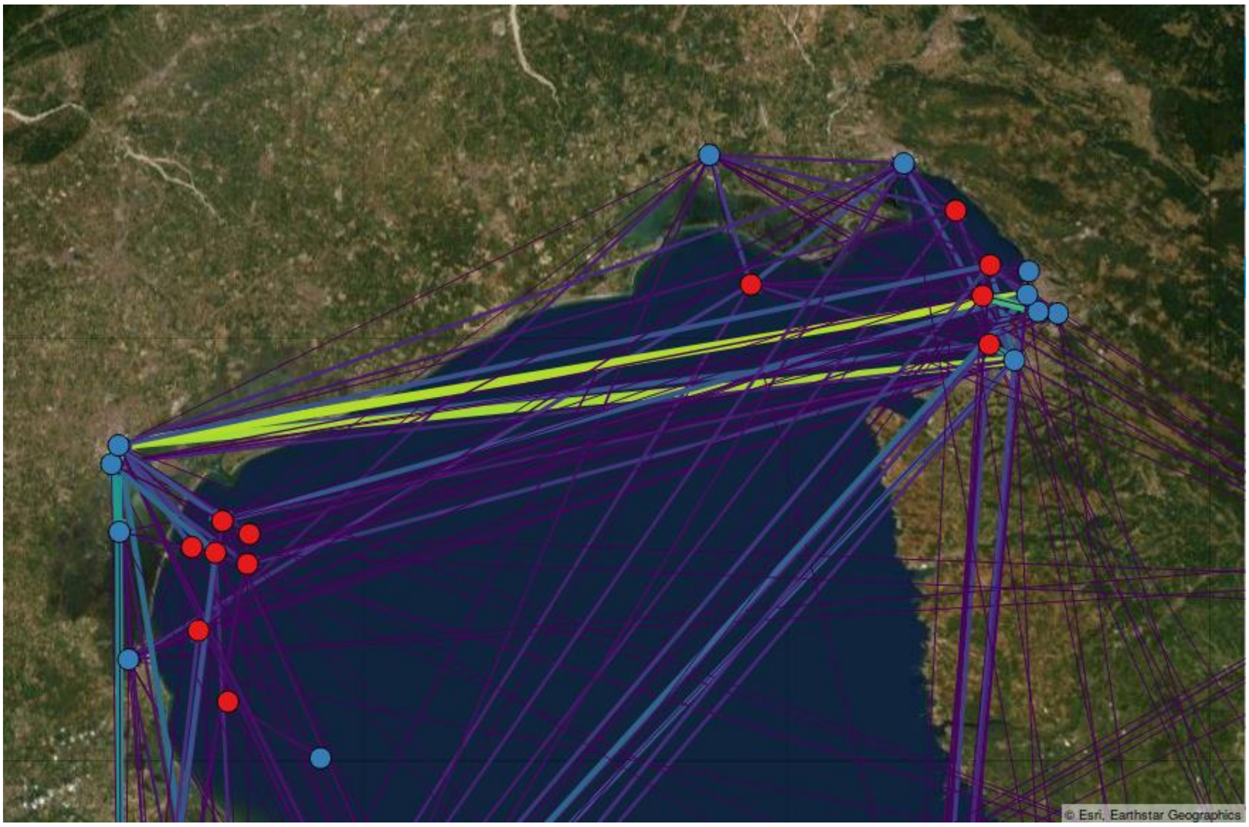


Figure 35. Vessels movements between ports in Slovenia, Italy and Croatia.[19]

To demonstrate the applicability of such analysis, the effect of the SARS-CoV-2 virus on the maritime traffic (Figure 36) is also analysed, which shows a clear decrease in traffic in the PPA passenger terminal. Unfortunately, no historical AIS data are available, so, no comparison of the traffic in March and April with previous years can be done, but still, a sharp decline is noticed, in comparison with previous months and it is not due to seasonality. The number is slowly decreasing, with more stringent measures being taken to slow down the spread of the virus.

[19] Centroids of anchoring areas are marked with red dots, terminals with blue. Lines represent movements between them. Wider line, more movements.

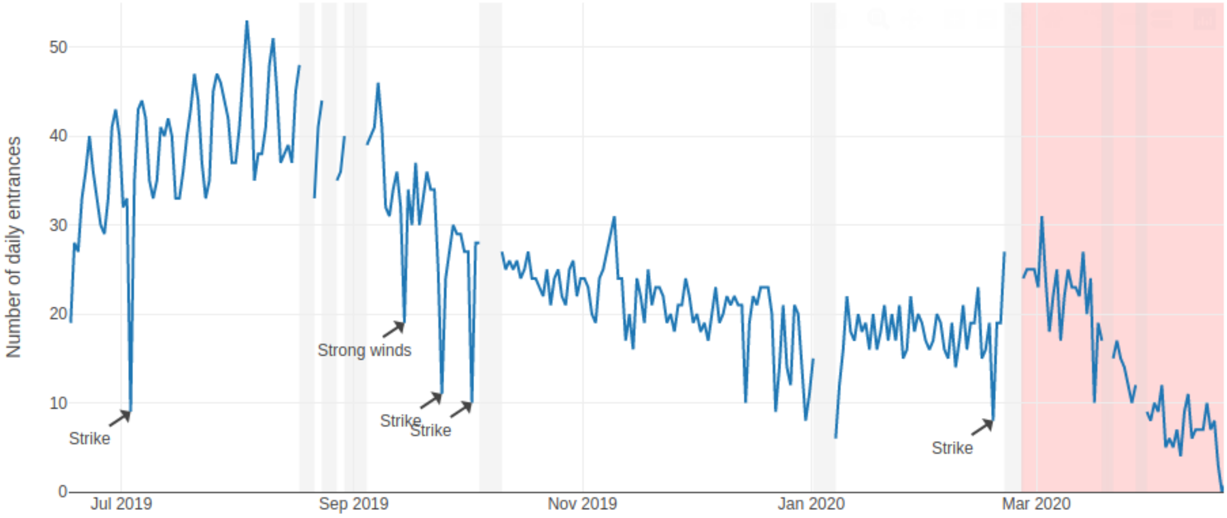


Figure 36. The number of daily entrances in the PPA passenger terminal.[20]

3.1.3. Data analytics of the port area (Event Detection)

The AIS maritime communication standard, allows static and dynamic operational information to be obtained from the ships. As has been seen throughout the document, this information can be used for different purposes, that may be of interest, to have a better insight into port operations. Besides, during the stop of the ship in the port, a series of events occur that are of interest to the port. Some of these events are the start of pilotage, towing, refuelling, as well as speeding in the port area. The early prediction of these events will allow logistics operations to be more efficient, as well as more secure, since, if it is possible to predict behaviour that will result in speeding or dangerous behaviour, early intervention could avoid this situation. If this behaviour is analysed upon prior detection of these events, such as speeding, a series of patterns can be detected, such as a continuous period with increasing speed or noticeable variation in acceleration, that will allow intervention before the occurrence of over speeding and therefore avoid a risky situation.

Currently, these events are detected manually, either when a ship approaches the refuelling area and remains on-site for more than a certain time, or even the manual operation of an operator are some of the triggers that launch and captures these events, generally operated through some logical rules like the detection of a ship within a specific geographical area or even a certain time without a change of position. This study aims to find out if it is possible to integrate the AIS information for the detection of such events.

The main events are the following:

- Speeding
- Start of pilotage
- Start of towing

Before developing any prediction algorithm, these events will be analysed for some useful information or pattern that allows an investigation, in greater depth, of the characteristics behind them. Data of the events detected comes from applying to AISHub data the logical rules existing within Posidonia Operations tool, that allows the suite to detect these events. Some of the logical rules are as described above, some time without change in position, sudden change in direction, etc. Posidonia Operations[21] is a complete system of

[20] Grey areas are marking days with missing data, due to data collection problems. Area in red is marking the period since the first COVID-19 case in Greece. There are visible seasonal trends (more entrances during summer) and the virus influence as the number of active vessels significantly decreased since the outbreak in Europe.

[21] https://www.prodevelop.es/en/ports/posidonia/posidonia-operations

real-time monitoring of ship activity that detects multiple events of the life cycle of a ship in port and allows to automate actions and assist a port operator in controlling the visit of the ship to the port. This port operations management system allows a port to optimize maritime activities related to the flow of ships within the port's service area, the integration of all the actors involved and all the relevant information systems. A history of more than two years of AISHub data of the Algeciras Port area has been used and more than 35.000 events types described above where detected.

By knowing the geographical position in which these events have been detected, it can be detected whether they are reported in a very localized area or if the geographical position is not very indicative.

The start of pilotage does not seem to occur in a completely defined area (Figure 37). Remember, that the start of pilotage takes place when a pilot from the port leads the ship to the estimated dock, either because of the size of the ship or because it is the first visit in that port.

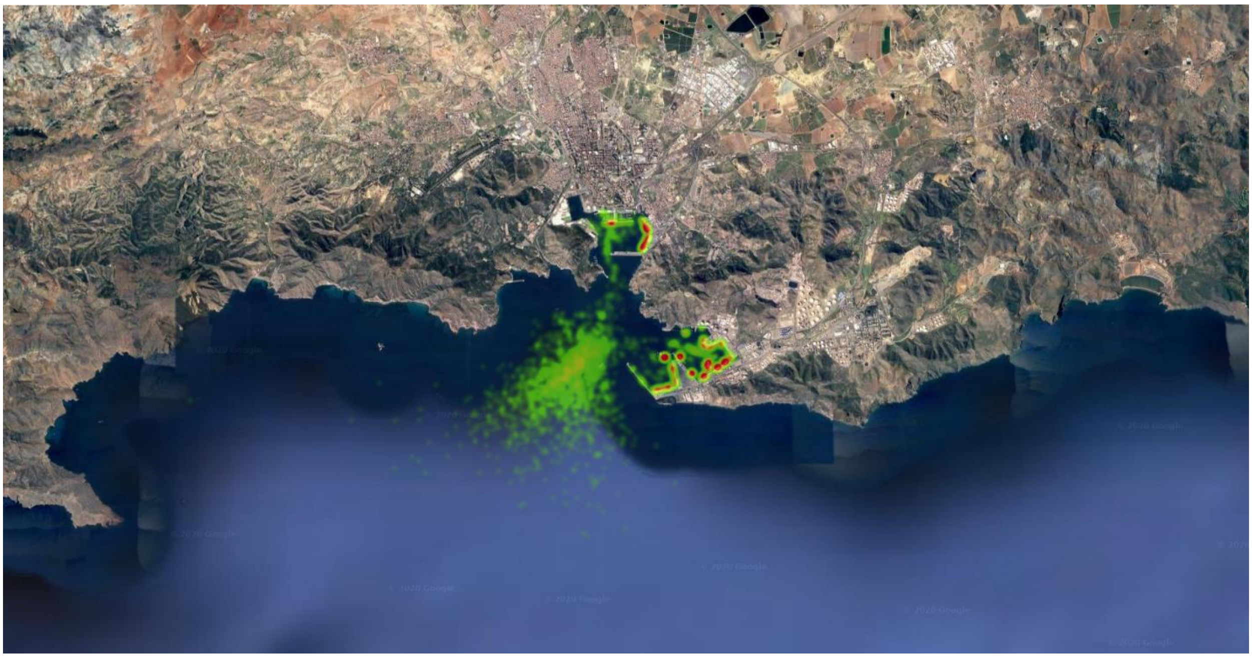


Figure 37. Pilotage starts areas.[22]

Moreover, the towing seems to be more geographically concentrated (Figure 38), in comparison with pilotage. This type of event takes place when a ship requires a smaller one to tow it to the desired area, when it is logical to rule out a port area that may be designed for smaller ships, than those that require this type of service.

[22] Satellite Images obtained from Google Map Services with the help of Airbus, European Space Imaging.

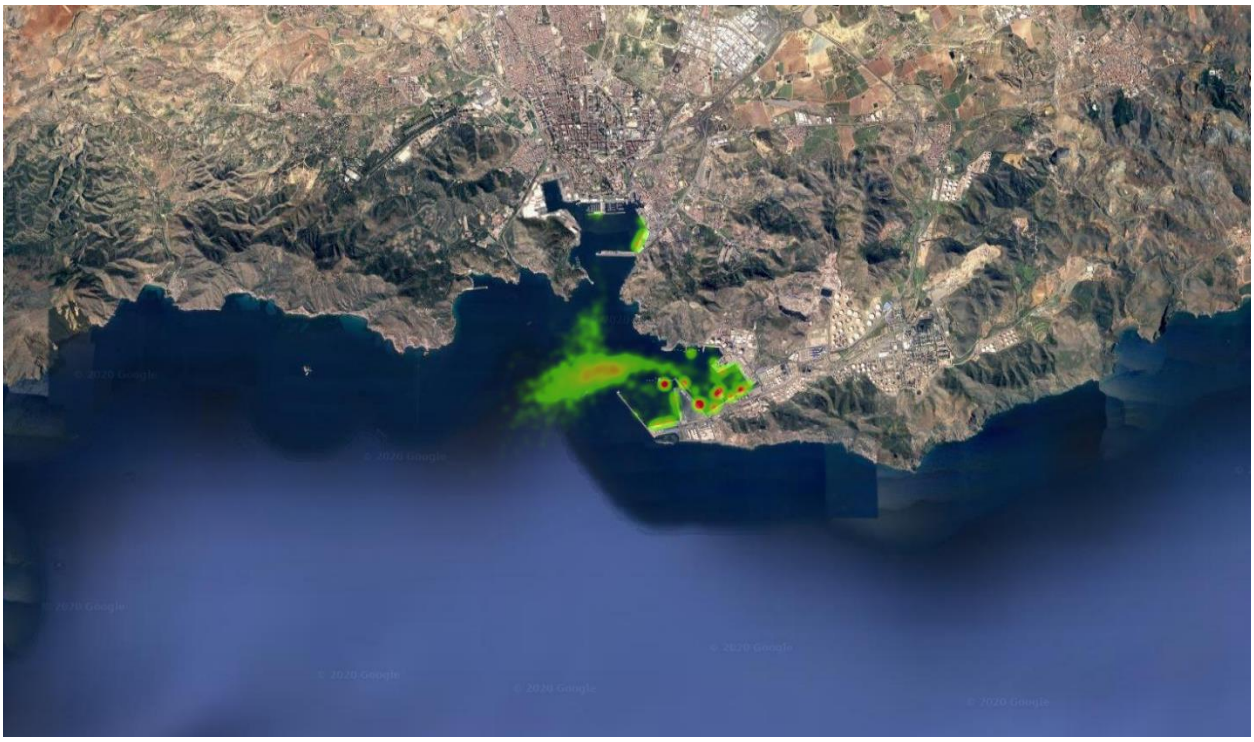


Figure 38. Towing starts areas.[23]

Unlike previous events in which there was no specific zone, two zones can be distinguished. These are the areas where the speed of ships is restricted because they are in areas of influence of the port (Figure 39).

[23] Satellite Images obtained from Google Map Services with the help of Airbus, European Space Imaging.

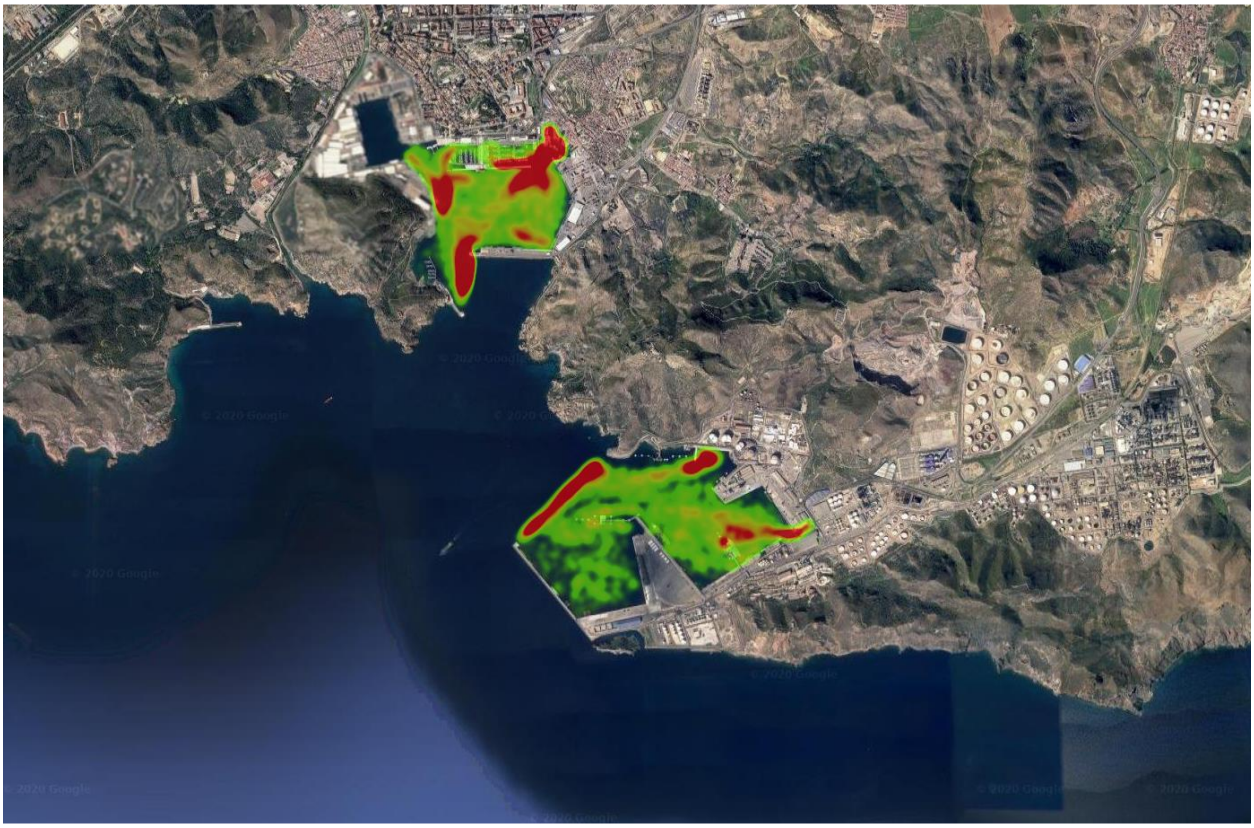


Figure 39. Speed exceed areas.[24]

With zones being so defined, it will be possible to allow these variables to have a quite important weight in the prediction algorithm used.

Next, the frequency of these types of events in ports activity (Figure 40) can be determined. Speeding represents most records compared to the other two, explained to some extent by the fact that pilotage and towage start are manoeuvres that apply to a specific subcategory of vessels than meet certain characteristics of size and capacity. An increase in speeding can be observed in summer, caused by an increase in port activity.

[24] Satellite Images obtained from Google Map Services with the help of Airbus, European Space Imaging.

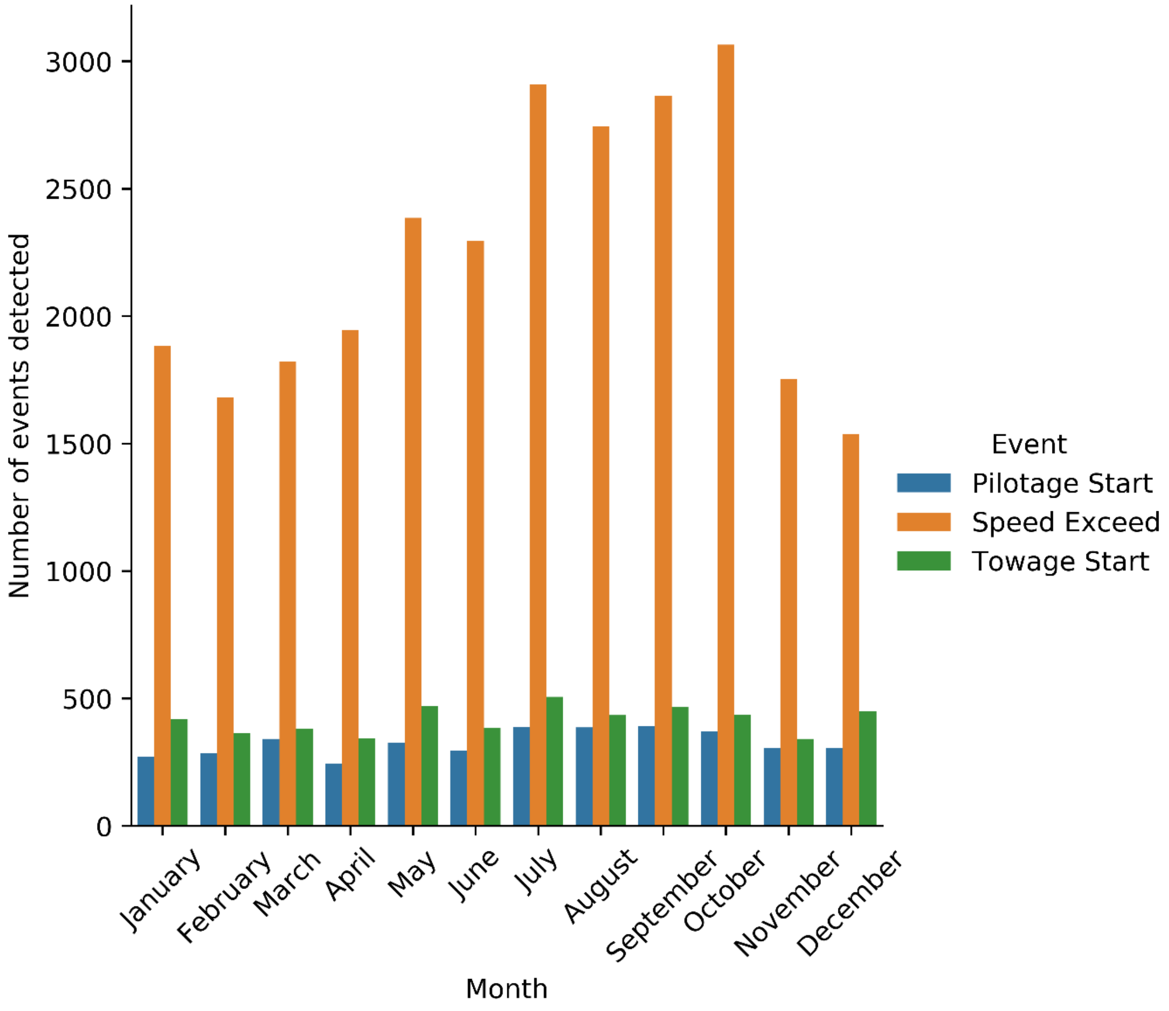


Figure 40. The number of event types detected.

Concerning the argument shown above, the size characteristics of each of the ships that launched these events can be analysed. The length and the beam are represented, obtained indirectly through the static information parameters present in the AIS messages, through scatter plots, in which the distribution of both variables in the axes will also be shown.

As it can be seen in Figure 41, it is confirmed that the different events present ships with very evident characteristics. While for the first two (i.e. pilotage and towage), ships usually have magnitudes corresponding possibly to cargo ships, with large dimensions since there are lengths between one hundred and three hundred meters.

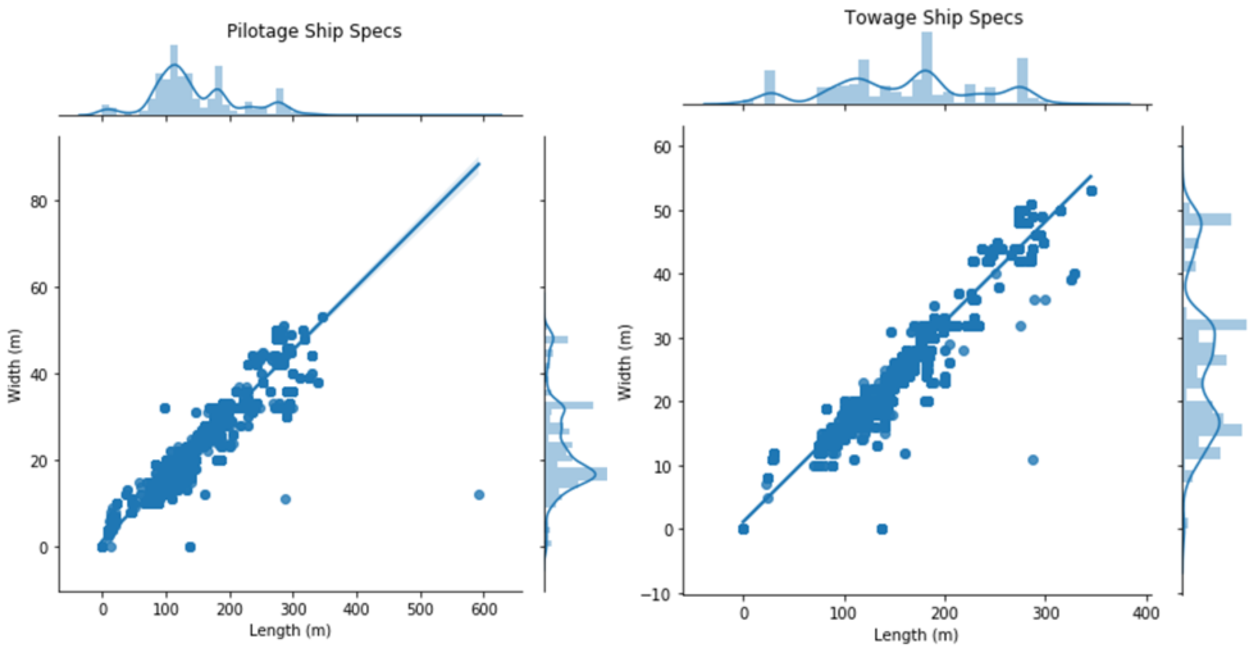


Figure 41. Scatter plots of ships characteristics.

Speeding (Figure 42, left) is mainly carried out by small-scale boats, boats that represent a greater number in proportion to the previous ones, since they include from fishing boats, ferries, and even recreational boats.

Finally, in the right part of Figure 42, the three events are represented together, where the speeding event can be differentiated from the other events, in terms of the vessel size.

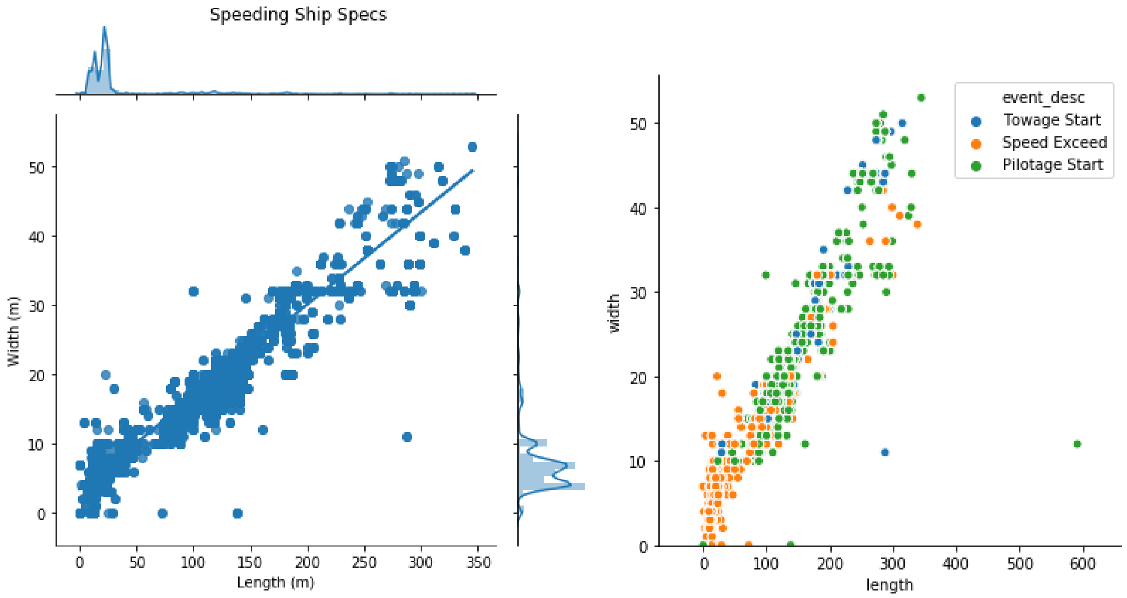


Figure 42. Scatter plots of ships characteristics.

Once some of the properties of the boats that have launched the events have been analysed, it is time to develop techniques that allow, given an event, to classify it in the corresponding group or, given the sequence of AIS messages, to predict the event which occurs. The events captured by the tool described above have been taken as ground truth, by dividing the data into a train (75%) and test sets (25%).

Random Forest and K-Nearest Neighbours algorithms for event classification were used. Random Forest is a decision tree-based algorithm and is part of the ensembled methods, offers good scalability and is easy to implement. Secondly, KNN (k-nearest neighbours algorithm) is one of the most famous algorithms for classification in which training is discriminatory, that is, it is trained by memorizing the data and its procedure is to find the closest K neighbours and determining their label, as the most frequent among their neighbours.

Feature selection and generation represent an important part and the attributes initially considered have been length, beam, speed, latitude, longitude, MMSI and the type of a vessel. These attributes are collected from the AIS message as welll. A series of feature selection and extraction techniques have been applied to these attributes to obtain which ones are the most important in the classification procedure, to improve classification performance.

Regarding the feature selection, classification with Random Forest allows obtaining the Gini values that have allowed separating the branches of the trees and therefore giving a magnitude of their importance. As presented in Figure 43, the most important attributes are mainly speed and length, while MMSI and the type of boat are discarded due to their low importance.

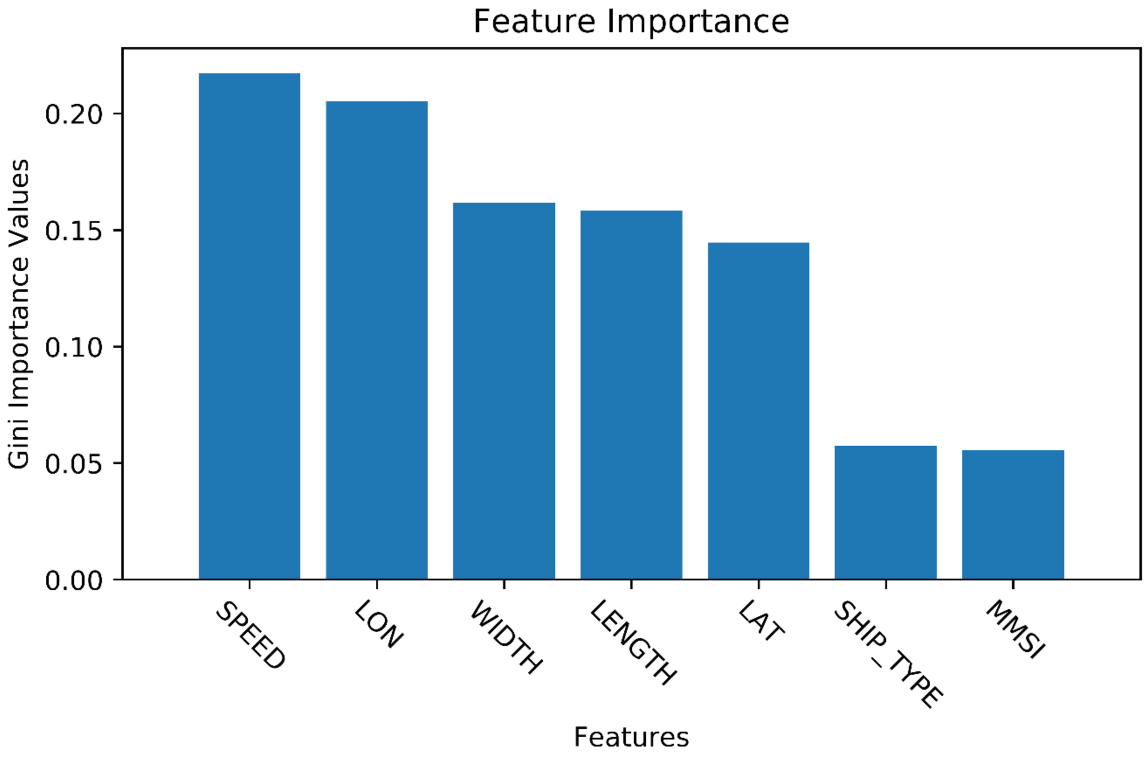


Figure 43. Features importance.

Secondly, concerning the extraction of characteristics, the principal component analysis has been used to obtain the directions that express the greatest variability in the data (Figure 44). The three main directions were chosen because they explain 90% of the variability. These methods allow to reduce the number of dimensions of the data, so the computational efficiency of its processing can be increased. Linear discriminant analysis was also used, in which instead of obtaining the directions of maximum variability, those that optimize class separability are sought after.

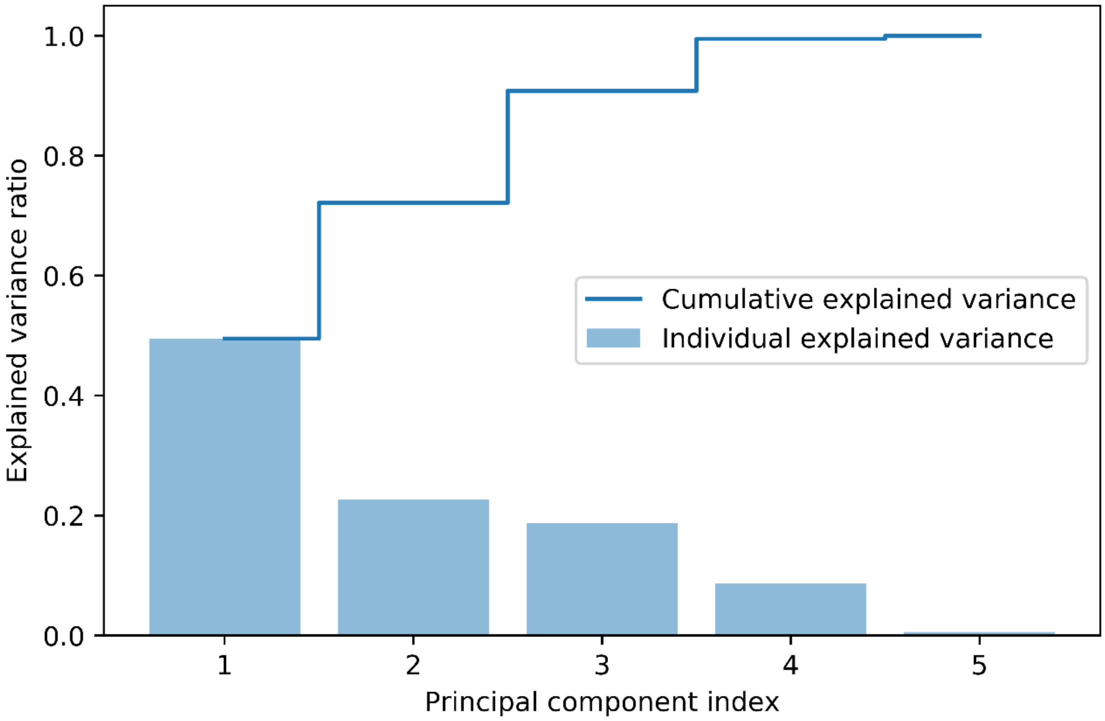


Figure 44. The variance of directions after PCA.

Finally, for the comparison between the different techniques and algorithms, accuracy and F1-score metrics have been used. While the first refers to the correct classification of the data, the second is more appropriate to our case in which the classes do not seem to be balanced. Metrics represent an average across all the three classes of the events.

Table 6 shows how the most optimal values have been achieved with KNN without any transformation, i.e. dimensionality reduction with PCA (principal component analysis) / LDA (Linear discriminant analysis), while Random Forest allows to almost matching its potential using PCA dimensionality reduction. It was concluded that using all the generated features, generates the best results and given the low number of attributes, dimensionality reduction does not play a major role, given the computationally efficient choice of predictive models (i.e. Random Forest and KNN).

Table 6. Classification accuracy and F1 score of our proposed event detection methods.

	Random Forest		**KNN**	
	Accuracy	**F1 - Score**	**Accuracy**	**F1 - Score**
Base	0.93	0.89	0.92	0.92
PCA	0.91	0.91	0.88	0.72
LDA	0.85	0.84	0.85	0.85

3.1.4. Short term ETA prediction for the ports

Accurate estimations of estimated times of ship arrivals are particularly important, as it has an influence on the whole logistic chain, not only on the operations in the port. ETA information is provided in AIS messages but is often unreliable. Because of the bad AIS receiver coverage, a methodology was developed, for a short-term ETA prediction for the GPMB area, to have an estimation of the time the vessel arrived at the Bassens terminals. The ETA predictive algorithm could also be used as a vessel arrival notifying system for any port, as the methodology is general and applicable when only AIS data is available. AIS data from

AISHub is captured and vessel calls data from VIGIEsip. AIS receiver is placed at the entry in the Gironde estuary and captures every vessel arriving or departing from the estuary, but has problems receiving messages that are closer to the Bassens terminal. All AIS messages sent from the Gironde estuary were enriched with the actual time of arrival from vessel calls data so that the proposed approach could be evaluated. This produced a dataset with features, such as vessel's location, speed, direction, size and the actual time of arrival. Vessels location was used to calculate its distance to the terminal. In addition to current speed, its moving average was also calculated. With this kind of generated features, linear regression and CatBoost[25] predictive model was fitted. Modified cross-validation was used as an evaluation method and different error metrics were calculated (e.g. MAE, MAPE, RMSE) and aggregated based on vessels' distance to the terminal. MAE provides easily directly interpretable absolute error value, MAPE provides relative errors, relative to the ground truth data and a combination with RMSE metrics provides insights into error distribution, as large errors have a higher impact in this metric.

AIS messages locations heatmap is presented in Figure 45. Note that ETA predictions were done, only for vessels that are already in the Gironde estuary, so messages that are outside of the estuary were not used in our case.

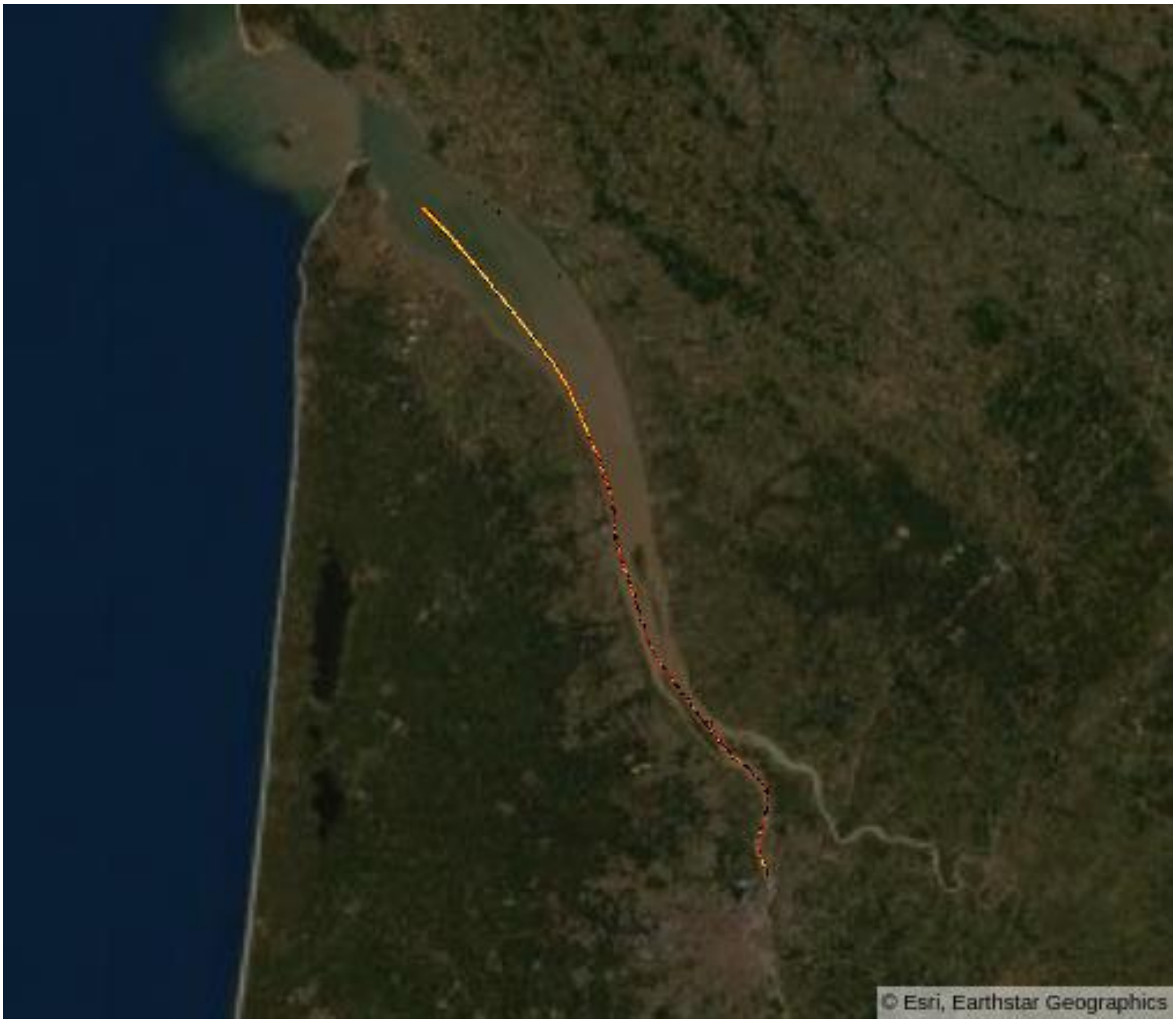


Figure 45. Heatmap of received AIS messages over the Gironde estuary.

Histogram of AIS messages based on their distance to Bassens terminal is presented in Figure 46. As already explained, the AIS receiver connected to AISHub does not have good coverage over the Gironde estuary.

[25] http://www.aishub.net/

Histogram of time needed to Bassens terminal is presented in Figure 47. Average time to Bassens for vessels that are more than 60 km away is 3 hours and 18 minutes.

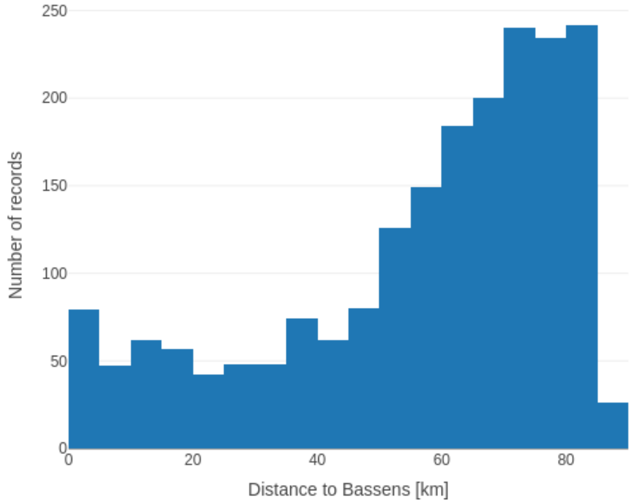


Figure 46. Histogram of distances of AIS messages to the Bassens terminal.

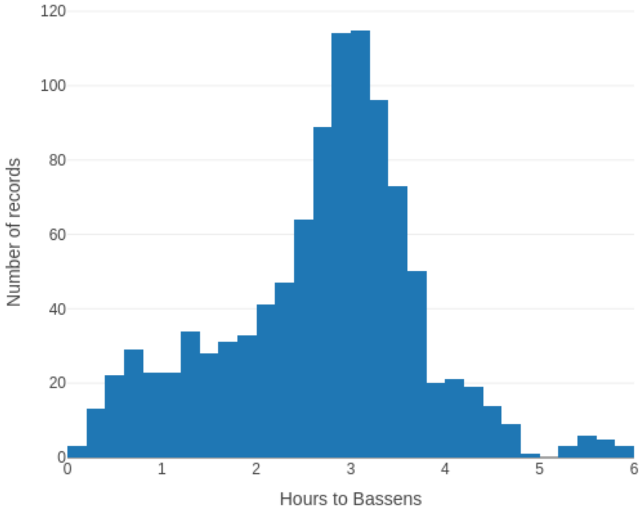


Figure 47. Histogram of times vessels needed from message to the terminal.

Keep in mind, that the amount of AIS messages is inversely proportional to the proximity of Bassens terminal, as there is bad coverage. Distance to Bassens was transformed into 5 km bins and the sum of means and standard deviations of needed time to Bassens in each bin was used for the normal time threshold computation. Messages with time to Bassens over the threshold are outliers and they were removed (Figure 48).

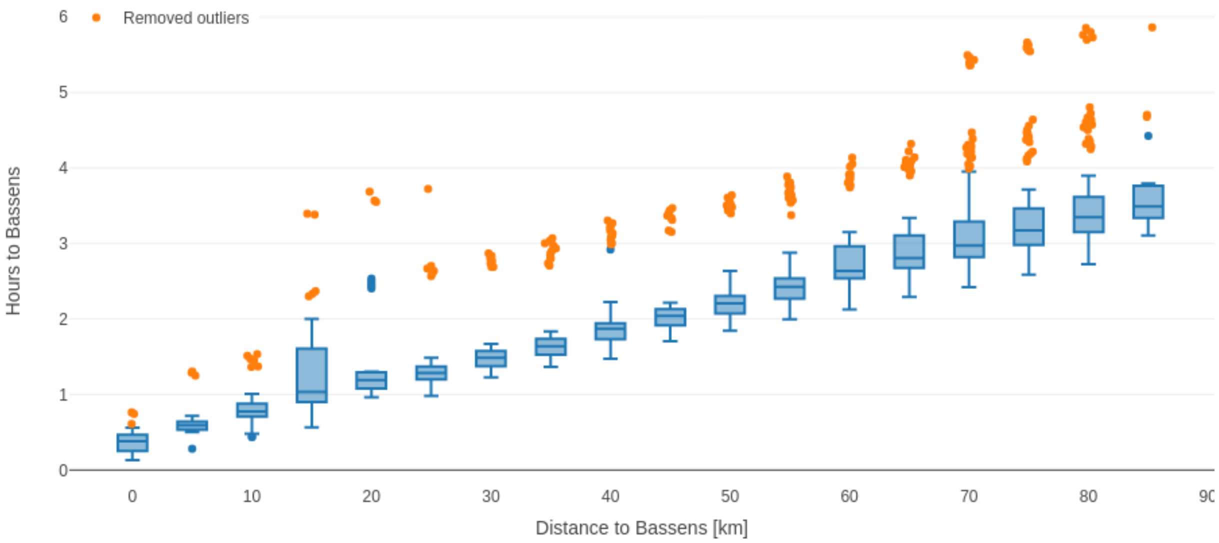


Figure 48. The time vessels need to reach the Bassens terminal after sending the AIS message[26]

Two different machine learning models were fitted and evaluated, linear regression and CatBoost. Best results overall were reached with a simple linear regression model. The reason is that the problem in our case was close to linear. Results are presented in Table 7.

Because the AIS receiver is based far from the terminal (at the estuary), there is a smaller amount of data available closer to the terminal, which results in larger errors as ships approach the terminal. For MAPE, there is also another reason. As the vessels are closer to the terminal, ground truth value gets smaller and absolute error is relatively bigger related to to the ground truth. In MAPE, 1 min absolute error has a higher impact if a vessel is only 5 min away than if it is multiple hours away from the terminal.

Table 7. Error metrics based on distance from the location of the message to the Bassens terminal.[27]

Distance to Bassens [km]	Linear regression			CatBoost		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE
[0, 10]	5.90	7.76	35.38	4.35	6.17	24.59
(10, 20]	13.44	21.56	20.8	15.56	22.00	25.36
(20, 30]	14.65	27.64	14.63	18.70	29.05	20.87
(30, 40]	7.69	10.03	8.85	7.53	8.60	8.23
(40, 50]	8.91	13.55	7.70	9.73	14.07	8.16
(50, 60]	10.34	13.25	7.61	10.48	13.15	7.57
(60, 70]	13.16	16.04	7.89	13.9	16.33	8.42
(70, 80]	14.28	17.42	7.65	14.42	17.37	7.76
(80, 90]	14.32	17.53	7.14	14.31	17.29	7.11
All	**12.48**	**16.93**	**10.41**	**13.01**	**17.05**	**10.62**

[26] Data is aggregated in 5 km bins. Orange dots present the removed outliers.

[27] Error metrics of linear regression and CatBoost models aggregated into bins based on distance from location of message to the Bassens terminal. Error is calculated in minutes.

4. Use of satellite imagery

Observing Earth from space presents a new dimension of information that offers an unprecedented global view for various domains and industries. Earth observation capabilities were till recently, mostly in the domain of the governments that could afford to put satellites into space. Technological advancements have made it possible to put more satellites into space, at a much lower cost and at the same time offering much higher spatial resolution and revisit times. This is nowadays called "democratization of the space" as more commercial providers are offering satellite imagery and at the same time governments or its institutions are opening its satellite constellations to the public, as open data. The most prominent example of open data satellite imagery providers is the Copernicus program from ESA (European Space Agency), offering constellations of satellites (Sentinels) for different domains with a free and open data policy[28].

The biggest factor opening this field, both for operational use cases and general use, is the exponential growth in the number of satellites in space. The so-called "SmallSat" revolution, has driven the prices of the satellites down by using commercial-off-the-shelf components in a much smaller form factor. Compared to traditional satellites that were in the size of a bus, with a price tag usually in the hundreds of millions of euros, these small satellites can be as small as 10x10x10 cm (CubeSats) or even smaller. Their cost is a fraction of the money required by traditional satellites. The cost connected with launching the satellites went also down by reducing the size and weight of the satellites and by the emergence of ride-sharing capabilities, offered by commercial providers, such as SpaceX. As of the end of 2018, there were 1900 satellites in space, 1200 of them used for EO (Earth Observation), while 3000 are to be launched between 2016 and 2022.[29]

One of the main impacts of the exponential growth in the number of satellites is that they are usually part of a larger constellation, offering much higher revisit times compared to traditional satellites, which are usually deployed alone or in small constellations due to costs. Traditional satellites offer a revisit time in the range of at least a few days, or even weeks, compared to current or planned commercial constellations that are offering daily revisit times or even better[30][31][32]. This opens plenty of potential use cases that were before not practical or feasible, due to insufficient frequency or spatial resolution of the data.

The amount of imagery that is captured with such a large amount of satellites is vast and it is increasing exponentially. The data that is captured also needs to be analysed, as the focus is shifting towards obtaining useful operational insights, which can be gathered from satellite imagery, compared to obtaining raw imagery directly. These operational insights need to be found automatically, using Machine Learning techniques. These are increasingly recognized as the main value of satellite imagery.

The maritime domain is at the forefront of the utilization of satellite imagery, especially for ship and oil pollution monitoring. SAR imagery is predominately used in the maritime domain for ship detection. The most popular approaches are based on CFAR (Constant False Alarm Rate) methods[33]. With CFAR based methods, all pixels brighter than the local threshold are regarded as pixels belonging to the ship. Majority of the research work is focused on ship detection in open waters, omitting the need for reliable ship detection in port and harbour areas. In such areas, there is a presence of multiple objects onshore as well as on the sea, which are causing strong SAR backscatter centres, which can cause a lot of false alarms with CFAR

[28] https://sentinel.esa.int/documents/247904/690755/Sentinel_Data_Legal_Notice

[29] https://www.geospatialworld.net/blogs/key-trends-in-earth-observation/

[30] https://www.planet.com/products/hi-res-monitoring/

[31] https://www.planet.com/products/planet-imagery/

[32] https://www.capellaspace.com/technology/

[33] Greidanus, Harm, et al. "The SUMO ship detector algorithm for satellite radar images." *Remote Sensing* 9.3 (2017): 246.

based methods[34]. Optical satellite imagery, especially medium resolution, which is particularly underutilized in the maritime domain[35] and can provide additional contextual information about the ship and its surroundings.

Medium resolution satellite imagery obtained from ESA Sentinel-2 (10m resolution) and Planet Labs Dove (3m resolution) have been used to develop a machine learning pipeline for ship detection. This satellite imagery is available free-of-charge in case of Sentinel-2 imagery or presents the cheapest commercial solution available on the market, with unique daily availability of new imagery for every point on Earth in the case of Planet Dove. To automatically annotate the data, a procedure was developed that combines openly available Automatic Identification System (AIS) data with obtained satellite imagery and cloud masks. With this approach, the exact positional matching of AIS GPS data and satellite imagery is captured, needed to evaluate ship detection performance. AIS data is also used in a novel way, to prepare weakly annotated ship detection dataset out of positional information and information about the ship length. With such a novel combination of existing VHR (Very-high-resolution) datasets and weakly annotated additional data, which can be obtained easily in large quantities, state-of-the-art detection results are delivered, as well as detection performance across different lengths. To the best of our knowledge, this represents the first application of state-of-the-art deep learning-based methods for ship detection on this kind of satellite imagery in the research literature.

4.1. Predictions and analytics for PIXEL

Advancements in Earth Observation (EO) capabilities, together with advancements in the AI domain, especially with the advent of Deep Learning, has opened new ways to gather operational insights from remote sensing data. The goal of this task in the PIXEL project was to utilize all these advancements for the benefit of the ports. The focus was placed on monitoring ship traffic in and around the port. Port and bay area were analysed in terms of number, types, and sizes of the ships, with the help of AIS data. Medium resolution imagery, of which Copernicus Sentinel-2 is particularly well suited, was utilized, however, there was no research work, which utilized it in any way for ship detection. There was no research work, that would utilize state-of-the-art deep learning-based methods for ship detection using Sentinel-2 imagery, which is especially challenging in the port area, not only for medium resolution but also for VHR imagery. The use of state-of-the-art methods is mostly limited, due to the limited availability of annotated data, which is hard to obtain and needed to train the methods. This is especially the case for medium resolution imagery, for which there are almost no annotated datasets of large enough quantity for ship detection. The goal was to fill this gap, with the application of existing very-high-resolution (VHR) imagery to medium resolution imagery and by a novel approach of data fusion with AIS data.

A novel procedure of AIS and satellite imagery fusion was used to train state-of-the-art models for ship detection, which were evaluated on large scale satellite imagery in the port of Long Beach and the greater San Francisco Bay area, based on 2 years of satellite imagery data. The satellite imagery from those two regions was used due to the availability of historical AIS data from U.S. Coast Guard and the use of commercially higher satellite imagery from Planet Labs. The commercial satellite imagery was freely available for research purposes for the whole area of California and served as a comparison against ESA Sentinel-2 imagery. Unfortunately, this research program was discontinued in October 2019 and the satellite imagery is longer available[36]. The complete data for ESA Sentinel-2 is freely available through their data portals[37]. Eo-learn library[38] was also used, together with Sentinel Hub trial accounts, available for research

[34] Zhi, Li, et al. "Ship detection in harbour area in SAR images based on constructing an accurate sea-clutter model." *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 2017.

[35] http://emsa.europa.eu/news-a-press-centre/external-news/item/3025-copernicus-maritime-surveillance-product-catalogue.html

[36] https://www.planet.com/pulse/sun-setting-on-open-california/

[37] https://scihub.copernicus.eu/

[38] https://github.com/sentinel-hub/eo-learn

purposes[39]. The use of an eo-learn library greatly simplifies the retrieval and processing of the ESA Sentinel satellite imagery, including state-of-the-art approaches for cloud detection, which was also utilized in our AIS data fusion pipeline.

In the next section, some of the main results of our proposed pipeline for ship detection out of satellite imagery are listed, which was presented at the OCEANS 2019 conference[40], where further details can be found.

4.2. Results

In this section, the experimental setup and methodology are presented, as well as the results of the automated system for ship detection. Detection performance is reported, across different locations (Port of Long Beach, San Francisco Bay area), satellite constellations (Planet Labs Dove, ESA Sentinel-2) and ship lengths. First, a description of the AIS data fusion procedure, which was used to create a novel PIXSAT (PIXEL Satellite Dataset) dataset and conclude by presenting main results of the proposed pipeline, together with some qualitative results on the challenging satellite imagery from port areas.

4.2.1. PIXSAT dataset

PIXSAT dataset is a large-scale dataset for ESA Sentinel-2 and Planet Dove satellite constellations of optical imagery. Compared to existing datasets which are based mostly on VHR optical satellite imagery (presented in D4.3), this presents the first attempt to build a large-scale dataset on medium resolution optical imagery. Existing large-scale datasets such as HRSC2016[41] are developed from Google Earth imagery, which does not represent real-life conditions due to the selection of best possible imagery. Such satellite imagery, without any clouds and other image distortions due to sensors and different atmospheric conditions, is not realistically processed in an operational environment. Other large-scale datasets, such as Kaggle Airbus ship detection dataset[42] are mostly captured on the open sea, which greatly simplifies ship detection and does not represent significant gains over SAR imagery in terms of ship detection performance. PIXSAT dataset is in comparison to existing datasets, constructed from operational satellite imagery, capturing regions of Port of Oakland (San Francisco Bay area) and Port of Long Beach for the years 2016 and 2017. A procedure was also developed, to automatically combine AIS data from the ships with satellite imagery and to automatically annotate ship positions in satellite imagery, while enriching them with metadata that is available in AIS messages. Visually this procedure is presented in Figure 49. This approach was used on all the available satellite imagery for the selected regions, for the years 2016 and 2017. Extracted patches (yellow patches) were used to train ship detection models (in combination with existing openly available VHR ship detection datasets). Altogether, 2420 satellite images were obtained from Planet Labs and 148 from Sentinel-2. The large difference is due to much higher - daily revisit times at Planet and due to different implementation of API for satellite imagery retrieval, given the provided region of interest. Cloud masks were also available through Planet API, for Planet imagery and through the eo-learn library, for Sentinel-2 imagery. In Table 8, the number of satellite images that were retrieved for specific locations is reported, satellite constellations and years.

[39] https://www.sentinel-hub.com/

[40] Štepec, Dejan, Tomaž Martinčič, and Danijel Skočaj. "Automated System for Ship Detection from Medium Resolution Satellite Optical Imagery." *OCEANS 2019 MTS/IEEE SEATTLE*. IEEE, 2019.

[41] Liu Z., Yuan L., Weng L. and Yang Y. (2017). "A High-Resolution Optical Satellite Image Dataset for Ship Recognition and Some New Baselines". *ICPRAM.*

[42] https://www.kaggle.com/c/airbus-ship-detection

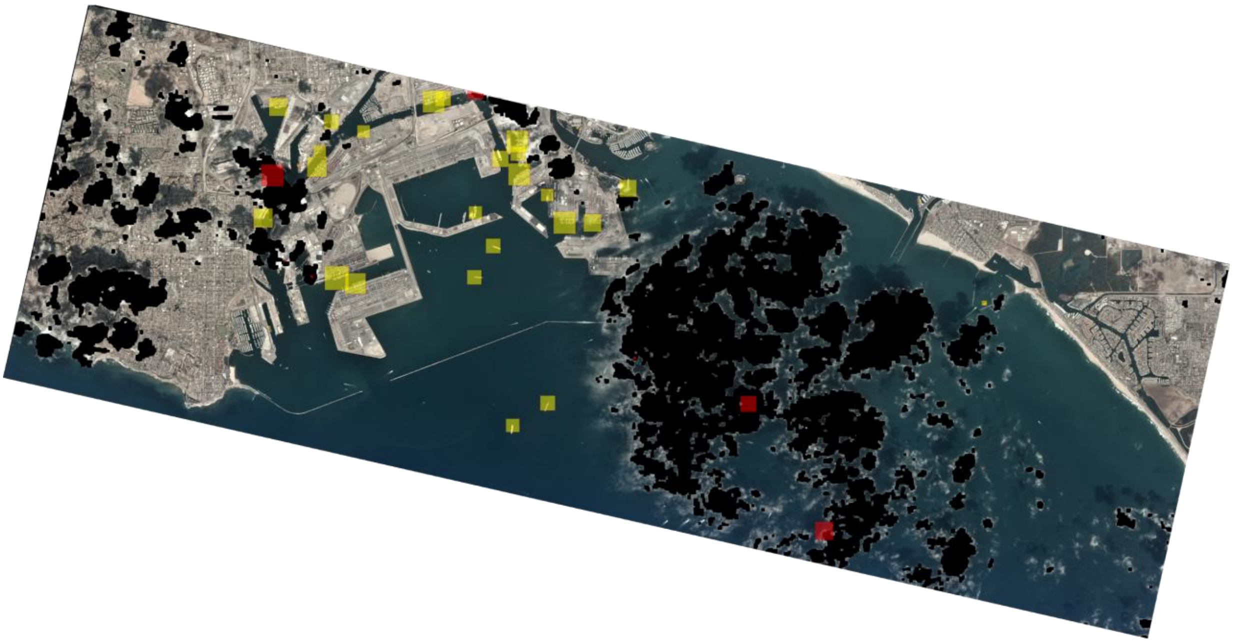


Figure 49. Satellite imagery of Port of Long Beach and matched AIS data.[43] [44]

Table 8. PIXSAT statistics - number of available satellite imagery for selected regions of interest.

	San Francisco Bay		Port of Long Beach	
	2016	**2017**	**2016**	**2017**
Planet	192/78	1000	212/117	1016
Sentinel	31/24	41	27/22	49

Satellite imagery was combined with AIS data to automatically annotate satellite imagery with ship positions and additional information, such as ship length and navigational status. The number of ships (as reported by AIS), that were correlated with satellite imagery for each region and year is presented in Table 9. Altogether 34894 ships were matched with Planet imagery and 5251 with Sentinel imagery. Only the ships of length greater than 30m were used for matching, due to our interest in commercial ships and spatial resolution limitations of used satellite imagery. Ships whose reported navigational status in AIS messages is not “underway using engine” were the ones that were matched, to have a direct positional matching, without the need of interpolation. A 5-minute window was used, for AIS positional report averaging of such stationary ships to ensure better positional accuracy. Matched ships might be covered with clouds or distorted due to different atmospheric conditions or image distortions. The data from 2016 was used for evaluation and 2017 for training purposes. To have valid ground truth annotations, all the satellite imagery from 2016 and its AIS matches, were manually inspected. The second number for the year 2016 in Table 8 and Table 9 provides the number of images containing at least one valid AIS matching and the number of valid ships, respectively. AIS matches that were not covered by clouds were retained and can be recognized by a human.

[43] Cloud mask is visualized in black and rectangles covered in clouds by more than 20% are visualized in red. Yellow rectangles are used as weakly annotated data for PIXSAT.

[44] Contains modified Open California Satellite Imagery ©2019 Planet Labs Inc. licensed under CC BY-SA 4.0.

Table 9. PIXSAT statistics - number of matches AIS positions, with the available satellite imagery.

	San Francisco Bay		Port of Long Beach	
	2016	2017	2016	2017
Planet	2085/446	10481	4076/1267	18252
Sentinel	669/199	957	1229/504	2396

An additional dataset was prepared out of the original PIXSAT dataset, that was used for training the ship detection method. Besides using AIS data for evaluating ship detection performance, ship length information was used, to provide weakly annotated data for training ship detection method. Patches out of the original satellite imagery of sizes 800x800 pixels were created, with 200 pixels of overlap and merged it with AIS data. Rectangles with the reported ship length were centred around the reported ship positions. All the 800x800 patches with annotations were manually inspected, to discard patches covered with clouds or patches where some of the AIS matches were false. The number of patches and the number of annotated ships for each data provider, location and year are reported in Table 10. Only the data from 2017 was used for training purposes, as 2016 data were used for evaluation, as reported in Table 8 and Table 9.

Table 10. PIXSAT statistics - number of patches and valid annotated ships, used for training.

	San Francisco Bay		Port of Long Beach	
	2016	2017	2016	2017
Planet	137/231	740/1417	601/1282	1773/5350
Sentinel	33/188	39/278	40/708	102/1629

4.2.2. Ship detection in port areas

In this section, the experimental setup and methodology is presented, as well as the results of the automated system for ship detection. Detection performance across different ship lengths is reported, different locations and satellite constellations. State-of-the-art Mask R-CNN (Region-based Convolutional Neural Networks) is used and object detection framework[45] from Facebook, which was adapted for ship detection. Different experiments were performed, with regards to training data. First, the existing dataset from Airbus[46] was utilized, which represented a baseline. An extensive augmentation on this existing dataset was also performed, by reducing its resolution and as such, adapting it to our lower resolution ESA Sentinel-2 and Planet Labs satellite imagery. Then these results were compared to the model that was learned solely on PIXSAT dataset, as well as a combination of Airbus and PIXSAT dataset. Ship detection (retrieval) rate is reported, which is considered successful if the ground truth position from AIS data is inside the reported detection from our method. Furthermore, detection rates are reported across different satellite constellations, different locations and different ship lengths. Multiple experiments were conducted to test different augmentation and training strategies, presented in Table 11.

[45] https://github.com/facebookresearch/maskrcnn-benchmark

[46] https://www.kaggle.com/c/airbus-ship-detection

Table 11. Overall ship detection performance (retrieval rate) across all ship lengths.

	San Francisco Bay		Port of Long Beach	
	Planet	Sentinel	Planet	Sentinel
baseline	52%	54%	54%	41%
baseline (Aug.)	50%	56%	52%	45%
PIXSAT	57%	86%	72%	92%
baseline + PIXSAT	61%	87%	76%	84%

Ship detection rate across different ship lengths, for all mentioned experiments, locations and satellite constellations, are also reported in Figure 50, Figure 51 and Figure 52.

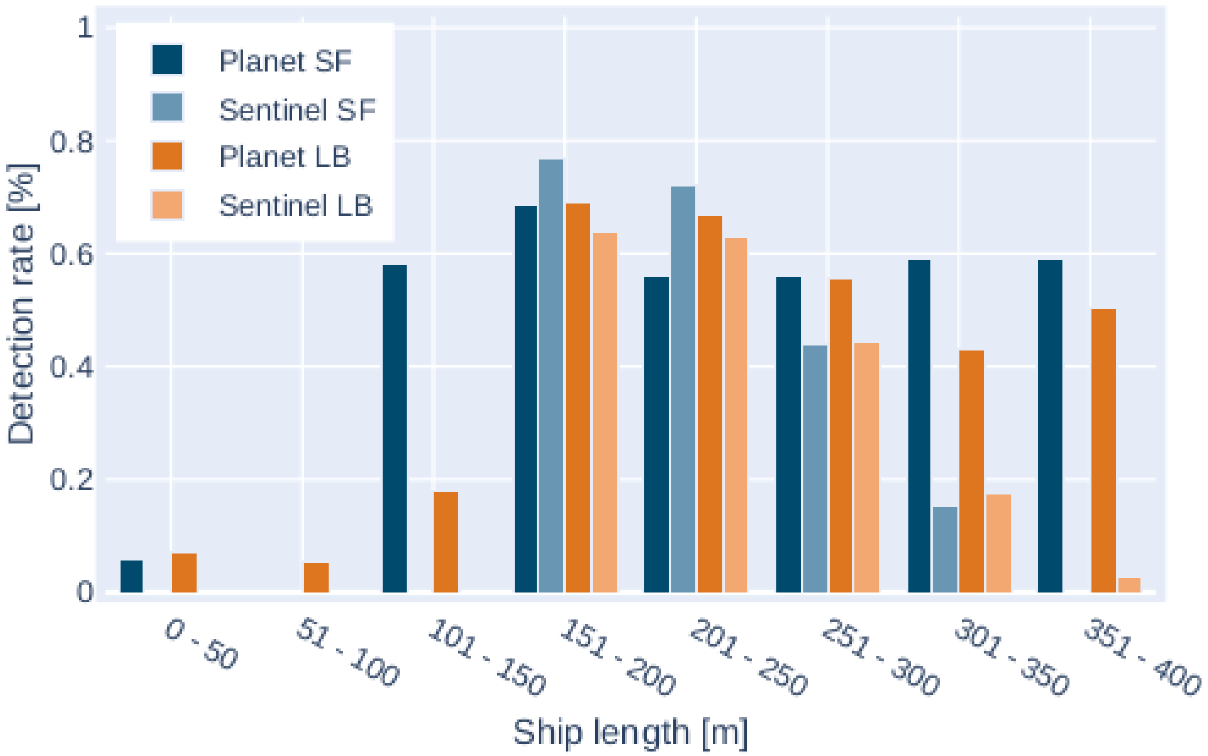


Figure 50. Ship detection performance across different ship lengths with the baseline (Airbus) model.

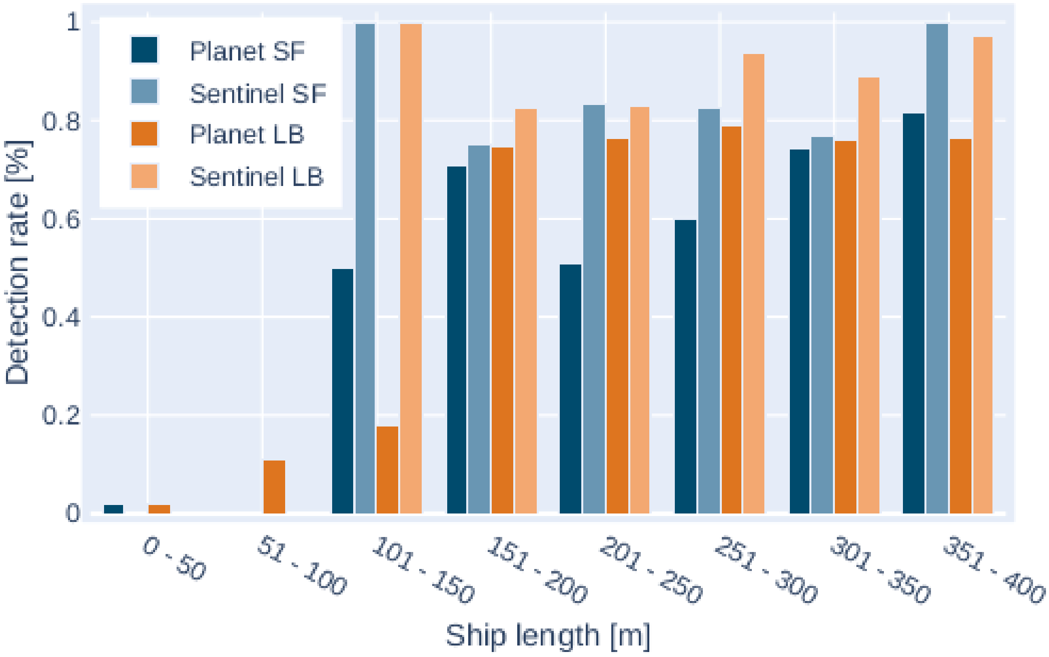


Figure 51. Ship detection performance across different ship lengths with the PIXSAT dataset.

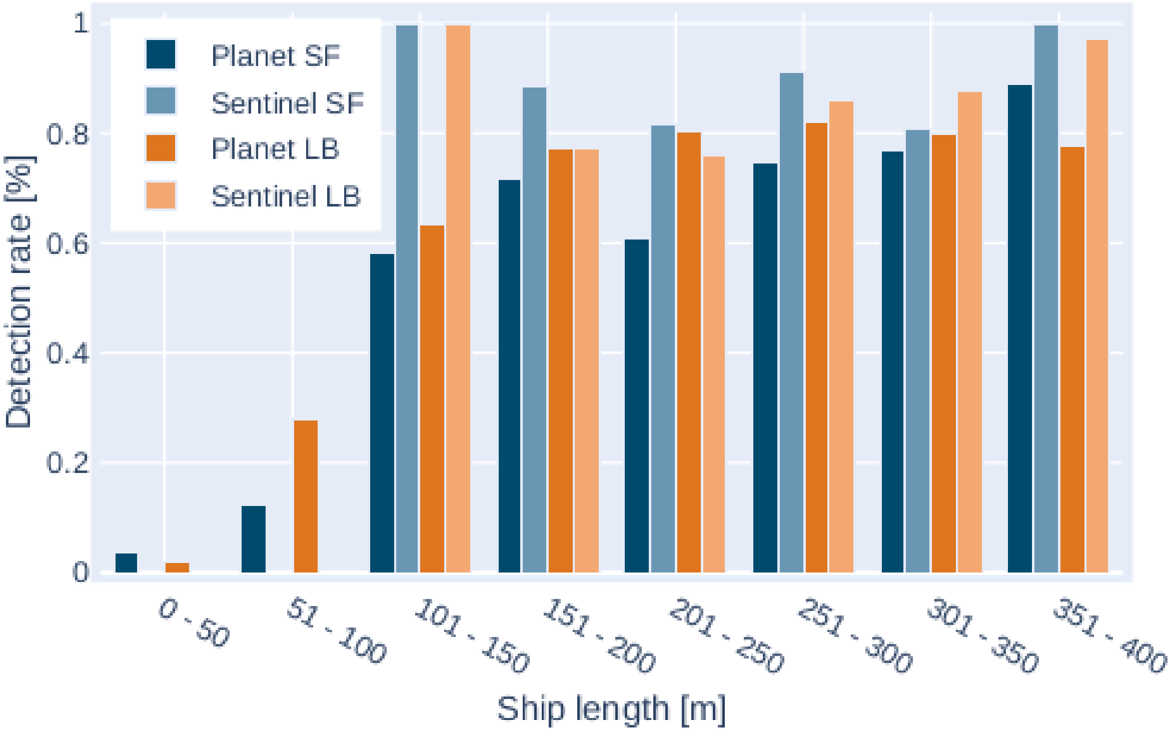


Figure 52. Ship detection performance across different ship lengths with combined datasets.

Qualitative results on operational satellite imagery used in the experiments are presented. Figure 53 presents the results of the best performing model on Planet Dove satellite imagery, which can be directly compared with baseline results, presented in Figure 54. Similar results of the best performing model for ESA Sentinel-2 are presented in Figure 55. It is clear, that detections are robust, in the port area, as well as in the area that is covered with clouds. One can notice that the detections do not capture the whole area of the ship. This is due to our training data from PIXSAT, where rectangles of ships lengths were fitted, to the reported AIS positions, which are usually captured in the ship bridge area. The selection of rectangles size and the influence of background on the detection performance is still to be investigated for future work. Results of the baseline model capture ship dimensions much more accurately, but it fails in the port area, due to lack of training imagery in such an environment. The use of heading information in AIS data was also investigated to make annotations much more accurate, but it did not prove to be reliable enough, especially for stationary ships.

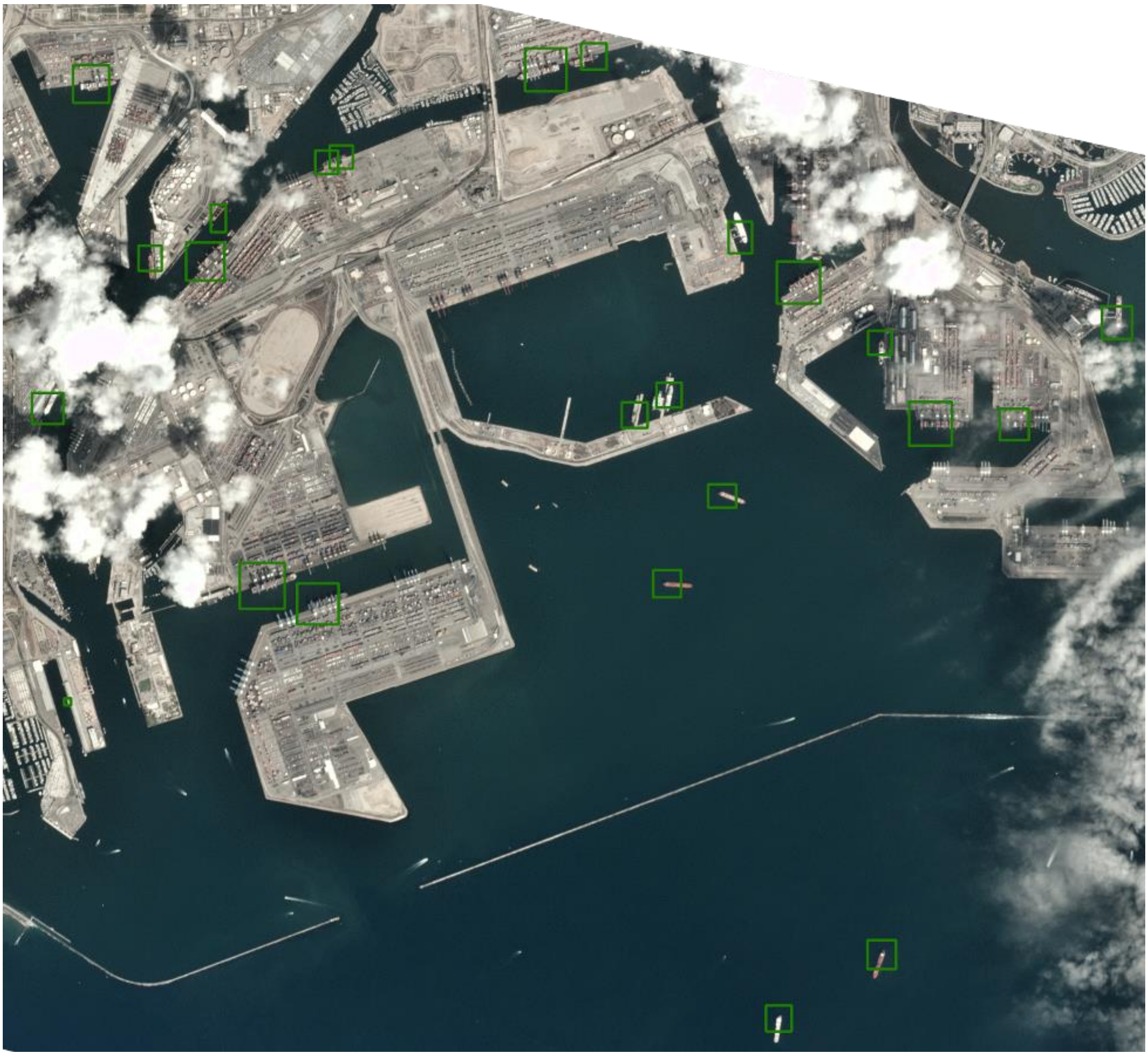


Figure 53. Results of the best performing model (baseline + PIXSAT) on Planet Labs Dove imagery.[47]

[47] Contains modified Open California Satellite Imagery ©2019 Planet Labs Inc. licensed under CC BY-SA 4.0.

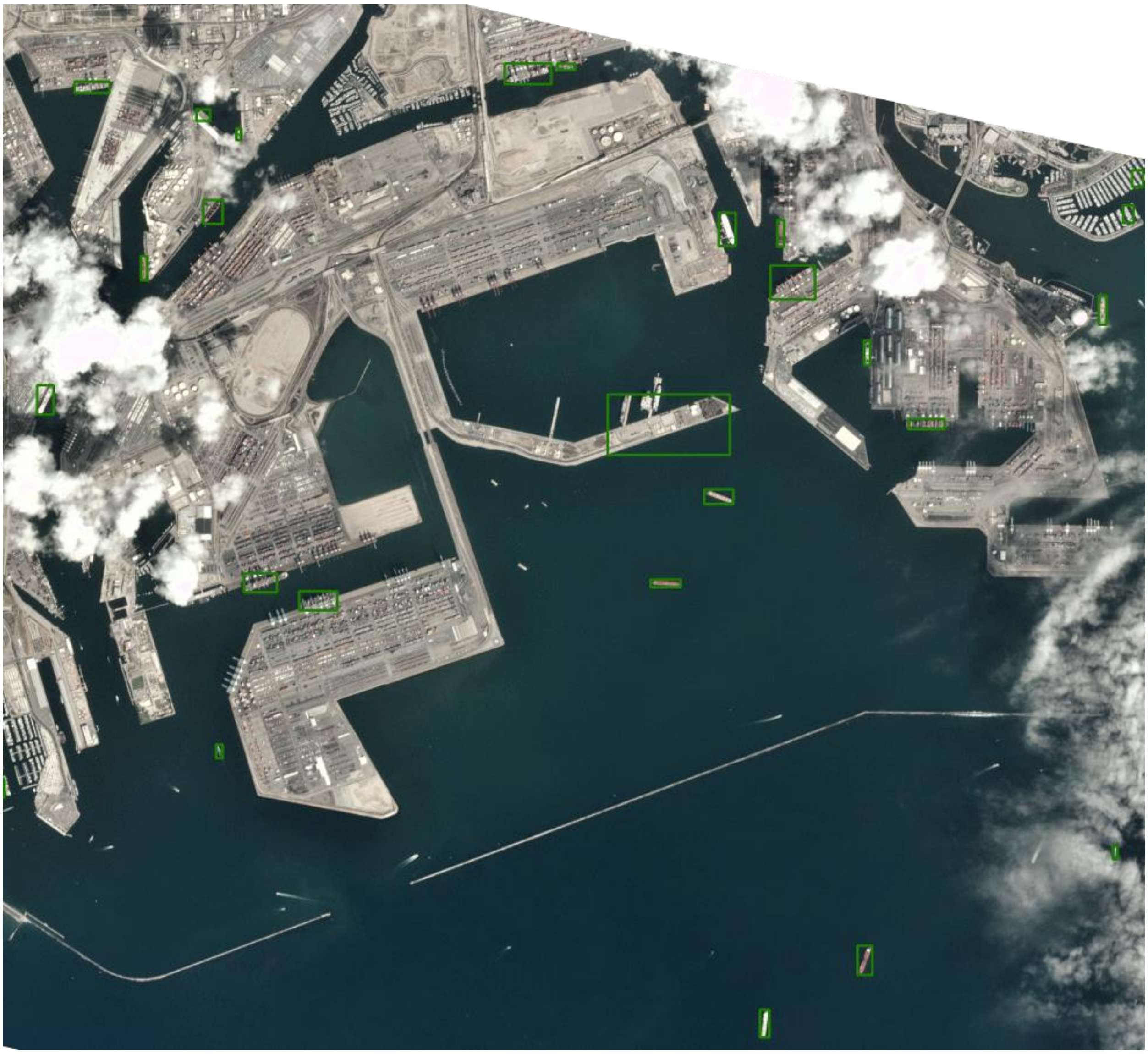


Figure 54. Results of the baseline model on Planet Labs Dove satellite imagery.[48]

[48] Contains modified Open California Satellite Imagery ©2019 Planet Labs Inc. licensed under CC BY-SA 4.0.

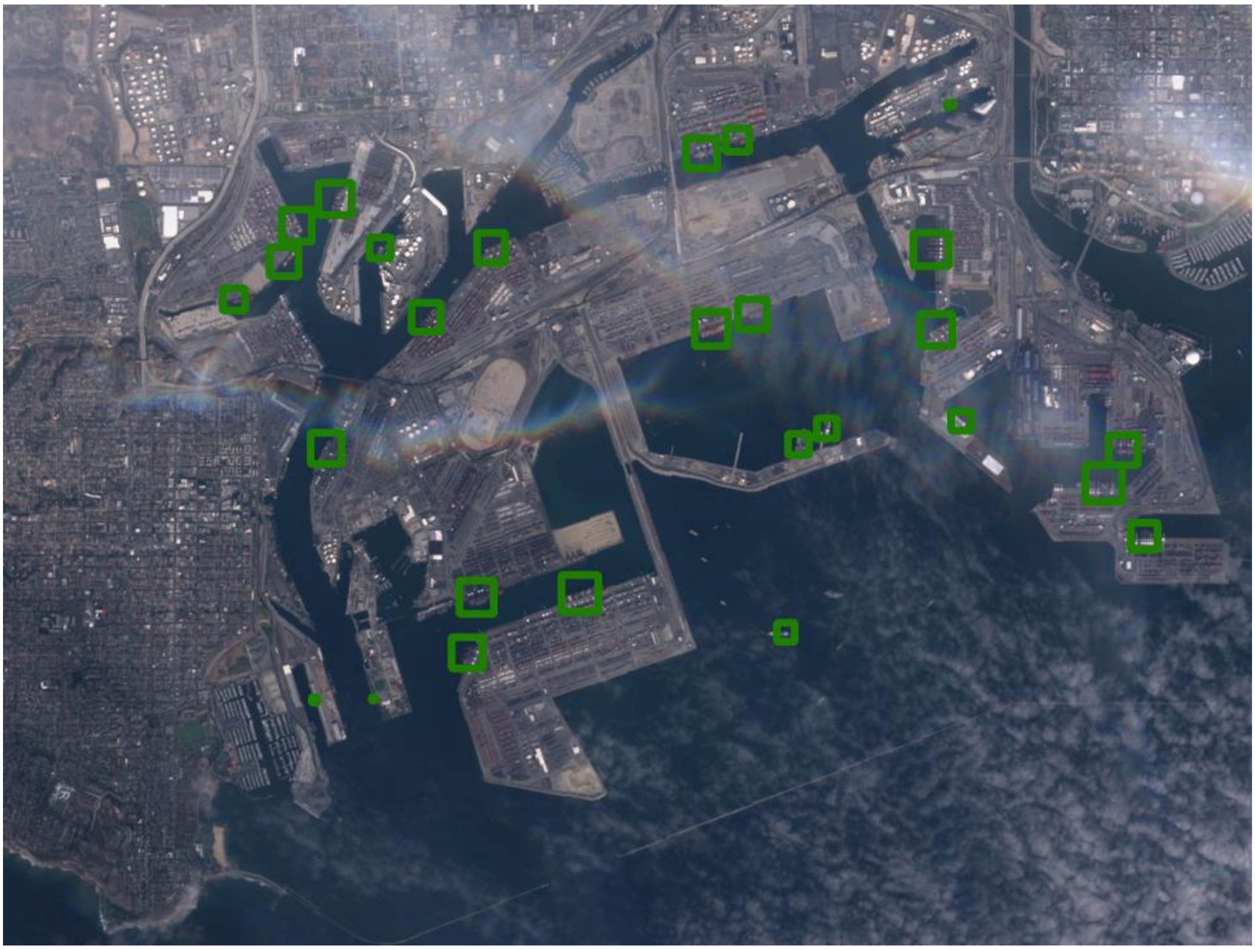


Figure 55. Results of the best performing model (baseline + PIXSAT) on ESA Sentinel-2 imagery for the Port of Long Beach area.[49]

[49] Contains modified Copernicus Sentinel data from Sentinel Hub licensed under CC BY-NC 4.0.

5. Analysis and prediction of road traffic conditions with connection to port operations

The focus of this chapter will be the analysis and predictions regarding traffic around the port area or in the regional road network. The results of this task will provide additional operational insights and forecasts, which will help operators at rerouting trucks to the inland terminals, gather information about road conditions in general or to simply provide operational insights about traffic out of historical data, which can assist in making ports more sustainable. Road traffic data will be correlated with the vessel call data in the port, to explore possible correlations between them, as well as with any other external data that might prove useful, such as weather information or traffic events. The predictive problem is framed as a short-term traffic volume prediction problem, which is a well-studied research domain, with established approaches and frameworks. The investigation goes beyond state-of-the-art time series forecasting frameworks, by demonstrating the superiority of the approaches, while of course trading the ease of implementation and generality to other ports. The amount and kind of data that was analysed is diverse and collected across different ports, thus making the proposed methods widely applicable.

5.1. Predictions and analytics for PIXEL

Traffic data around the three ports in the PIXEL consortium (i.e. ASPM, PPA and ThPA) was used for traffic analysis and predictive modelling. Although data comes from different sources and in different formats, a common data format was agreed, for all three use cases and almost the same methodology was followed, thus providing common analytics and predictive modelling pipeline for any port with similar data. First, a common methodology is described and later, specifics of each use case separately.

The data format agreed, as was common across all the ports, is combined out of three attributes: *location id*, *timestamp* and *value*. Value is the aggregation of values in a specified period (e.g. 1 hour). In the ASPM/SDAG and ThPA case, the value represents the number of vehicles that passed the location in a certain period. The value in PPA represents the average speed of vehicles passing the location in each period, as there was no data about the traffic volume available.

The next step after preparing the data in the right format was exploratory data analysis (EDA). Different visualizations were prepared and performed the statistical analysis, that provided insights into traffic dynamics and different influences on volume or average speed of the vehicles. Different seasonality's and trends were discovered, as well as features, that influence the amount/speed of the traffic. Weather, holidays and port activities were considered as major contributing factors. Specific data, such as arrivals of cruise ships were also considered in certain ports, as this can reflect in a higher number of cars, taxies and buses around the port area. Cargo vessels could cause many trucks driving through city roads and entering the port and were also considered in the ports, where vessel call data was available and could be combined with road traffic data.

After performing EDA, data was cleaned and new attributes were generated, that were describing the data and circumstances that were affecting the target variable (i.e. traffic volume or speed). The need for data cleaning and methods are specific for each use case. Attributes had to be adapted to different ML (machine learning) algorithms. General-purpose time series forecasting library, Facebook Prophet[50], was used, which provides a general framework for time series data analysis and predictive modelling and can be easily integrated to the ports and the data, that is commonly available there. A custom predictive model was also prepared, by transforming time series problem into a classical structured supervised ML problem, where state-of-the-art gradient boosting methods were applied, which outperformed Facebook Prophet by a significant margin but require more effort and data, to be integrated into port systems. All the models were evaluated on left-out datasets using different error metrics, used in time series forecasting domains, presented in the result sections.

[50] https://facebook.github.io/prophet/

5.1.1. ASPM/SDAG methodology

As already explained in D4.3, the data for ASPM/SDAG was collected from the SILI (Sistema Informativo Logistico Integrato) system. The acquired dataset contains almost 95 million records of vehicles passing the gates from March 2015 until August 2019. Data were collected from 11 locations around the FVG (Friuli Venezia Giulia) region. Five stations represent the port gates (one in Port of Monfalcone and four in the Port of Trieste), four stations are located on the highways and regional roads, one station at the SDAG (Stazioni doganali autoportuali di Gorizia - autoporto) parking area and one at Interporto di Cervignano. Locations are geographically presented in Figure 56 and the actual cameras from the SILI system and the gates at Port of Monfalcone, in Figure 57.

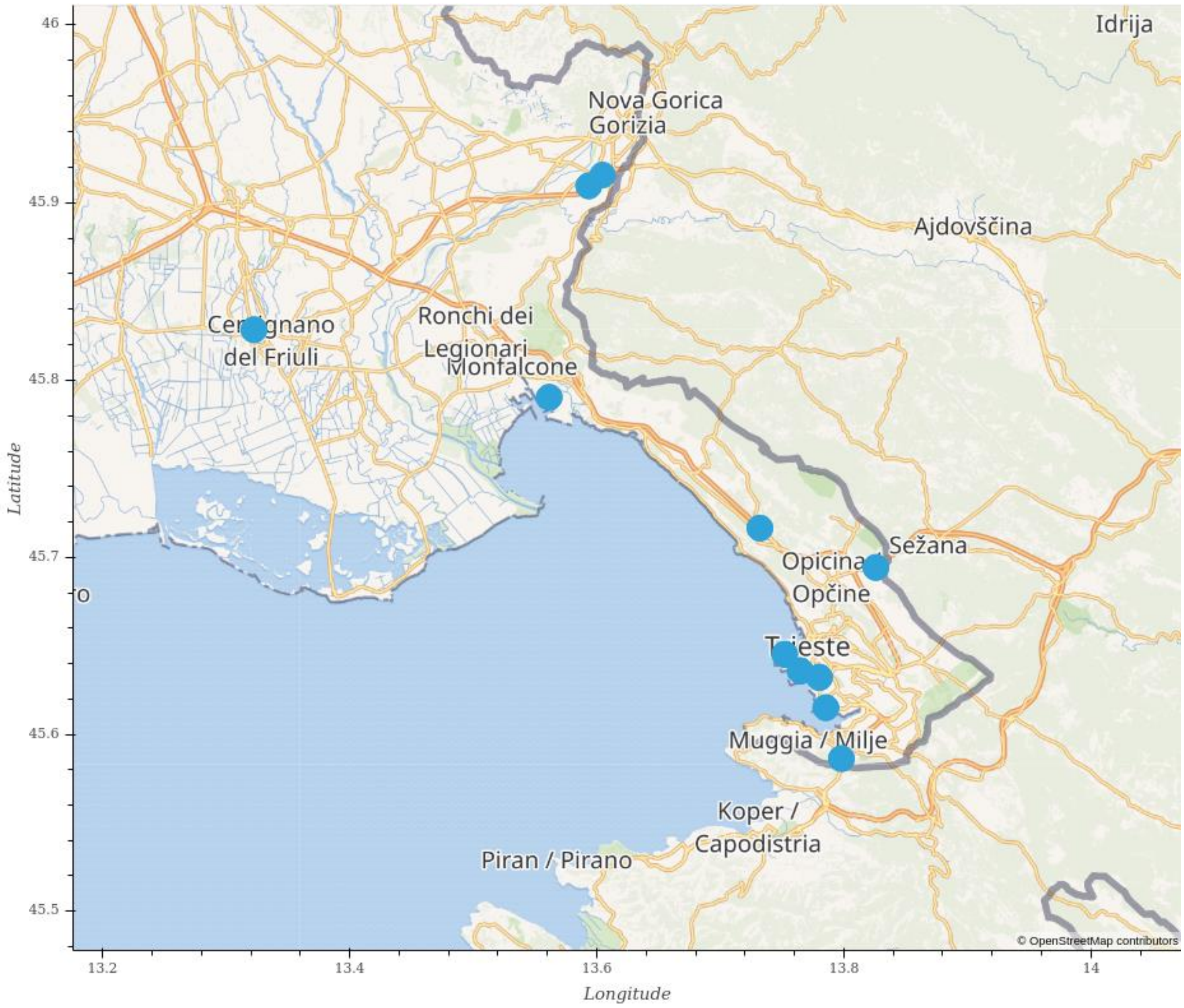


Figure 56. Map of traffic sensors locations from the SILI system.

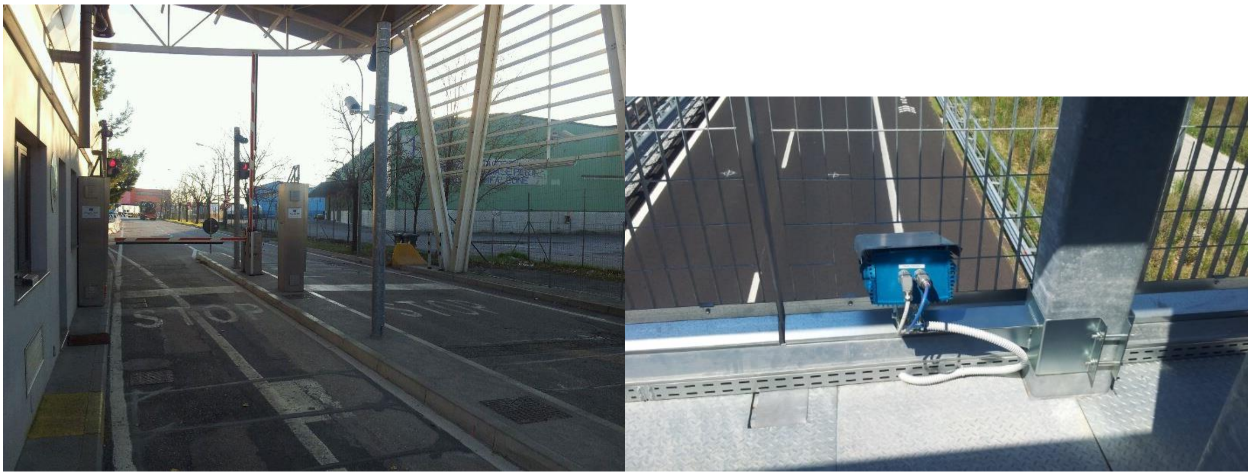


Figure 57. Gates at the port of Monfalcone.[51]

Our goal was to develop a predictive algorithm for short-term traffic volume prediction on regional roads, to support the decision-making process, regarding rerouting the trucks to the inland parking premises, such as SDAG, as well as to better predict the inflow of the trucks to the port area.

In D4.3 it was promised to use vessel call data to improve traffic volume predictions, but after the discussions with the port, they noted, that enough warehouse capacity is available in the port, thus there isn't a spike in the number of trucks entering the port at vessels arrivals, because there is no direct correlation, due to long term storing of the cargo. Enough vessel call data for ASPM was also not available for this kind of analysis.

Data cleaning is an especially important step in machine learning since data from real case scenarios is rarely without noise or false information. Because Facebook Prophet can handle missing data, it was important to remove false zero values and mark them as missing. No vehicles may pass the gates of the port in some time, especially during the night, weekends or holidays. To remove false zero values and set them to missing (i.e. SILI sensor is down), a threshold of 24 hours was chosen. If there is no vehicle in 24 hours, that time is marked as a missing value. Moreover, the upper threshold had to be set, which is different, depending on the location of the sensor. Given the results of statistical analysis and physical limitations of the trucks and gates, the upper threshold for port gates is set at 150 vehicles per 15 minutes. For highways, the threshold is set to 750 vehicles per 15 minutes. Any value over the upper threshold is set to the threshold value (150 for port gates and 750 for highways). Note that the SILI system captures traffic flow via visual cameras, which are prone to errors, thus miscounting the traffic.

For evaluation of the models, cross-validation (included in the Prophet library) on a horizon of 24 hours was performed. Initial training data consisted out of 1200 days in the first cut-off and then predictions for every 10 days were conducted. This corresponds up to 37 folds (there may be less in case of the missing data). Different error metrics were calculated, such as MAE, MdAPE and RMSE. Error metrics will be presented for specific locations. MdAPE is Median Absolute Percentage Error and was used instead of MAPE, due to problems of calculating MAPE where ground-truth value is 0.

Also, a different approach to solving time series forecasting was considered. Infinite time-series data was transformed to the classical structured form, used for supervised machine learning, by extracting lagged features and encoding timestamps to different features, such as the year, hour of the day and the day of the year, which were encoded as cyclical features via sinus and cosine transformation. For machine learning algorithm, XGBoost[52] Python library was used. The model was evaluated using cross-validation for time series data. It was not possible to directly use the Prophet's cross-validation method, so a new one was developed. As initial training dataset of 20.000 records were used and the remaining (around 10,000

[51] http://sili.regione.fvg.it/area/cms/portale/servizi-portuali/controllo_accessi/

[52] https://xgboost.readthedocs.io

records) were split into 50 folds and incrementally added (respecting chronological order). Same evaluation metrics were used, as for the Prophet method.

5.1.1. PPA methodology

The data from provider Telenavis, described in D4.3 was not available, so an alternative solution had to be found. There are various services capable of providing real-time traffic information from a location. Some of the providers considered were TomTom, HERE and Waze. The main problem was that most of them do not have historical data available and, additionally, those who did have were too expensive. Therefore, two main candidates were finally considered, TomTom and HERE, and finally it was decided to use HERE Traffic API for the total number of free calls that it offers, which would allow us to collect more data, as well as the amount of information in each request.

Once the service that was going to supply the information was decided, it was time to choose the area of interest for which to collect the data. Moreover, the procedure for calling, collecting, and storing the data was still pending.

For the Piraeus Port Authority, there was a defined area around the port for which there was a special interest. For this, according to the operation of the API, which returns traffic information from a bounding box, it was decided to select the same area that the required area that was validated by the port, presented in Figure 58.

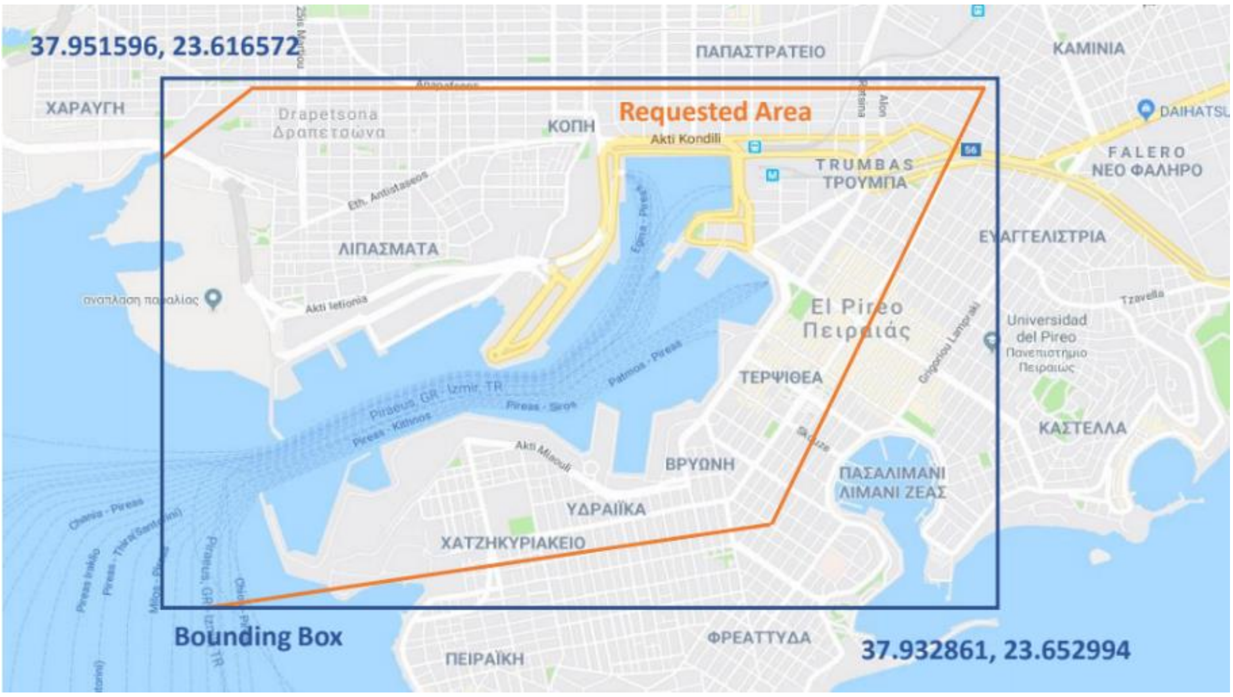


Figure 58. PPA area of interest for traffic prediction.

Considering that there are 250,000 free daily calls, 125 different traffic status could be collected for each day and each point inside the area. This allowed the accuracy of up to 15 minutes to be achieved between two consecutive records. After various tests, the collection began at the end of July 2019, reaching in April 2020, almost six million records.

To store the information, a Linux machine was mounted in the PIXEL environment, which through a service, allowed to execute a Python script every 15 minutes which called the API, collected the information and stored it in an SQL (Structured Query Language) database.

Based on a future aim, in which to search for the relationship between the variation in speed and other attributes, such as weather, port activity and others, it was decided to replicate this procedure with weather provider information services, such as OpenWeather[53]. It was carried out in such a way to allow the same time precision of 15 minutes as the traffic.

Although the use of these services allows access to this information and even the creation of history, it is important to comment some of the drawbacks presented during the use of these.

Firstly, the HERE service returns traffic information inside a bounding box requested, so one of the main drawbacks has been that, the addresses for which information was given have not been constant over time. So, if we focus our predictions on a single location, we can stop receiving information after a while. Also, some of the addresses go outside the required area.

Secondly, the service has been down several times during all these months, preventing the maximum possible data collection and, therefore, no traffic behaviours have been registered that may have been of great interest to the prediction algorithm (Figure 59).

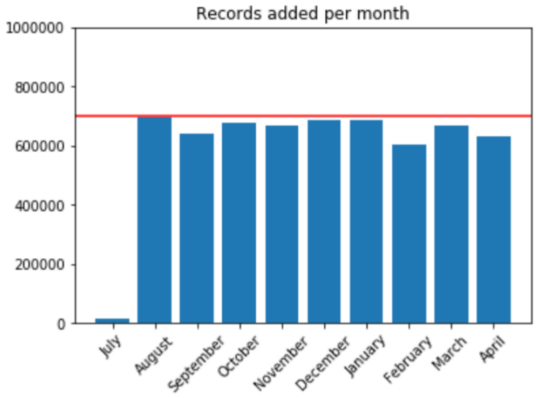
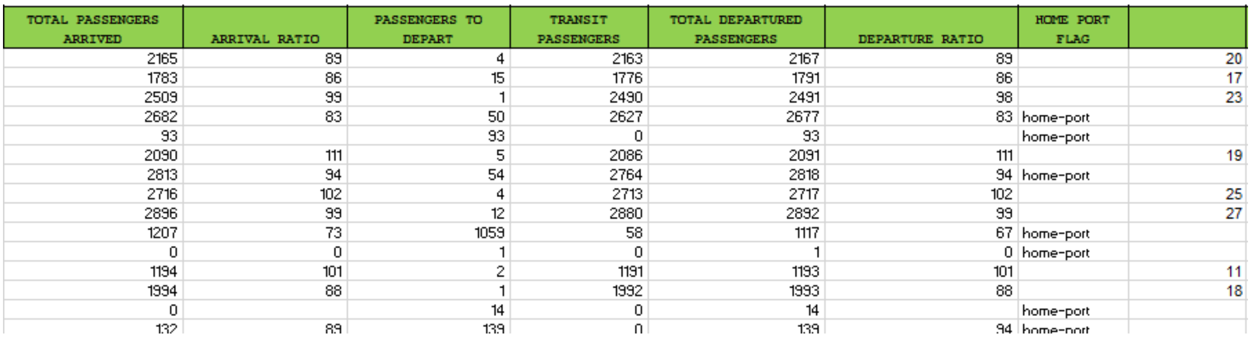


Figure 59. Traffic records added since July 2019

Finally, the additional attributes that have been considered, were in terms of meteorological information: temperature, probability and intensity of precipitation, and wind speed. For example, it is plausible to think that in rainy days the use of the vehicle increases, as well as during the days of low temperatures or high wind speeds, in which transportation on foot or bicycle is unsafe or unpleasant. This would reflect in lower average speed, as there are more cars on the roads and surface is wet and slippery.

Besides, in terms of port activity, the types of vessels that according to the port authority have the most impact on traffic (i.e. cruise ships), have been considered as well. So, the number of cruises and their capacity, the number of buses they require and the number of total passengers arriving at the port have been used. While the weather data is obtained from OpenWeather API, the port data comes from manual extractions in Excel format, provided by the port authority (Figure 60).



TOTAL PASSENGERS ARRIVED	ARRIVAL RATIO	PASSENGERS TO DEPART	TRANSIT PASSENGERS	TOTAL DEPARTURED PASSENGERS	DEPARTURE RATIO	HOME PORT FLAG	
2165	89	4	2163	2167	89		20
1783	86	15	1776	1791	86		17
2509	99	1	2490	2491	98		23
2682	83	50	2627	2677	83	home-port	
93		93	0	93		home-port	
2090	111	5	2086	2091	111		19
2813	94	54	2764	2818	94	home-port	
2716	102	4	2713	2717	102		25
2896	99	12	2880	2892	99		27
1207	73	1059	58	1117	67	home-port	
0	0	1	0	1	0	home-port	
1194	101	2	1191	1193	101		11
1994	88	1	1992	1993	88		18
0		14	0	14		home-port	
132	89	139	0	139	94	home-port	

Figure 60. Sample of cruise activity data

[53] https://openweathermap.org/

The data processing has been carried out mainly in Python with the Pandas library. However, other libraries have been required for visualizations such as Matplotlib, Seaborn, and Geoviews.

5.1.2. ThPA methodology

After having analysed in D4.3 the global scope of the traffic prediction for ThPA, during the period M12-M24 the technical team has used the different data to achieve those objectives. Thus, a methodology was followed.

According to D4.3, the aim for ThPA operators (terminal, environmental) is to have a tool supporting decision making that can be done in the port with regards to the congestion at the gates, using a Gate Congestion Index. The HMI (Human-Machine Interface) for such a tool was considered of paramount importance, as it will be, after PIXEL, the way the personnel in the port would interact with and use the information gathered (and created) by the traffic prediction model.

Additionally, regarding the data, a prioritization scale was done: (i) RFID (Radio-frequency identification) traffic data at the gates, (ii) weather, (iii) traffic at the city, (iv) vessel calls. With this, the procedure would focus first in analysing the most prominent data source and use it for the prediction and, later, once this had been achieved, continue with other less relevant (for the prediction) data sources.

Several meetings took place with the responsible team in the port (ThPA) to define the HMI: the quantity of information, the periodicity, the visual aspects of the results, prediction horizons, the way of representing results and the layout. API for access to the traffic data was exposed by the ThPA. Historical data since April 2018 is accessible for 4 different locations: gate 10A entry, gate 10A exit, gate 16 entry, and gate 16 exit.

For the case of ThPA, a code was developed retrieving data of all trucks passing the gates and it was digested to be converted in a time series of the volume of cars at each gate in 60 minutes. Aggregated data was exported to a CSV (Comma Separated Values) equivalent format, which had 62193 rows. Despite the fact, that the RFID sensors at the gates are reliable, sometimes experience missing vehicles, false repetitions and other outliers. Facebook Prophet handled those cases properly, to some extent.

After the most important data source, the sensors at port’s gates, was covered, ThPA team tried to compose advanced predictions taking into the account the additional information provided. At this point, the less-prioritised data was recovered for using it as inputs for the model. By order, this was the set of data transformed and to which format each:

- Weather data – daily aggregations
  - Available data since September 2018
  - timestamp, temperature, wind_speed, precip_intens
- Traffic city data
  - Available data since September 2018
  - timestamp, avg_nearbies_speed
- Vessel calls data
  - since April 2018
  - timestamp, number of vessels in the port

Models were fitted and evaluated on different subsets of features in the dataset. First, only basic features were used, and later learning dataset was enriched with weather, vessel calls and city traffic information. After all the executions, the last step was the reflection. When executing the different notebooks, both for EDA and for the prediction itself, certain insights were gained. At this point, and as it is reflected in section 5.2.3, different conclusions were extracted that will be, with no doubt, of help for ThPA for decision making.

5.2. Results

5.2.1. ASPM/SDAG

5.2.1.1. Exploratory Data Analysis

Port of Monfalcone gates

There was no way to associate entry and exit of the same vehicles, due to unavailability of the unique identifier, so most of the EDA for port gates sensors will be done only on entry data.

Data was acquired for 4 and a half years, from March 2015 to August 2019. Unfortunately, there are periods of missing data. The longest period of missing data is almost 4 months long (Figure 61).

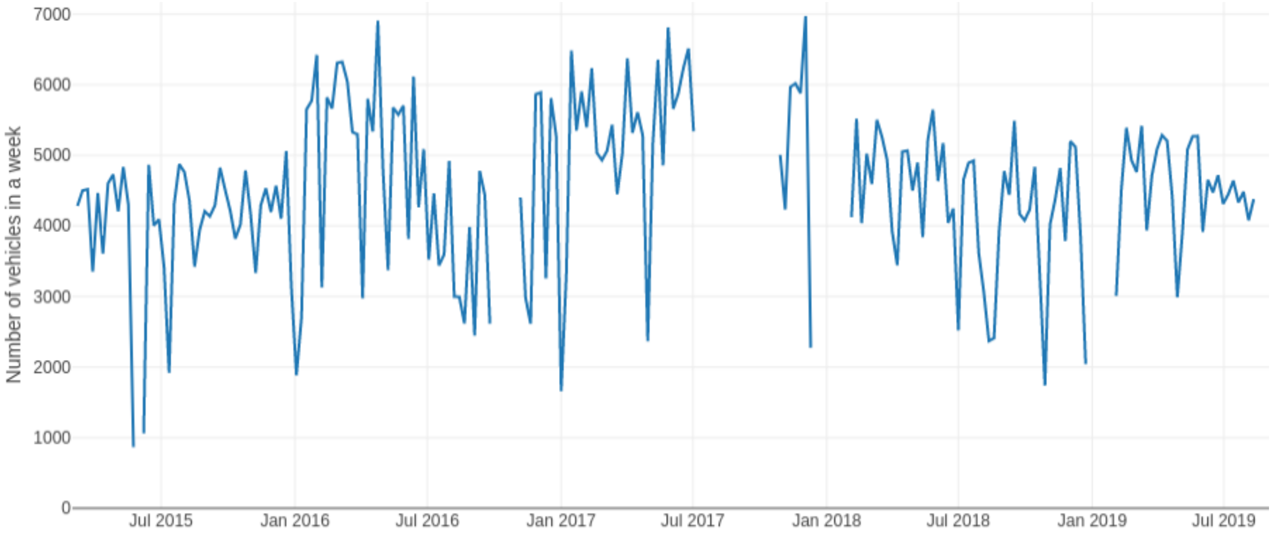


Figure 61. The number of vehicles entering the port of Monfalcone in a week.

The average number of entries per day (Figure 62) during the working days is 717 (with highs on Thursdays, 757 on average and lows of 674 on Fridays) and 108 during weekends (168 is average on Saturdays and 48 on Sundays). The average number of entries during holidays is even lower - 37 vehicles per day.

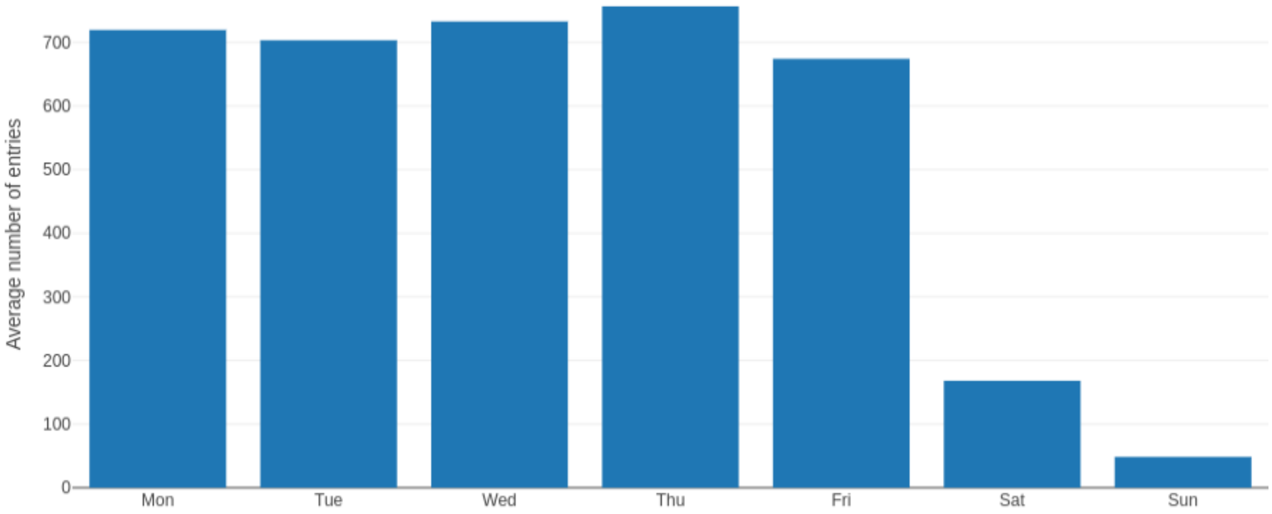


Figure 62. The average number of entries in the port of Monfalcone per day of the week.

Vehicles start entering the port at 6 AM, with a peak at 7 AM and continue throughout the day, stopping at 5 PM (Figure 63, left). Exits start later around 8 AM and continue until about 8 PM (Figure 63, right). Most of the activity happens during the working days, less on Saturdays and almost nothing on Sundays.

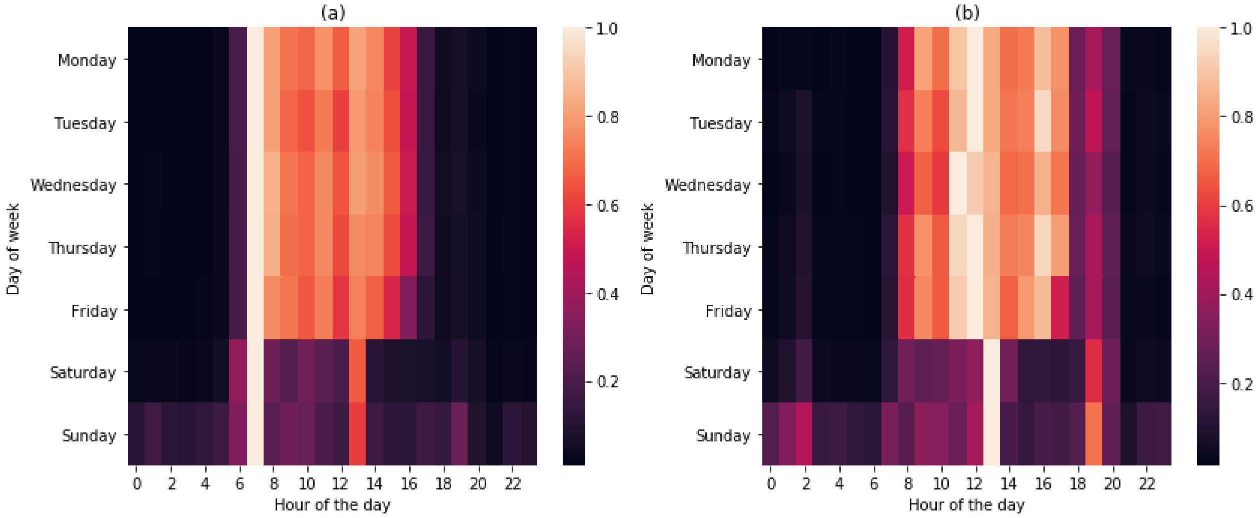


Figure 63. Daily seasonality of traffic volume at Monfalcone port entry gates (a) and exit gates (b).[54]

Even though the average number of vehicles is changing throughout the months (Figure 64, left), seasonality has a similar characteristic for all the months (Figure 64, right). Values are normalised by dividing all values in by months' maximum values.

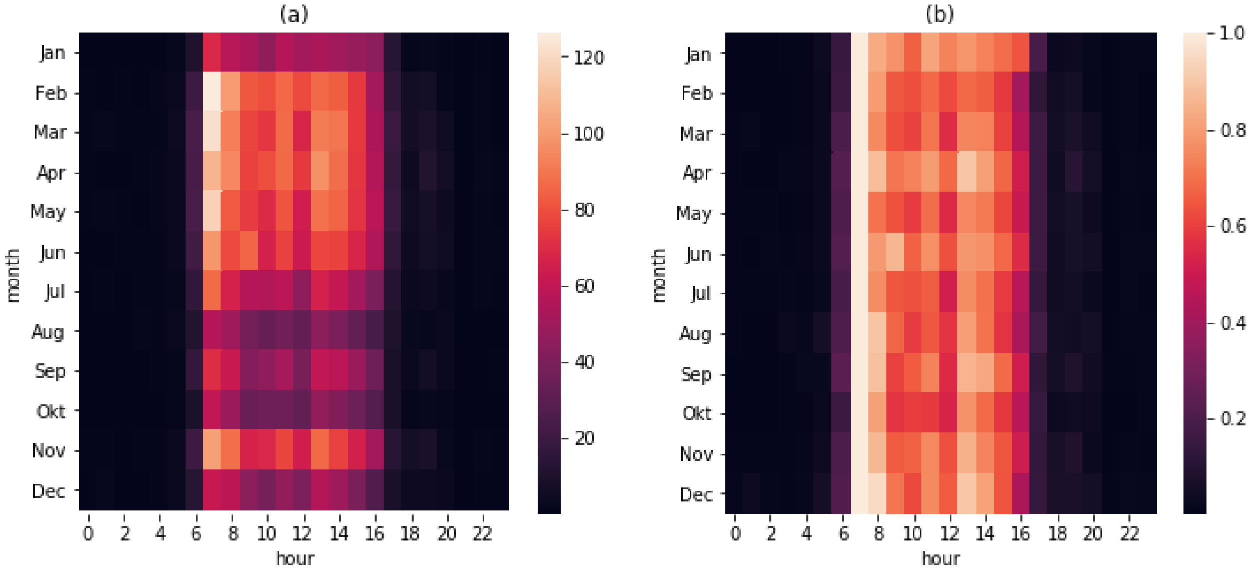


Figure 64. The average number of vehicles by month and hour (a), normalised values on (b).

Regional roads

Although there are missing data, yearly seasonality can be observed in Figure 65, Figure 66, Figure 67, and Figure 68. Peaks appear during the summer, around July and August, because of the tourist season.

[54] Values are normalised. Value 1 is daily maximum, 0 is daily minimum.

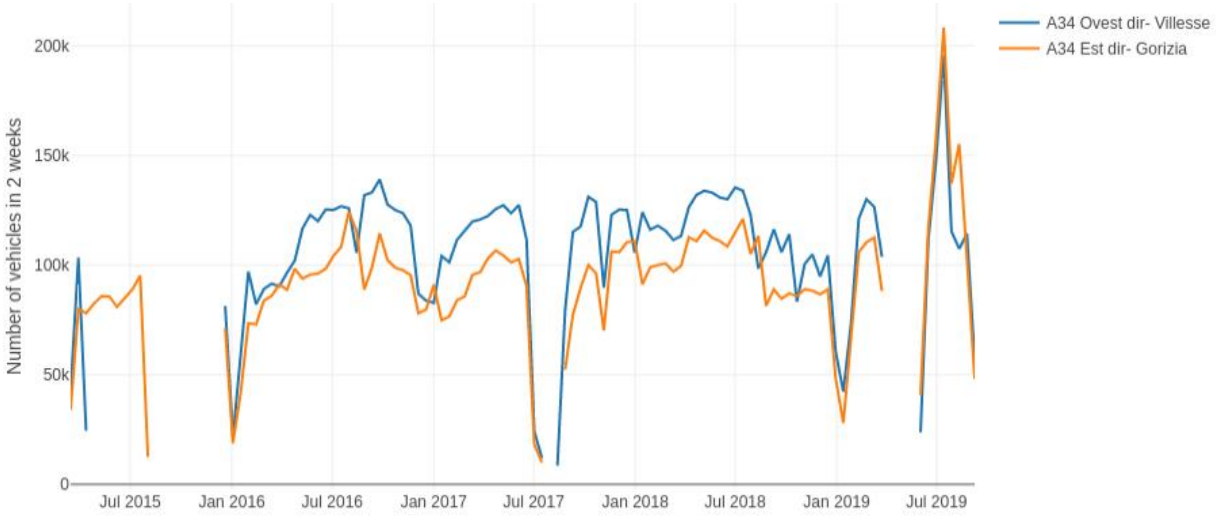


Figure 65. Traffic volume at location A34.

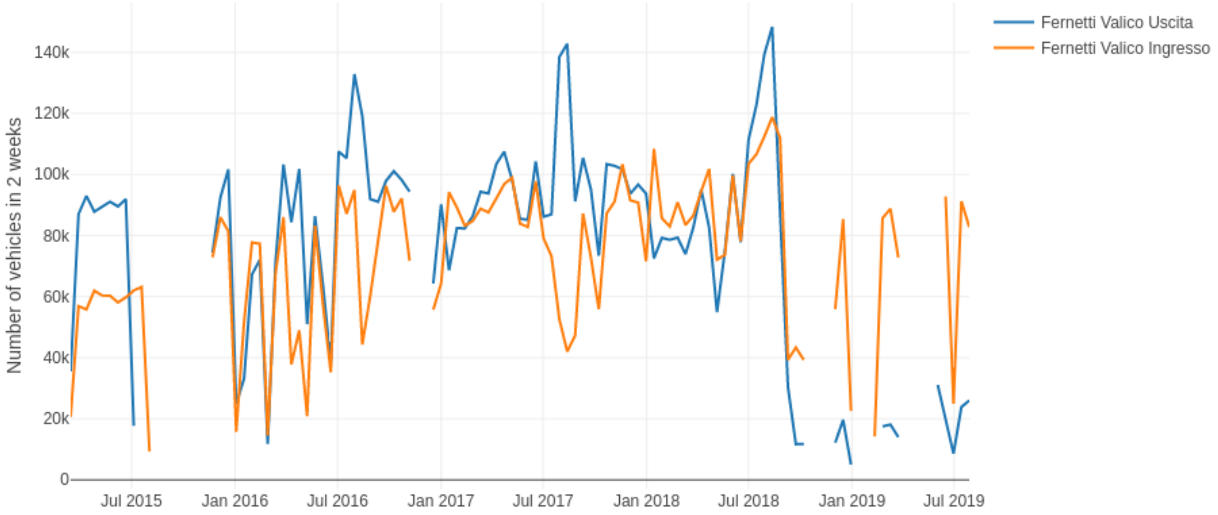


Figure 66. Traffic volume at location Fernetti Valco.

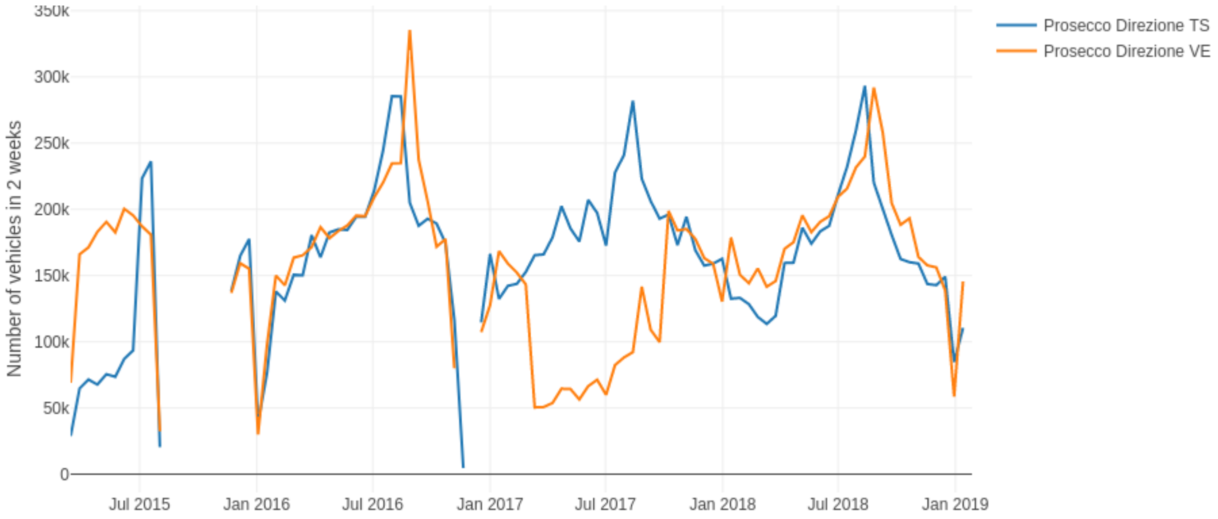


Figure 67. Traffic volume at location Prosecco.

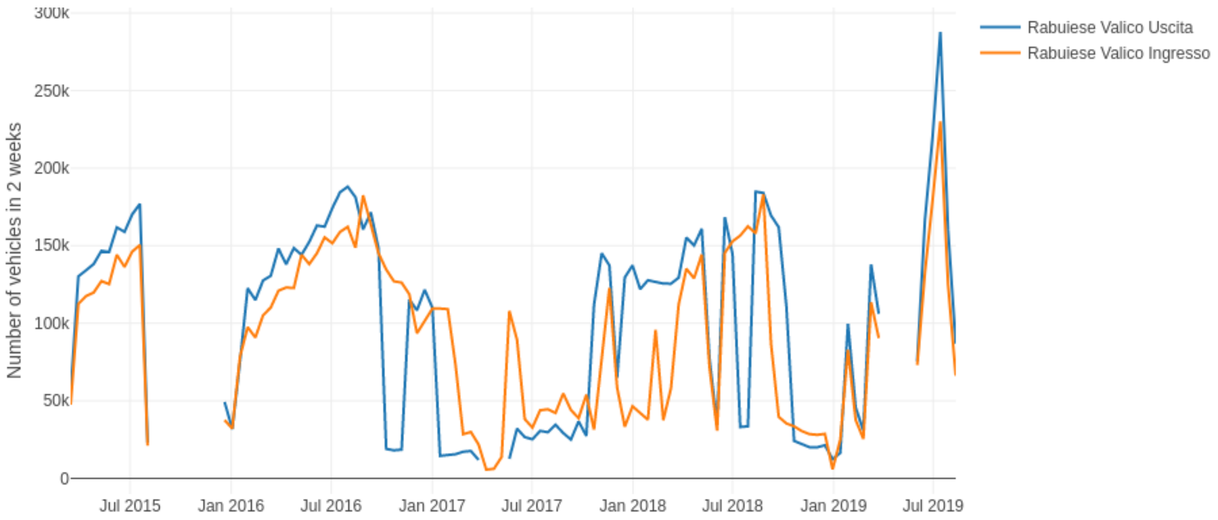


Figure 68. Traffic volume at location Rabuiese Valico.

Average daily number of vehicles based on the day of the week and location are presented in Figure 69. Fridays and Saturdays have the highest average of daily vehicles. At some locations, Tuesday and Wednesday have the lowest traffic volume, on other locations the lowest traffic volume is on Sundays.

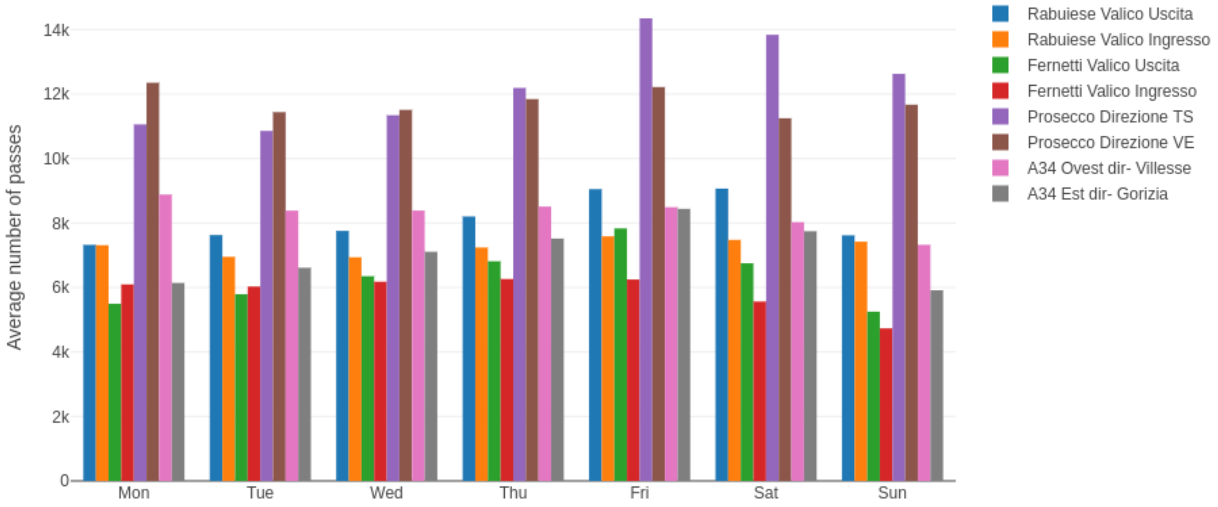


Figure 69. Average daily number of vehicles on regional roads.

Daily seasonality for all the days of the week is presented in Figure 70. Some of the locations have peaks in the morning (b, e) and others in the afternoon (a, c, f, h). Some of the locations have peaks both, in the morning and the afternoon (d, g), which is due to the daily migration of workers. Traffic volume throughout days is similar for all working days, except Friday, that has higher traffic volume later in the day, as well. Daily seasonality of weekends is different; traffic starts later in the morning or it even lasts all day.

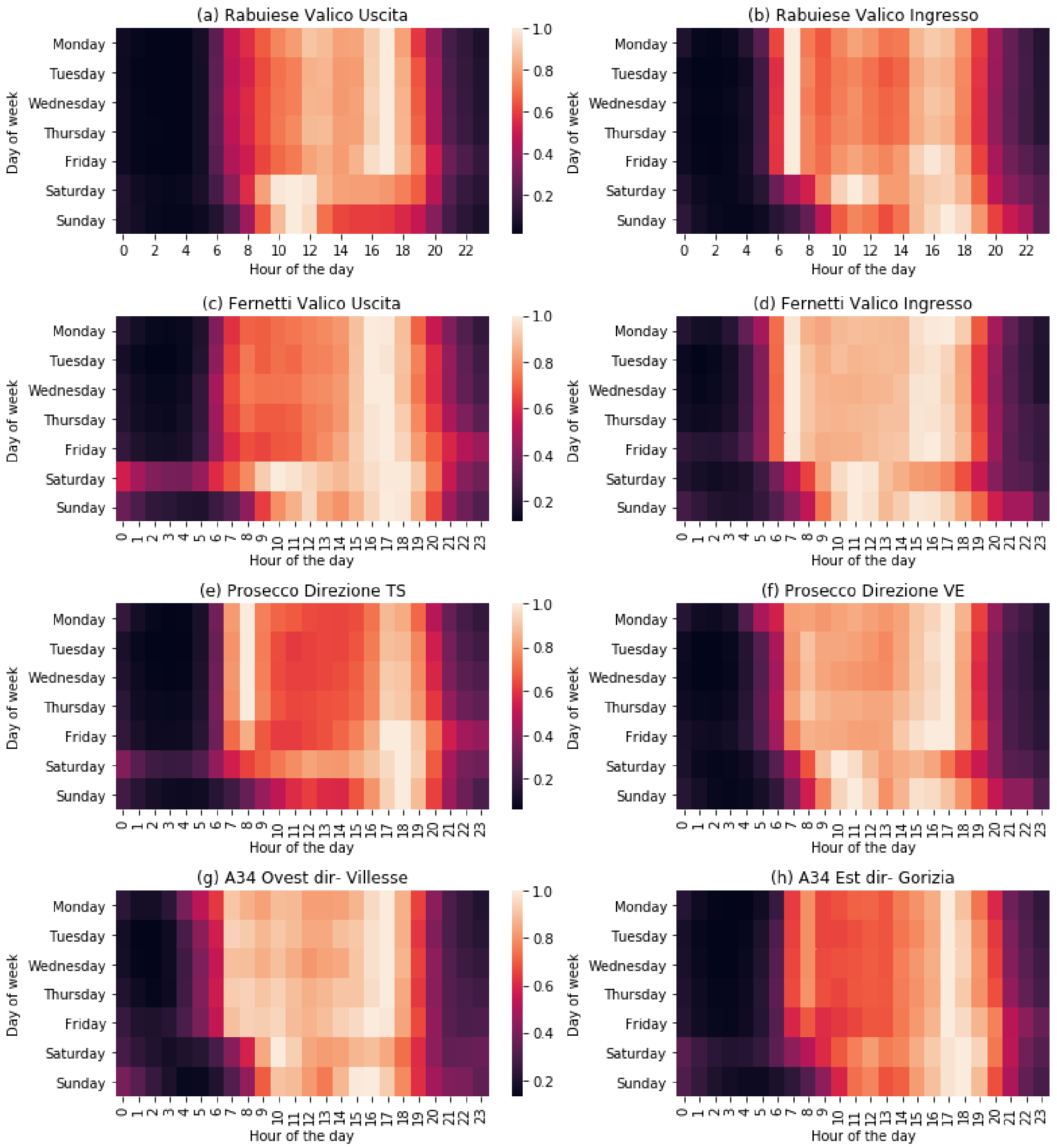


Figure 70. Daily seasonality of traffic volume.[55]

Daily seasonality of traffic volumes is changing throughout the year. During the summer, there is more traffic in the late hours of the days and at nights (Figure 71).

[55] Values are normalised per day.

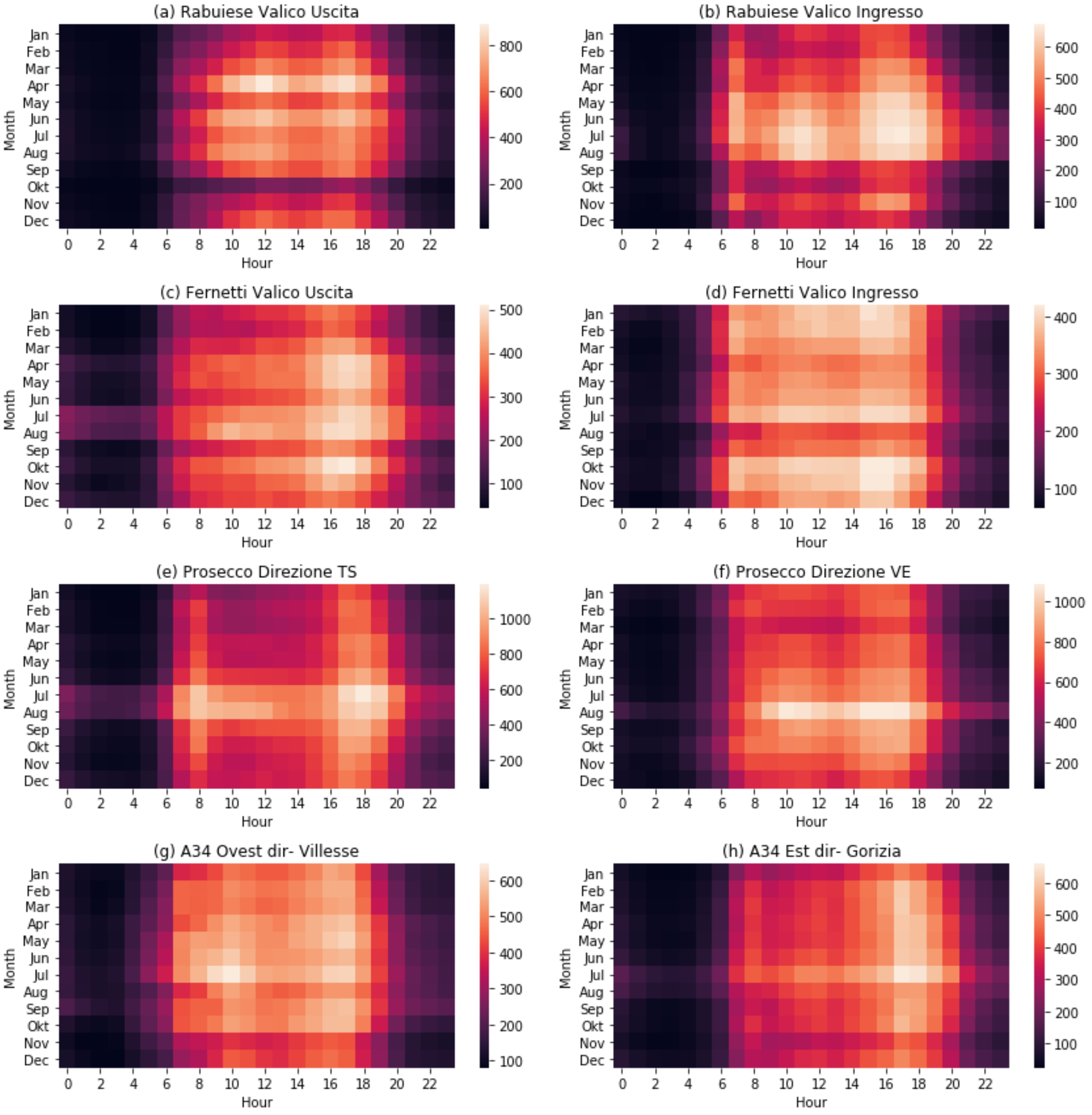


Figure 71. Average daily seasonality of traffic volume per month.[56]

5.2.1.2. Facebook Prophet predictions

Custom seasonality for each day of the week was added (Table 12) with Fourier order of 7 and a period of 1 (Facebook Prophet method input parameters), which resulted in 7 additional Boolean features in the input dataset.

Table 12. Facebook Prophet input data sample with daily seasonality.

ds	y	Mon	Tue	Wed	Thu	Fri	Sat	Sun
2015-03-01 00:00:00	4	False	False	False	False	False	False	True
2015-03-01 01:00:00	1	False	False	False	False	False	False	True
2015-03-01 02:00:00	3	False	False	False	False	False	False	True

[56] Values are normalised per day.

Example of Facebook Prophet model components is presented in Figure 72. The model was successful in learning seasonality, as it was proven, during the EDA. Most of the traffic appears on Friday and Saturday, and less on Sunday and during the beginning of weeks. Yearly peaks take place during the summer, in July and August. Also, daily seasonality has two peaks, one in the morning and one in the afternoon, as it was proven during the EDA. Weekends' daily seasonality is also different than daily seasonality of the rest of the days, as found in the EDA.

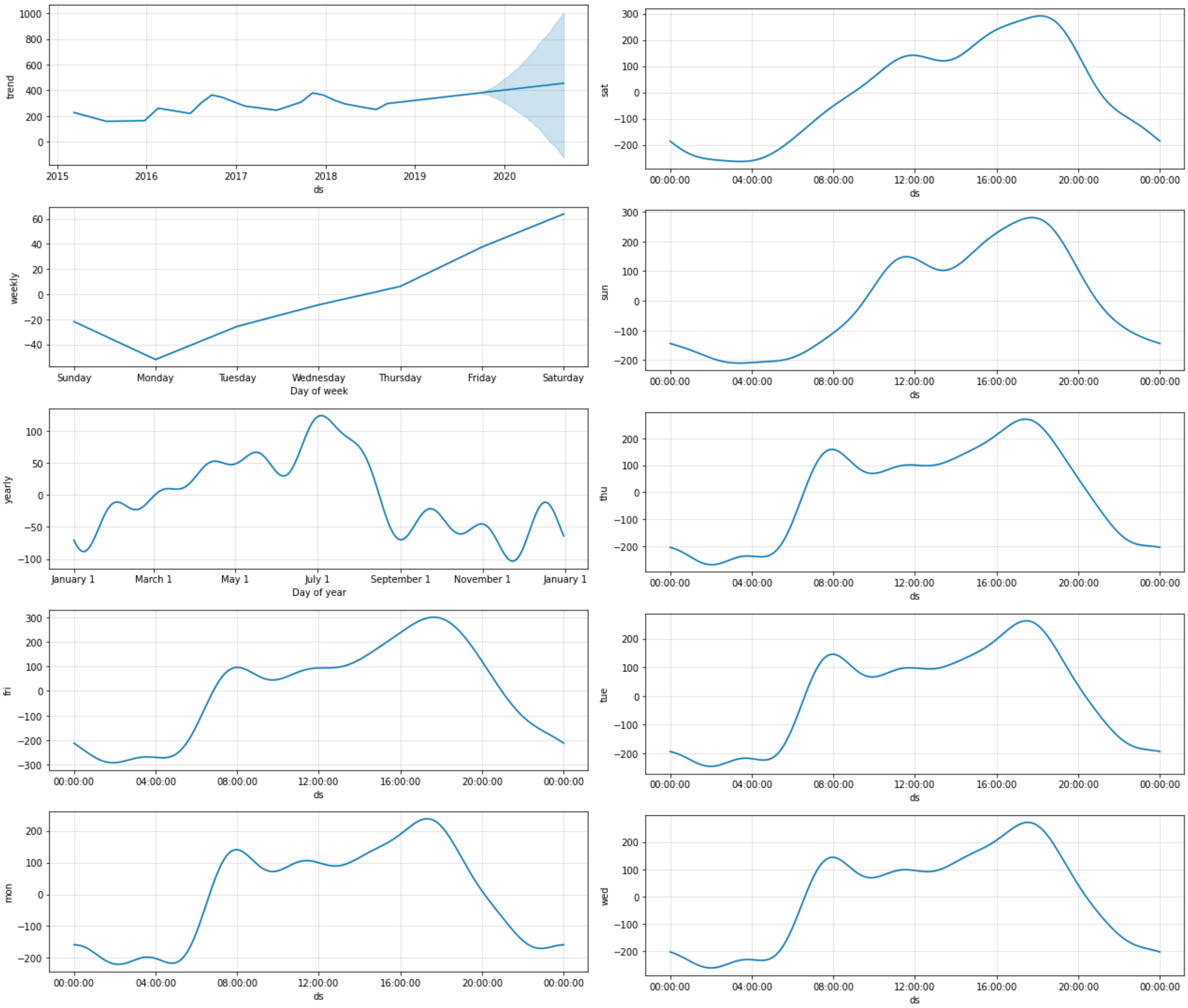


Figure 72. Facebook prophet model components for A34 Est dir- Gorizia model.

Error metrics for regional roads locations are presented in Table 13. Some of the locations have higher error because data quality for those was worse (more missing data).

Table 13. Error metrics for Facebook Prophet traffic volume predictions, 24-hour horizon.

Location	MAE	RMSE	MdAPE [%]
Rabuiese Valico Uscita	182.29	241.04	89.36
Rabuiese Valico Ingresso	121.06	171.73	41.01
Fernetti Valico Uscita	80.41	98.7	73.49
Fernetti Valico Ingresso	74.78	114.67	19.48
Prosecco Direzione TS	89.46	129.86	13.29
Prosecco Direzione VE	119.46	190.51	14.83
A34 Ovest dir- Villesse	59.35	91.04	14.39
A34 Est dir- Gorizia	59.58	89.59	16.45

To better understand the model and predictions, error metrics were calculated as presented in Table 14, related to the location and hour of the day. Mean absolute errors are higher during the times of the day with higher traffic volume.

Table 14. MAE for Facebook Prophet traffic volume predictions, 24-hour horizon.

hour	A34 Est dir. Gorizia	A34 Ovest dir. Villesse	Fernetti Valico Ingresso	Fernetti Valico Uscita	Prosecco dir. TS	Prosecco dir. VE	Rabuiese Valico Ingresso	Rabuiese Valico Uscita
0	47.4	44.8	40.1	33.6	53.4	98.8	56.2	66.5
1	42.4	44.4	31.2	31.0	70.7	118.4	65.5	67.7
2	39.4	41.6	31.3	29.0	58.9	105.2	76.2	73.9
3	45.9	40.3	40.0	26.1	73.1	106.9	79.7	76.0
4	46.5	36.9	33.2	23.8	72.4	99.1	66.2	67.3
5	43.4	44.2	50.1	31.7	76.5	93.0	70.5	75.0
6	45.2	51.8	65.1	41.8	78.2	101.2	63.3	79.3
7	50.9	79.9	105.1	49.3	123.6	115.8	86.2	114.6
8	68.2	81.5	88.6	69.7	137.9	147.3	91.8	167.8
9	55.2	71.7	98.6	72.1	105.0	99.6	107.4	205.4
10	49.8	69.2	108.6	80.1	73.1	131.2	143.9	249.7
11	61.1	64.4	101.7	71.6	93.8	90.8	151.4	264.0
12	58.8	76.4	110.8	83.0	105.9	82.2	129.8	252.9
13	66.2	71.8	112.3	71.2	54.7	134.7	118.5	233.8
14	73.4	80.6	99.1	85.8	70.4	132.0	129.0	228.2
15	73.7	85.3	111.8	94.8	79.2	150.2	153.8	218.5
16	95.2	80.5	107.0	110.3	111.8	158.1	154.2	256.3
17	90.3	74.9	102.5	111.0	135.8	167.1	149.3	261.3
18	92.8	63.3	85.3	105.5	126.0	142.8	126.4	227.3
19	70.1	40.7	71.3	80.6	90.6	132.4	101.0	181.6
20	71.6	41.9	59.3	51.8	72.2	108.7	77.4	128.4
21	49.4	43.2	44.9	42.8	65.5	124.2	72.0	70.6
22	42.2	44.6	52.0	42.4	64.1	117.4	70.7	63.1
23	45.2	48.5	44.8	38.8	55.5	103.8	51.5	67.6

5.2.1.3. XGBoost predictions

Structured dataset for XGBoost method[57] is assembled from the lagged features, cyclical encoded timestamp features via sinus and cosine transformation, Boolean (0/1) features and classical numerical features. Lagged features are all values from the last 24 hours and values from 2, 3 and 7 days ago at the same time. The month of the year, day of the week and hour of the day are encoded using cyclical encoding. Weekends and holidays are represented using Boolean features. A year and our target value are classical numerical features.

Example of 168 hours horizon (7 days) with actual traffic volume and prediction is presented in Figure 73. Prediction is good and catches the trends and daily seasonality. For this example, MAPE is only 10% for 7 days horizon.

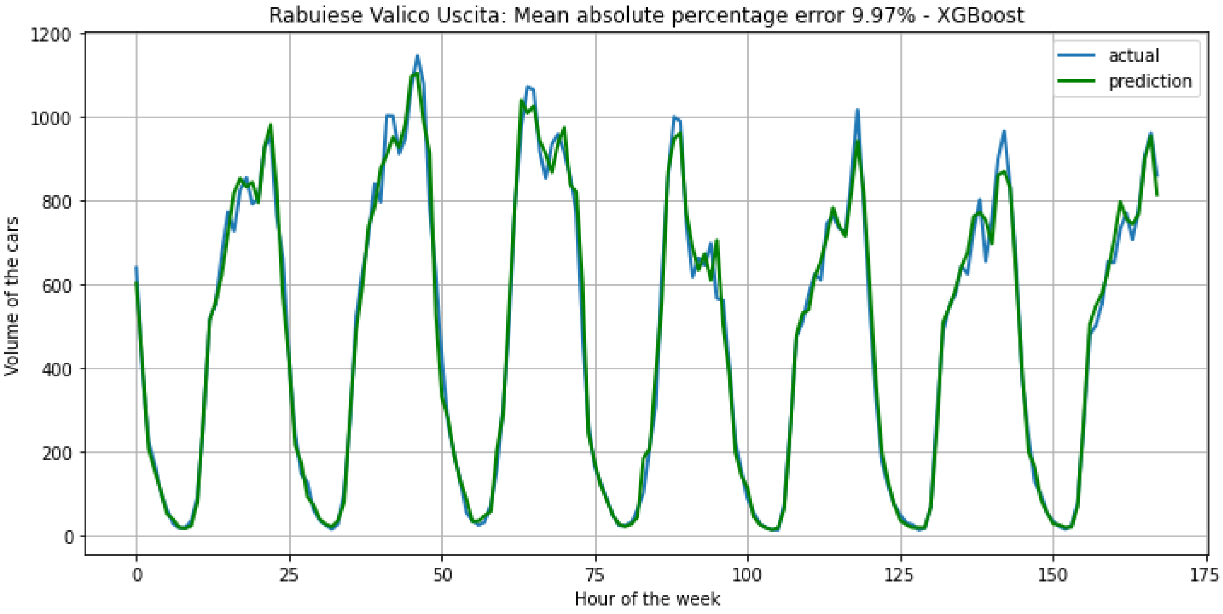


Figure 73. Example of 7 days horizon forecast by XGBoost method with 9.97 % MAPE.

MAE, RMSE and MdAPE error metrics for XGBoost predictive algorithm on horizon 24 hours, are presented in Table 15. Results are significantly better than with Facebook Prophet.

Table 15. Error metrics for XGBoost traffic volume predictions on regional roads, 24-hour horizon.

Location	MAE	RMSE	MdAPE [%]
Rabuiese Valico Uscita	27.97	52.3	9.19
Rabuiese Valico Ingresso	29.83	56.4	10.33
Fernetti Valico Uscita	23.51	35.55	11.95
Fernetti Valico Ingresso	29.15	52.8	8.03
Prosecco Direzione TS	42.09	66.97	7.19
Prosecco Direzione VE	39.06	62	6.23
A34 Ovest dir- Villesse	28.37	44.48	6.69
A34 Est dir- Gorizia	26.76	43.79	7.47

MAE for the location and hour of the day is presented in Table 16. As expected, the error is lower during the night, because there are fewer cars on the roads.

[57] https://xgboost.ai/

Table 16. MAE for XGBoost traffic volume predictions on regional roads, 24-hour horizon.

hour	A34 Est dir. Gorizia	A34 Ovest dir. Villesse	Fernetti Valico Ingresso	Fernetti Valico Uscita	Prosecco dir. TS	Prosecco dir. VE	Rabuiese Valico Ingresso	Rabuiese Valico Uscita
0	16.5	18.8	13.0	14.6	25.1	21.5	11.9	9.5
1	13.0	19.2	15.0	8.7	19.6	21.7	8.4	6.4
2	12.8	18.6	11.9	11.5	17.3	18.1	6.0	5.9
3	9.3	17.8	13.9	10.2	13.6	17.8	6.1	6.0
4	14.8	16.2	17.4	11.0	15.7	26.9	8.3	8.7
5	15.3	19.1	21.6	15.7	21.0	25.9	14.4	11.5
6	19.0	31.5	31.2	22.2	34.3	39.5	18.0	25.6
7	27.3	40.9	34.9	26.9	49.0	44.8	31.2	36.5
8	37.0	43.9	42.6	25.1	62.2	46.4	35.1	33.2
9	40.8	34.2	40.3	27.3	50.1	49.5	49.1	45.2
10	37.4	36.7	33.0	25.6	45.7	62.4	53.4	45.4
11	26.2	36.3	40.0	25.6	54.1	41.6	39.5	42.6
12	28.4	36.2	40.6	27.7	52.8	45.0	53.1	58.9
13	35.0	33.9	44.3	28.2	51.0	57.0	43.2	37.3
14	28.2	29.1	32.7	32.6	41.7	56.8	44.2	36.5
15	37.2	30.6	35.5	29.7	46.9	51.2	39.4	31.2
16	41.7	35.0	34.0	31.9	64.8	45.8	40.5	37.5
17	40.4	34.4	48.2	36.8	72.0	55.6	38.0	36.5
18	38.7	30.5	39.3	34.6	56.1	47.0	46.0	30.5
19	30.0	32.8	24.9	29.5	64.4	48.6	32.7	45.4
20	24.7	25.2	21.7	25.1	40.7	32.2	29.4	27.2
21	27.8	22.2	22.7	23.7	39.9	28.7	26.8	17.2
22	20.8	17.6	23.9	21.8	39.6	31.5	19.8	13.8
23	19.7	20.5	14.3	14.9	32.7	21.9	18.5	11.9

5.2.2. PPA

5.2.2.1. Exploratory Data Analysis

Before making any predictions, data was analysed. The objective was to better understand the typology of the data, internal patterns, seasonality, peaks and others. For this, the base information collected by HERE is used.

A standard data format allows to reuse part of the analysis, performed on ASPM/SDAG use case and being easily applicable to others. All the scripts necessary to process the data and to develop the analysis have been developed in Python and with the help of Jupyter Notebooks, such a way that development can be exported in a more didactic way.

It should be stressed that in the use case of PPA, there are many different locations within the area of interest. Some of the charts have been applied to a specific location, while others go through all locations. It is important to highlight that the rest of the locations can present behaviour, but in general, they present the same patterns internally. The location for this analysis and for which the forecasting is done in "Δραπετσώνα", close to the port and affected, as it was proven, by port activity.

One of the first analyses that took place, was to represent the information of the time series in a linear graph that included the entire period for which the record is available, as can be seen in Figure 74.

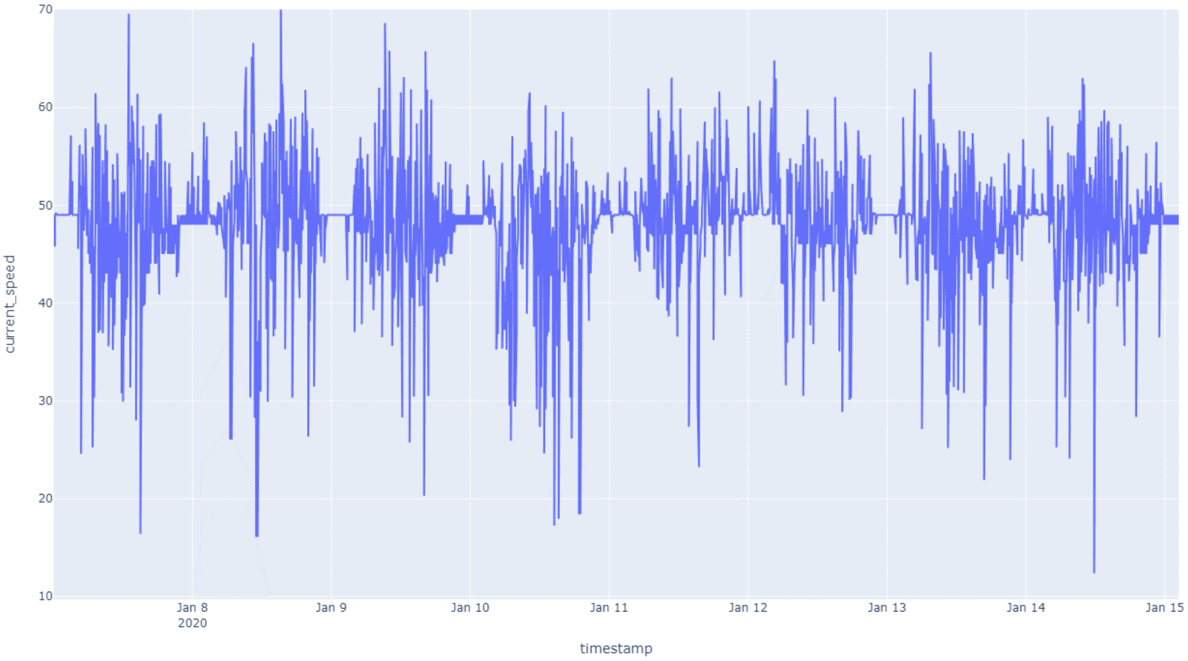


Figure 74. Line plot of speed per 15 minutes recorded for our location.

Figure 74 has been generated with Plotly, allowing to dynamically select and deselect locations, as well as to increase the temporal precision, in this case up to the maximum available of 15 minutes. In this representation is impossible to see the seasonality of the data, so a series of heat maps and line graphs will be presented below, that will allow it.

In Figure 75, information from all locations is represented, the mean speed for each hour of each month. In the case of having more than one record for the same hour, these values were grouped.

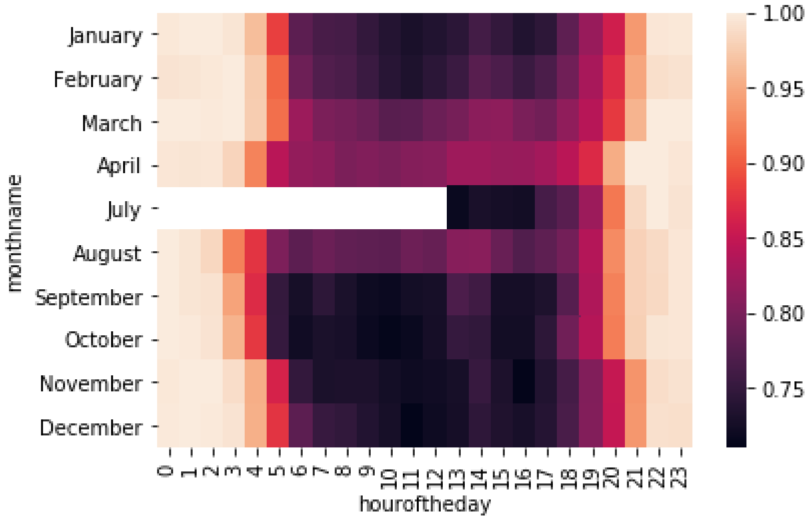


Figure 75. The traffic average speed across different months and hours of the day.

As can be seen on the heat map (Figure 75), at the corresponding hours between 5 AM and 7 PM, there is a noticeable reduction in speed, corresponding to working hours. Also, there is a difference throughout the months, with greater variance in the summer months.

Beside variability throughout the year, with the use of a similar heat map, but grouping the information for the days of the week, the conclusions of seasonality are even clearer, as can be seen in Figure 76.

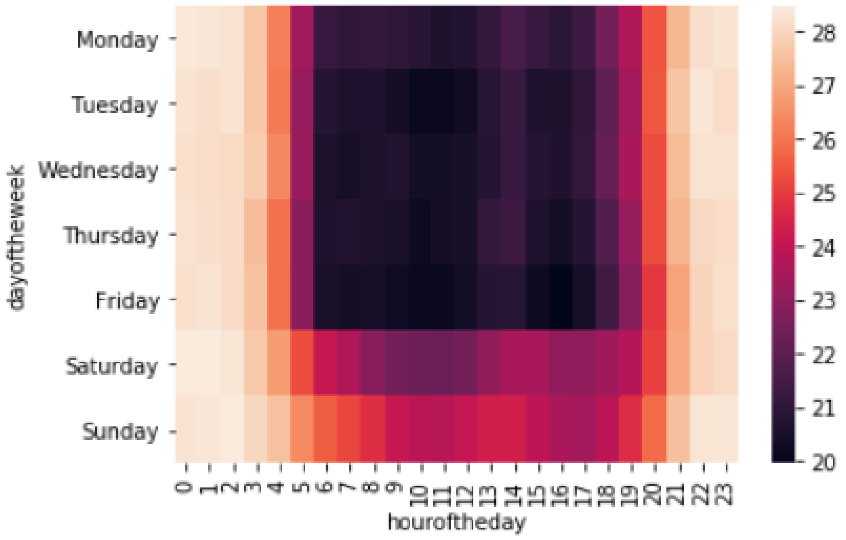


Figure 76. the traffic average speed across different days of the week and hours of the day.

On this occasion, it is evident how on working days, the traffic density is much higher compared to the weekends, when fluidity is practically total. Also, on weekdays, it is visible how the hours between 7 AM and 9 AM and 4 PM and 6 PM, are the times of greatest congestion, normally corresponding to job migrations.

Next, all the graphs and analysis presented will be from the previously selected location. Another similar way in which to be able to observe the seasonality of the data, both weekly and yearly, is by representing the information in a line graph, with time as the horizontal axis. In Figure 77 you this behaviour is reflected, that had been presented in the general heat map, depicting how working days generally show less fluidity compared to weekends. Also, seasonality throughout the day is observable, even in which the working hours between 7 AM and 7 PM, are again those that represent lower average speed values.

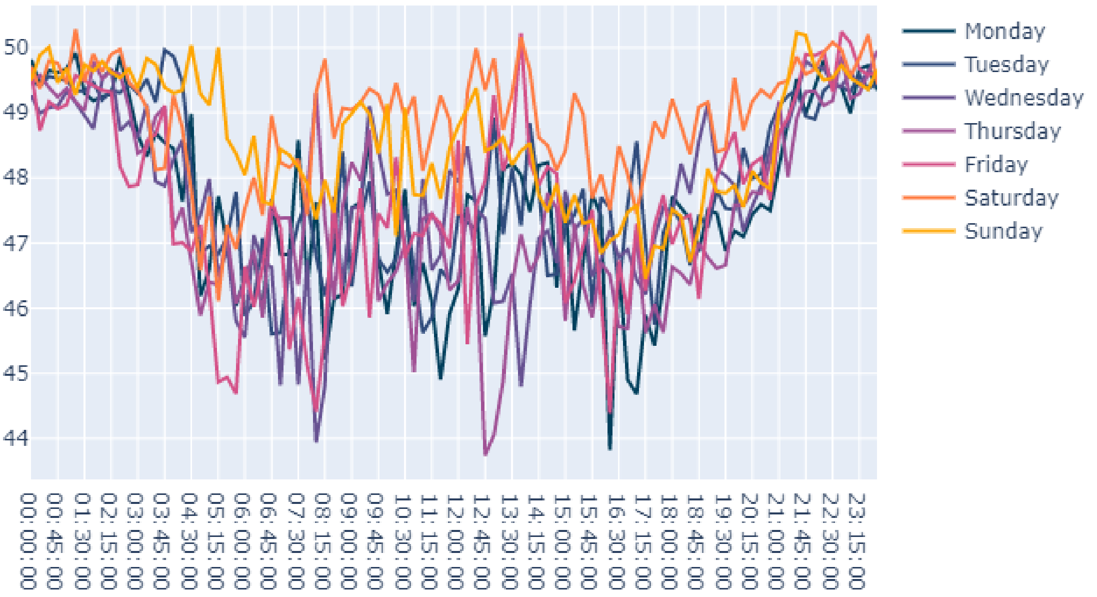


Figure 77. Speed variations per hours for different days.

If weekly information is grouped into two categories, business days and weekends, the behaviour described above becomes much more evident, as can be seen in Figure 78.

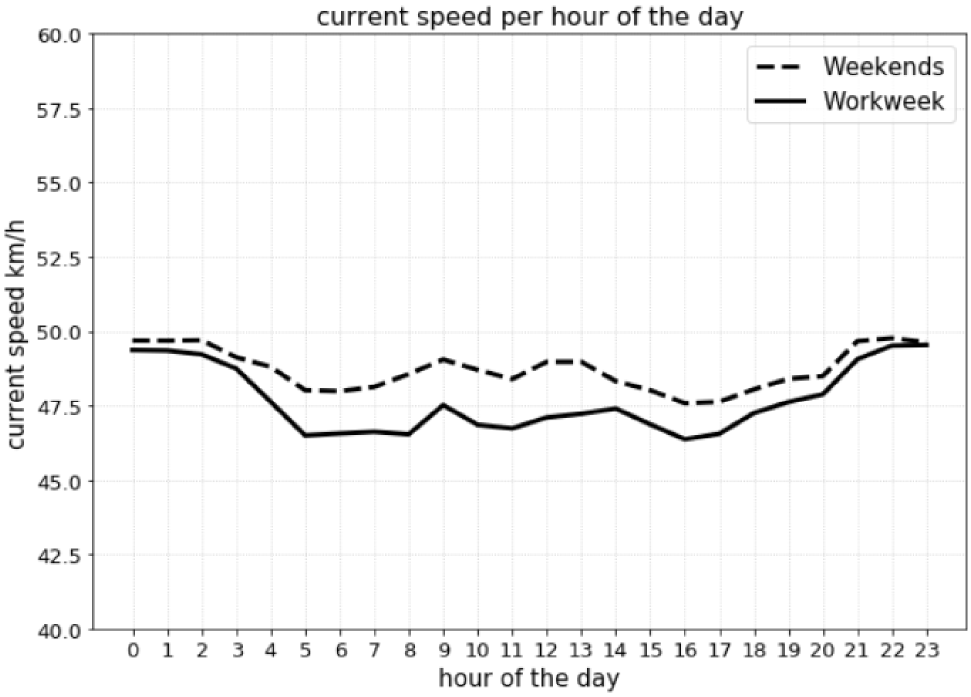


Figure 78. Speed variation per hour of the day between work and weekend days.

Finally, the last part of the analysis is to describe the behaviour of the traffic, for one day of the week throughout a year. It is clear how a Monday in January, presents a different behaviour than one in August, mainly as it is a holiday month, in which road traffic is usually less than the rest of the year. Figure 79 shows the different behaviour for Tuesday and throughout the year.

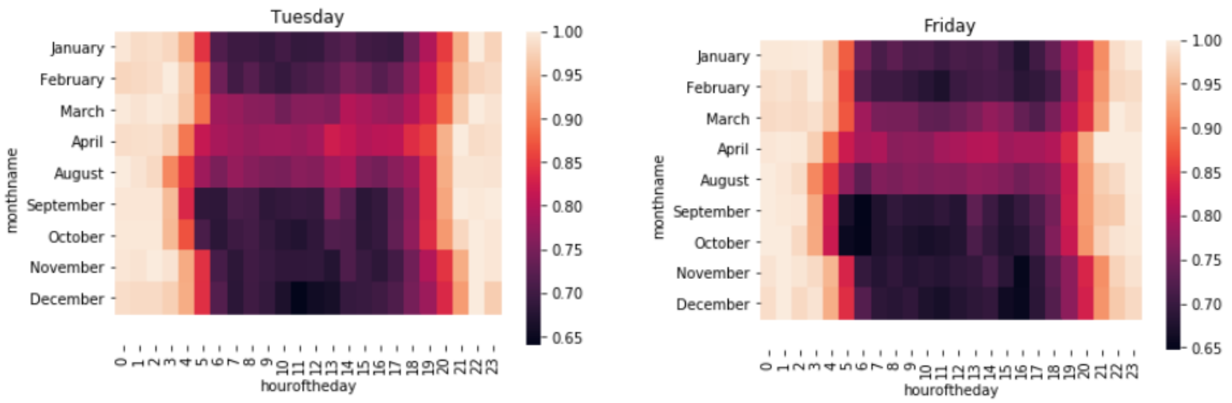


Figure 79. Heatmaps of speed per hour of the day for one specific day across the year.

5.2.2.2. Facebook Prophet predictions

Once all the base information has been analysed, the predictions made with Prophet, time series forecasting algorithm, will be described.

Three different sections will be considered depending on the type of data used:

a. Baseline traffic data
b. Baseline traffic data plus weather data.
c. Baseline traffic data plus cruise ships data.

First, it should be noted that national holidays have been manually entered as input to the model, since in this way if a variation is observed in behaviour on those dates, the model will associate the cause with said variation and perform better estimations. Also, a daily seasonality has been added, to be able to observe the weekly and daily seasonality described above, but not using statistical analysis, as previously performed. For each of the use cases, different training and test datasets have been chosen due to the availability of the data, but all the procedures have been applied for the different sections.

Baseline Traffic data

For this occasion, the available historical data from August 2019 to April 2020, was used. It was decided to use the last month as a test, while the first seven months as train data. It is important to highlight that, with this type of data, time series, it is important to have a history of all periodicities, whether daily, weekly or even annual, since the situation may arise, that having used the last month of April for testing and not having the information for that month for training, then the seasonality of that month is not learned and therefore the predictions are of poorer quality.

Figure 80 represents the division described above. The black dots are real traffic data points that have served to train the model and the blue line is the model once trained, with their respective upper and lower margins. In this case, a prediction horizon of the same length as the test horizon has been selected.

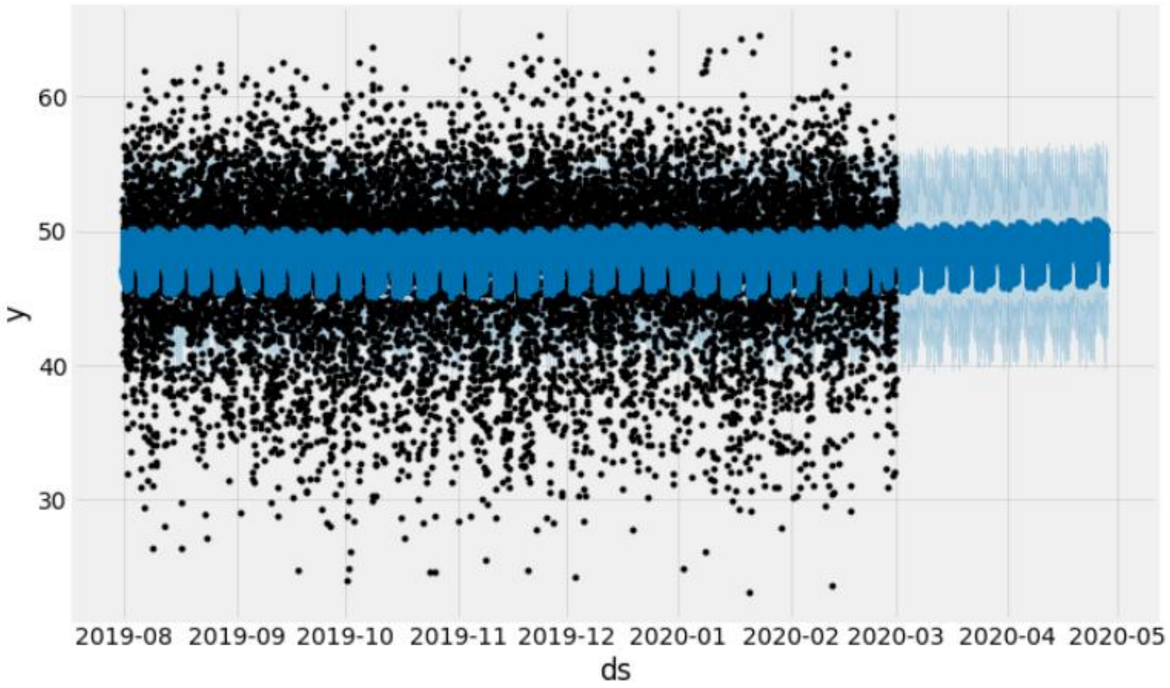


Figure 80. Baseline data forecasting.

Figure 81 describes the particularities learned by the model. Firstly, the general trend of speed, which even though it has some variability, is minor. Anyway, the trend is in a decline since the beginning of the year. Secondly, the impact that holidays have, is also presented. Additionally, weekly seasonality is presented. It was also noticed that on average, speed is lower during working days.

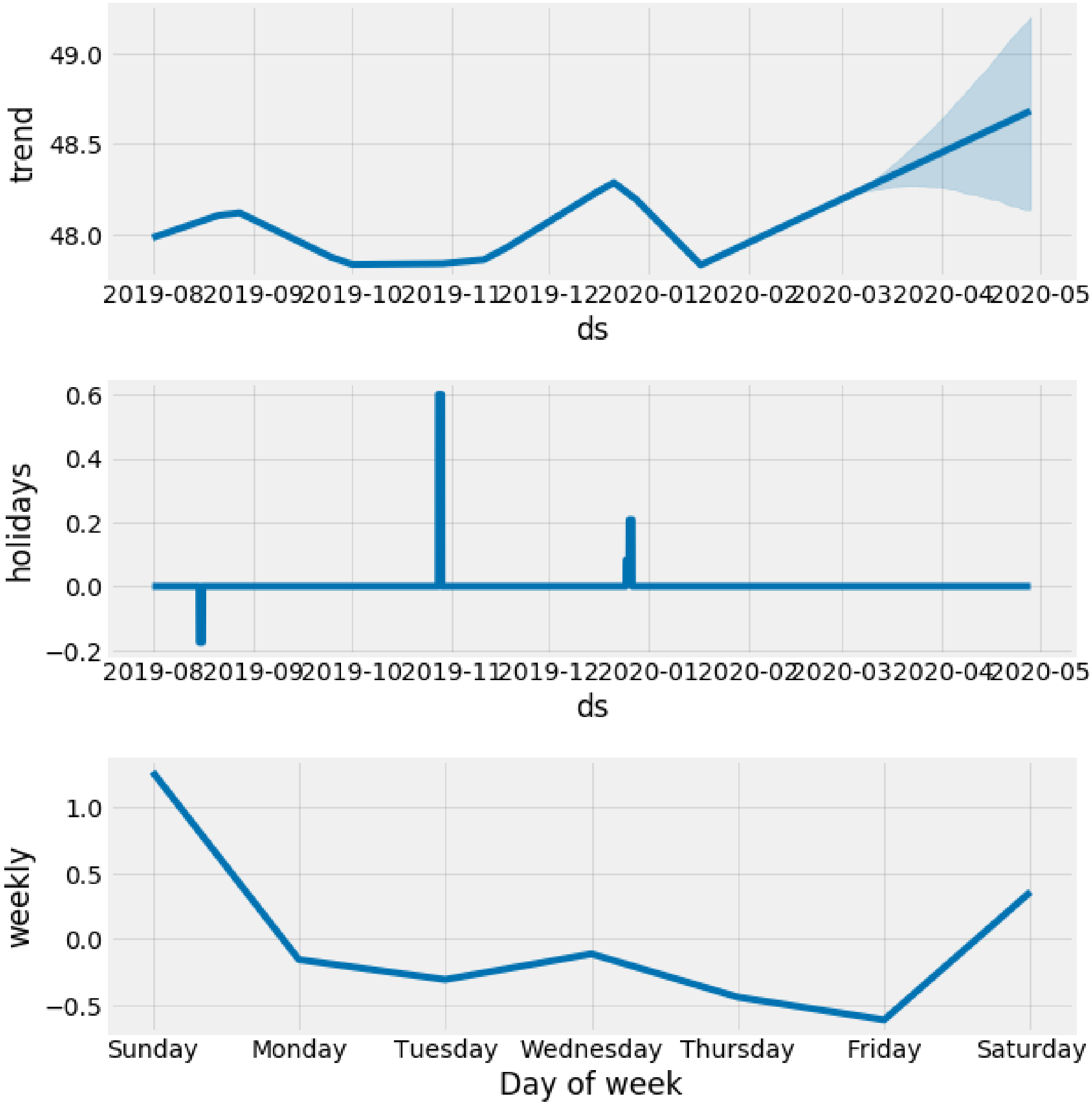


Figure 81. Components of the trained model.

Finally, daily seasonality is also shown. The behaviour reflected in the previous analysis is seen, in which, for the working days the flow of traffic is considerably reduced, compared to the rest of the days. The daily seasonality is matching with results in the EDA section, emphasizing the viability of Prophet for predicting time series, due to its ability to detect seasonality in the data.

Figure 82 shows the previously described seasonality. It is important to mention that, the temporal precision of the predictions, is the same as that of original data - 15 minutes. For all these graphs the timeline is presented on the horizontal axis, while for the vertical km / h.

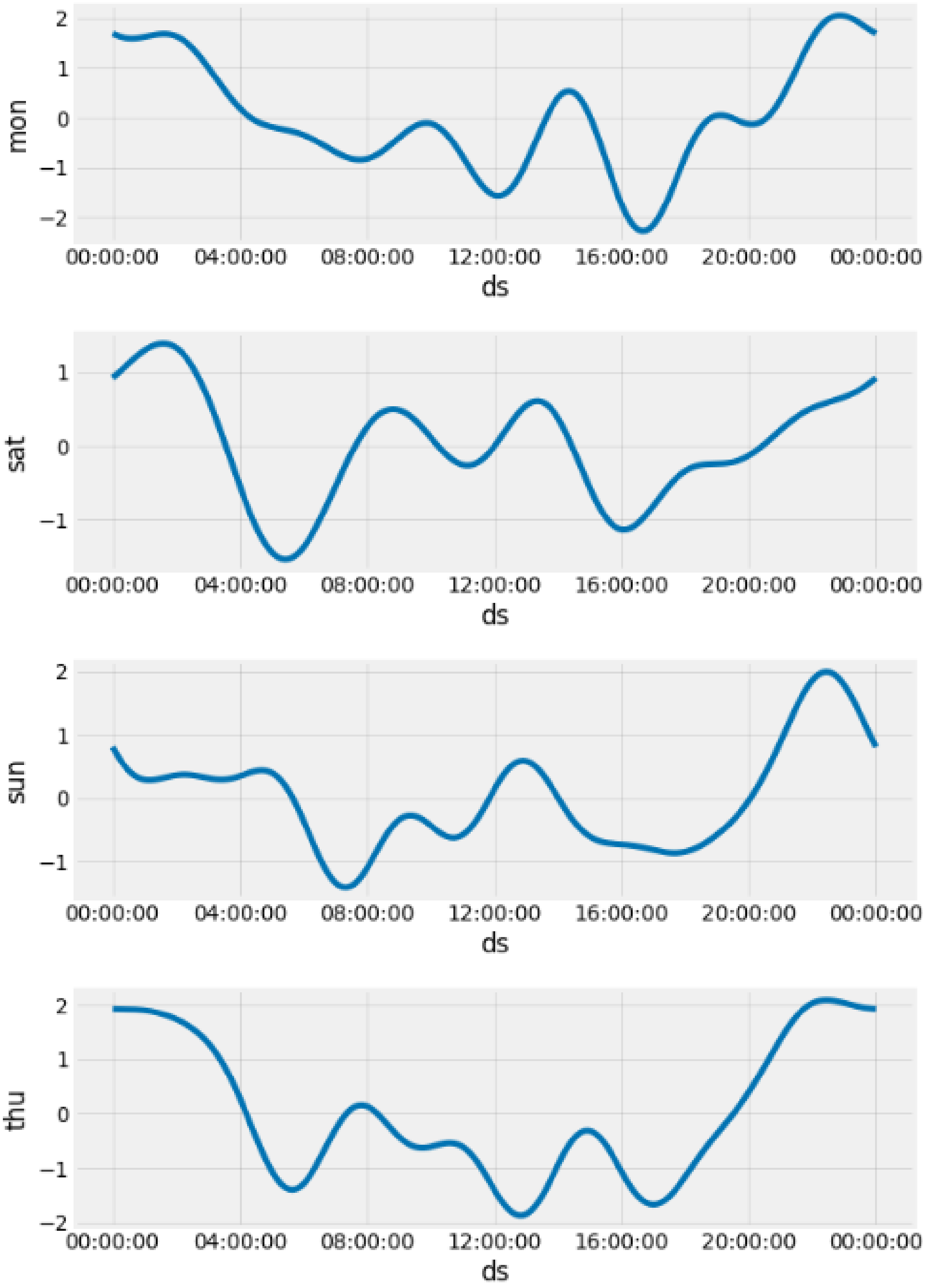


Figure 82. Daily seasonality.

Finally, in Figure 83 and Figure 84, there are two joint plots, the bivariate plot with marginal univariate plots, to represent through scatter plots of the predicted values and the respective real values. In addition to each of the variables, their distribution is presented. It should be noted that the *yhat* value corresponds with the predicted values and *y* with the real values. The two graphs correspond to both the training and test data.

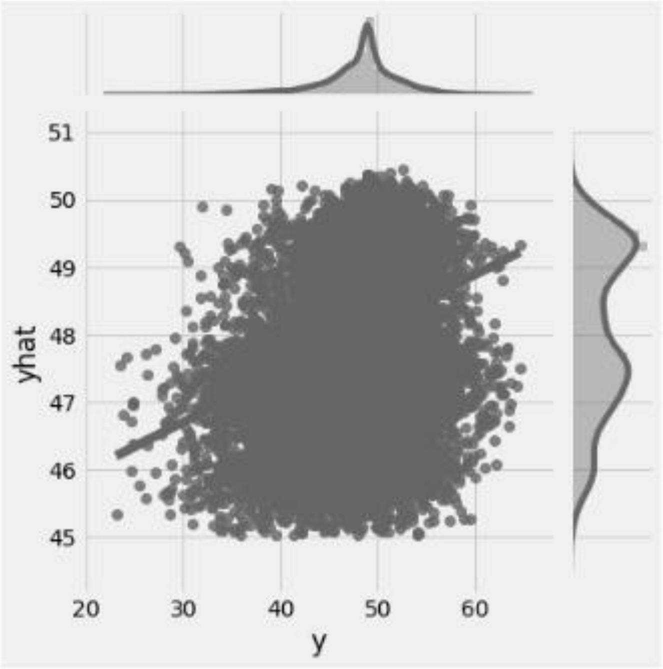


Figure 83. Joint plot train set.

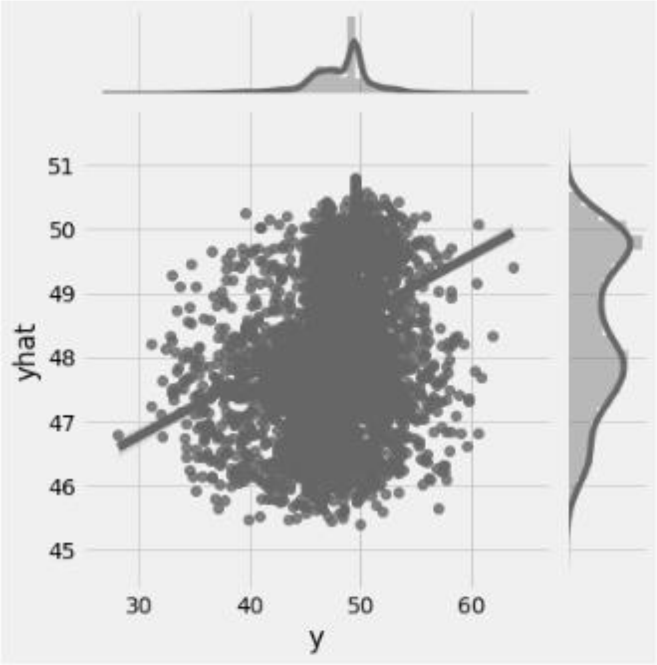


Figure 84. Joint plot test set.

As can be seen in both graphs, both variables present most of the records concentrated around 30, so the predicted and actual values agree.

Besides, Figure 85 depicts the distribution of differences for test data, between predicted and actual values. In this case, speed in km / h is presented in the horizontal axis, while the normalized frequency, in the vertical axis. As observed, the distribution is centred around zero and flattens on the sides, hinting that errors tend to zero.

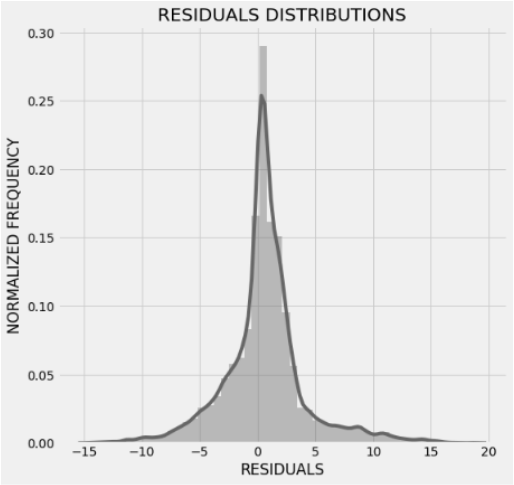


Figure 85. Residual distribution.

Finally, in Table 17, a series of metrics are presented, that will allow comparing the different approaches. As it can be seen the values obtained in training, are better than in the test part, one of the consequences of entering data, for which the model has no previous records, such as March data.

Table 17. Summary of principal metrics of the baseline forecast.

Metric	Train dataset	*Test dataset*
Daily average MAE	3.00	2.41
MSLE	0.0091	0.0060
MSE	19.325	12.989
MAPE	6.62%	5.32%
Residuals median	-0.528	
Residuals mean	-0.618	

It should also be noted that cross-validation has been carried out for the evaluation of the model. Two have been performed, one with the 1-day forecast horizon and the other with a 4-day forecast. Figure 86 and Figure 87 represent these configurations.

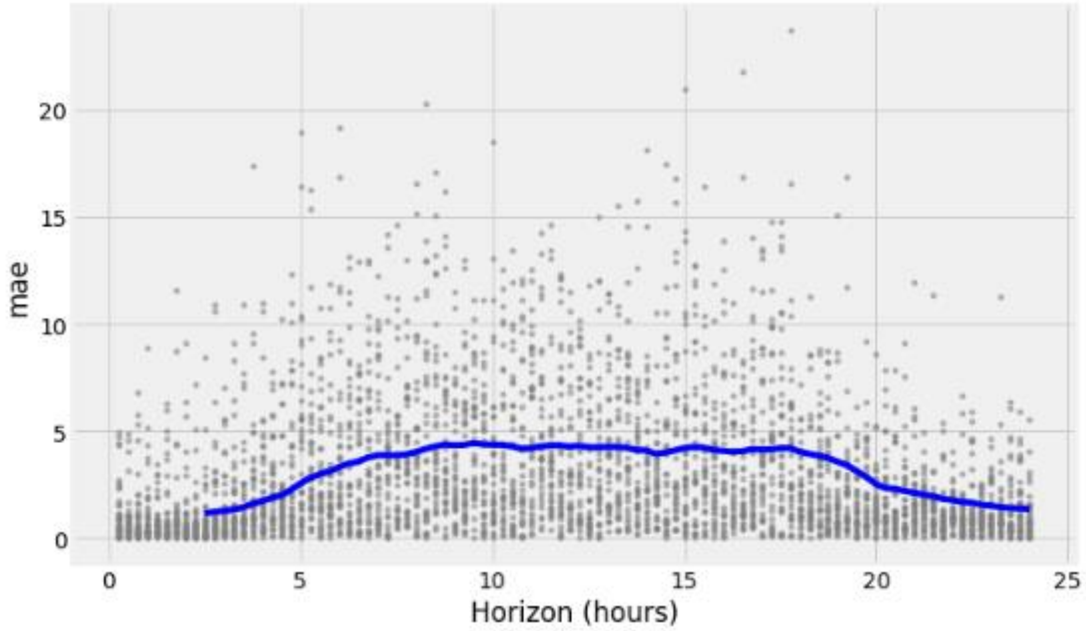


Figure 86. Cross-validation for the 1-day horizon.

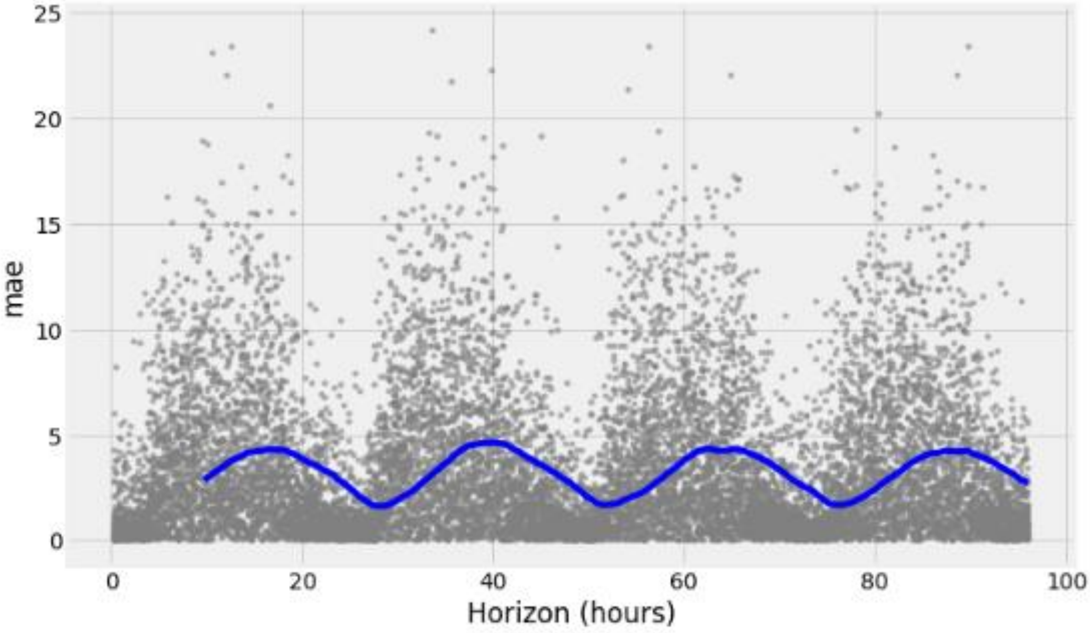


Figure 87. Cross-validation for the 4-day horizon.

Baseline Traffic data plus weather data

Once analysis with the base traffic data is finished, it is time to add new attributes to the model that allow explaining part of the variability of the data and therefore achieve a more accurate model. One of the main data sources to work with, are meteorological ones. It is logical to think that when climatic conditions are adverse, the flow of traffic is reduced, as insecurity, congestion and even the risk of accidents increase.

This information, as previously described, has been obtained through a procedure like traffic information gathering, through services that allow via API to obtain an optimal number of records.

This data also required transformation to the proper format, that allows it to be entered into the algorithm. The procedure followed has been the same as with the base information, both cross-validation and division of the data between training and testing. Specifically, the division has been the same as with the base information since the same history is available and both Linux services were launched at the same time.

The attributes considered are the following: intensity and probability of precipitation, temperature, and wind speed. The reason for each of these is simple. Precipitation intensity because it is a real value, the amount of water hinders the speed of traffic so if it is expected to have a real impact. As for the probability of precipitation, since it is a probabilistic value, it may have an inference in the population; if people observe precipitation values of 90% for the next day, they may dispense with public transport and it may happen that finally, precipitation does not occur. It has also already been considered that low and high temperatures can prevent the use of public transport. On very cold days, people may decide to do without the bicycle or public transport, as well as on particularly hot days. Finally, airspeed is also incorporated, as it is also a factor that can determine the choice of transport to move. The result after the inclusion of weather attributes is presented in Table 18, Figure 88 and Figure 89 for 1 and 4-day horizons.

Table 18. Summary measurements of prediction with weather data.

Metric	Train dataset	*Test dataset*
Daily average MAE	3.03	2.54
MSLE	0.0093	0.0067
MSE	19.628	14.358
MAPE	6.68%	5.62%
Residuals median	-0.30	
Residuals mean	0.50	

Indeed, the precision has decreased. Although these additional attributes may explain exceptional behaviour at specific times, they generally add more noise to the model, preventing it from learning base traffic behaviour, reducing its accuracy. Some further analysis would be to test these new attributes.

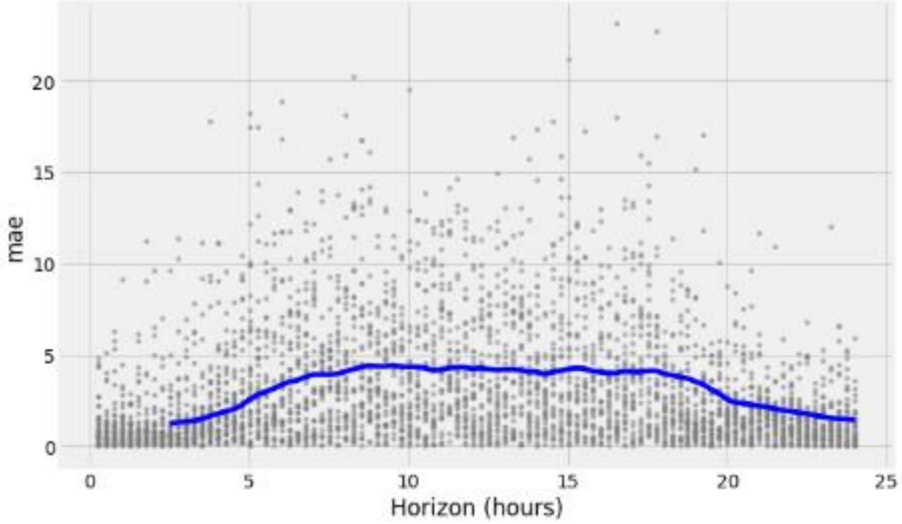


Figure 88. Cross-validation for the 1-day horizon.

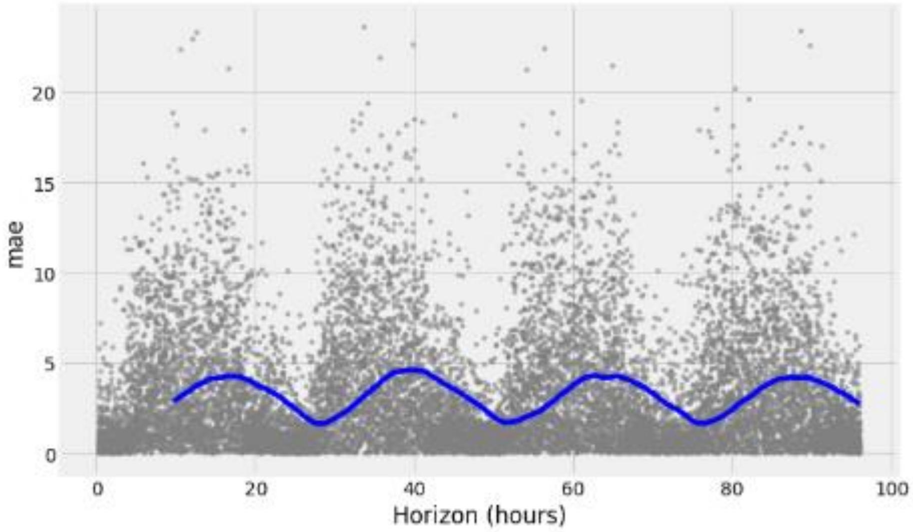


Figure 89. Cross-validation for the 4-day horizon.

Baseline traffic data plus cruise ships data

Finally, data of passenger ships that arrive at the port of Piraeus will be included. As it was described in D4.3, one of the objectives of the port, was to estimate the impact that tourist buses have on city traffic. It should be noted, that the port is one of the busiest in cruise ships in the world, so, to displace these passengers, the high number of buses that can suddenly enter the circulation, can cause more than significant reductions in traffic flow.

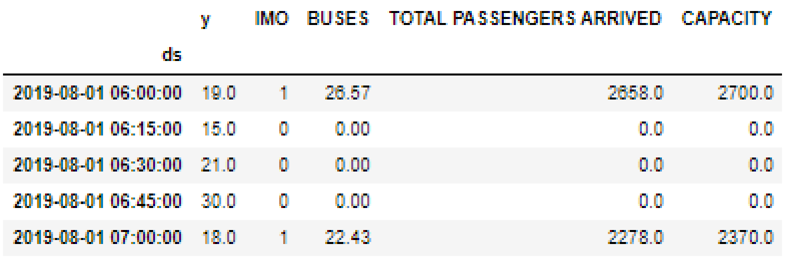
In this case, the information comes from a manual extraction carried out by port operators, which includes each of the passenger ships that arrive at the port, as well as a series of descriptive fields, such as the number of passengers, capacity, as well as the number of buses needed. It is important to highlight, that this is the estimated number and not the actual data since this parameter has been obtained by dividing the number of passengers arriving by 100, it is estimated to be the average capacity of the buses used. This aspect must be taken into consideration since as there is a linear relationship between these two variables, they will not contribute any relevance to the model as both contain the same information, only one is proportional to the other. However, it has been decided to incorporate both into the model to estimate their impact anyway.

The following attributes have been retained: number of passengers that have arrived, the capacity of the ship and number of buses on the ship, as well as a column called IMO. In this column, the number of ships arriving at the port in the relevant time slot is filled in.

At this point it is important to highlight the research work that has been carried out to study the distribution of passengers, leaving once the cruise ship arrives at the port. If a cruise ship requires fifteen buses, these will not immediately be distributed once the ship has arrived, but there will be a period between 1 and 2 hours in which the number of buses necessary will be distributed among traffic. The most plausible behaviour was found to be the same, as a chi-square distribution of 4 degrees of freedom, thus approximating the spectrum of the curve to the considered interval of 90 minutes in which, all passengers have disembarked. Being the mass function of the distribution that indicates the percentage of passengers, who have already disembarked.

However, although this approach to the treatment of information is more plausible, it had considerably greater errors than the direct inclusion of the data, and therefore it was unfortunately discarded.

In Figure 90 there is an example of how the information has been incorporated together with the speed data.



	y	IMO	BUSES	TOTAL PASSENGERS ARRIVED	CAPACITY
ds					
2019-08-01 06:00:00	19.0	1	26.57	2658.0	2700.0
2019-08-01 06:15:00	15.0	0	0.00	0.0	0.0
2019-08-01 06:30:00	21.0	0	0.00	0.0	0.0
2019-08-01 06:45:00	30.0	0	0.00	0.0	0.0
2019-08-01 07:00:00	18.0	1	22.43	2278.0	2370.0

Figure 90. Sample of cruise ship data integrated

For the first record, a drawback of proportionality described above between the number of buses and the rest of the parameters can be noted, which are practically proportional.

Another aspect to consider is the amount of data available, which covers the period from June to January, so the sizes of the training and test sets have had to be modified to fully match the current availability of the additional attributes. Specifically, the period from July to December has been used for training, while the test month of January has been used for testing.

One of the drawbacks described above is the lack of data history. Time series analysis is exclusively about detecting seasonality in the data and not providing a history that covers the entire period, accurate predictions cannot be made. A correct prediction of traffic behaviour in January cannot be expected if the model has not yet learned characteristics of these data. It is important to take this into account; use data that is representative of the population is vital, as it is the cornerstone of the supervised algorithms.

There are approximately 10.000 records for training data with the 15-minute time precision. Table 19 presents the results, together with Figure 91 and Figure 92 for 1 and 4-day horizons, respectively, as well as the approximations making use of the cross-validation for the different prediction horizons.

Table 19. Summary measurements of forecasting with cruise data.

Metric	Train dataset	*Test dataset*
Daily average MAE	3.15	3.01
MSLE	0.0098	0.0089
MSE	20.812	18.987
MAPE	6.964	6.62%
Residuals median	-0.315	
Residuals mean	-0.171	

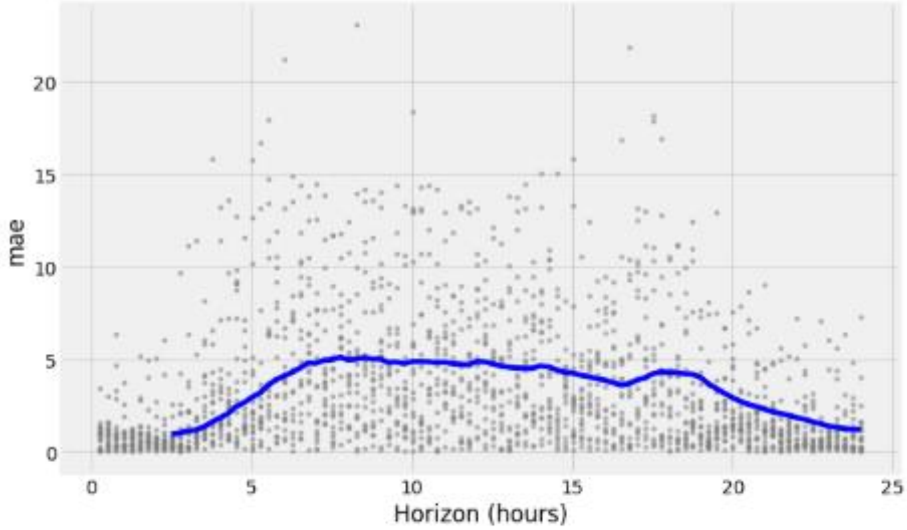


Figure 91. Cross-validation for the 1-day horizon.

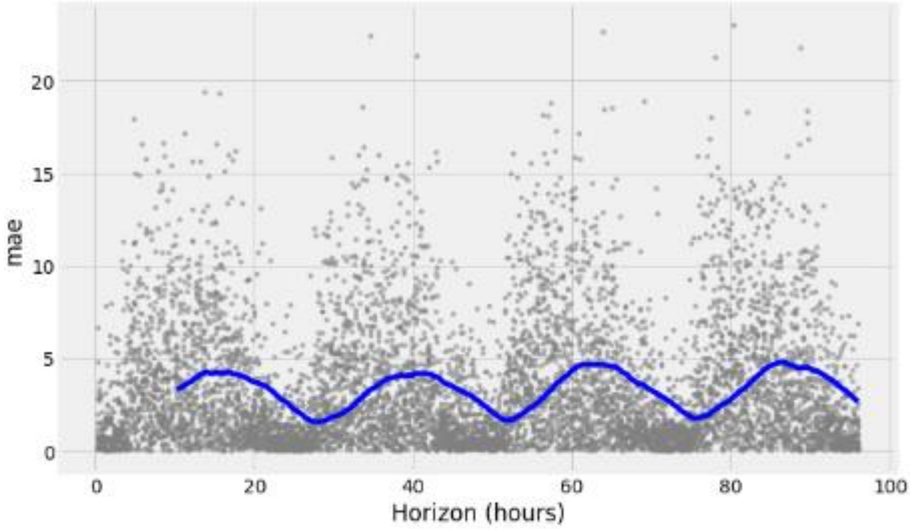


Figure 92. Cross-validation for the 4-day horizon.

Reported error metrics are slightly higher than the other approximations. In this case, there is an explanation similar to that of weather; the model is capable of predicting the impact that cruises have on the port, which is positive, since it is possible to obtain the behaviour of traffic hours before the arrival of a cruise, although nevertheless, it contributes a noise that, in general terms, as it has been verified, reduces the precision of the model.

The approximation that presents better results, is that with the base values. Although it can predict the behaviour and impact that passenger ships have, the error they contribute to the model can be considerable.

Finally, this work path must be fully applicable to any port, independent of the technology that is implemented, since it does not require sensors installed in traffic gates as in the rest of the use cases, since it only uses the free part, of different data providers. To achieve the predictions, the first step is data collection and then follow the process described.

. This first approach can allow different ports to see the value that this information can have in their operations and, therefore, motivate them to install devices that allow data to be collected in the areas of interest, instead of deploying technology, without knowing if it will be possible to obtain a value response.

Regarding this specific use case, it should simply be noted that, the approximation has been positive with exceptionally low error values, but that the availability of a greater historical amount of data is simply necessary, compared to the rest of the use cases that are working with more than one historical year of data (both in terms of traffic and cruise ships, in which the lack of data has not permitted the use of all the available history).

5.2.3. ThPA

5.2.3.1. Decision support tool

In cooperation with the Port of Thessaloniki visualization was prepared, that covers their needs to perform decision making, leveraging the volume prediction done by the Prophet model deployment with the historic traffic data. During that step, the technical team developed in Vue.js and using AmCharts graphics (with sample data) a tentative visualization composed of 4 equally sized sub-screens in a homogeneous layout to contain all the agreed information.

The following order maps to the Figure 93 layout with a left-right, top-down fashion:

- **Observation of the traffic in real-time**

Over the map of the Port of Thessaloniki, two circles corresponding to gates 10A and 16 are marked coloured following the GCI (Gate Congestion Index) rationale. ThPA will define certain thresholds that will be used to colour these points. Depending on the current volume of cars (considering values of the last 60 minutes) trespassing each gate, the GCI will be red (high congestion), yellow (average congestion) and green (light volume) and the circles will be filled accordingly.

- **Observation at a first glance of the current and predicted congestion levels**

In the second space, the GCI colour code is maintained but using a bar diagram. This will be used to analyse, briefly, how many hours in that day (model is executed every morning and predicts daily traffic) one gate will have congestion. This diagram will have a selectable to pick a location (there are 4 for Thessaloniki – Gate 10A entry, Gate 10A exit, Gate 16 entry, Gate 16 exit).

- **Trend (evolution) of the traffic (with real past data and predicted data)**

A combined view in a linear graph of the data measured and the predicted. The dividing line will be advancing as the day will advance.

- **Horizon 1 day of the traffic congestion per hour (semaphore-like)**

As mentioned, the model will be scheduled to run every morning to predict the traffic of that day. Initially, for ThPA, the horizon is 60 minutes as this has been the selected baseline granularity. In this last screen, evolution is seen as a snake that changes its colour depending on the probability of high congestion.

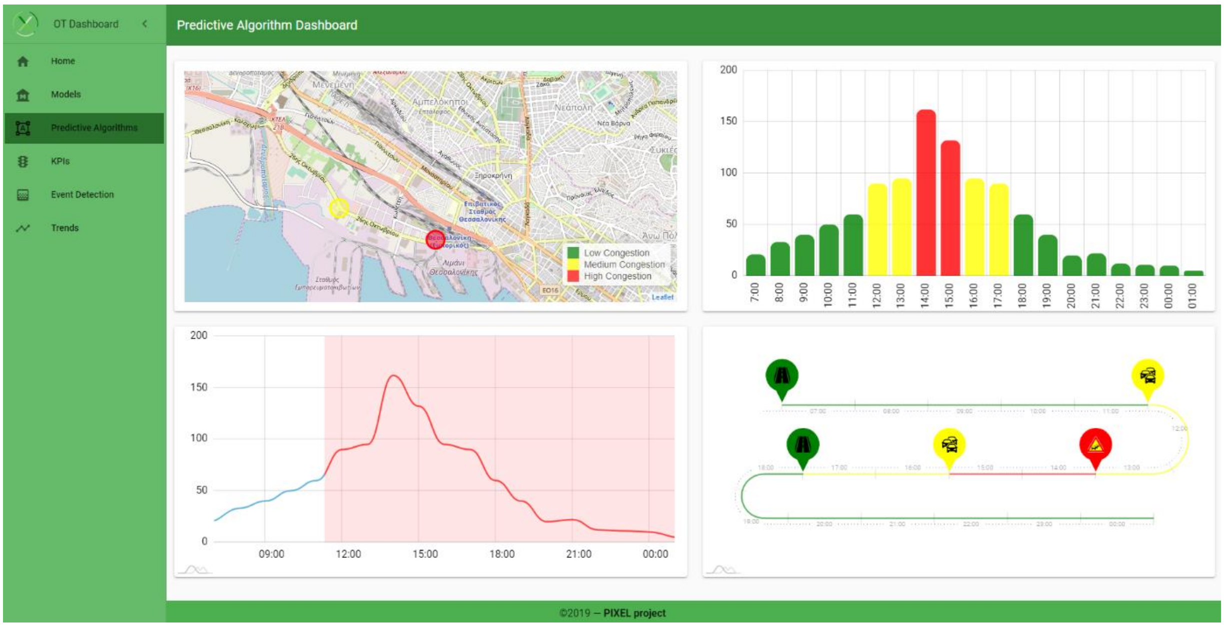


Figure 93. Visualization for the decision-support tool from predicted traffic volume in ThPA.

5.2.3.2. Forecasting with baseline dataset

Exploratory Data Analysis

Data were acquired from two gates in the Port of Thessaloniki from 28th April 2018 to 5th February 2020 and aggregated as count of vehicles in 1-hour periods. The maximum number of vehicles passing the gates at a single location in a single direction is 138, the average per hour is 70 vehicles and certain seasonality peaks per month (September and January).

The technical team (aligned with the other implementations in ASPM and PPA) embarked in a deeper analysis observing the data:

- Global variations per month.
- Global variations per day
- Depending on location (gate, entry or exit), analysing the variations on the average per day.
- Behaviour on the same day of the week during a year

For this analysis, the technical team assumed that there is no major difference among locations (or gates of the ports) on their behaviour within a different month. It means, it was assumed that whatever effect happening in a certain month (e.g. holidays in August) would affect the same way to Gate 10A exit than it would to Gate 16 exit.

With that in mind, there was no need to analyse separately the data per month and location, and it was enough with making a sum of the traffic of all 4 locations and the average of this sum in the different repetitions of the month, along the whole dataset period.

It was decided to use a heatmap to represent the variations of volume/hour per month. According to what was depicted in Figure 94, there were some interesting reflections to be extracted: (i) January, May and September are the months with more traffic, both in volume and in the hours in which the volume is higher (till 5 p.m.). (ii) Traffic starts at 5.30 a.m. and lasts till 9 p.m. (iii) the centric hours of the day are the ones in which the traffic is higher, especially between 12 and 2 pm, the maximum rates of traffic are concentrated. (iv) traffic at the gates is higher during the active morning hours than during active afternoon working hours. (v) Months with less traffic are April and November.

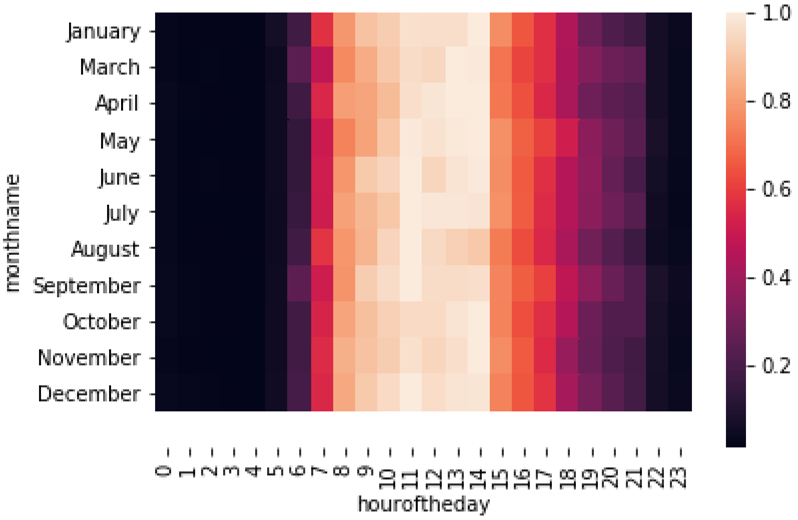


Figure 94. Global variations of volume/hour per month.

According to what Figure 95 showed, the following facts were realised: (i) Sundays are mostly inactive with regards to traffic, which was logical to expect, as it is a generally free labour day. (ii) Similar applies to Saturdays, which experiences more traffic for different reasons, such as some office works. (iii) Tuesdays and Fridays are the days with more traffic, possibly due to the need to process with more urgency, the pending cargo/activities before finishing the labour week. (iv) for the rest, it is noted that the days within the week, keep an almost identic distribution per **hour considering the sum of the traffic of the two gates at both directions**. To extract more interesting conclusions, it is mandatory to drill down the graphs and observe the behaviour per gate and direction of traffic.

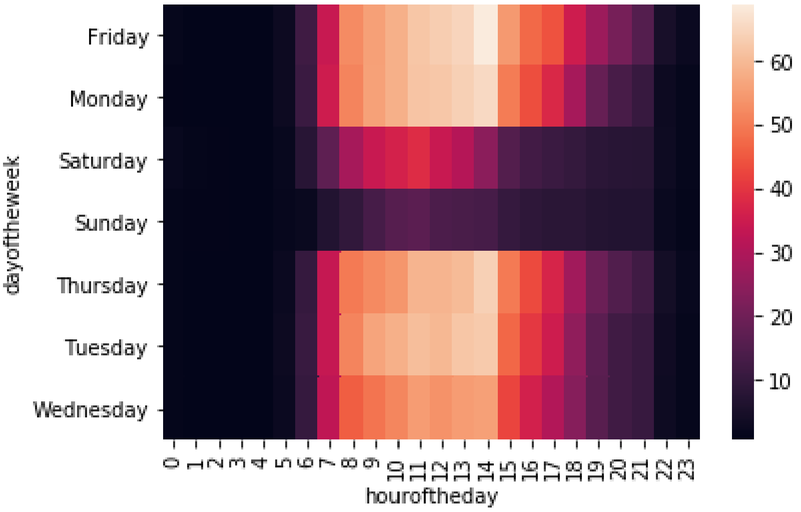


Figure 95. Global variations of volume/hour per day of the week

A) Gate 10A – entry

Figure 96 represents the average value of traffic entering each day of the week through gate 10A. As it is observed, the fluctuation in days is only sensible for labour days versus weekend and the hour range where more vehicles enter the port is between 9 AM and 11 AM. The decrease in entries is abrupt at 3 PM. The heatmap shows the same distribution, indicating the "hot zone" in the morning. It will be more probable to have congestion in that direction and gate, at the beginning of the workday and very unlikely to happen in the afternoons.

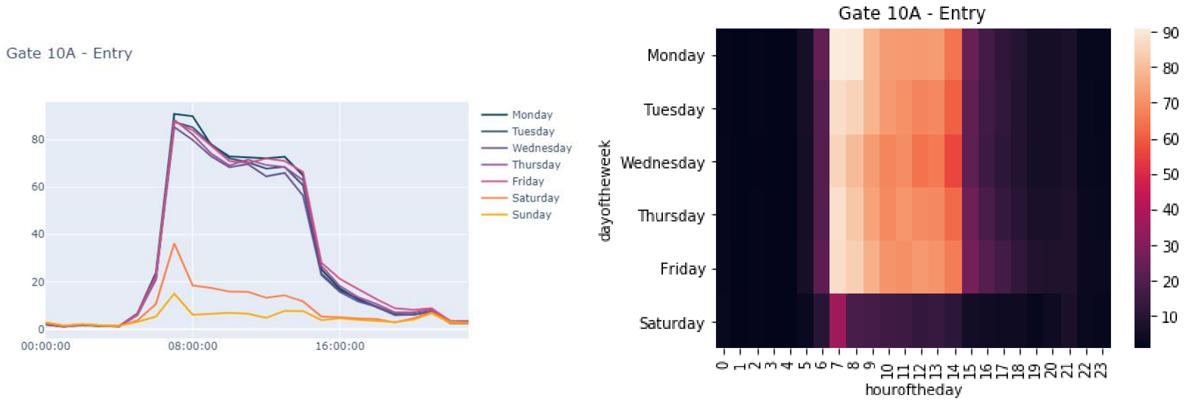


Figure 96. Gate 10A – entry: traffic volume through days in a week

B) Gate 10A - exit

In contrast to the entries, the traffic volume per hour at Gate 10A (Figure 97) is higher as the morning comes to its end. The peak is reached in averaged every day at 2 PM when the morning shift ends, and the trucks have picked the cargo and wish to leave the port towards their destination. There is almost no variability of this distribution among the labour days of the week. This is since this gate is mainly used by trucks destinated to the conventional cargo terminal.

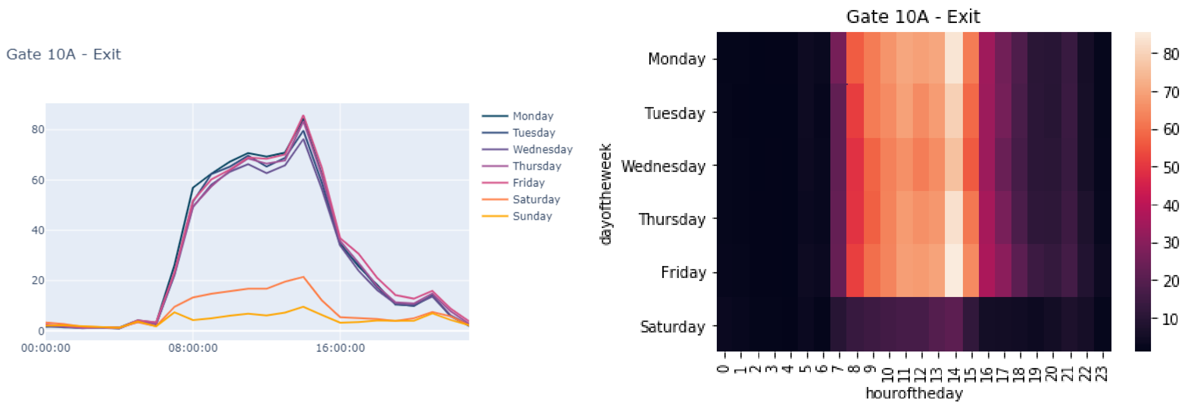


Figure 97. Gate 10A - exit: traffic volume through days in a week.

C) Gate 16 – entry

Gate 16 is a different case. Whereas Gate 10A is used mainly for the commercial side of the port, especially focused on the conventional cargo, gate 16 is a three-liner, multi-use gate which experiences heterogeneous traffic types and sizes. This is the first indicative sign that one appreciates while checking Figure 98. Weekends experience traffic entering and exiting that gate, as trucks use more frequently that gate on Saturdays, while 10A is used for private vehicles. Other gates of the port experience the same effect, but due to another cause. Specifically, gates G6 and G11 are normally used on weekends by citizens to cross from one part to another of the coastal side of the city avoiding semaphores. This takes place mostly on the afternoons (after work) and during weekends.

Regarding the traffic that can be associated to the port itself in gate 16 under study, the gates are similarly busy (at average) during the whole day, with slightly more intensity in the afternoon (after 3 PM) due to the arrival of trucks and vehicles for the afternoon shift.

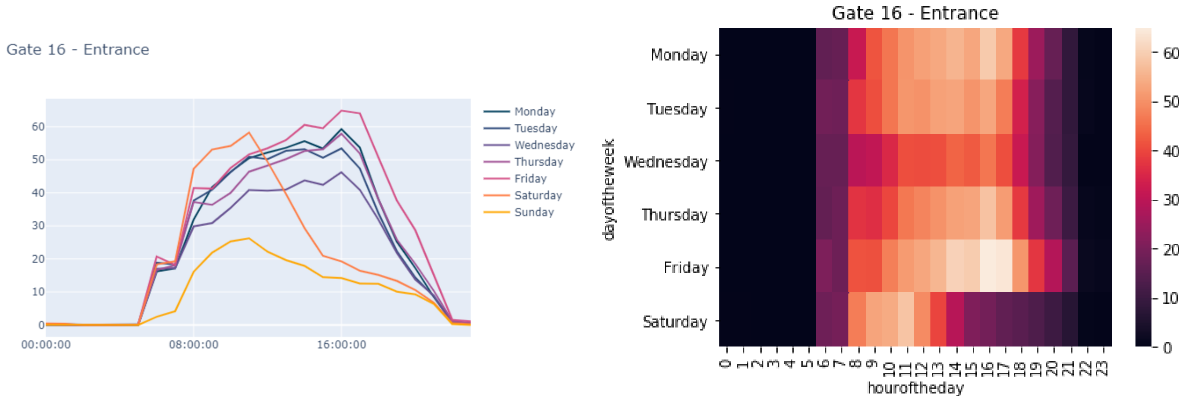


Figure 98. Gate 16 - entry: traffic volume through days in a week.

D) Gate 16 – exit

Regarding the exits on Gate 16, the same reflection as the entries apply. It was concluded that the most likely time frame for eventual congestion of the gate 16, is from 2 PM to 5 PM, just in contrast to Gate 10A. Gate 16 will have more flexibility in terms of opening/closing more gates if needed in case of congestion (Figure 99).

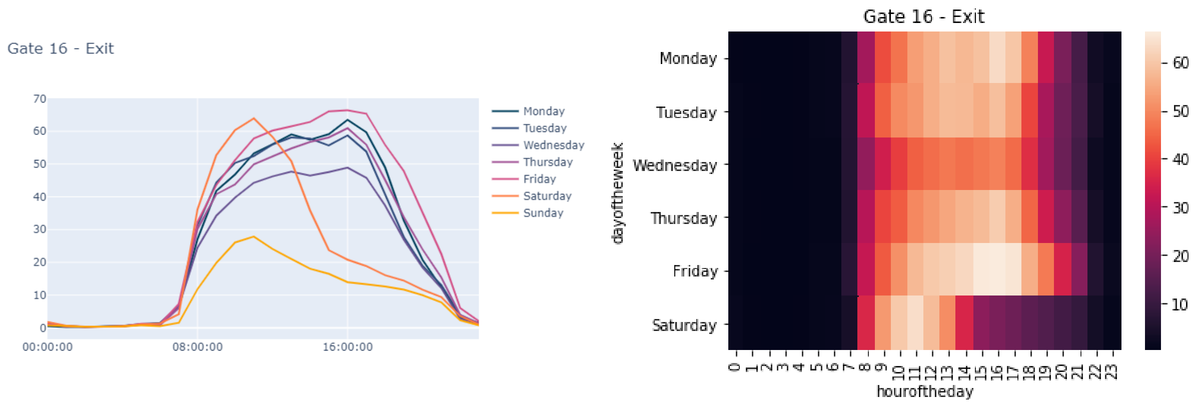


Figure 99. Gate 16 – exit: traffic volume through days in a week.

As seen in Figure 100, Monday is a non-much-variant day of the week, with regards to the average volume per hour along a whole year. Only June and October show more traffic than usual on Mondays, and especially during the mornings, which is probably due to the conventional cargo terminal, that shows more activity at spring and autumn.

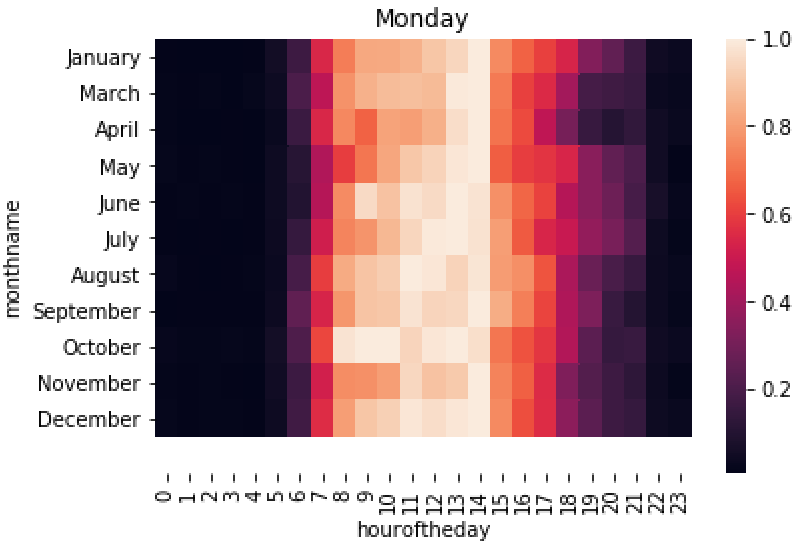


Figure 100. Traffic volume on Monday throughout the year and different hours of the day.

Tuesday (Figure 101) presents some additional variations compared to Mondays. It is especially remarkable: (i) the increase of activity on the afternoons of September (likely linked to the high quantity of cargo processed that month) and the intensification of the traffic during the mornings of December.

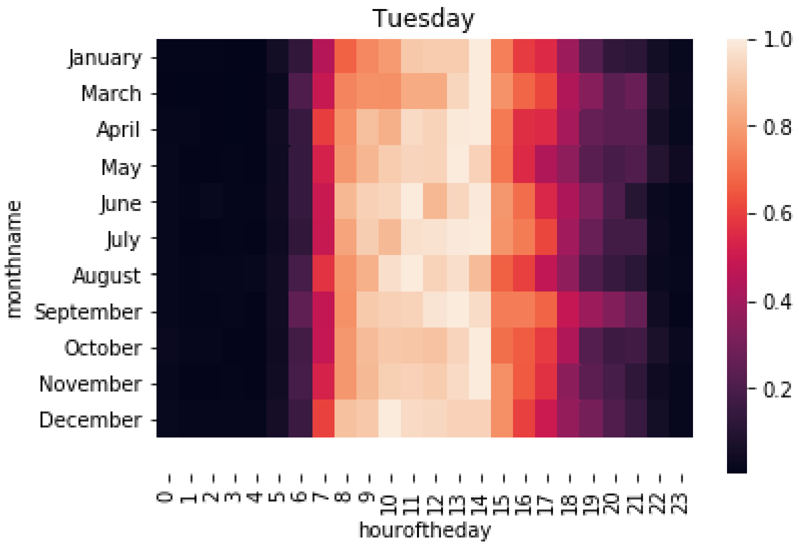


Figure 101. Traffic volume on Tuesday throughout the year and different hours of the day.

Wednesdays (Figure 102) show to be equally distributed throughout the year. Keeping 1 to 2 PM as the busiest range on average, the periods where the days are longer (June), the traffic tends to increase during the afternoon. December and April are the months with less traffic in general, coinciding with the conclusions extracted before during the global monthly analysis.

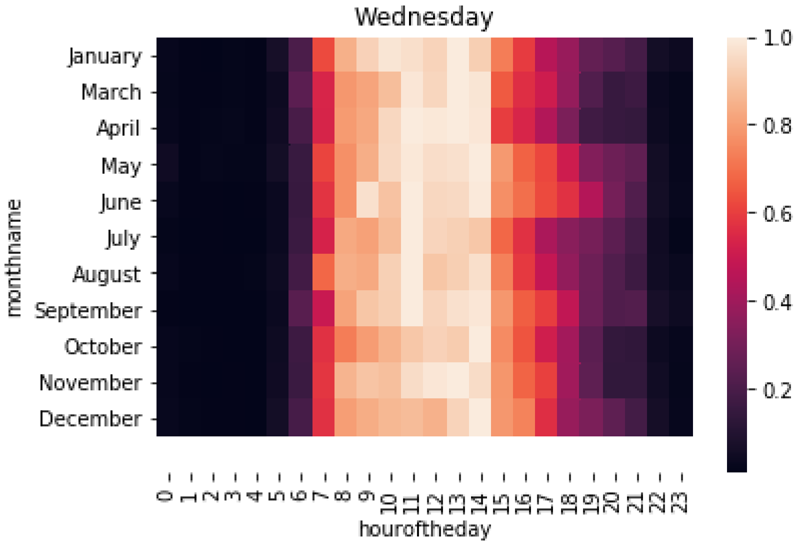


Figure 102. Traffic volume on Wednesday throughout the year and different hours of the day.

Thursdays and Fridays (Figure 103) are sustainably maintained with a high level of traffics volume per hour during the "central" hours of the day: 10 AM to 2 PM Afternoons are equally distributed during the year and no specific seasonality are noticed checking the heatmap for the day.

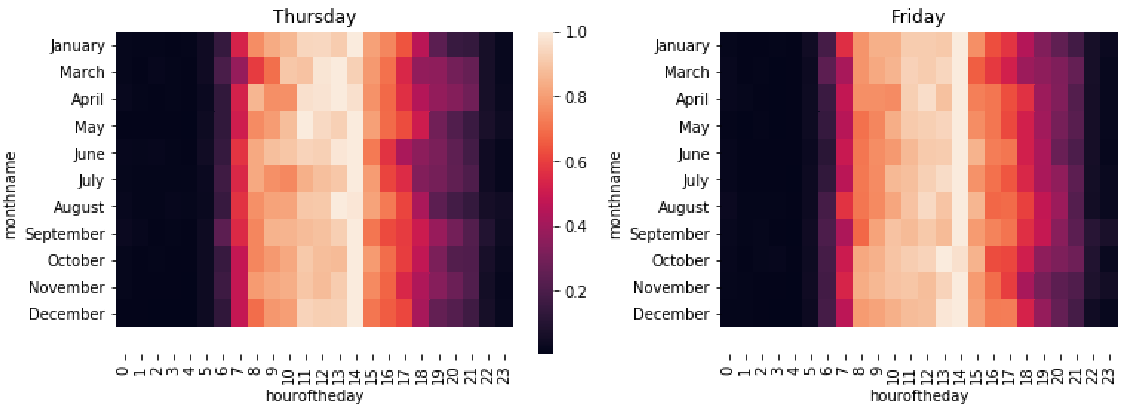


Figure 103. Traffic volume on Thursday (l) and Friday (r) throughout the year and hours of the day.

Weekends (Figure 104) are the most heterogeneous days with regards to traffic, highly depending on the number of vessels visiting the port. This is also due to the diverse moments and types of traffic that use the gates of the port (It must be considered that the reference “1” in this case (maximum value) is inferior to the usual average values during the labour week. The timeframes most used are the mornings, except for the spring and autumn months, during which the traffic in afternoons is also relevant.

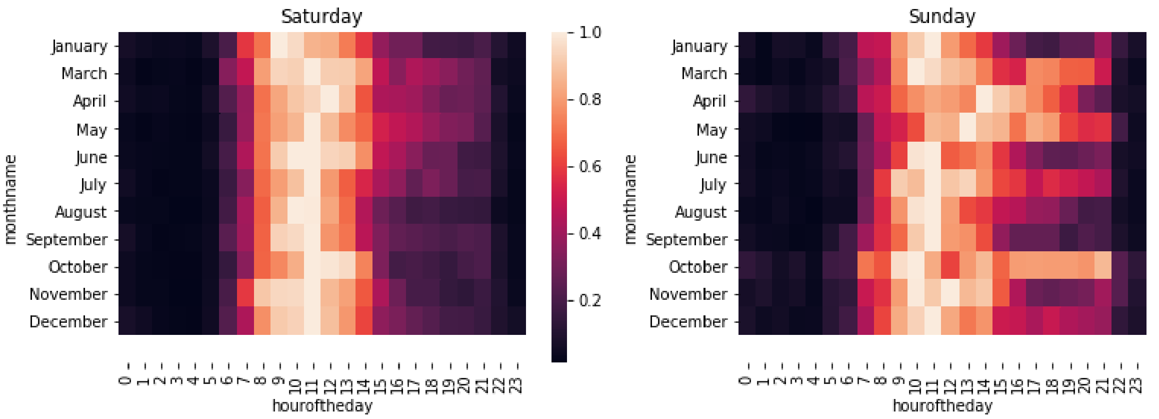


Figure 104. Traffic volume on Saturday (l) and Sunday (r) throughout the year and hours of the day.

Facebook Prophet predictions

After analysing the baseline data, according to the methodology (section 5.1.2), the team proceeded to build an ML time-series forecast regression model using the Facebook Prophet framework for Python, as it was decided so for ASPM and PPA as well.

First, data on holidays in Greece was introduced. Prophet has the capacity of considering the variations in volume in the holidays of previous years and takes them into account in the forecasts for the future.

Second, daily seasonality has been added. As it has been analysed in pages before, traffic varies considerably depending on the day of the week, and it follows certain patterns among months per day. Therefore, daily and weekly seasonality has been included in the model-building input parameters, making use of the features of the Prophet framework.

. As Prophet framework only admits forecasting time-series for a single point at a time, the dataset was divided accordingly, for each gate and direction, thus having 4 datasets (Gates 10A entry and exit and Gate 16 entry and exit). Additionally, the labels and columns were prepared to fit the Prophet model inputs needed. Besides, a sliding window filtering process was applied, to reduce the “amplitude” of the signal and to minimize noise, while the obtention of an “average line” was also useful for analysing a “global trend” of the traffic volume, before applying the prediction. This would serve as an overview reference to check whether the prediction abides by the “theoretical evolution line”.

From this point on, the same exercise was repeated for all the 4 datasets, but for the sake of clarity on the deliverable, **only Gate 10A – entry is fully documented**. Along with the next pages, all the graphs and main explanations will be referred to this location. After each sub-section, **only a brief reference to the results and conclusions for the other 3 gates** and directions is depicted.

The last thing before applying the training and forecasting was to re-check the data, confirming that everything was in place and the prediction would be valid.

In the first two graphs below (Figure 105), the reader can have a perspective of how one (out of the 4) modified dataset for Prophet looks like. As expected, values respect the EDA conclusions before. An average of 400 to 600 cars per day per gate and direction were measured, and the distribution per day of the weeks is more or less similar with more cars concentrated (in average) in the central days of the labour week, while weekends are less crowded.

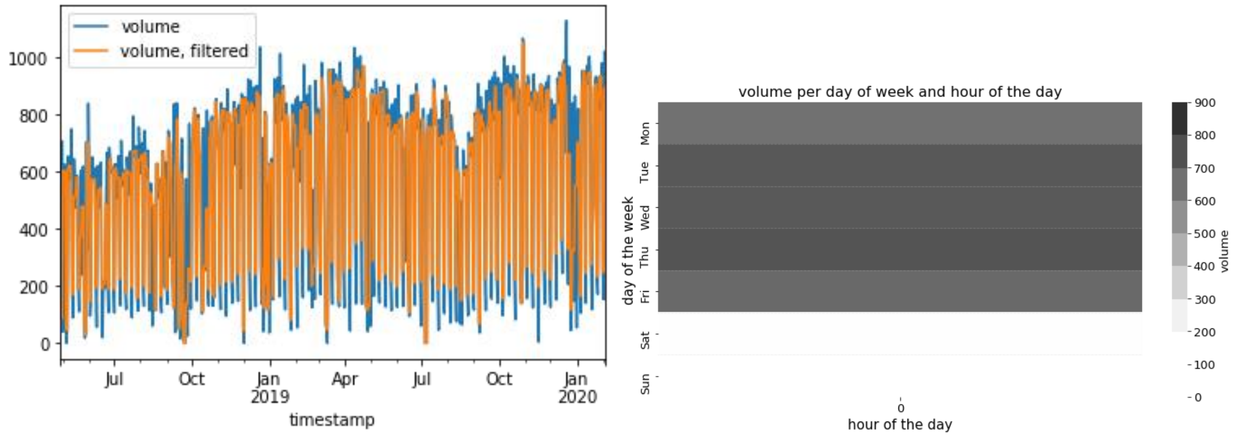


Figure 105. Traffic volume at the Gate 10 per day.

In Figure 106 the team used different visualization options, to get a global idea of the average line of the evolution of the traffic volume (at the Gate 10A) per day, throughout the whole dataset period. As it can be observed, 2019 was a year with more average daily traffic in ThPA than 2018, with a positive gradient in the trend curve towards 2020 that only experienced a bay during December-January, presumably due to Christmas.

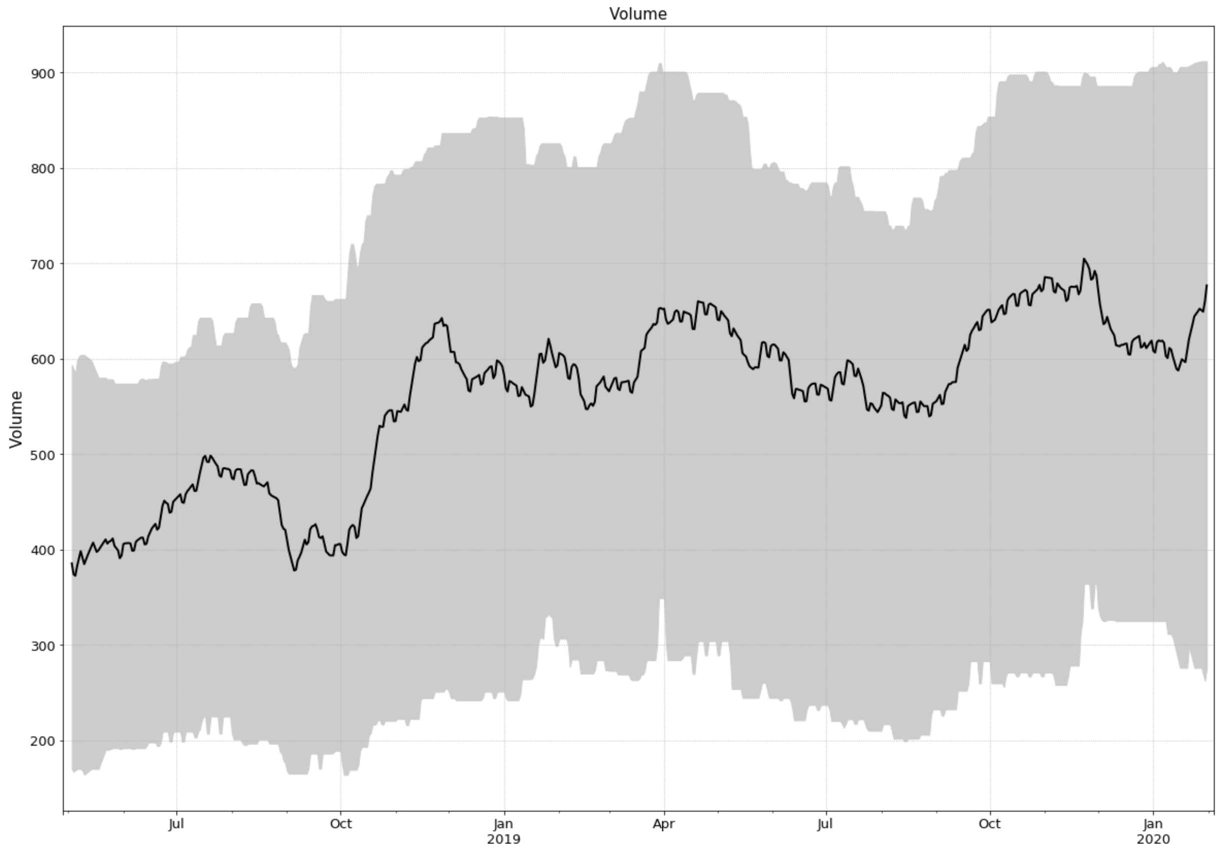


Figure 106. Traffic volume Gate 10A - Daily values and average daily line.

Figure 107 aims at representing the same information as Figure 10 but smoothing the average line depicting a monthly value instead of daily. As it is seen, changes are less abrupt and global tendencies are better interpretable. No new remarkable discovery, but the team confirmed the validity of data for training.

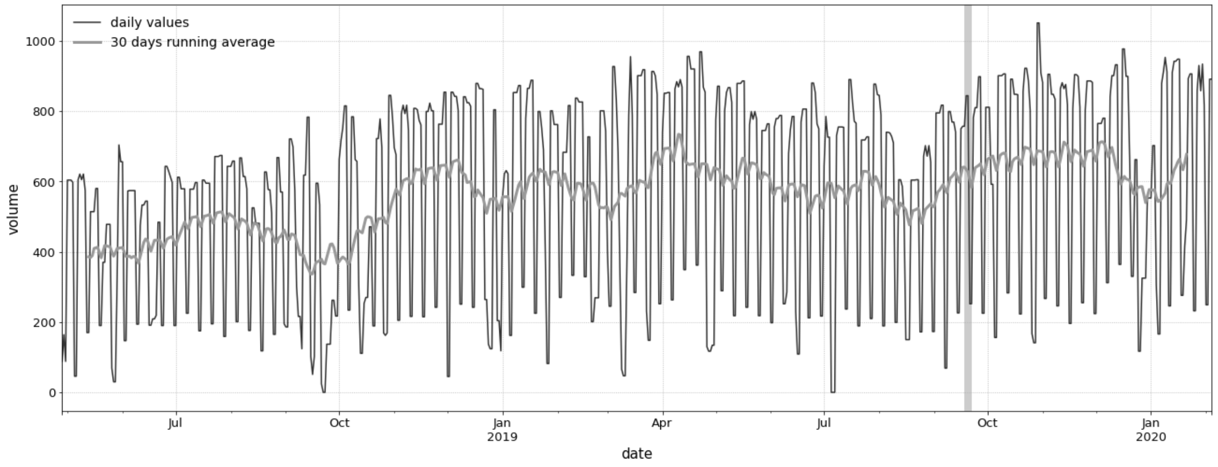


Figure 107. Traffic volume Gate 10A - 30-days average line and daily values.

Other gates followed the same pattern whenever comparing this tiny EDA with prepared data, compared to the detailed one explained before.

Finally, the team proceeded with the model training and forecasting. The strategy followed was two-fold:

A) **To use the whole dataset to train the model and to forecast the next few days using that model.** This was thought as a "test" of the future real use of the model in PIXEL: having past data and applying the model to infer the traffic volumes per location in the next days (24 hours as requested by ThPA is initially needed).

The team decided to use a prediction horizon of 14 days with a granularity of 60-minutes (following the same as for the dataset). The idea in the graphs below is to show the results of the prediction and compare them with the average trend lines, EDA values and the other conclusions and assumptions made till this point.

Note for interpretation: Due to data preparation and how Prophet models provide its outputs, the following graphs have considered the window technique for representation of the Y-axis. Traffic volume is never negative (unless it may seem so observing the figures). This numbering must be understood considering 30 vehicles is the baseline value (0 – reference level in the graphs).

In Figure 108, the traffic volume (per hour) is plotted for all the dataset and the period calculated. According to the original dates, a prediction was made using the Prophet model from 6th to 19th February 2020. In the graph, it can be noted that the blue line (predicted values by the model) keeps the shape and trends for the original datasets. Black dots are the original measurements. Prophet only predicts what is indicated for the horizon but plots a "prediction" of the whole set.

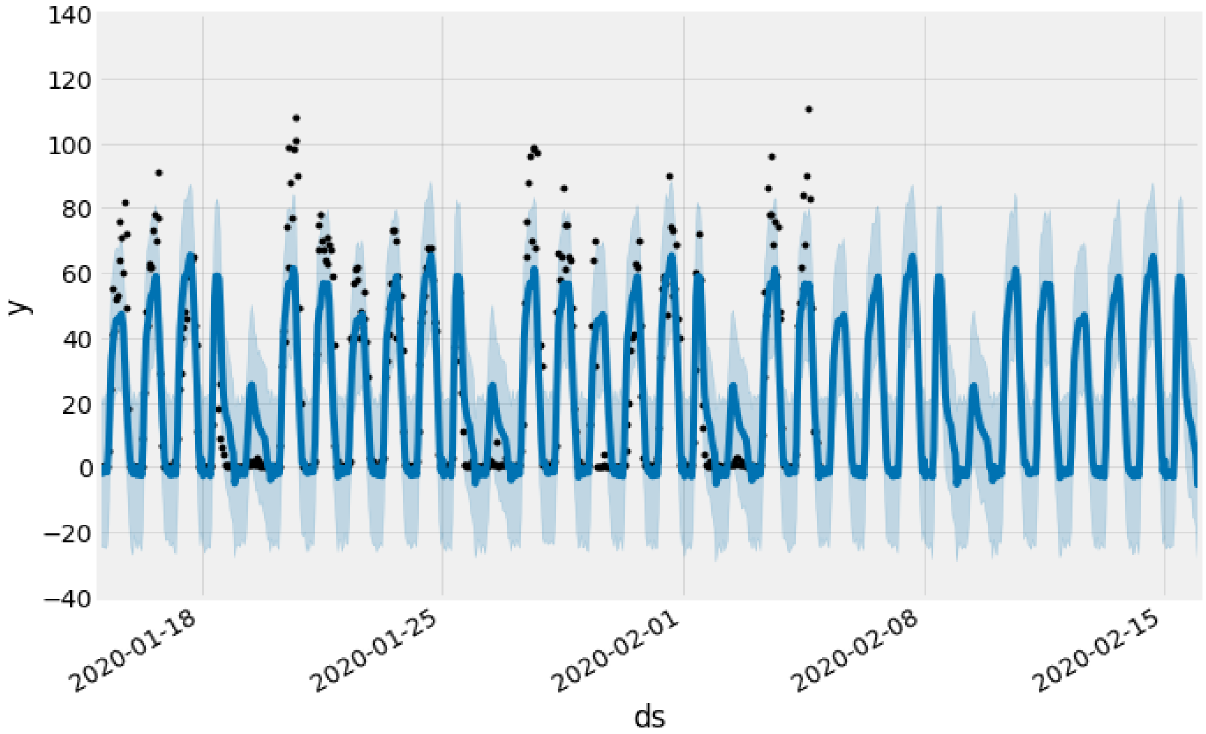


Figure 108. Baseline data and 14-days prediction.

The three graphs in Figure 109 show different information that confirms the alignment with the dataset and gives a perspective of the appropriateness of using Prophet as valid forecasting model. As expected, during the whole prediction holidays were considered, where (almost) no traffic was experienced. The most interesting graph here is the weekly prediction evolution. The bottom graph represents the forecasted traffic (per hour) for the first week of the horizon. Averages, hour values and differences among labour days are evaluated and keep full consistence with the EDA realised.

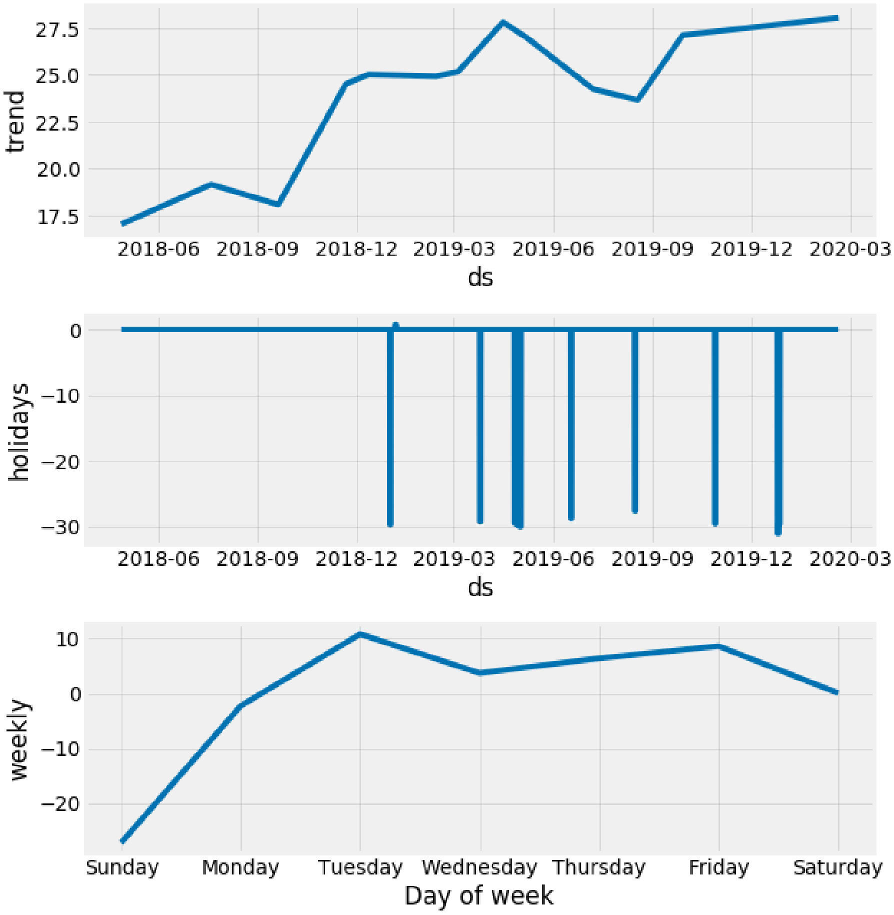


Figure 109. Graphs of 14 days prediction baseline Prophet (I).

Figure 110 provides results of the prediction for each day of the week of the first week of the horizon (6th-12th February 2020). For the labour days of the week (examples Monday and Friday at the two graphs, at the top), the volume line matches with high accuracy the average values per day, according to the dataset. There the range 62-85 vehicles per hour were registered for the timeframe 7-15h, which as it is seen is like the forecasted values in Gate 10A for Monday and Friday, of the forthcoming week. Saturday and Sunday are also feasible to follow the predicted line, especially considering the minimum and maximum values per hour (30, 40 at most) and the drawing of the curve.

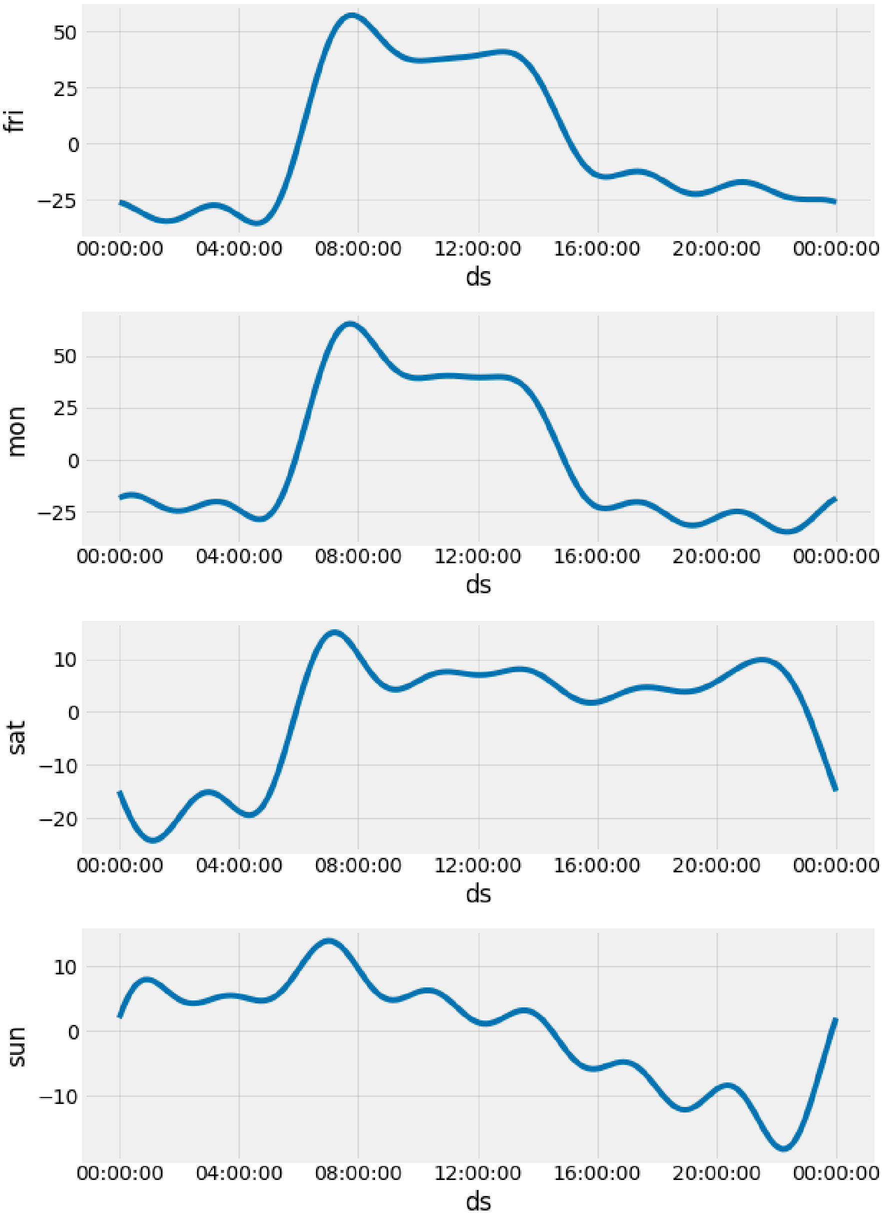


Figure 110. Predicted traffic volumes for different days of the week.

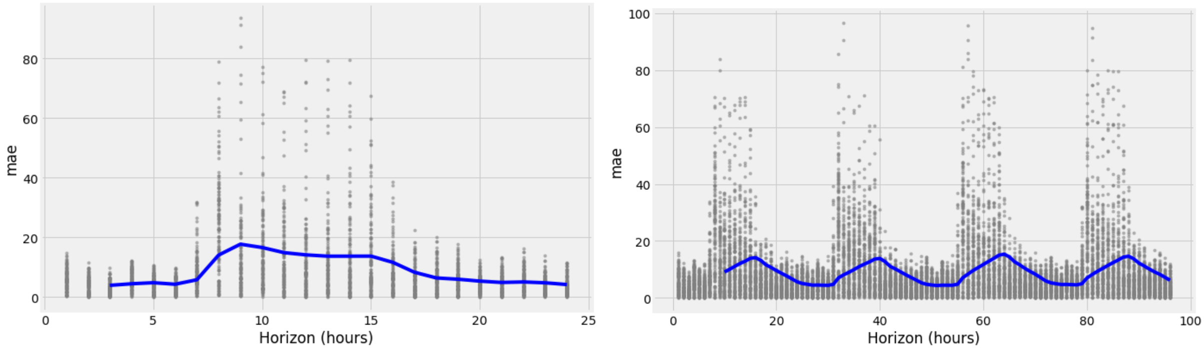


Figure 111. Mean absolute error (MAE) Prophet all baseline.

Fig X: Mean absolute error (MAE) Prophet all baseline

As it has been explained, this action a) only aimed at simulating a real prediction of scenario, without a clear strategy of ML model validation. However, taking advantage of the whole dataset prediction by the

Prophet model, the accuracy of the forecasts of this model, can be globally analysed using the mean absolute error (MAE) – presented in Figure 111. In that sense, the team proceeded to create two analysis: the average MAE of all days of the dataset (first graph). This is particularly useful as ThPA expressed they will aim at having a one-day prediction. According to this, in the most active labour hours of the day (7-15) the error in the number of vehicles (by mean) was less than 20, which is considered a good accuracy. This pattern is repeated with daily periodicity.

For **other gates and directions**, the results were equivalent, with less MAE in 10A-exit (average:10), less MAE in 16-entry (average: 11) and more MAE in 16-exit (average: 17).

B) **To validate the model following usual procedures**. To make this, **the baseline (gates traffic) dataset** was divided into “training” and “test”. This way the team was able to discover the accuracy and effectiveness of the Prophet model for prediction without needing to wait for new data.

The procedure consisted of dividing the dataset following the criteria:

- Train: 28th April 2018 till 20th September 2019 (78,5% of the dataset)
- Test: 21st September 2019 5th February 2020 (21,5% of the dataset)

The results were the following:

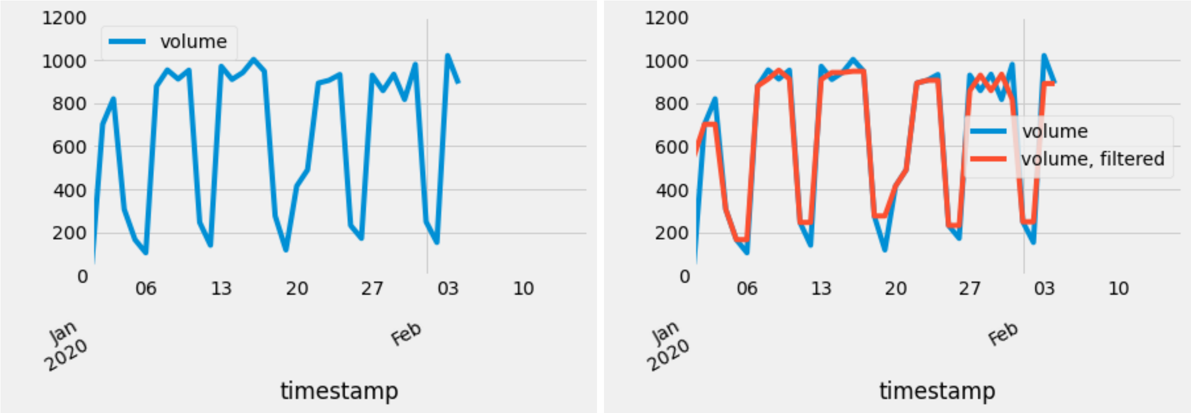


Figure 112. Predicted values per hour baselined divided.

The trend is consistent with past data and with the curve for the 14 days horizon (February 2020) obtained in Figure 109. In the first graph of Figure 113, it can be observed that predicting for all the dataset period but using only a portion for the training, makes the trend curve less accurate to the reality, with lower capacity for predicting more abrupt changes over the mean. In this case, we have not opted for the windowing technique and the values are not biased in the Y-axis. Predicted values, in this case, adopt a descending gradient whereas using the whole dataset they were, in general, ascending. Holidays are well predicted due to Prophet understanding.

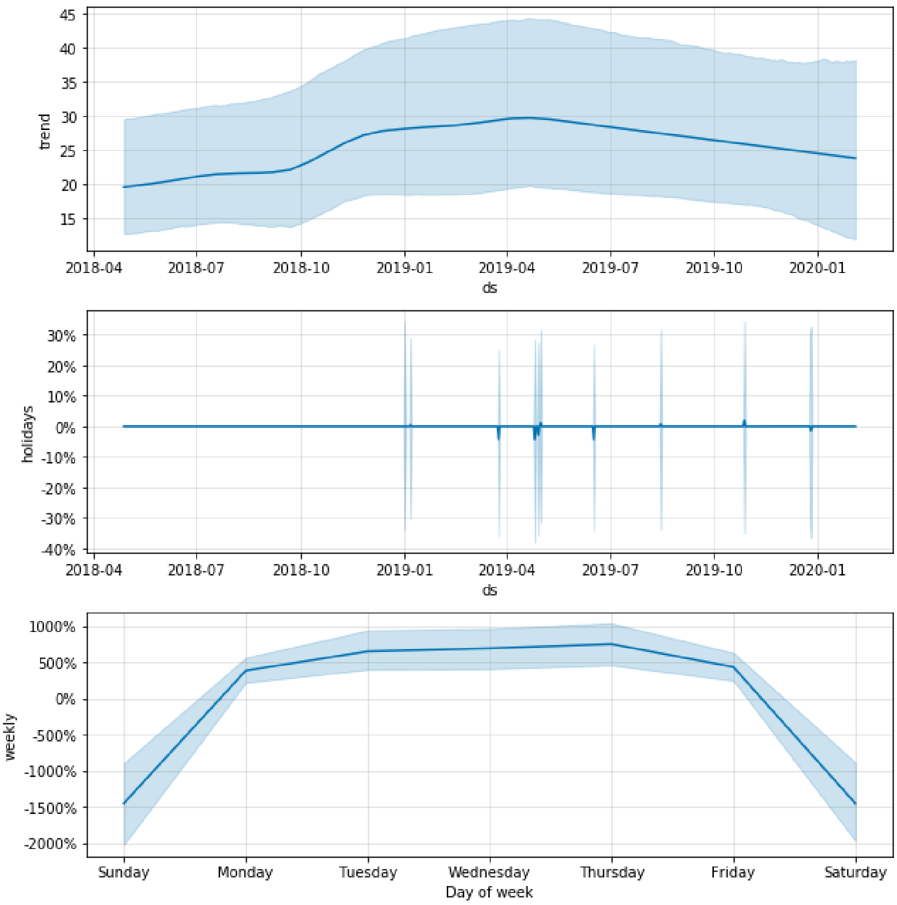


Figure 113. Detailed graphs baseline data.

Accuracy analysis of the model:

A comparison between the predicted values and the real values of the test data set is needed to analyse the performance of the Prophet model trained with 78,5% of the dataset. For doing so, the mechanism decided by the team was to create a joint plot including cross-validation of the *y* (values of the test dataset) and the *yhat* (predicted values) per day, to check at a first glance, the overall appropriateness of the model. According to the nature of the joint plot (see), in perfect prediction scenario, the dots (*yhat*, predicted values) were situated exactly over the diagonal line. This would mean that, for each day, the value predicted would coincide with the number of vehicles (values of both axes) trespassing Gate 10A during the period 20th September 2019 to 5th February 2020.

This representation is known as a bivariate plot with marginal univariate plots.

Additionally, the graph was built to include extra information: on the top of the image, there is the volume distribution per days of the real values of the test dataset. This graph means that a peak can be observed in the number of days that the traffic volume was circa 150 cars (presumably Saturdays, Sundays and holidays) and the second peak is around 850 vehicles per day, which is the average number on labour days of the week. In a scenario of perfect prediction, this graph should match exactly with the one at the right of the outer square, that represents that exact information but upside down for the predicted values.

In Figure 114, prediction correlation joint plot Gate10A baseline provides a clear reflection: the discrepancy between the values measured after the separation of the dataset with the predicted values for that period is greater when the number of vehicles is lower. This would mean that for the labour days of the week this model will predict better than for the weekend.

This discrepancy is called: residuals and, as it can be observed in Figure 115, most of the times that difference is between 0 and 200 vehicles per day. The more frequent number is 150 and the second more frequent residue is 50 vehicles. Considering the average per day is about 850, a deviation of 75 is 8.8% fewer vehicles predicted per day.

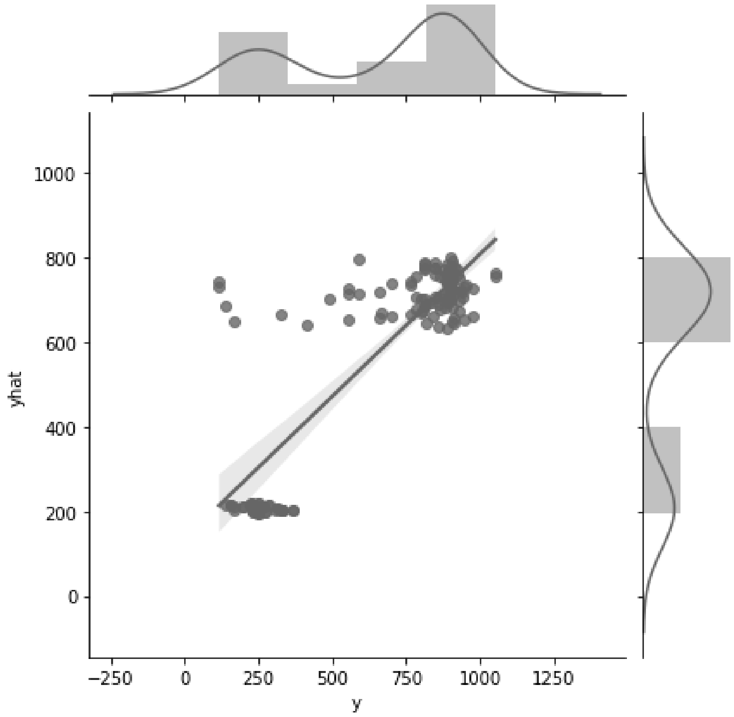


Figure 114. Prediction correlations join plot Gate 10A baseline.

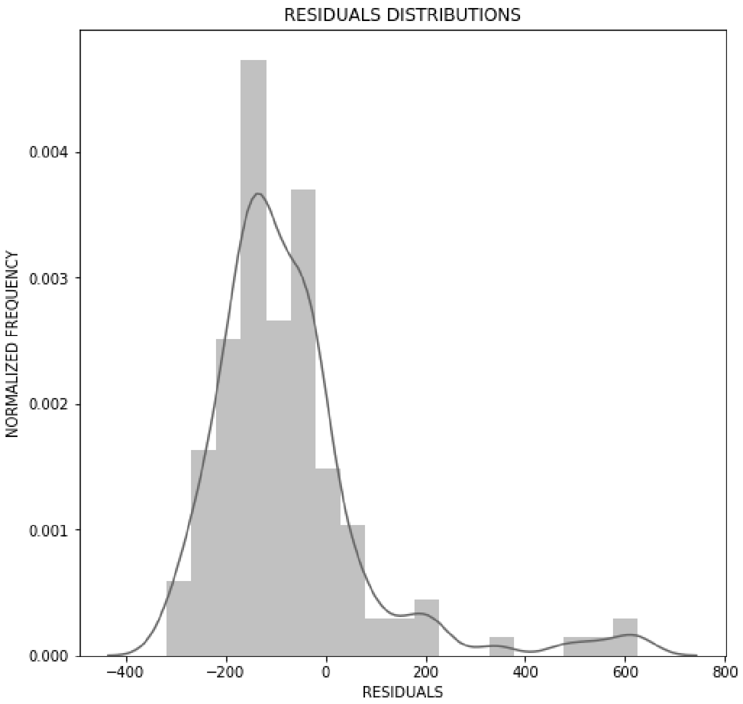


Figure 115. Residuals distribution baseline divided.

Different error metrics were computed to further compare this baseline model, with other models – presented in Table 20.

Table 20. Metrics Prophet model baseline data divided train and test.

Metric	Test dataset	Train dataset
Correlation value y (values), yhat (predicted)	**0.851736**	0,826675
Daily average MAE	**97,488**	139,289
Residuals median	24,4823	
Residuals mean	-3.6398	
Skew	-2,496	
MAPE	29,65%	64,09%
MSE	24754,485	45425,14
MSLE - Mean squared logarithmic error regression loss	0,1336	0,6308

For **other gates and directions**, the results were equivalent to the explanation done for the non-divided dataset.

5.2.3.3. Forecasting with included weather data

Next step, after having reached a usable model, was to introduce additional regressors that may influence the volume of traffic at the gates, at a certain timeframe.

Is the congestion at the gates influenced by heavy rain? Is more likely to have a congested Sunday if the weather is favourable? As designed in the methodology, it is an objective to correlate and include in the regression model weather information.

For tackling this objective, the team used the free web service provided by Stratus Meteo of Greece[58]. Different sensors are installed throughout Greece and for this case, we made use of the one closest to the Port of Thessaloniki (Presented in Figure 116):

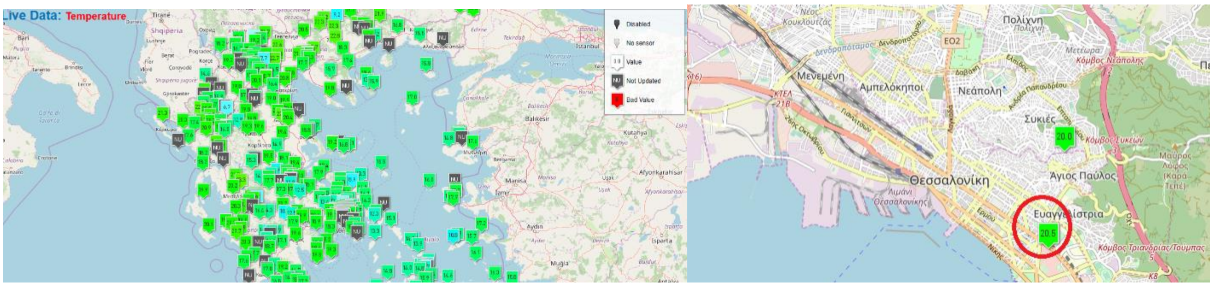


Figure 116. Selection of sensor data source for Weather correlation Gate 10A.

The issue at this point was that the historical weather information retrieved from this external service was only served with a daily granularity. In that regard, a preliminary action needed from the team was to re-shape the baseline data of traffic at the gates (each 60-minutes, as selected) and have the same data frame for the accumulated quantity of one day in every single row.

Additionally, the dataset of weather was also available from September 2018, therefore we reduced the quantity of data for training and validating the model. Thereafter, a single data frame was constructed, including data from the baseline and weather: average temperature per day, average wind speed per day and precipitation intensity per day.

Facebook Prophet predictions

Prediction with the weather (both baseline data and weather have a daily granularity) using it as an additive feature to the already built Prophet model:

[58] http://stratus.meteo.noa.gr/front

Next action performed was **to replicate the b) phase of the prediction procedure for baseline data** but including in the base data frame, three additional regressors: weather information. As done before, data was prepared using the medfilt filter to smooth the data.

Equally, for dividing the dataset the same delimiter was used: 20th September 2019. Therefore, the dataset was this time divided into 73,7% for training and 26,3% for testing, with less training data.

The results were the following:

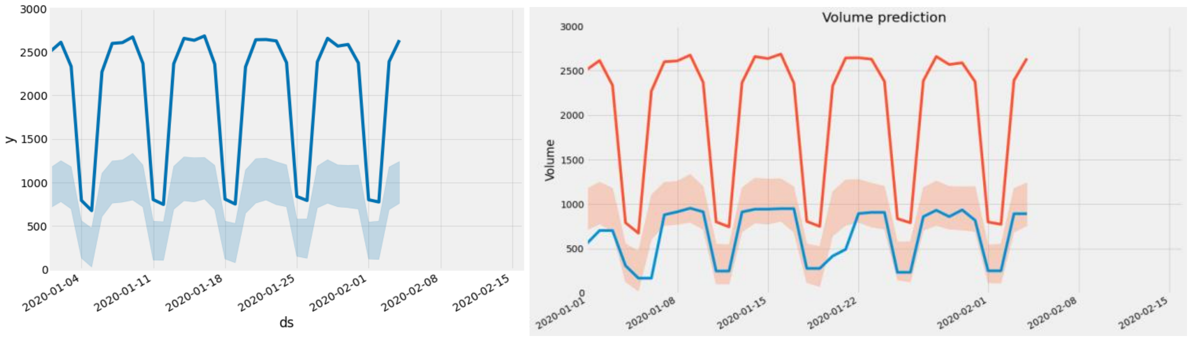


Figure 117. Predicted volume Gate 10A per day – divided dataset including weather information.

Figure 117 presents the results after the weather data included in the Prophet model predicted for the “test” dataset time the values and for the rest of the timeframe, the black dots represent real traffic volume measurements. As it can be observed, the trend is in an ascending gradient reaching little by little the 1.200 vehicles per day on average for the 14 days horizon (February 2020). At first glance, these results seem to be too high and not very consistent with the trend curves observed in the most accurate prediction done with the data (see Figure 109).

The intuitions behind the previous reasoning were confirmed, when analysing more detailed graphs about the prediction made by the recently trained model (Figure 118).

Looking at the top graph at the right, the trend curve was far from the real one, having, as a result, an almost linearly ascending gradient in time. This would not work on the mid- and long-term, and most likely will not be accurate for the real volumes in February 2020.

Holidays were reasonably well predicted, with circa 30% of deviation (less volume).

Regarding the average on days on the week for the whole dataset period, the prediction of the model in this occasion was (at average) 15% deviated during the labour week and between 40% and 60% on weekends.

Within the same day, the curve is not abiding by the logical curves of deviation above/below the means at different hours. Additionally, this graph must be discarded, from a logical analysis viewpoint, as the samples were daily, not per hour.

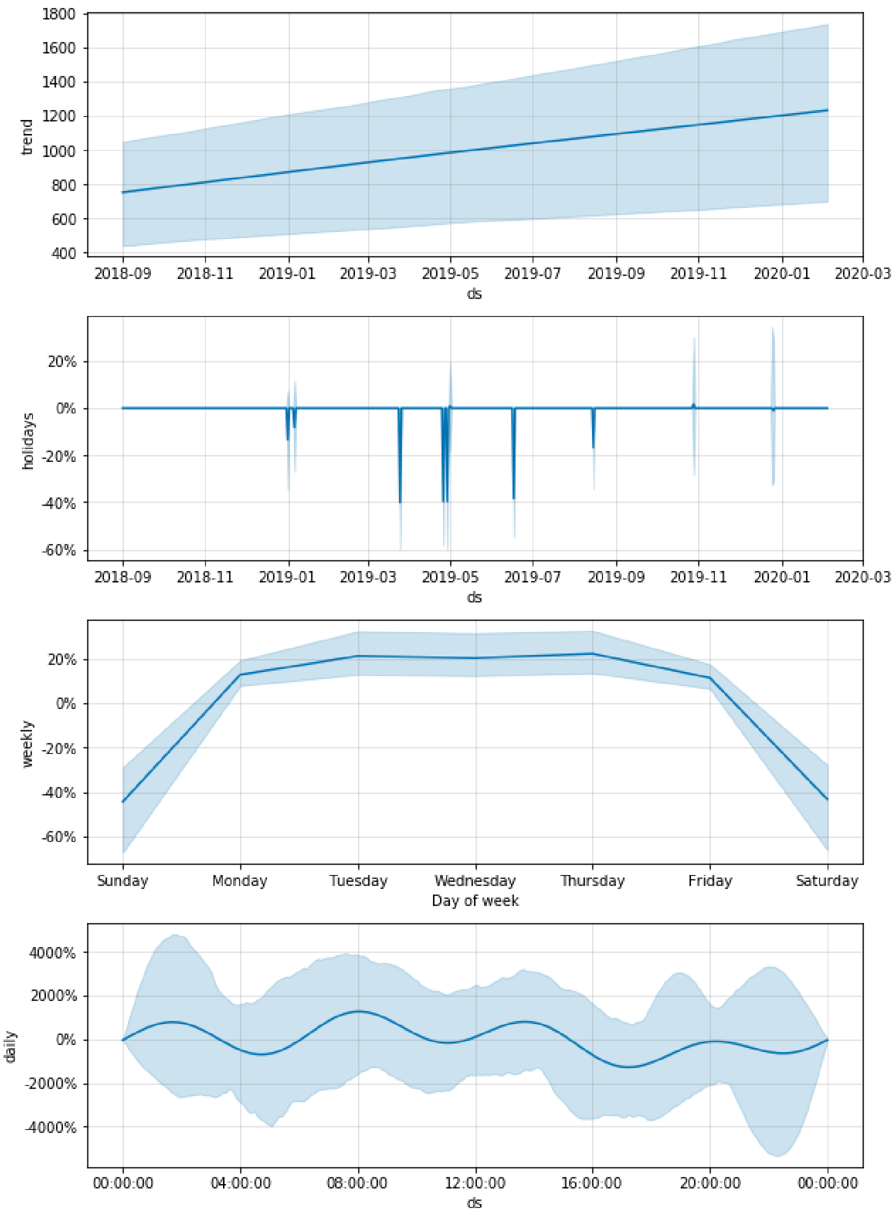


Figure 118. Detailed graphs Prediction including weather.

What is interesting in this case is to check how the prediction has changed considering the values of the extra regressor added (in this case: weather). Figure 119 shows two things: (i) a clear seasonality on the summer months, having the weather a negative influence on the predicted values (100 less than the mean), and (ii) that as per average, the weather contributes to predicted values by 50 per day, reaching peaks of 150.

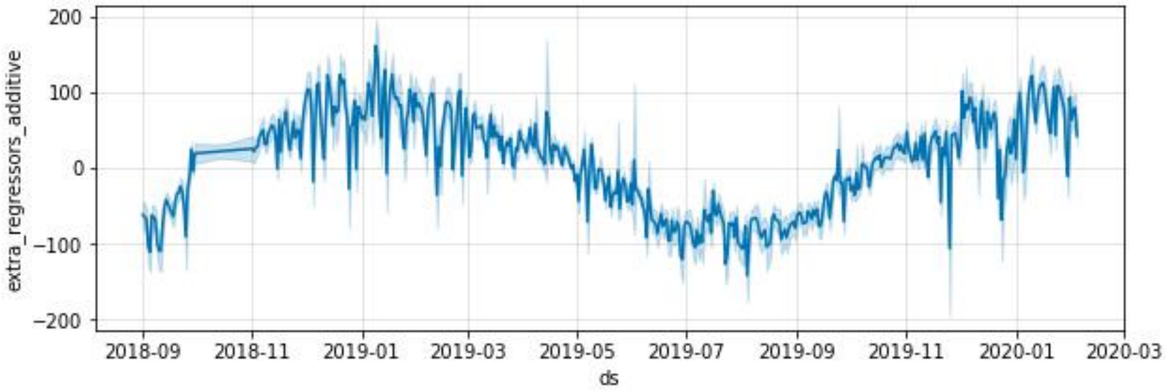


Figure 119. Weather regressor addition effect in the prediction of traffic volume Gate 10A.

The graph in Figure 120 shows the correlation between the predicted values (test data set - dotted line; train data set - solid line) and the ground truth. While for most of the months there is a strong correlation, February, October and December have a relatively low value, therefore the model would have to be improved for the winter period.

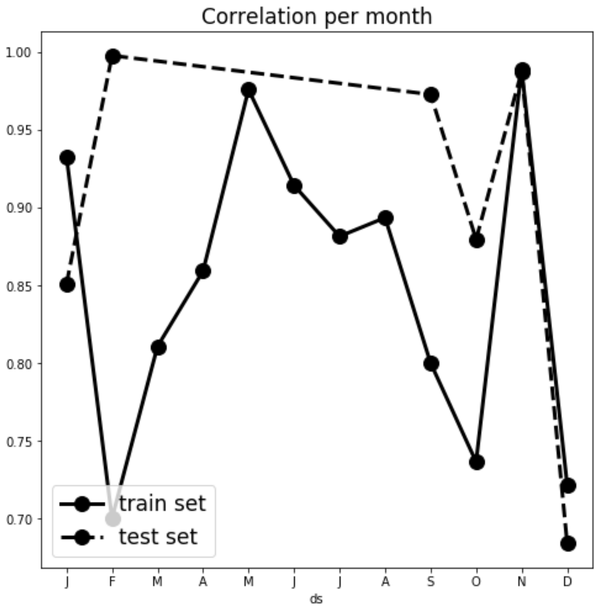


Figure 120. Correlation between predicted data sets per month – weather.

Accuracy analysis of the model:

Table 21. Metrics Prophet model baseline and weather data divided train and test per day.

Metric	Test dataset	Train dataset
Correlation value y (values), yhat (predicted)	**0.844357**	0,81169
Daily average MAE	**1252,049**	926,858
Residuals median	-1460,94	
Residuals mean	-3.6398	
Skew	-1252,049	
MAPE	237,9%	245,61%
MSE	1864896,86	1039664,086
MSLE - Mean squared logarithmic error regression loss	1,3606	1,8458

Compared to the baseline prediction, including the weather, has not improved the metrics of correlation nor the average of mean absolute error (Table 21). Although the comparison, in this case, cannot be directly done because this model has been trained with fewer samples (September instead of April and one per day instead of one each 60 minutes), this model will not be recommended to ThPA to be used because of (mainly) the ascending trend curve.

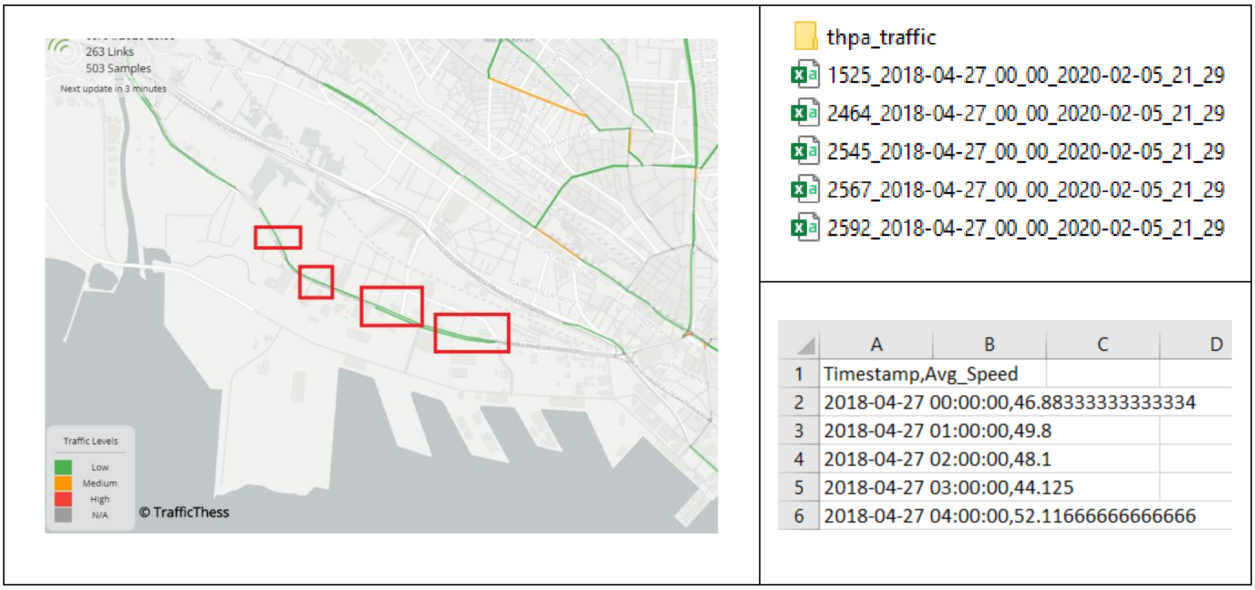
Nevertheless, with this exercise, the team was able to discover different correlations between meteorological factors and the traffic volume at the gates of the port. For the future, gathering more historic data with more granularity may correct these conclusions.

5.2.3.4. Forecasting with traffic data from the city

How much influence does the external traffic in the city have to the queues in the port? Is it true that with lighter congestion in the city, the more agile passing of vehicles through port gates is? Can the already tested model have a better performance, if it is coupled with surrounding traffic data? As designed in the methodology, it is an objective to correlate and include in the regression model the information about the traffic in the city.

For doing so, the technical team used the data source identified since D4.3 of open data (GPS-based) of the city of Thessaloniki. This data (with interest in the historic offering) is served by courtesy of CERTH-HIT (a member of PIXEL) via the website TrafficThess[59]. This website has varied information about the different roads of the city (referred to as links) updated every 15 minutes, which is perfectly aligned with the used baseline dataset for this task.

The procedure followed was: (i) selecting the surrounding links to the gates of ThPA, (ii) extracting the data for those 5 links from 28th April 2018 till 5 February 2020, (iii) selecting the interesting fields of the information provided, (iv) building a .csv with the average of the average speeds of the 5 surrounding links. This is briefly illustrated through the images in Figure 121.



thpa_traffic
1525_2018-04-27_00_00_2020-02-05_21_29
2464_2018-04-27_00_00_2020-02-05_21_29
2545_2018-04-27_00_00_2020-02-05_21_29
2567_2018-04-27_00_00_2020-02-05_21_29
2592_2018-04-27_00_00_2020-02-05_21_29

	A	B	C	D
1	Timestamp,Avg_Speed			
2	2018-04-27 00:00:00,46.88333333333334			
3	2018-04-27 01:00:00,49.8			
4	2018-04-27 02:00:00,48.1			
5	2018-04-27 03:00:00,44.125			
6	2018-04-27 04:00:00,52.11666666666666			

Figure 121. City road traffic data origin and procedure.

In this case, only an easy conversion was needed, to align both datasets. As commented, TrafficThess provides historical data with samples every 15 minutes. As per our convention, the data was averaged and condensed to have a 60-minutes granularity per row. Thereafter, a single data frame was constructed including data from the baseline and the traffic of the city.

Weather is too varying to find clear patterns on the evolution within a dataset timeframe of (almost) 2 years. Then, no EDA was conducted for that data information. However, finding seasonality, patterns and "hot

[59] https://www.trafficthess.imet.gr

periods" of the traffic of a city can be useful, to understand at a glance, if effects on the city traffic affect directly the congestion at the gates of the Port of Thessaloniki.

For making this EDA, a simpler approach was followed including only two graphs that could serve the team (and the port personnel) to realise about relevant conclusions: (i) patterns per month and (ii) patterns per day of the week.

The results, presented in Figure 122, led the team to certain conclusions: (i) the month in which the traffic is, per average, lighter (higher speed in the roads) are December and January, most likely caused to the Christmas holidays. Port roundabouts are not very frequented in holidays. (ii) More congested months are November, followed by February, March and June. These reflections coincide with the traffic at the gates concluded before: the slower the average speed, the fewer vehicles at gates.

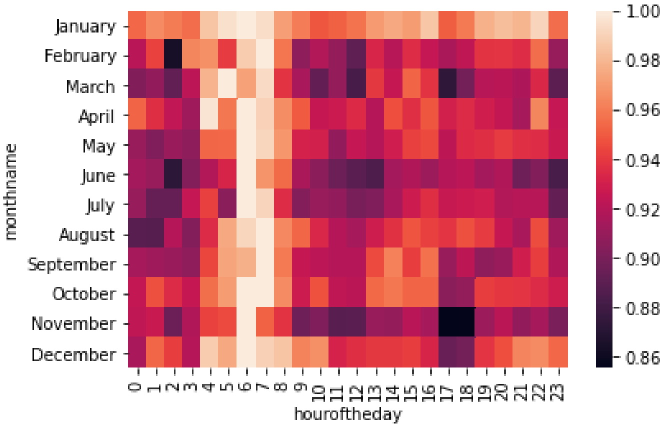


Figure 122. The average speed of the traffic in the city per month and hour of the day.

Regarding the common information that can be extracted from weekly day seasonality (Figure 123) was the following: (i) as expected, the moments of the day the traffic is faster is 4 AM to 6 AM, (ii) average speed of the traffic in the surrounding area of ThPA is stable from 8 AM to 1 PM, (iii) a peak of congestion is normally experienced at 5 PM and the moments during the daylight that is less congested is 2 PM to 3 PM

Saturdays and Sundays are in general stable and pretty much less congested than labour week.

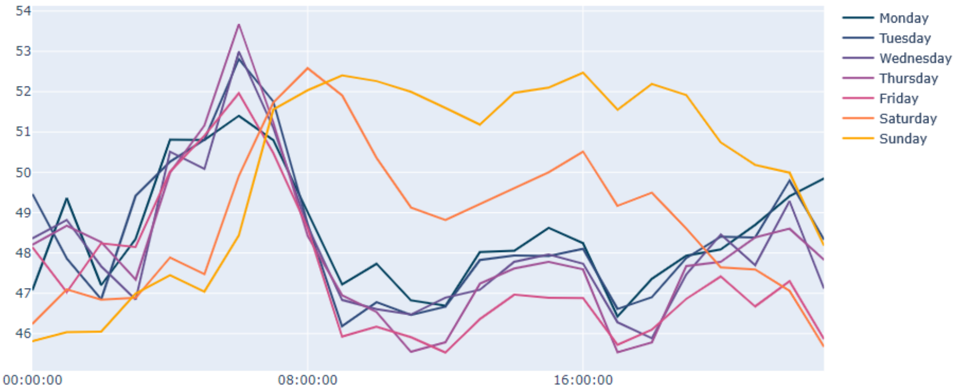


Figure 123. Average speed per hour of each day of the week (traffic of the city).

Thereafter, a single data frame was constructed including data from the baseline and the average of averages speeds, of the 5 roads links surrounding the port.

Facebook Prophet predictions

Predictions with traffic at the city (both datasets have a 60-minutes granularity) using it as an additive feature to the already built Prophet model:

Next action performed was **to replicate the b) phase of the prediction procedure for baseline data** but including in the base data frame, one additional regressor: the information about traffic in the surrounding area. As done before, the data was prepared using the medfilt filter to smooth it.

Equally, for dividing the dataset the same delimiter was used: 20th September 2019. Therefore, the dataset was this time divided into 78,5% for training and 21,5% for testing, as for the baseline try.

Results were the following:

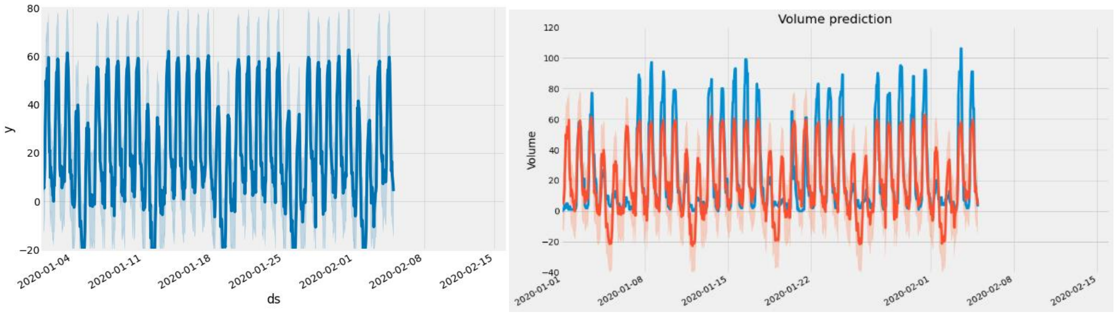


Figure 124. Predicted volume Gate 10A per day hour - dataset includes city traffic information.

Figure 124 presents the results with the city traffic data included in the Prophet model, predicted for the "test" dataset, time the values and for the rest of the timeframe, the black dots represent the real traffic volume measurements.

As it can be observed, the trend is consistent with past data and with the curve for the 14 days horizon (February 2020) obtained in Figure 109. This is, at this point, a good sign as it seems to be a valid model.

The intuitions behind the previous reasoning were confirmed when analysing more detailed graphs (Figure 125) about the prediction made by the recently trained model. Looking at the top graph at the right, the trend curve seems close to the baseline prediction. Additionally, a dynamic range of max-min deviation in the predictions of the whole dataset (light blue shadow) looks very constrained. Holidays were reasonably well predicted, with circa 30% of deviation (less volume). However, the most reassuring graph was a daily prediction. As it can be seen, the curve is tightly adjusted to Figure 5, with higher values in the range of 7-9 AM., little decrease but stability around the mean till 3 PM and then decreased again. This gives a good picture of the validity of this model.

However, the team wished to extract more information about the differences, correlations and statistic values between the predicted values (yhat) and the measurements (y) for the test set timeframe.

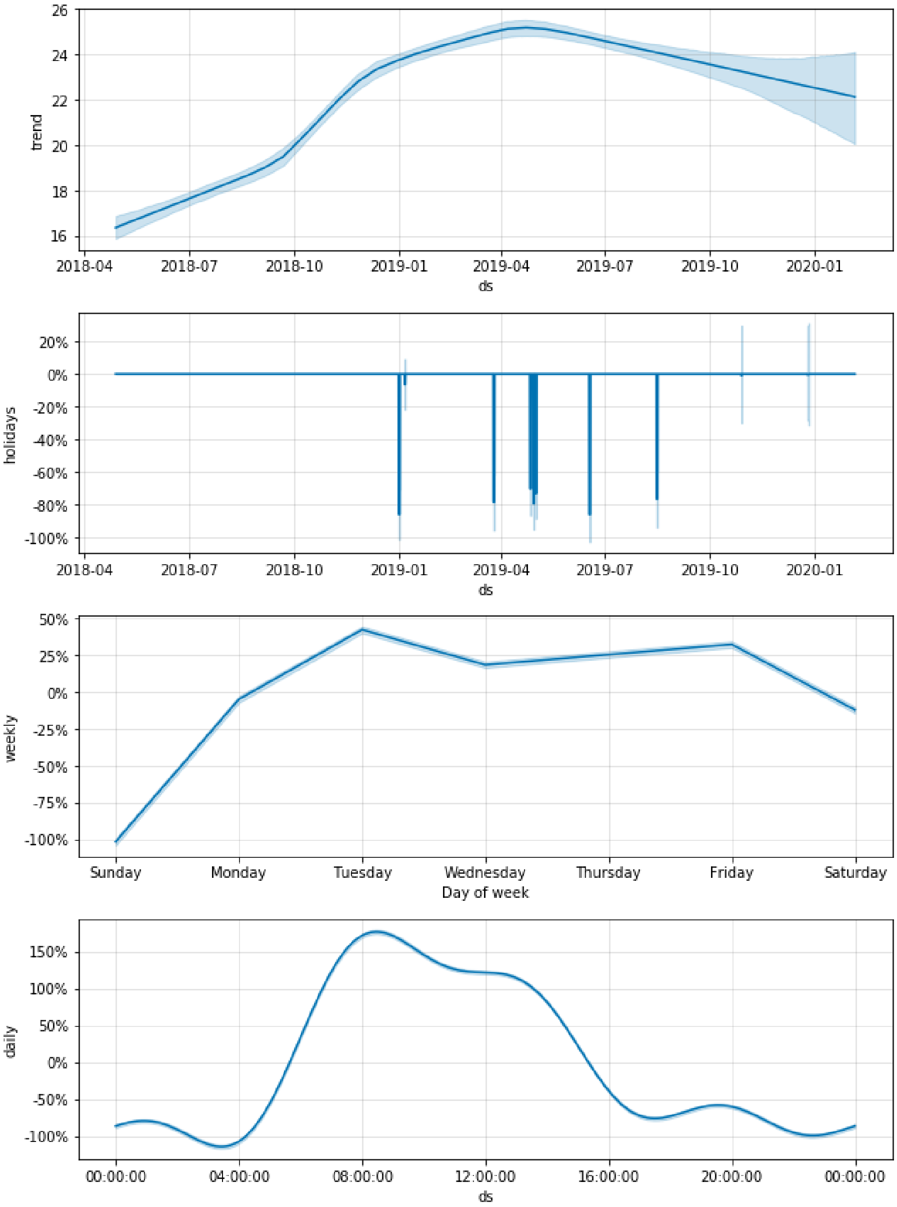


Figure 125. Detailed prediction graphs including city traffic.

In line with that wish, in the following picture, the Pearson's R correlation value was calculated for every single predicted point (traffic volume at Gate 10 A in a certain hour of the dataset timeframe). The results were satisfactory, as during almost all the period R was around 0.8 – Figure 126.

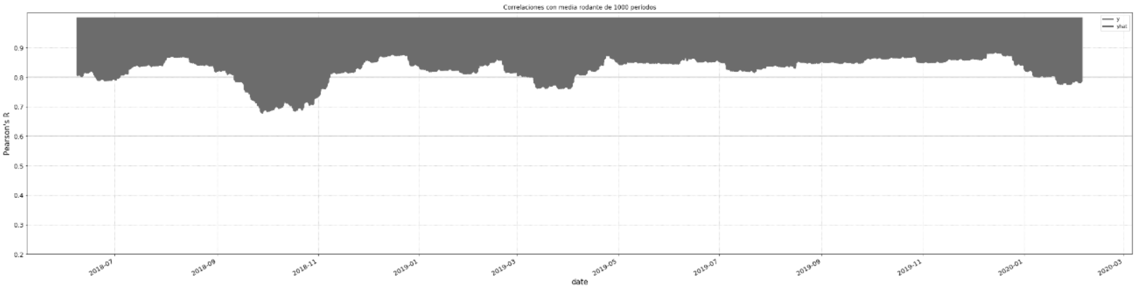


Figure 126. Pearson's R correlation value predicted baseline and traffic data of the city.

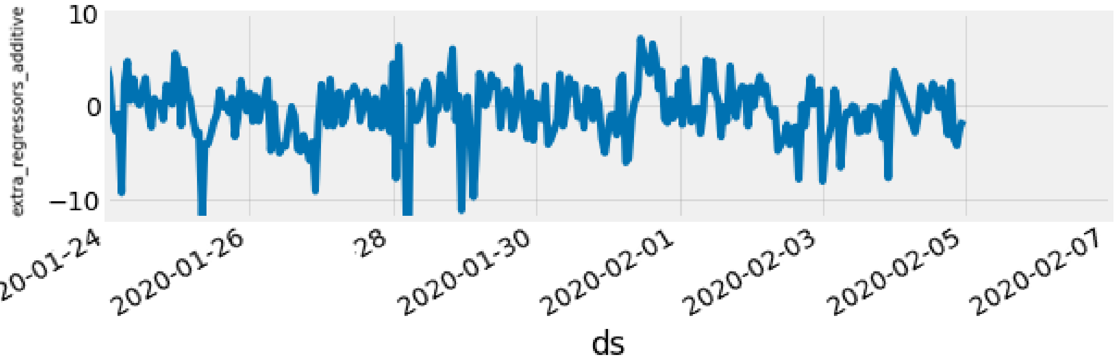


Figure 127. City traffic regressor addition effect in the prediction of traffic volume Gate 10A.

What is interesting in this case is to check how the prediction has changed considering the values of the extra regressor added (in this case: traffic) – Figure 127. As can be seen, t**his regressor had a lot of effect on the predicted values**. Despite the average is about 1 vehicle more per hour to be added to the expected value, the values swing between +22 and -20, which could mean even up to 29,3% of vehicles predicted, coming from the influence of vessels being berthed at the port that time.

The graph in Figure 128 shows the correlation between the predicted values (test data set - dotted line; train data set - solid line) and the ground truth. What can be concluded is that (except March and September) it is much better than the previous one, ranging from 0,78 to 0,88.

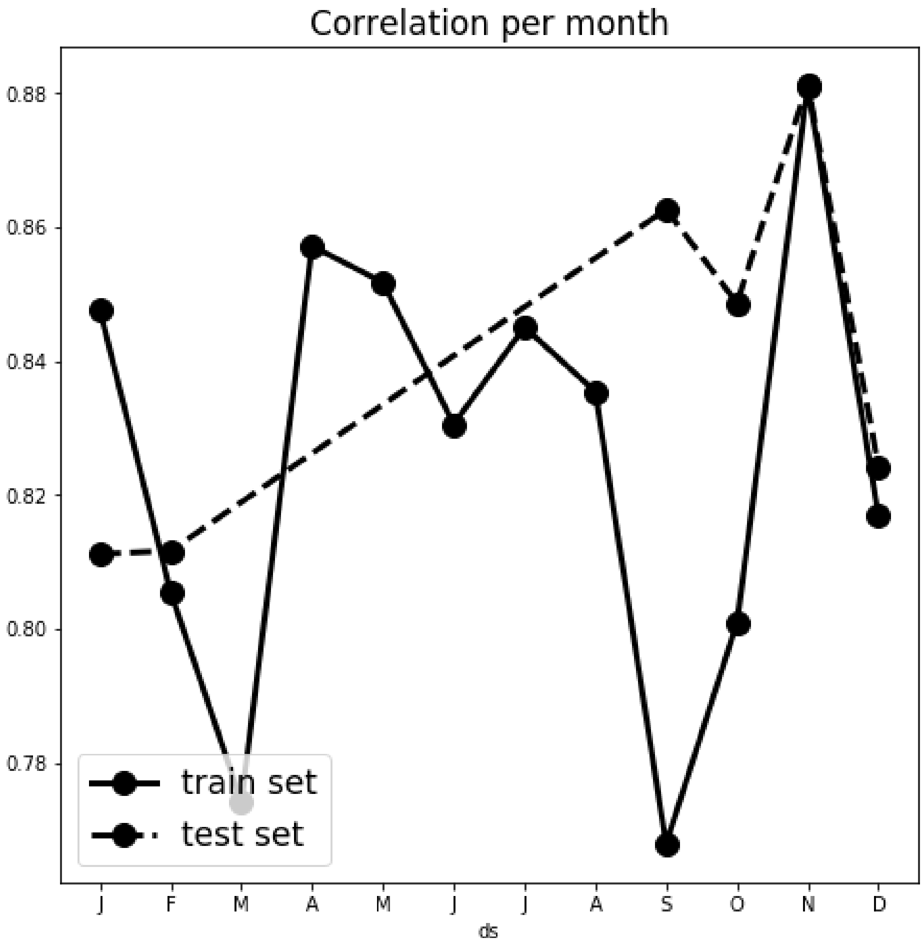


Figure 128. Correlation between predicted datasets per month- city traffic.

<u>Accuracy analysis of the model:</u>

Compared to the weather prediction, including traffic has much improved the metrics of correlation and average of mean absolute error. Although the comparison, in this case, cannot be directly done because the previous model was trained under disadvantageous conditions.

Table 22. Metrics Prophet model baseline and weather data divided train and test per day.

Metric	Test dataset	Train dataset
Correlation value y (values), yhat (predicted)	**0,858281**	0,840023
Daily average MAE	**12,16**	10,9701
Residuals median	1.678	
Residuals mean	3.2575	
Skew	-0,09359	
MAPE	184,52%	216,75%
MSE	253,35	210,963

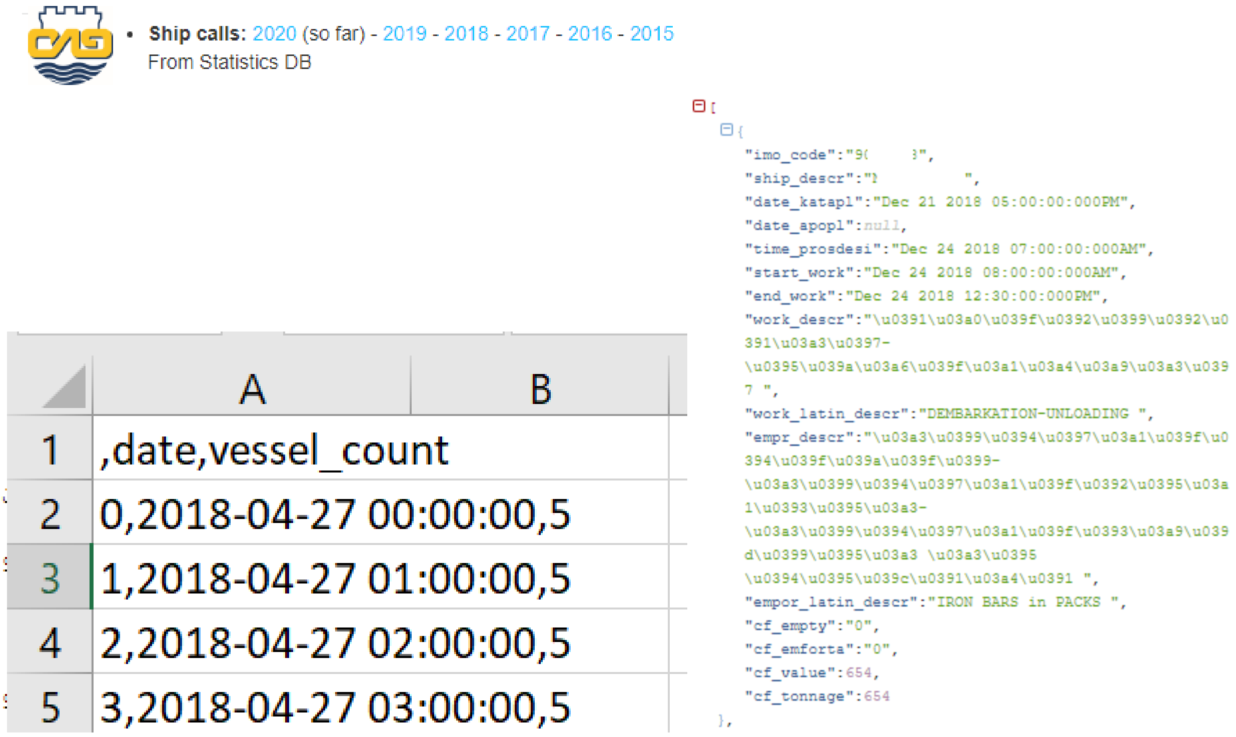
Concerning the baseline data, this model has improved all metrics. Therefore, it can be concluded that including traffic is beneficial. This model can be recommended to ThPA to be used in PIXEL.

5.2.3.5. Forecasting with vessel calls data

How much influence does the number of vessels being currently operated in the port to the queues in the ports for entrance/exit? Is it true that the more vessels (more business growth) the more trucks will be causing congestion at the gates of the port? Can the already tested model have a better performance if it is coupled with the number of ships? As designed in the methodology, it is an objective to correlate and include in the regression model the information about the vessels being operated.

For doing so, the technical team used the data source identified since D4.3 of the vessel operated by ThPA. This data (with interest in the historic offering from 2018) is served by the Information Technology department of the Port of Thessaloniki via a REST (Representational state transfer) API web created explicitly for PIXEL project. The query to that API returns all the vessels that were operated (one API per year) in a JSON (JavaScript Object Notation) format including rich information and details of every single vessel in a year. This information is only timestamp-referred by including fields of "start_work" and "end_work" of each ship, therefore certain data pre-processing after the acquisition was needed.

The procedure followed was: (i) downloading every data of all vessels processed in 2018-2020, (ii) fine-tuning the timeframe (28th April 2018 till 5 February 2020), (iii) grouping, filtering and counting the vessels to have a .csv prepared with the proper info: number of vessels at berth/manoeuvring in the port separated by periods of 60 minutes This is briefly illustrated through the images in Figure 129.



- **Ship calls:** 2020 (so far) - 2019 - 2018 - 2017 - 2016 - 2015
  From Statistics DB

	A
1	,date,vessel_count
2	0,2018-04-27 00:00:00,5
3	1,2018-04-27 01:00:00,5
4	2,2018-04-27 02:00:00,5
5	3,2018-04-27 03:00:00,5

```
[
  {
    "imo_code":"9(      3",
    "ship_descr":"\          ",
    "date_katapl":"Dec 21 2018 05:00:00:000PM",
    "date_apopl":null,
    "time_prosdesi":"Dec 24 2018 07:00:00:000AM",
    "start_work":"Dec 24 2018 08:00:00:000AM",
    "end_work":"Dec 24 2018 12:30:00:000PM",
    "work_descr":"\u0391\u03a0\u039f\u0392\u0399\u0392\u0
    391\u03a3\u0397-
    \u0395\u039a\u03a6\u039f\u03a1\u03a4\u03a9\u03a3\u039
    7 ",
    "work_latin_descr":"DEMBARKATION-UNLOADING ",
    "empr_descr":"\u03a3\u0399\u0394\u0397\u03a1\u039f\u0
    394\u039f\u039a\u039f\u0399-
    \u03a3\u0399\u0394\u0397\u03a1\u039f\u0392\u0395\u03a
    1\u0393\u0395\u03a3-
    \u03a3\u0399\u0394\u0397\u03a1\u039f\u0393\u03a9\u039
    d\u0399\u0395\u03a3 \u03a3\u0395
    \u0394\u0395\u039c\u0391\u03a4\u0391 ",
    "empor_latin_descr":"IRON BARS in PACKS ",
    "cf_empty":"0",
    "cf_emforta":"0",
    "cf_value":654,
    "cf_tonnage":654
  },
```

Figure 129. Vessels count per hour, data origin and procedure.

In this case, only an easy conversion was needed to align both datasets. The data of vessels were mostly ignored except for counting the number of vessels in a timeframe. Checking the fields of start and end work and making use of pandas' features, the dataset was built. Thereafter, a single data frame was constructed, including data from the baseline and the vessel count per hour.

Exploratory data analysis

Like it was done for the traffic of the city, before including this dataset within the Prophet model training, it was interesting to look at the vessel count data separately, trying to find seasonality, patterns and "hot periods", potentially to understand at a glance, if more vessels being operated in the port affects directly to the congestion at the gates of the Port of Thessaloniki.

For making this EDA, a simpler approach was followed including only one graph (Figure 130) that could serve the team (and the port personnel) to realise about relevant conclusions: the patterns per month. Vessels stay a varying number of hours in the port and no relevant knowledge is extracted from weekly/daily seasonality. However, different periods (e.g. months) may repeat certain behaviours, as the Port of Thessaloniki holds certain usual schedules of vessels, with a fixed yearly route.

The maximum rate (1.0 at the colour scale) was an average of 6 vessels berth at the same time during a timeframe of one hour. The moments of the year when fewer vessels are berthed in ThPA, are June and July, especially at nights. The vessels/hour rate is kept stable at 5,5-6 at central hours of the day (10 a.m. – 4 p.m.), while the less productive months during the day are March, June, July and August.

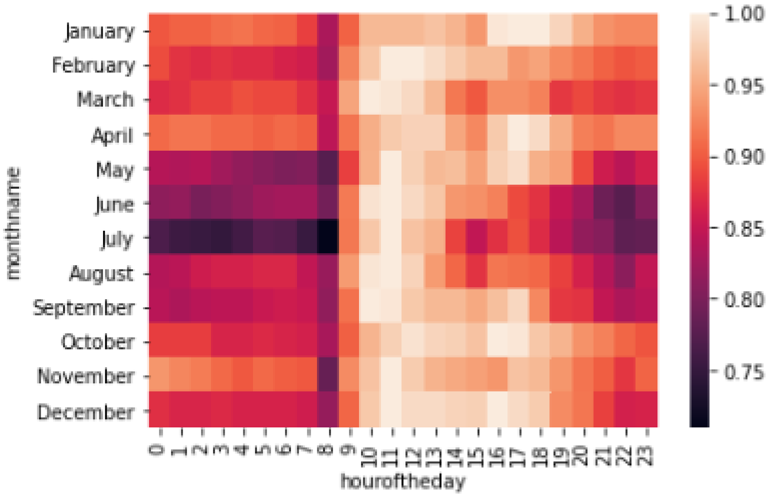


Figure 130. Heatmap number of vessels per hour (on average) analysed per month.

It was also observed that the moment of the week (on average) when fewer vessels are berthed in the port, is the night from Sunday to Monday. Besides, by average, the number of vessels berthed in ThPA per hour, decreases in more than 2, between one day and its corresponding night. This means that one-third of the vessels operated in ThPA, do not remain more than one day in the port.

Thereafter, a single data frame was constructed, including data from the baseline and the number of vessels berthed at the docks of ThPA, at each timeframe.

Facebook Prophet predictions

<u>Prediction with vessel calls (both datasets have a 60-minutes granularity) using it as an additive feature to the already built Prophet model:</u>

Next action performed was **to replicate the b) phase of the prediction procedure for baseline data** but including in the base data frame, one additional regressor: the vessel counts per hour. As done before, data was prepared, using the medfilt filter to smooth it.

Equally, for dividing the dataset the same delimiter was used: 20th September 2019. Therefore, the dataset was this time divided into 78,5% for training and 21,5% for testing, as for the baseline try.

Figure 131 presents the results with the vessel count data included the Prophet model, predicted for the "test" dataset time the values (and for the rest of the timeframe the black dots represent the real traffic volume measurements).

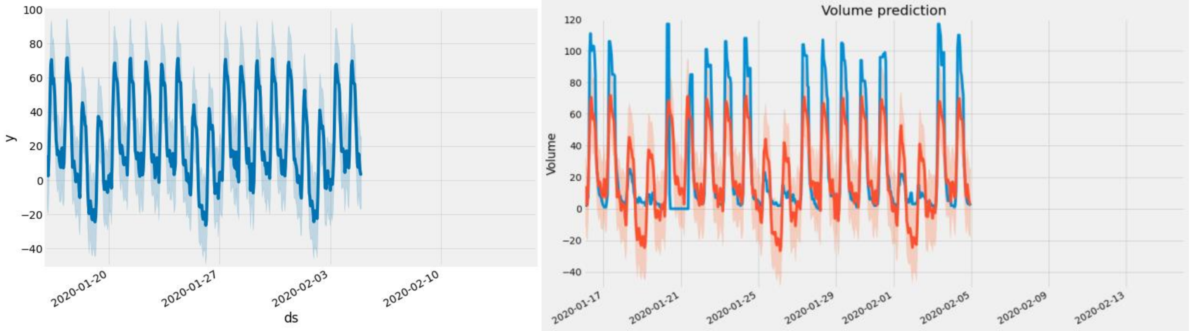


Figure 131. Predicted volume Gate 10A per day - divided dataset including vessels information.

As it can be observed, the trend is consistent with past data and with the curve for the 14 days horizon (February 2020) obtained in Figure 109 and pretty similar to the prediction obtained, while using traffic city data in the prediction as well. This is, at this point, a good sign as it seems to be a valid model.

The intuitions behind the previous reasoning were confirmed, when analysing more detailed graphs (Figure 132) about the prediction made by the recently trained model.

Looking at the top graph at the right, the trend curve seems close to the baseline prediction. Additionally, a dynamic range of max-min deviation in the predictions of the whole dataset (light blue shadow), looks very constrained.

Holidays were not very well predicted though, with circa 50% of deviation (less volume).

However, the most reassuring graph was a daily prediction. As it can be seen, the curve is tightly adjusted to the Figure 5 and Figure 31, with higher values in the range of 7-10 a.m., little decrease but stability around the mean till 3 p.m. and then decrease again.

This gives a good picture of the validity of this model.

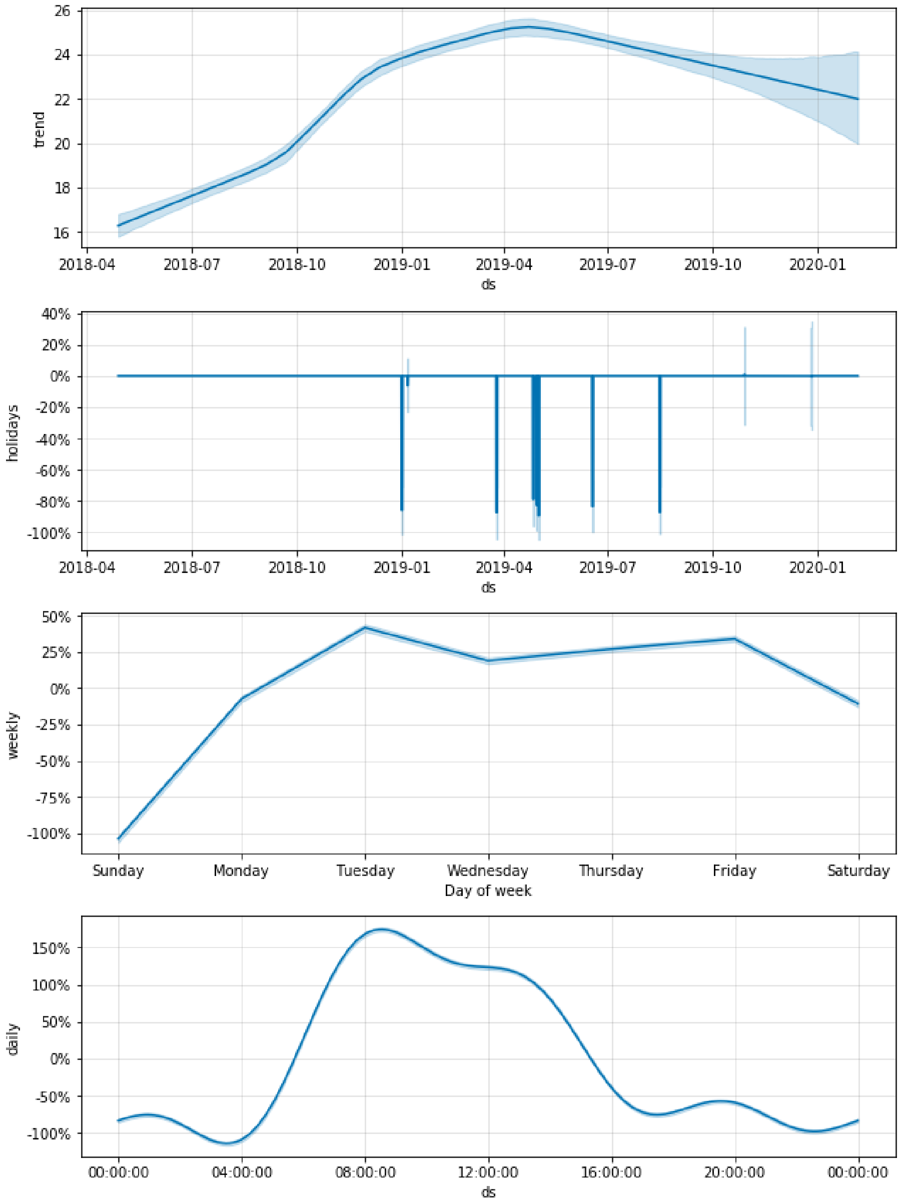


Figure 132. Detailed prediction graphs including vessel count per hour

However, the team wished to extract more information about the differences, correlations and statistic values between the predicted values (yhat) and the measurements (y) for the whole dataset timeframe. For doing so, in the following picture, the Pearson's R correlation value was calculated, for every single predicted point (traffic volume at Gate 10 A in a certain hour of the dataset timeframe). The results were satisfactory, as during almost all the period R was around 0.8 – Figure 133.

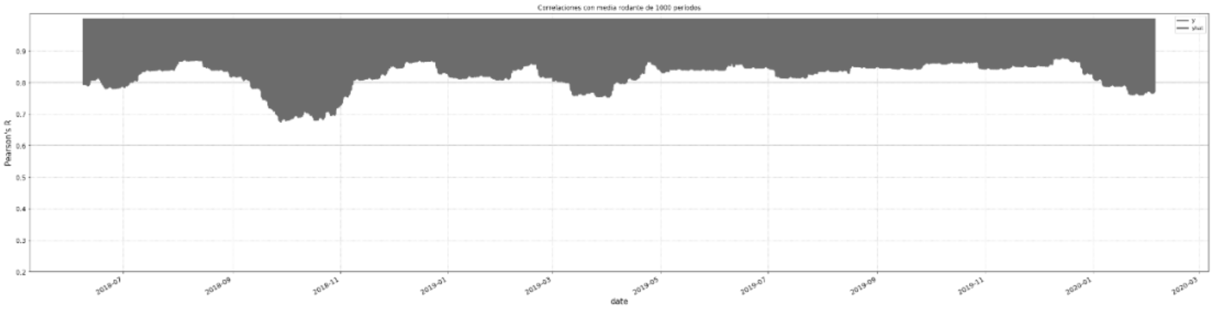


Figure 133. Pearson's R correlation value predicted baseline and vessel count per hour data

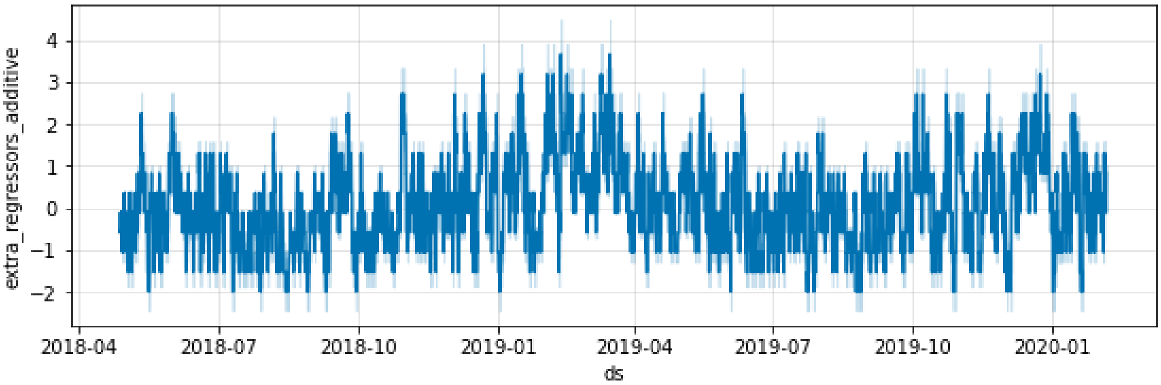


Figure 134. Vessels count regressor addition effect for the predicted values.

It is also interesting to check how the prediction has changed considering the values of the extra regressor added (in this case: vessel count per hour) – Figure 134. As can be seen, this regressor did not have too much effect on the predicted values, with an average of 0,4 vehicles more per hour to be added to the expected value. Maximum: +4 vehicles, minimum: -2 vehicles. This means less than 1% of vehicles predicted coming from the influence of vessels being berthed at the port at that time.

From another point of view, the monthly correlation between the predicted and real values (Figure 135) both test and train timeframe, emanates that (except for March and September) the accuracy is aligned with the traffic case and much better than for the weather, having values between 0,79 and 0,89. These values support the recommendation of this model to be used in PIXEL.

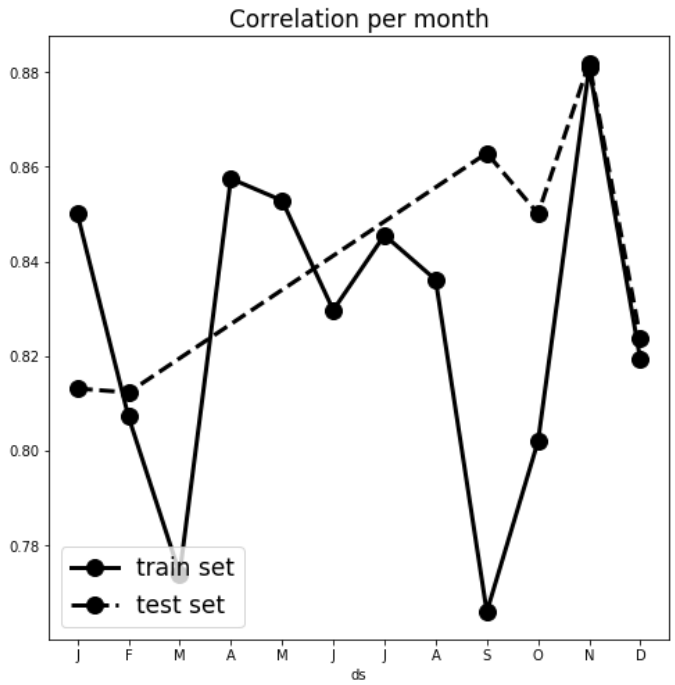


Figure 135. Correlation between predicted datasets per month – vessel count.

Accuracy analysis of the model:

Compared to the weather prediction, including vessel information has much improved the metrics of correlation and average of mean absolute error. Although the comparison, in this case, cannot be directly done because the previous model was trained under disadvantageous conditions.

Concerning the baseline data, this model has improved all metrics.

Table 23. Metrics Prophet model baseline and weather data divided train and test per day

Metric	Test dataset	Train dataset
Correlation value y (values), yhat (predicted)	0,843181	0,826586
Daily average MAE	14,655	12,8322
Residuals median	-0,2545	
Residuals mean	3,3173	
Skew	0,0823	
MAPE	200,8%	247,58%
MSE	378,958	288,985

However, in comparison with the traffic of the city influence model, the team discovered that dependency and accuracy values are a little worse. Therefore, it can be concluded that including vessel count is beneficial, but not as relevant as using the traffic of the city information. This model can be used by ThPA in PIXEL, but T4.5's team still recommend using the city traffic inclusion one.

For **other gates and directions**, the results were equivalent, thus every rationale above can be extrapolated. Metrics were similar, no remarks.

6. Prediction of renewable energy production

As described in the deliverable D4.3, ports must decarbonize their energy supplies. Moreover, they also need to find new business opportunities, to be able to diversify their activities. The installation of photovoltaic panels on the numerous warehouses' rooftops existing within the ports, is an interesting opportunity to produce carbon-free energy, while potentially creating a new source of income. This has been described by GPMB in the deliverable D3.4 "Use cases and scenarios manual v2". GPMB wants to estimate if the investment in solar panels on their warehouses is valuable. This is fully described in the port manager scenario (GPMB-PM-1). The following work contributes to this optimization of energy management.

In the previous deliverable, a state-of-the-art review has been done and four main approaches were highlighted, to predict solar radiation and thus PV system production:

- Numerical weather prediction: not explored in PIXEL, because it required to much computational power and a huge need for fine-tuning of specific data.
- Sky image: not explored in PIXEL, since it provides a too short-term prediction with a high equipment cost for sensors.
- Image satellite: used in PIXEL through the PVGIS (Photovoltaic Geographical Information System) tool. It provides good results for a typical day, month, or year.
- Time series prediction with statistical learning methods: These methods have been tested, using dataset example of real PV system production.

The work carried out here, comes in two distinct but complementary approaches as described in the "Plan and future work" section of deliverable D4.3. The first one can be used directly by ports when they are not equipped with solar panels to assess the viability of investing in such a system. The second one will be used by ports equipped with a PV system, when they want, for energy management purposes, to predict their energy production, for the following day or week.

The first approach consists in allowing the evaluation of the photovoltaic production potential of a port, according to its location. For this, PVGIS was used, an already existing tool developed by the European Commission Joint Research Centre. Using PVGIS and its database on irradiance and weather condition, a typical day, month, or year can be described, for the one-site irradiance. PVGIS also provides the ability to use well-known physical models of solar panels (based one technical specification). The main work here has been to interact with web services like PVGIS, to obtain historical data and extract a typical irradiance. In the following, PVGIS's use in PIXEL is explained, show results about the potential PV production at GPMB and compared it with its typical energy consumption. The same methodology can be used for every port, asking itself about investing in a PV system.

The second approach uses historical production data and weather conditions to predict energy production for the following days. In this approach, there is no need to have a technical description and specification of the PV system. As written in D4.3, since no ports in PIXEL are equipped with solar panels yet, this approach has been developed and tested using real data coming for PV output (data source described in D4.3). In the following, the different methods that have been tested are presented and provide a full description of the one that gives the best prediction. The full methodology and results can be easily transferred and adapted to port.

6.1. Data analysis of the photovoltaic production potential with PVGIS

In the following, PVGIS is used for studying a PV installation design, based on the GPMB use case. The same methodology can be used for every port, asking itself about investing in a PV system.

6.1.1. PVGIS presentation

PVGIS[60] has been developed at the European Commission Joint Research Centre, to contribute to the dissemination of knowledge and data about irradiance and PV system. It is completely free to use, with no restrictions on what the results can be used for, and with no registration necessary. It provides several features, from PV plant production calculation for multiple time resolutions (from annual to hourly), to PV electricity cost calculation (including the investment amortization through a "Levelized Cost of Energy" method). Furthermore, PVGIS provides both an online web application and an API (automatic requesting without user interface) with exportable data (CSV or JSON). For all these reasons, it appears logical to use it in PIXEL as a tool for assessing potential photovoltaic production of port.

PVGIS allows user to get data on solar radiation and photovoltaic (PV) system energy production, at any place in most parts of the world. This solution is a combination of two components. The first part is an irradiance and weather reconstituted database[61]. Values for a specific location can be accessed and a specific time (or averaged). The second part of PVGIS uses well-known physical models of PV installation. This tool allows users, to estimate the potential of energy production of a PV system (estimation of what the installation would have produced, for similar weather condition to the ones recorded in the database). PVGIS is a tool with a robust academic background, legitimizing its value. A demonstration follows, of how PVGIS can be useful to ports in the PIXEL context, to estimate a "typical" PV production.

6.1.2. Application to GPMB use-case

In this section, the typical PV production, provided by PVGIS is investigated and port electrical consumption of GPMB evolution. This provides valuable knowledge that can be used to evaluate if an investment in solar PV is profitable. The methodology that has been followed, can be used by any port. PVGIS provides both annual average values of energy production and in-plane solar irradiation. It also provides the monthly values of energy production, which is useful, to provide a good estimation of the potential of investing in a PV system. Results consider the different losses in the PV output, caused by various effects.

6.1.2.1. Methodology

Objective

PVGIS provides automatic recommendation about panels orientation (slope and azimuth), to maximize the PV installation total production. But it does not provide an automatic sizing (number of panels, e.g. installation nominal power) recommendation, to optimize the matching between PV production and electrical consumption. Here a trial and error approach has been used to evaluate PVGIS usefulness for installation design, regarding electrical needs.

Electrical consumption

Currently, the only available data regarding GPMB electrical consumption, are monthly data covering January 2014 to April 2017. It is planned that electrical sensors will be integrated into the PIXEL platform during WP7. These future data could be used to have a more precise analysis. In the following, only monthly data is presented.

[60] https://ec.europa.eu/jrc/en/PVGIS/docs/usermanual

[61] Different databases are available through PVGIS. In the following work PVGIS-SARAH database is used. This database includes solar radiation from 2005 to 2016, with a spatial resolution of 5 km. https://ec.europa.eu/jrc/en/PVGIS/docs/usermanual#fig:default_db

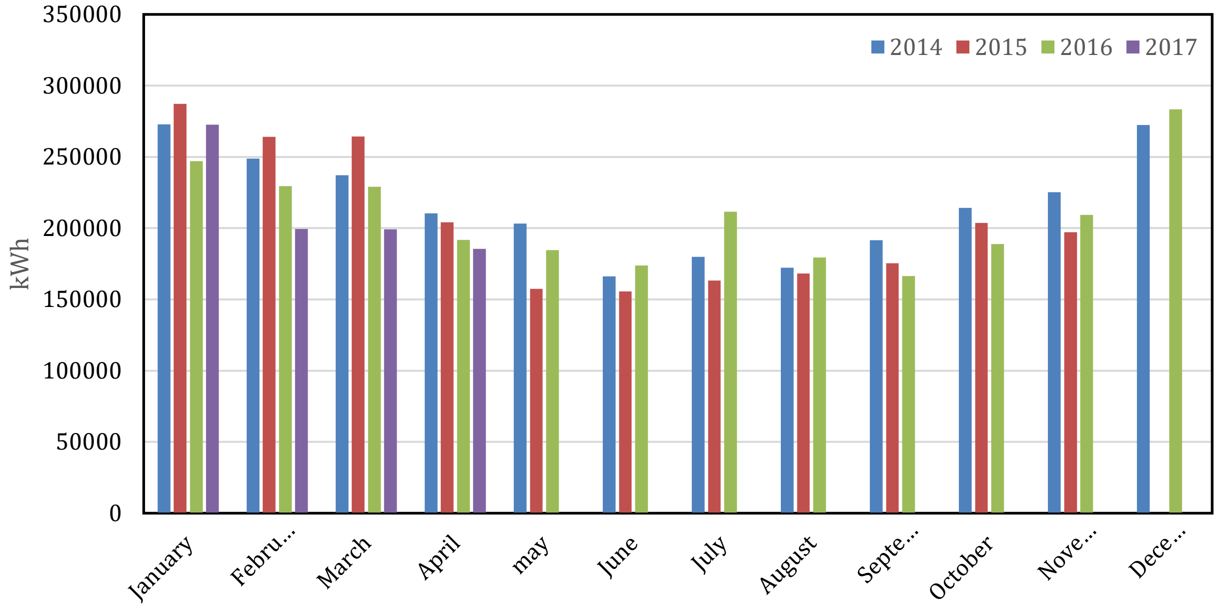


Figure 136: Monthly electrical consumption of GPMB

PV system production

Some available information is used to design a hypothetic PV system in GPMB:

- PV panels would be installed on the warehouse rooftop, which induces constraints about the PV installation weight. This may limit installation to a fixed panel since tracking systems are heavier.
- The available rooftop surface is about 30 000 m². Without information about slope and azimuth, PVGIS optimum calculation will be used. This can significantly overestimate real production capacity.
- GPMB has several terminals along the Gironde estuary. For warehouses in the north of the estuary, it can be assumed that no shadow will decrease the PV production. But for others, shadows might have an impact. However, shadows have not been considered in calculations, which also might overestimate PV production.

6.1.2.2. Results

Monthly scale

PVGIS can be used in two ways. A first approach is to start from a hypothetical PV installation to get a typical production. A second approach is to start from a targeted PV production, to deduce the PV installation sizing.

Table 24. Design parameters of the PV system and potential production

		PVI			Average consumption
		'Uncapped'	'Equal Total '	'Equal June'	
Operational consideration	Objective	Maximize production	Total production = total consumption	June production = June (min) consumption	NA
	Area covered	100 %	46 %	29 %	
	# panel	23 585	10 966	6 900	
PVGIS parameters	Nominal Power	4 245 kWp	1 974 kWp	1 242 kWp	
	Losses	14 %			
	Slope	38 ° (auto)			
	Azimuth	1 ° (auto)			
	Shade	(auto)			
Production (kWh)	Total	5 397 128	2 509 760	1 579 089	2 509 347
	Mean (monthly)	449 761	209 147	131 591	209 112
	Min (monthly)	255 230	118 687	74 675	165 199
	Max (monthly)	577 080	268 352	168 842	277 919

To begin with, a PV installation is considered, with the purpose to get the estimation of maximal PV production potential in GPMB. The whole rooftop area is used for panels deployment. This configuration is denoted as "Uncapped" PV installation in the following and has a total production estimated[62] to 5.4 GWh, which is higher than (by a factor 2) the total GPMB electrical consumption (2,5 GWh). If the overproduction cannot be sold to an external energy provider, then the PV installation should be downsized to meet GPMB electricity consumption. This first result just provides the total energy that can be produced, but there is no way of knowing if energy production can always be higher than energy consumption.

To have a total production equal to the total consumption, the PV installation nominal power should be decreased by the same factor, 2. This leads to a decrease in the number of panels to 10.966 (corresponding

[62] This is a maximal potential value, since as exposed in the methodology section, parameters provided to PVGIS are optimistic.

to 13.948 m², 46% of the total rooftop area). This second set of inputs denoted hereafter as ‘Equal Total” PV installation, has been used in PVGIS and the corresponding production is presented below.

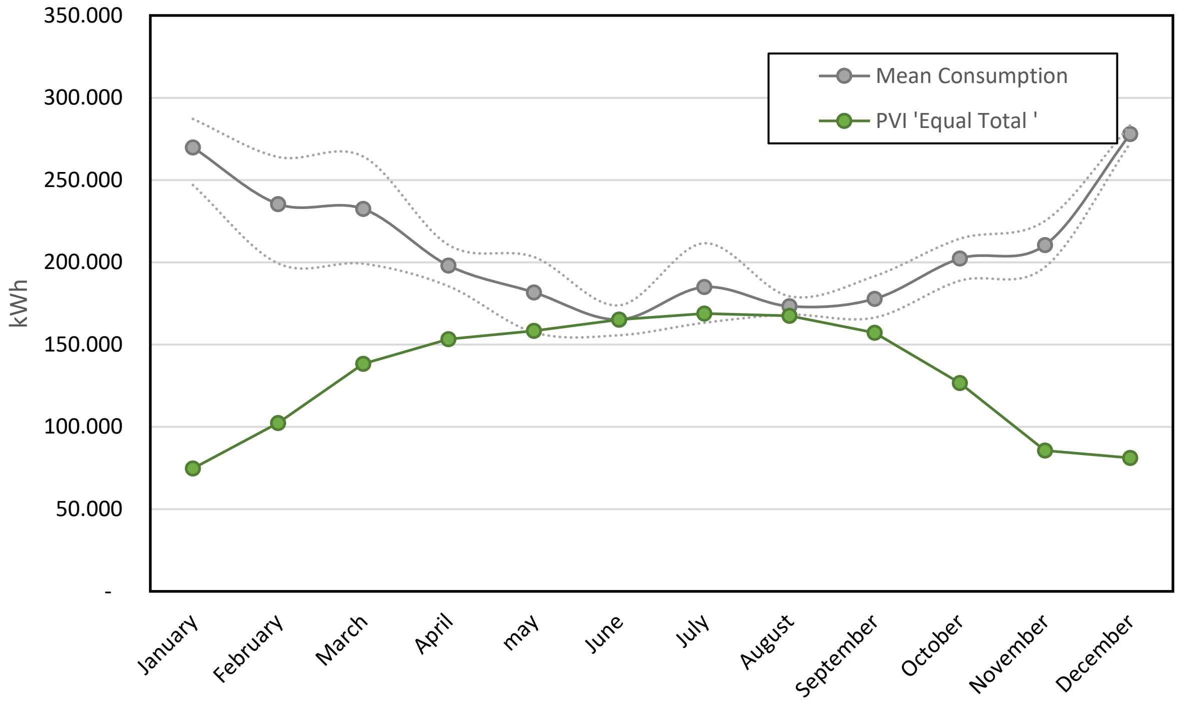


Figure 137: Comparison between electrical consumption of GPMB and its potential PV production

Such focus on total amount over a whole year is not enough for a PV installation design. It is necessary to consider the evolution of the balance between production and consumption across time. Indeed, the PV production has a usual seasonal variability, with a maximal production during summer. But GPMB's electrical consumption also has a seasonal periodicity. Moreover, this periodicity is roughly inverse to the production’s one, as shown in the figure below (this can be explained using light during winter).

Consequently, “Equal Total” PVI (Photovoltaic Installation) shows a balanced alternation of overproduction (about 0,5 GWh during summer) and underproduction (about -0,5 GWh during winter). Such behaviour at a daily scale could be managed through batteries. The noon overproduction would be accumulated, to get redistributed later in the afternoon and the next morning. But at the annual scale, both the required amount of battery and the energy preservation during months may be an issue.

Thus, GPMB may desire to design its PV system to have maximal production, without facing overproduction. Considering the periodicity, this can be achieved, by targeting a production during June equal to the consumption in June, as it is the yearly minimal consumption. Such a PV system is designed as “Equal June”.

As shown in the figure below, the PVI Uncapped is the only configuration that can cover the monthly electrical consumption every month. The PVI Equal Total configuration cannot cover consumption in winter months. The PVI Equal June design will only cover electrical needs during the summer months. Knowing this, GPMB can decide which is the better strategy: self-consumption, self-consumption and electricity sold when over-production, self-consumption with help of batteries, … To help this kind of decision a deeper economic study must be conducted by GPMB.

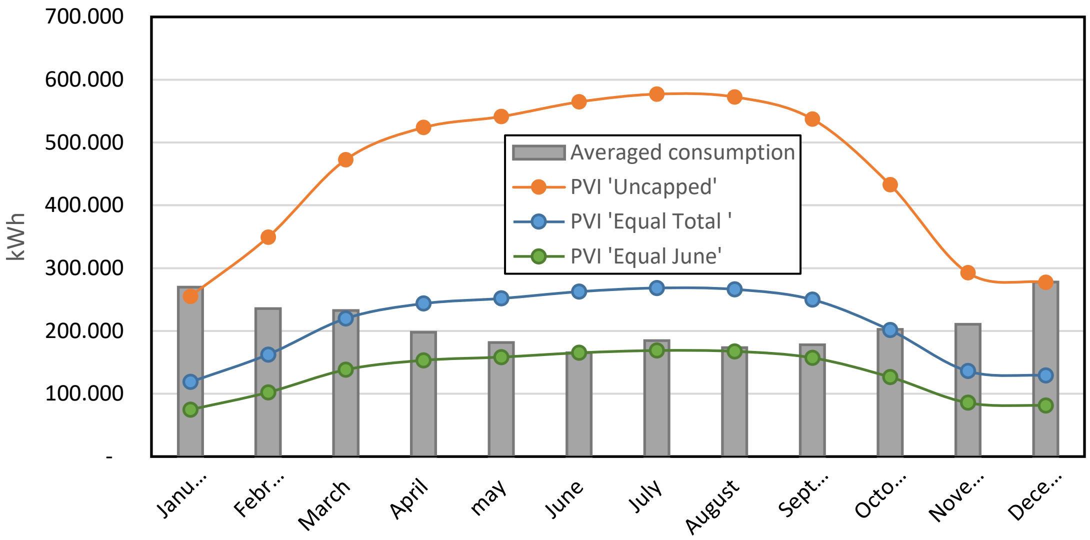


Figure 138: Production of a PV system compared to the average GPMB electrical consumption.

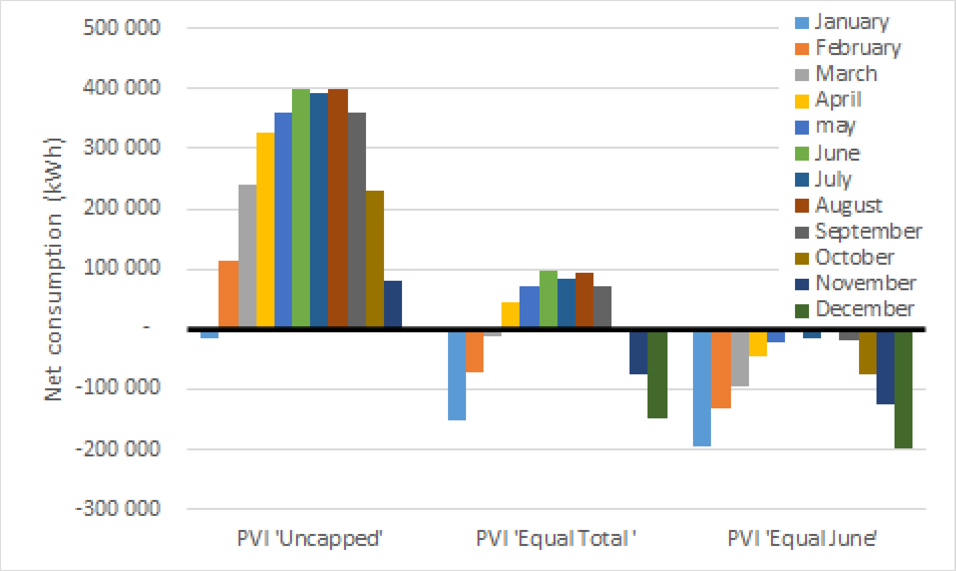


Figure 139: Balance between production and consumption for the different PV systems.

Hourly scale

Since there is currently only monthly data, no data exploration can be provided for daily balance. But as a short illustration, one may consider a situation where the production peak (around noon) corresponds to a whole in consumption (lunch break). This can be an important consideration for battery sizing. In the same way, if it appears that the increased electrical consumption during winter may be significant due to early and late hours, this could suggest that action on the lighting system may be the most effective level for energy management improvement.

General consideration

Note that apart the number provided by PVGIS, each iteration of the trial and error approach can provide valuable contextual information that should be considered. As an example, the less the rooftop area is covered with panels, the more the hypothesis of ideal panel orientation gets credible. Firstly, because only the most suitable warehouse can be used, secondly because heavier panel racks (instead of building integration) could be considered.

6.1.3. PVGIS usefulness conclusion

The underlying assumption in the use of PVGIS is that averaged past irradiance conditions in the database are representative enough of the future condition that will be encountered. Nevertheless, PVGIS provides a useful estimation of the possible PV installation output in the PIXEL context. As illustrated for GPMB use-case, PVGIS proposes a suitable balance between estimation accuracy and inputs requirements. Then it seems to be a great tool for ports, that want to get a first estimation of their photovoltaic potential or sketching the suitable installation's size.

Furthermore, this tool is based on the solid scientific background, richly documented and has long-term support through European Commission Joint Research Centre. Finally, by allowing to build a scenario and test it, this tool meets the PAS (Port Activity Scenario) modelling approach to complete it. Numerical tools as PVGIS and PAS modelling can provide valuable information about the port's energy needs and use. If they are combined with the port agent's knowledge about operations and processes, more efficient energy management will be settled progressively.

6.2. Prediction of photovoltaic production based on historical production

In this section, historical production data is used and weather conditions to predict energy production for the following days. In this approach, there is no need to have a technical description and specification of the PV system.

As previously described, none of the ports in PIXEL has a PV system installed. Thus, it was decided to use open data that represent the production of a real PV system. Data have been obtained using PVoutput4, allowing us to have access both to live and historical data of the generated power, the efficiency (that is directly linked with weather conditions: high efficiency means a sunny day) and temperature, with a resolution between 5 and 10 min. An initial description of this dataset had been described in section 7.4 of the deliverable D4.3.

This data has been used to set up a methodology for forecasting solar PV production, knowing the date, type of weather conditions. Different models based on past data have been developed and tested. Readers must keep in mind that, the models are not expected to be able to predict solar production in GPMB, but the methodology will be transferable when the PV system will be installed.

6.2.1. Methodology description

For the prediction of photovoltaic production, the procedure followed 4 stages, described thereafter.

Stage 1: Data understanding

Descriptive statistics and visualization techniques are used, to understand the data content, assess data quality and discover initial insights about the data. Sample Entropy is chosen to evaluate the complexity of the time-series signal. Sample Entropy combines two advantages: data length independence and relatively trouble-free implementation. The next step is graphing the data using the techniques decomposition and autocorrelation to find if there are consistent patterns, a significant trend and a seasonality. Any outliers in the data have been checked and explained by domain experts.

Stage 2: Data pre-processing

In this stage, a dataset is constructed, that will be used in the subsequent modelling stage. Data preparation activities include data cleaning (dealing with missing or invalid values, eliminating duplicates, formatting properly) and transforming data into more useful variables.

Hourly data were downsampled to daily data, weekly data, and monthly data to make predictions of energy production at various levels.

Stage 3: Choosing and fitting models

In this stage, the aim is to compare the classical methods with a deep learning method. Because the dataset has a significant trend, a seasonality and more features should be included in the dataset, a linear method was chosen: SARIMAX (Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors)[63]. This method is an extension of the SARIMA model that also include the modelling of exogenous variables.

Another model to be tested is additive Holt-Winter Exponential Smoothing. Exponential smoothing is a time series forecasting method for univariate data that can be extended to support data with a systematic trend or seasonal component. It is a powerful forecasting method that may be used as an alternative to the popular Box-Jenkins ARIMA family of methods.

The prophet is an open-source library, developed by Facebook, utilizes a Bayesian-based curve fitting method to forecast the time series data. It is based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. The Prophet package provides intuitive parameters, which are easy to tune. It does not require much prior knowledge or experience of forecasting time series data since it automatically finds seasonal trends beneath the data. Hence, it allows non-statisticians to start using it and get reasonably good results.

Deep learning methods offer a lot of promise for time series forecasting, such as the automatic learning of temporal dependence and the automatic handling of temporal structures like trends and seasonality. The goal is to develop an end-to-end forecast model for multi-step time series forecasting, that can handle multivariate inputs. A model recurrent neural networks are built, like LSTM (Long Short-Term Memory network), to add the explicit handling of order between observations when learning a mapping function from inputs to outputs, not offered by MLP or CNN (Convolutional Neural Network). They are a type of neural network that adds native support for input data comprised of sequences of observations.[64]

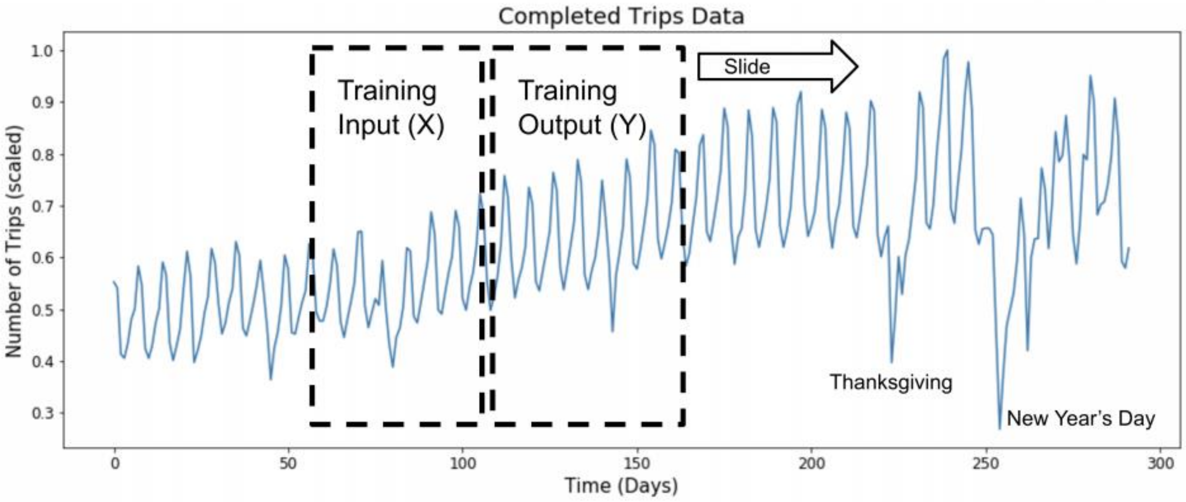


Figure 140: slide window approach used in LSTM model[65]

The training data is normalized between 0 and 1. The sliding window approach (as described in the figure below) is used, to construct the training dataset by splitting the historical data into sliding windows of input and output variables. The LSTM was defined, with 20 neurons in the hidden layer and 1 neuron in the output layer for predicting energy production. The input shape will be a one-time step with one feature at least because the performance of the model needs to be evaluated with a univariate training dataset and another multivariate training data including the weather. The number of neurons in the output layer can be

[63] https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/

[64] https://machinelearningmastery.com/deep-learning-for-time-series-forecasting/

[65] Laptev, Nikolay et al. "Time-series Extreme Event Forecasting with Neural Networks at Uber." (2017).

changed to get a multi-step prediction. The loss function is the mean squared error (MSE). The optimizer for finding the minimum of a function is Adaptive moment estimation (Adam). Adam is a popular algorithm in the field of deep learning because it achieves good results fast. To prevent overfitting, models should not get too complex. One way to penalize complexity is to multiply the sum of squares with another smaller called weight decay. Then this result is added to the loss function. Grid search is used to find good values for the hyperparameters.

For every model, a prediction interval for the forecast values is provided.

Table 25: Prediction model description.

Model	Type	Prerequisites	Input	Hyper-parameters
SARIMAX	Statistics	A stationary time series.	• Averaged weekly production • Averaged monthly production	Differencing order and seasonal order
Holt-Winter	Statistics	-	• Daily production • Averaged weekly production • Averaged monthly production	Type of trend Length of a season
Prophet	Bayesian	-	• Daily production • Averaged weekly production • Averaged monthly production	-
LSTM	Neural network	Split data into samples with input and output components. Data normalisation.	• Daily production	Number of epochs, Sequence length, Number of layers, Number of neurons by layer, Batch size, Weight decay
Multivariate LSTM	Neural network	Same as LSTM.	• Daily production • Daily weather	Same as LSTM

Stage 4: evaluating a forecasting model

Evaluation of the quality of the models is done by analysing the residuals. The “residuals” in a time series model are what is left over, after fitting a model. For many (but not all) time series models, the residuals are equal to the difference between the observations and the corresponding fitted values.[66] Residuals are useful in checking whether a model has adequately captured the information in the data. A good forecasting method will yield residuals has the following properties:

- The residuals are uncorrelated. If there are correlations between residuals, then there is information left in the residuals which should be used in computing forecasts.
- The residuals have zero mean. If the residuals have a mean other than zero, then the forecasts are biased.
- The residuals have constant variance.
- The residuals are normally distributed.

To evaluate forecast accuracy, the available data was separated into two portions, training and test data, where the training data is used to estimate any parameters of a forecasting method and the test data is used

[66] https://otexts.com/fpp2/residuals.html

to evaluate its accuracy. The sizes of training data and test data are 80% and 20%. Forecast accuracy is measured by using Root mean squared error.

6.2.2. PVOutput data analysis

As previously described, none of the ports in PIXEL has a PV system installed. Thus, it was decided to use open data that represent the production of a real PV system. Data have been obtained using PVoutput[67] allowing access to both live and historical data of the generated power, the efficiency (that is directly linked with weather conditions: high efficiency means a sunny day) and temperature, with a resolution between 5 and 10 min. An initial description of this dataset had been described in section 7.4 of the deliverable D4.3.

This data has been used to set up a methodology for forecasting solar PV production, knowing date and type of weather conditions. Different models based on past data of solar PV production and weather conditions have been developed and tested. Readers must keep in mind that the models are not expected to be able to predict solar production in GPMB, but the methodology will be easily transferable when the PV system is installed.

Data from Jan. 2013 to Feb. 2020 are used, which represents 2541 days. Days are split so that 80% are used as training data, for analysis and prediction.

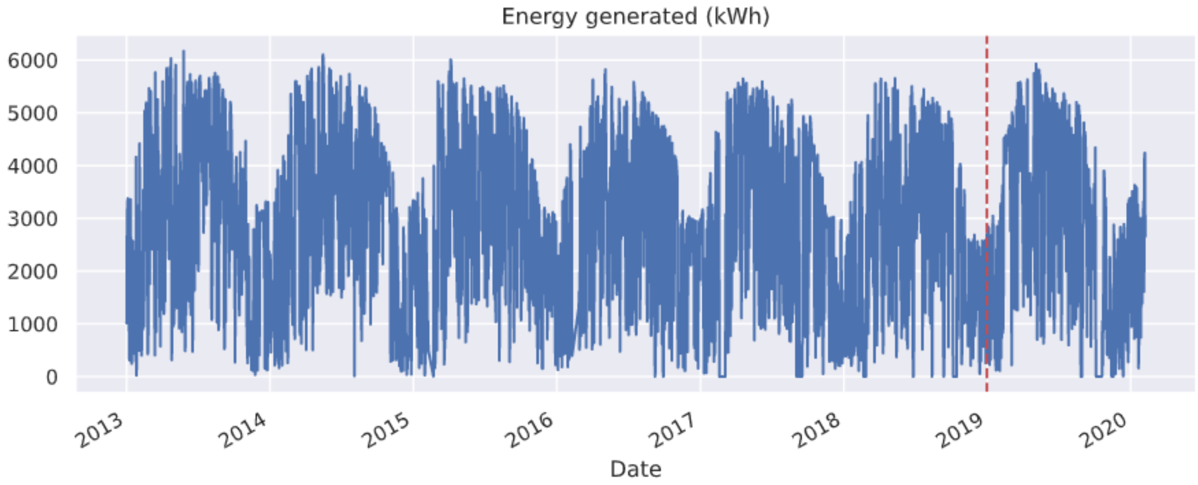


Figure 141: Photovoltaic production (PVOutput raw data).[68]

A decreasing trend is observed, as well as a clear yearly seasonality component. This trend can be understood as the decrease of solar panel system efficiency over time.

[67] https://pvoutput.org/

[68] Red dots represent the training/testing boundary.

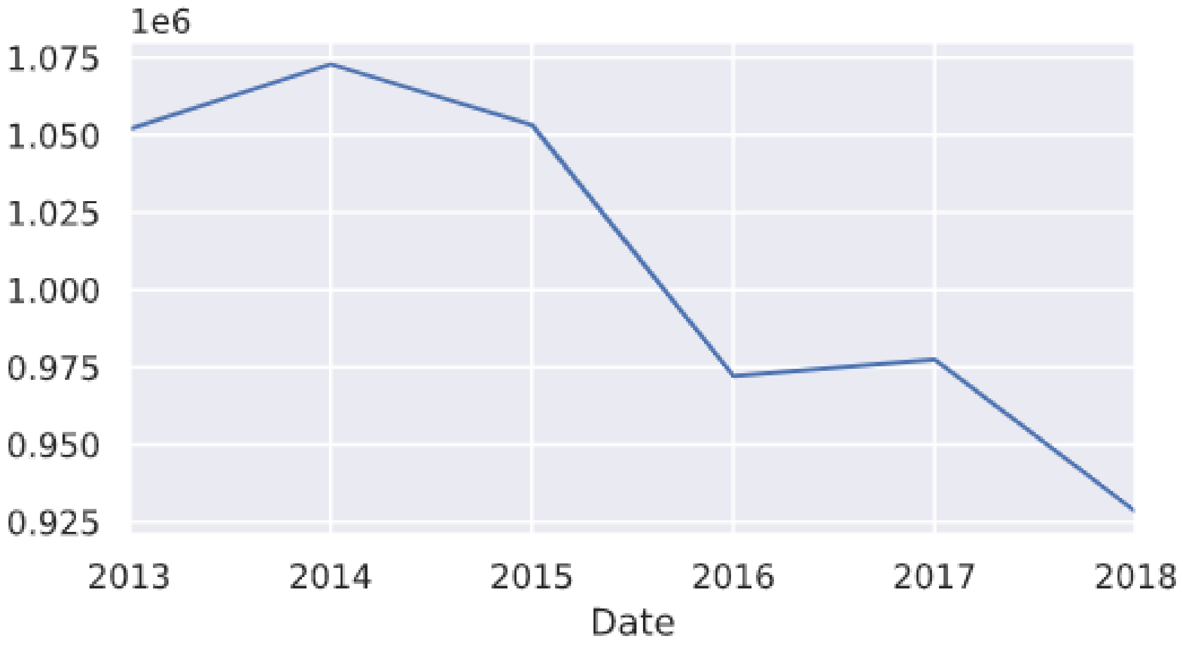


Figure 142: Yearly mean trend.

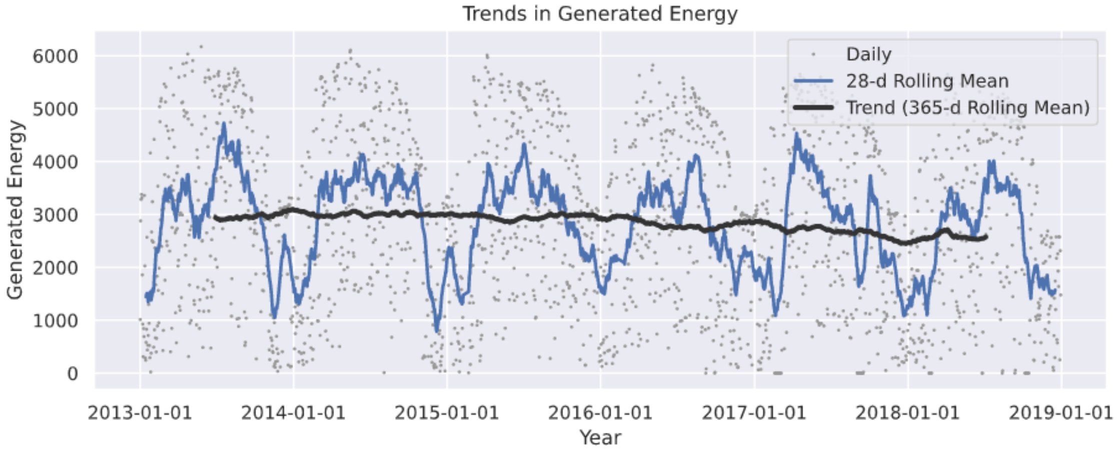


Figure 143: Mean trend details according to different smoothing.

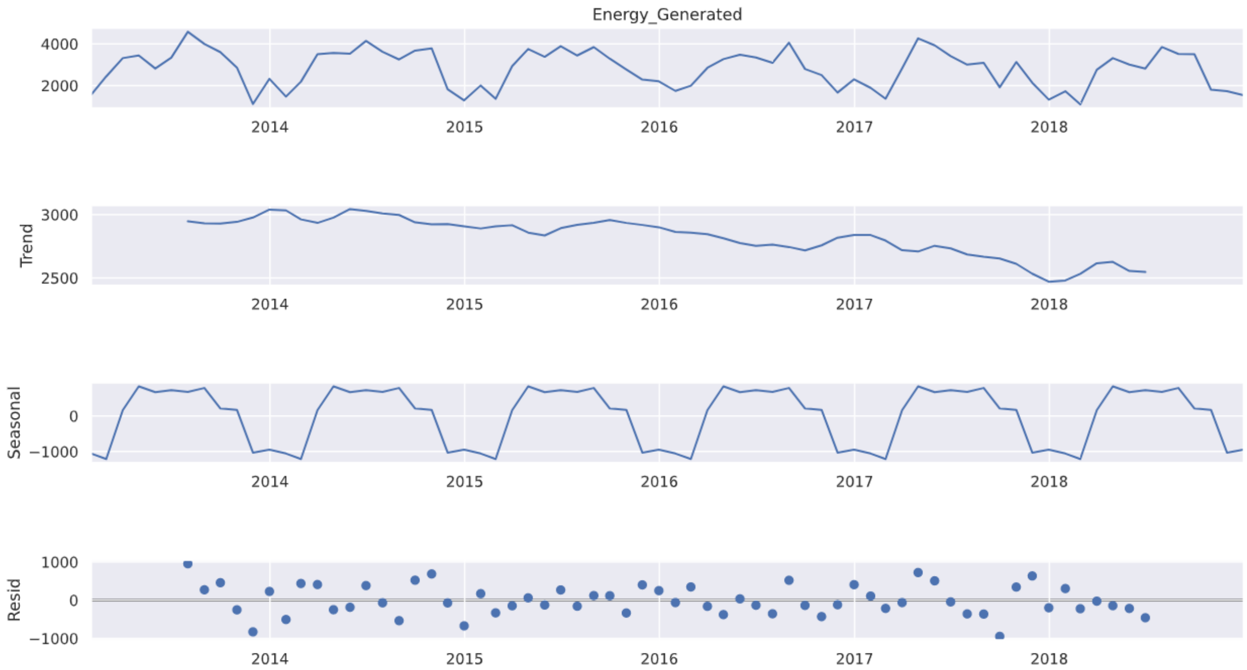


Figure 144: Seasonal decomposition of training data (monthly aggregation).

Dataset entropy is 1.67, compared to a random distribution whose entropy exceeds 2.

6.2.3. Models predictions

Models described in the previous section are fitted on the 80% trained data and used to forecast the remaining 20%. Data are either aggregated monthly (for SARIMAX, Holt Winter and Prophet models) or used daily (Holt Winter, Prophet, LSTM and LSTM with weather).

6.2.3.1. Monthly predictions

With data averaged by month, the model must forecast only 14 months. Results among the three models tested are similar, with 0-mean and normally distributed residuals and comparable RMSE values. Prophet model gives here slightly best results, with a smooth prediction that handles outliers well.

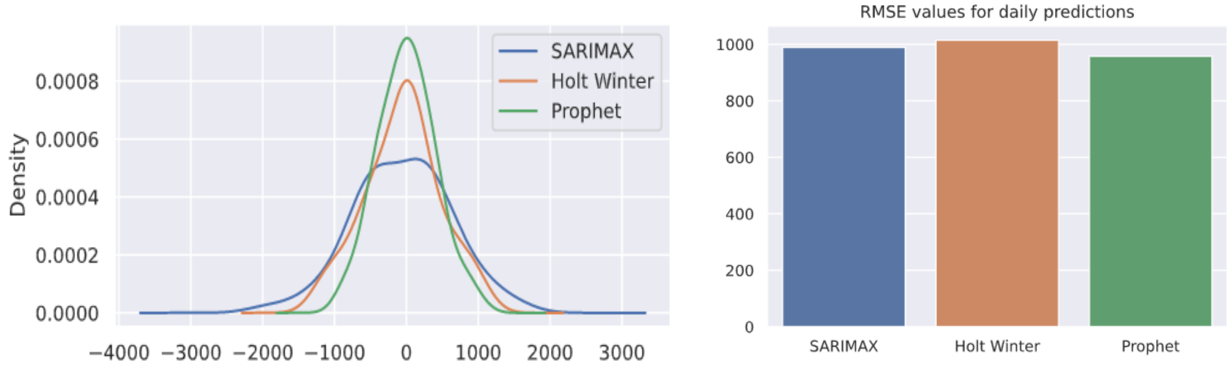


Figure 145: (a) Residuals distribution for the monthly fit, (b) RMSE for monthly predictions.

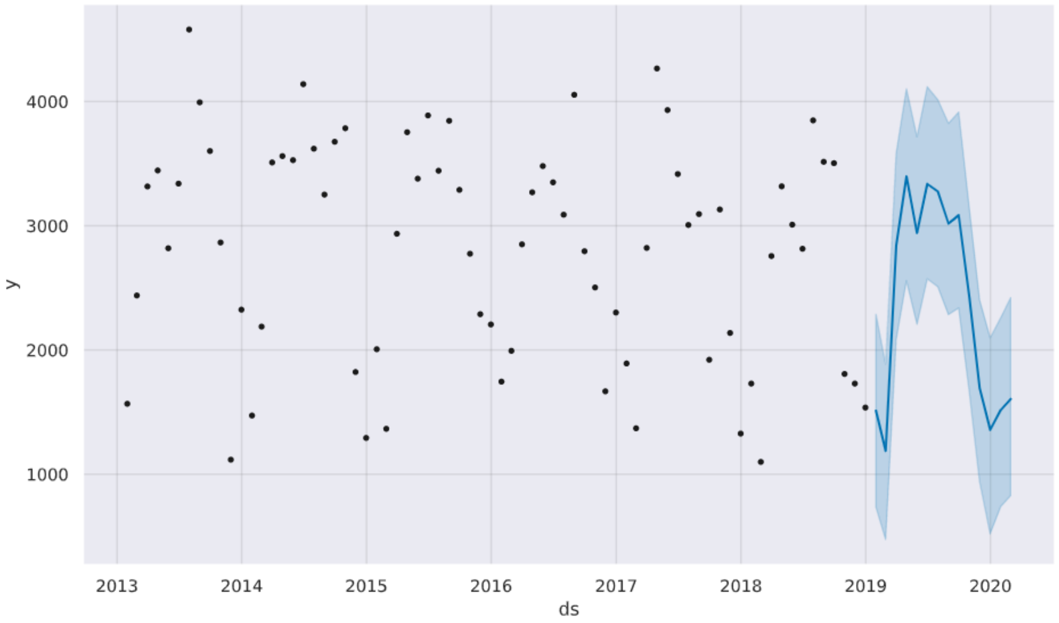


Figure 146: Prophet model monthly predictions.

6.2.3.2. Daily predictions

Neural network models cross-validation

LSTM models are at first cross-validated, to ensure results reliability. Forward chaining is used on the “80% training set”, with a one-year increase at each fold for training and one year for testing[69].

LSTM model results

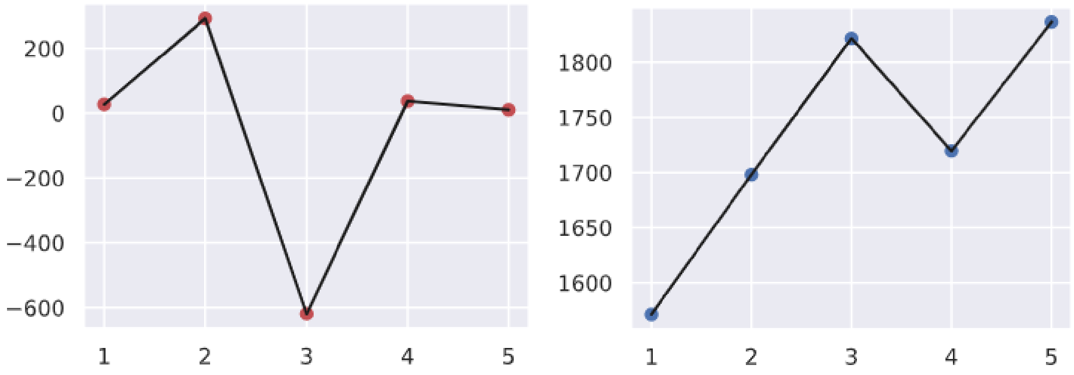


Figure 147: (a) Mean residual error according to fold training set length (in a year) for LSTM. (b) RMSE according to fold training set length (in a year) for LSTM[70]

[69] https://robjhyndman.com/hyndsight/tscv/

[70] 5 years of training: +17% RMSE

LSTM multivariate model

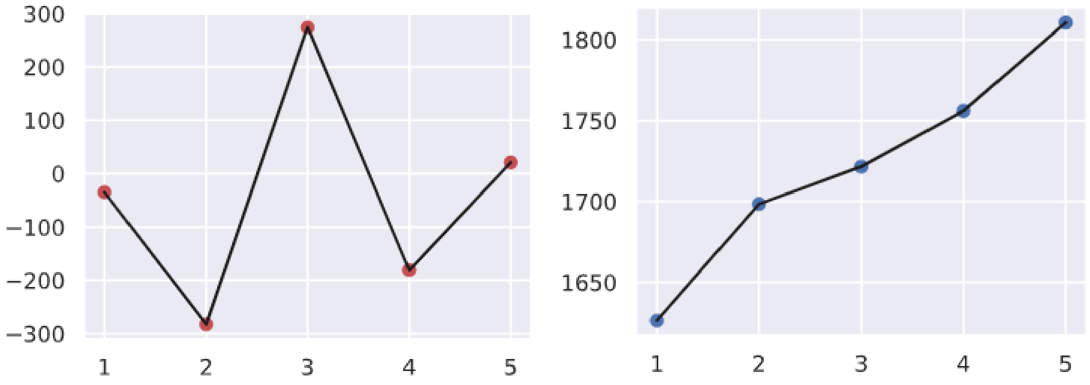


Figure 148: (a) Mean residual error according to fold training set length (in a year) for LSTM_weather. (b) RMSE according to fold training set length (in a year) for LSTM_weather.[71]

While RMSE increases with the length of the training set, the mean residual error draws near 0 when the training set is large.

Predictions on test data

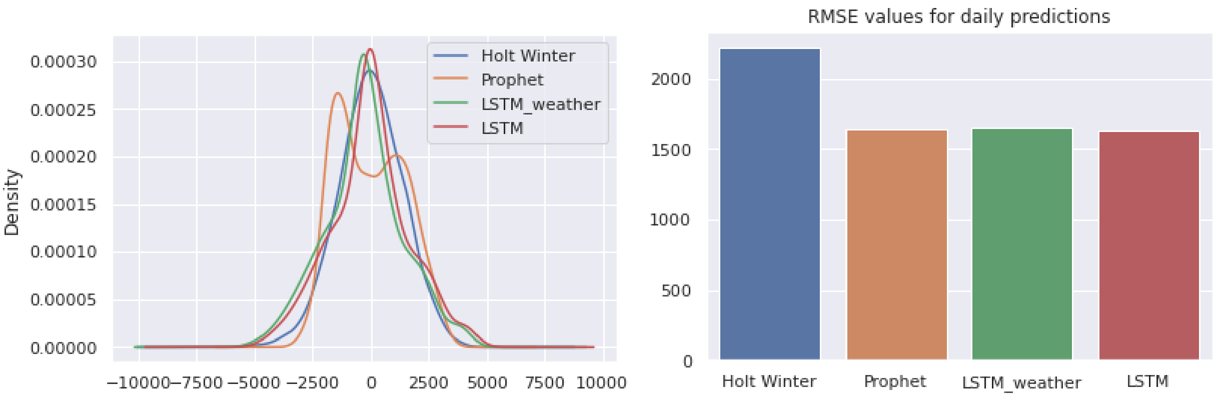


Figure 149: Residuals distribution for a daily fit of models. (b) RMSE for daily predictions.

A performance gap can be observed between Holt Winter and the three other models, and Holt Winter is thus disqualified. Prophet still offers good performances and its ability to generalize its prediction is high, but its training residuals do not show a normal distribution. Under-estimations of predictions (and to a lesser extent over-estimations) are indeed common.

Therefore, LSTM models are considered best for the daily prediction task.

Their principal inconvenience is the number of hyperparameters to define. Here an LSTM layer with 20 hidden states is run with the daily time series included and weather condition as parallel inputs. A batch size of 128 is used, and the model is trained for 400 epochs. The predictions are made using 90 days sequences. All those parameters were defined by empirical method but could be refined with a complete grid search analysis. A prediction interval is also computed using the method proposed by Pearce et al[72].

[71] 5 years of training: +13% RMSE

[72] https://arxiv.org/abs/1802.07167

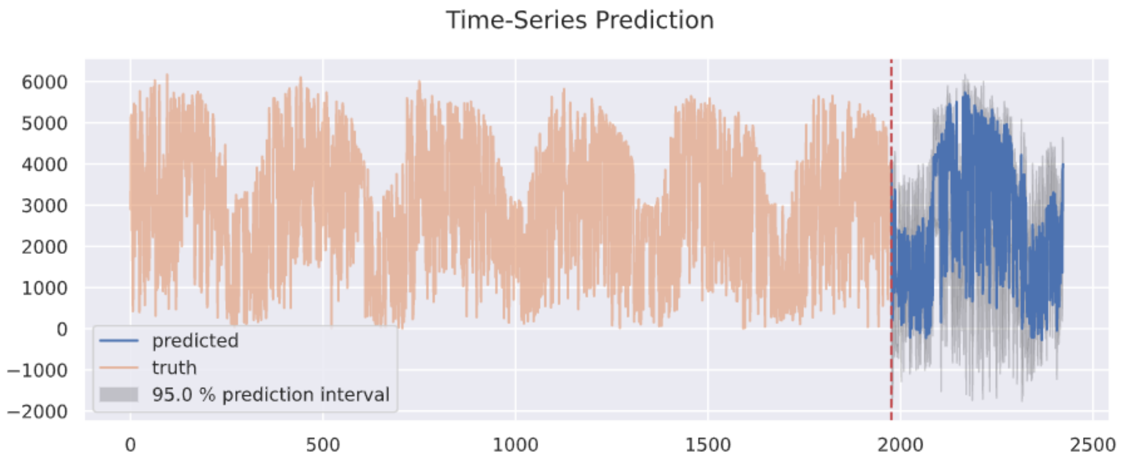


Figure 150: LSTM model daily predictions with a prediction interval

The multivariate LSTM model (with weather condition data) does not improve those results, even if this information is highly correlated to the production.

This kind of model is also more difficult to adapt to port data, as it requires a real weather forecast, which would add uncertainty to the model. Weather conditions in PVOutput are divided into six categories, meaning that this forecast should also be encoded as categories (ex: a 50% humidity forecast could correspond to “mostly cloudy”).

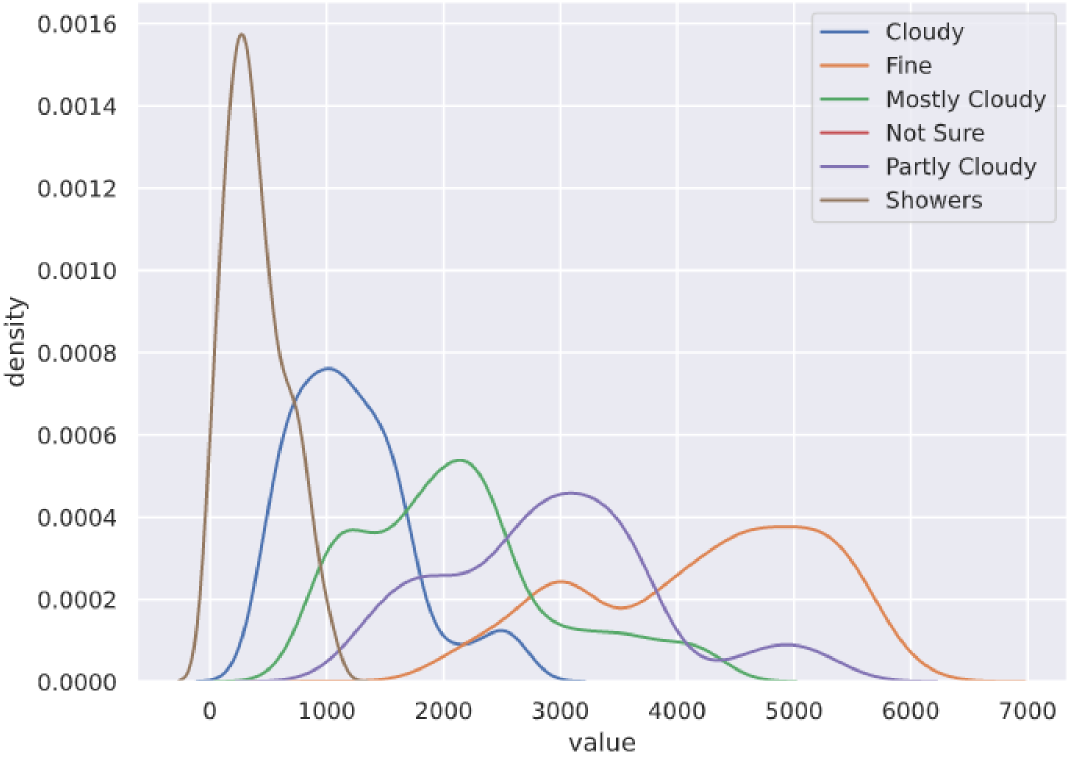


Figure 151: Production distribution according to weather

6.2.4. Predictive algorithms benchmark conclusion

The daily error is, of course, superior to the monthly error, but daily predictions are more useful for a photovoltaic installation. While the LSTM model obtains the best daily results, it comes with the cost of a significant number of hyperparameters. Hyperparameters tuning, such as the number of hidden states or the length of the training sequence, is an ongoing problem that depends on the data used. It can be solved case by case with grid search, but this step requires additional validation dataset. Since the addition of weather information to the model does not improve the results, a traditional LSTM would be enough for the port daily production prediction task.

7. Conclusions

In this deliverable, the results of the work that has been performed in T4.5 is presented. Overall, it was concluded that there is plenty of data captured in the ports for different operational processes; however, this data is underutilized or even not utilized at all, to provide any operational insights. What is even more important is, that there is a lack of awareness in smaller sized ports about the importance of properly capturing, storing and creating added value products from operational data. In T4.5 various domains of crucial importance for the ports and surrounding cities were addressed, such as vessel calls, maritime and road traffic. The presented results could help ports at optimizing their operational efficiency in different domains that create impacts on their operational performance, as well as on the environment.

The use of emerging technologies such as AI and data analytics is one of the areas where the gap between large and small ports is especially noticeable. Larger ports are increasingly investing in AI-based solutions. In addition to acquiring solutions that can already be found on the market and quickly applied, they also invest in research, to find new application areas and means to address them. The potential is huge and the problems that are addressed by larger ports are ambitious and challenging. AI has reached the world of transport, shipping industry and ports and is here to stay.

The solutions that were proposed in this deliverable, are more cost-effective when compared to solutions in larger ports, considering the limited resources of smaller ports and feasibility of implementation. The main goal is to utilize internal data that is captured, stored and available in ports or by other stakeholders. Vessel call data presents one of such data sources that all the ports store, or have information available to store, due to internationally standardized procedures that are in place, to capture this kind of information (FAL Forms). One of the main goals was, to show ports the importance of this data for their operational use, a thorough analysis of this data for gathering insights into port operations, as well through predictive modelling. A novel machine learning-based system for turnaround time prediction, based on standardized port call data was proposed and evaluated on the real-world example of the Port of Bordeaux. It is also demonstrated that port call data offers an insight into port operations, by exploring the collected data to obtain various business metrics to be used for strategic data-driven port operations and future development planning.

Furthermore, the use of AIS for data analytics and predictive modelling in the port area was explored. First, we demonstrated the importance of data validation and cleaning, which presents the most important part of all the derived products, especially the ones that rely on accurate navigational status. By accurately modelling navigational status of the ship, separate voyages can be distinguished and the amount of time spent in each particular state, which can be effectively used for modelling ship emissions, as well as for port business metric computation, especially in terms of the waiting times in the front and in the ports themselves. Different anomalies can be detected in the maritime traffic, as well as events classified, which can increase situational awareness of the ports. Ports can also increase their efficiency, by increasing the accuracy of accurate estimations of the arrival time, which can in fact, also complement and validate some of the vessel call data, captured through the FAL forms. Availability of the AIS data in the ports can be easily achieved, with the help of the low-cost hardware. Developed analytics and predictive models are general and applicable to any port, with available AIS data.

Satellite imagery can complement AIS data and it was demonstrated that with a novel use of AIS and satellite imagery data, ship detection across operational satellite imagery of medium spatial resolution can be performed. A novel procedure for large scale ship detection dataset construction was presented, which offers a novel approach of using weakly annotated data, available in abundance, for developing ship detection methods, without the need of human labelled data. It was proven that solely by using weakly annotated data, the results can be improved compared to using existing, human labelled datasets. The methodology was applied to the largely unused (in the maritime domain) medium resolution optical satellite imagery from ESA Sentinel-2 and Planet Labs Dove satellite constellations. The presented methodology is general and applicable also for other satellite constellations and opens opportunities for new use cases in the maritime domain.

Data analytics solutions and predictive models for traffic analysis and forecasting in port and city areas were also presented. A general workflow was followed, which was validated in multiple port and city-regions. Different visualizations and decision support tools were demonstrated, along with accurate predictive models, developed on top of state-of-the-art time series forecasting libraries, as well as custom made solutions, thus providing the ports with a variable set of tools, with different levels of performance and ease of deployment. Different traffic data sources were considered, such as the road traffic data available in the ports (i.e. gate data), regional traffic data, as well as traffic data from 3rd party providers (e.g. HERE). Road traffic data was also combined with external environmental data, as well as vessel call data from the ports.

Moreover, the methodology was provided, allowing the ports to estimate their potential photovoltaic production. To do so, the use of PVGIS tools is recommended. Based on the GPMB use-case, it was proven how ports can easily and quickly use PVGIS as a decision-making tool for investment in PV systems. The second approach that was followed, has been focused on benchmarking different predictive algorithms to forecast energy production based on past energy data. The presented models can be used and adapted to ports equipped with PV systems to help them predict their production and thus, have a better energy management plan.

According to this, predictive tasks were identified in D4.3 and solutions presented in D4.4, that support real-life scenarios and PIXEL objectives, while at the same time, closing the gap between small and large ports and helping them to make a step towards the Port of the Future.

Appendix 1: Data sources description

Overview

Identification of data sources to be used for predictive algorithms described in this document has been a crucial task for a successful fulfilment of needs and requirements put forward by ports, involved in PIXEL. This section lists, in Table 26, the data sources originally identified in D4.3 as well as data sources considered in advanced phases of task execution. In total, 32 different data sources were considered, and used 18 of those have been used for the predictive tasks in PIXEL.

Furthermore, each data source has a summary table provided in this appendix, outlining providing the following:

- **Dataset name:** Name of the dataset, as reported by the publisher.
- **Data source:** Either URL to the original data or explanation about how the data was obtained.
- **Description:** Description of the data source, including size, timeframe, attributes and similar.
- **Usage in PIXEL:** Intended usage in PIXEL (relation to specific tasks, scenarios and use-cases).
- **Algorithms:** Explanation of analytical algorithms applied to the data.
- **Sharing of results:** Rules for sharing of results obtained by data processing.
- **License and terms of use:** License and terms of use, with focus on specific PIXEL applications.
- **DPO assessment:** Assessment of the data source and any data processing restrictions concerning the processing of personal or other sensitive data. Note that, for each dataset, the assessments were performed before data processing.

List of data sources

Table 26. List of data sources.

Dataset name	Short description	Status[73]	Usage in PIXEL
GPMB vessel call data	Vessels call data for GPMB.	Yes	Analysis and prediction of vessel calls.
PPA Vessel call data	Vessels call data for PPA.	No	Models have been developed on GPMB data, as enough historical data is available. Availability and feasibility of PPA and ASPM datasets will be analysed in WP7.
ASPM Vessel call data	Vessels call data for ASPM.	No	
ThPA Vessel call data	Vessels call data for ThPA.	Yes	Correlation of port operations with road traffic.
DEBS (ACM International Conference on Distributed and Event-based Systems) 2018 Challenge dataset	Data used for DEBS 2018 Challenge that was prepared from AIS data for the Mediterranean Sea.	No	Implementing long-term ETA and destination port prediction. The focus of D4.3 and D4.4 was on short-term ETA prediction; thus, this data was not utilized.

[73] Usage status of the data source. Yes – identified in D4.3 and used for PA. No – identified in D4.3, but not used for PA (explanation provided). New – identified after the submission of D4.3 and used for PA.

Dataset name	Short description	Status[73]	Usage in PIXEL
Danish Maritime Authority AIS data	Historical AIS data for Danish waters.	Yes	ETA prediction, AIS data analytics around the port area, satellite imagery data fusion. Danish data was only used for visualization. AISHub data was the most complete data and presented the main source of AIS data. U.S. Coast Guard data presented the largest source of historical data and was used for satellite imagery data fusion.
U.S. Coast Guard AIS data	Historical AIS data for USA coastal area.	Yes	
AISHub data	Live AIS data from AISHub	Yes	
Thessaloniki car fleet data	Car fleet equipped with GPS (Global Positioning System) providing location and speed information.	No	Correlation of port operations, regional road network and weather data for analysis and prediction purposes. Thessaloniki car fleet data and Traffic data form the Thessaloniki stationary sensor network have been replaced with TrafficThess, a new, comprehensive, data source that was published after the publication of D4.3. Meteorological data has been added to explore the correlation of weather with traffic data.
Traffic data form the Thessaloniki stationary sensor network	Traffic data form the Thessaloniki stationary sensor network: traffic counters and speed sensors.	No	
Thessaloniki traffic data for vehicles entering/leaving the port	Traffic data for vehicles entering and leaving the port by gate and timestamp	Yes	
TrafficThess		New	
Stratus Meteo of Greece	Different sensors installed throughout Greece	New	
Traffic data from SILI system	Traffic information from cameras and gates connected to SILI system. The direction of traffic, lane number, exact times, license plates, nationality and type of vehicle.	Yes	Short-term traffic prediction and analysis.
Live data report from the police (traffic department) for Piraeus data traffic.	The data provides the traffic situation of the main Athens - Piraeus road arteries.	No	Datasets used for Piraeus Road Traffic Prediction are from HERE and OpenWeather. The aim is to be more consistent with datasets present in the other use cases. Subscription to Telenavis was ruled out due to high cost. Also, usage of those services allows a wider application to
Traffic reports from the Region of Attica observatory on traffic congestions.	These data reports are published every 6 months. These data are concerning to the traffic on the main streets of Attica.	No	

Dataset name	Short description	Status[73]	Usage in PIXEL
Data obtained from a third-party subscription database in CSV format.	These data are being prepared by the third party so that PPA can do their tests and evaluate the results. These will occur within May 2019.	No	other ports. Also, it is important to note that congestion information is already present in the HERE data source, while for the reports by the Piraeus police, a preliminary study with similar incidents offered by the HERE service was made, in which no observed existing relationship was found.
HERE Traffic Data	Traffic data from HERE Traffic API. Returned current speed, free-flow speed from different locations inside a bounding box provided.	New	PPA Road Traffic Prediction and Exploratory Data Analysis to search for patterns and stationarity.
OpenWeather Data	Current weather	New	As an additional attribute in road traffic prediction model of PPA.
ESA (European Space Agency) Sentinel satellite imagery.	Optical and SAR satellite imagery provided by Sentinel-1 and Sentinel-2 satellites.	Yes	To develop methods for ship detection and classification which will be further used for traffic analysis and predictions of port operations and intermodal transport forecasts. HRSC, xView, DOTA and Kaggle Planet Labs datasets were not needed due to enough annotated imagery in Kaggle Airbus ship detection dataset.
Planet Labs satellite imagery	Optical satellite imagery over California (OpenCalifornia) provided by PlanetScope (Dove) satellite constellation.	Yes	
Kaggle Airbus ship detection	Satellite imagery with annotated ships for ship detection from optical imagery.	Yes	
HRSC 2016	HRSC (High-Resolution Ship Collections) 2016 satellite imagery with annotated ships for ship detection and classification from optical imagery.	No	
xView dataset	Satellite imagery with annotated ships (and other classes) for object detection from optical imagery.	No	
DOTA (Large-scale Dataset for Object DeTection in Aerial Images)	DOTA satellite imagery with annotated ships (and other classes) for object detection from optical imagery.	No	
Kaggle Planet Labs	Satellite imagery from Planet Labs with ship/not ship imagery and annotations.	No	

Dataset name	Short description	Status[73]	Usage in PIXEL
PVGIS Data	Web applications to browse and query GIS databases of solar radiation and other climatic parameters.	Yes	To develop methods for forecasting solar energy production based on weather conditions using past data of production. The PVGIS data set has been used to evaluate the potential of photovoltaic production. The PVoutput data set has been used to develop and test different predictive algorithm for energy production prediction. OpenWeatherMap and OpenDataSoft have not been used, as an investigation of the PVoutput dataset has shown that using the weather conditions as an input of PA do not significantly improve the precision of the prediction.
PVoutput data set	Free online service for sharing and comparing photovoltaic solar panel output data.	Yes	
OpenWeatherMap	Current weather, daily forecast for 16 days, and 3-hourly forecast 5 days.	No	
OpenDataSoft data	Fields of analysis and forecasts in grid points, resulting from the atmospheric model Arome on the metropolis.	No	

Details for the used data sources

Dataset name	GPMB vessel call data
Data Source	Project partner CATIE (CATIE Centre Aquitain des Technologies de l'Information et Electroniques) provided the data owned by GPMB. Data has been shared through consortium document collaboration platform.
Description	7 years (2010-2017) of vessel call data for GPMB. Name of the ship, type of cargo, entry and exit times and amount of cargo is provided.
Usage in PIXEL	Analysis and prediction of vessel calls: (1) General statistical analysis and visualisations of the provided data; (2) Predicting call time; (3) Predicting vessel call.
Algorithms	Different supervised machine learning methods for regression, time series analysis and classification.
Sharing of results	Results initially available to project partners. Any publishing must be reviewed with the data owner.
License and terms of use	Usage allowed for the PIXEL consortium for research purposes as stated in the Grant Agreement.
DPO assessment (XLAB)	Considering the GDPR (General Data Protection Regulation) definition of personal data and considering that the dataset relates to cargo vessels (which are large ships, owned by legal entities and not individuals), this dataset only includes non-personal data: vessel name, cargo type, entry/exit times, cargo amount. These

	parameters cannot be used to, by reasonable means, directly or indirectly identify any data subject. The same holds for the 'IMO number' parameter. The dataset can be processed as indicated above without any limitation, considering that the processing is in line with the data-sharing agreement or licence of the dataset owner/provider.

Dataset name	ThPA vessel call data
Data Source	ThPA (data source). Sample data for 2-3 months was sent and are prepared to share more historical data.
Description	Provided fields are voyage code, vessel name (no IMO provided), origin and destination port, arrival and departure time, scheduled arrival and departure date.
Usage in PIXEL	Analysis of vessel calls: Impact of vessel calls on road traffic (congestion around the port).
Algorithms	Algorithms used will relate to road traffic but will consider vessel calls as an input parameter
Sharing of results	Results initially available to project partners. Any publishing must be reviewed with the data owner.
License and terms of use	Usage allowed for the PIXEL consortium for research purposes as stated in the Grant Agreement.
DPO assessment (XLAB)	Considering the GDPR definition of personal data and considering that the dataset relates to cargo vessels (which are large ships, owned by legal entities and not individuals), this dataset only includes non-personal data: vessel name, origin port, destination port, arrival date, departure date. These parameters cannot be used to, by reasonable means, directly or indirectly identify any data subject. The dataset can be processed as indicated above without any limitation, considering that the processing is in line with the data-sharing agreement or licence of the dataset owner/provider.

Dataset name	DMA (Danish Maritime Authority) AIS data
Data Source	Danish Maritime Authority https://www.dma.dk/SikkerhedTilSoes/Sejladsinformation/AIS/Sider/default.aspx# ftp://ftp.ais.dk/ais_data/
Description	Historical AIS data for Danish coastal waters. AIS data is already decoded and provided in CSV files with all the fields present. Compared to DEBS 2018 data, this includes IMO and MMSI numbers, as well as ship name and call sign.
Usage in PIXEL	This data will be used to support tasks regarding the use of AIS data, mainly in the port area. Usability for short-term ETA prediction will also be investigated.
Algorithms	Different supervised machine learning methods for regression, time series analysis and classification.
Sharing of results	Results initially available to project partners. External use to be decided.
License and terms of use	https://www.dma.dk/SikkerhedTilSoes/Sejladsinformation/AIS/Sider/AIS%20datamanagementpolitik.aspx http://ec.europa.eu/newsroom/document.cfm?doc_id=1262 (PSI act)

DPO assessment (XLAB)	By law, only large vessels (ships of more than 300 gross tonnage, passenger ships, and fishing vessels with a length of above 15 metres) are required to carry an AIS transponder. These vessels are owned by legal entities and not natural persons. **Hence, data from these large vessels do not contain any personal information.** Other, smaller vessels (including those, which could be owned by natural persons) are not required to be fitted with AIS. If they are fitted with AIS, they provide data to the maritime authorities on voluntary basis considering and accepting their terms of use and privacy policies (which explain that the collected data are available as open data: https://www.dma.dk/SikkerhedTilSoes/Sejladsinformation/AIS/Sider/AIS%20 datamanagementpolitik.aspx). Nevertheless, since directly no personal data relating to natural persons are included in the AIS calls (only vessel ID), re-identification of natural persons would require significant (financial, technological, timewise) resources. It can be assumed that, currently, such means would not reasonably likely to be used. According to the Recital 26 of the GDPR, it can be concluded that, even in the case of smaller vessels, this dataset is anonymized and the GDPR does not apply. The dataset can thus be processed as indicated above without any limitation if the processing is aligned with the licence of the data provider.

Dataset name	U.S. Coast Guard AIS data
Data Source	https://marinecadastre.gov/ais/
Description	Historical AIS data for USA coastal waters. AIS data is already decoded and provided in CSV files with all the fields present. Compared to DEBS 2018 data, this includes IMO and MMSI numbers, as well as ship name and call sign.
Usage in PIXEL	This data will be used to support tasks regarding the use of AIS data, mainly in the port area. Usability for short-term ETA prediction will also be investigated.
Algorithms	Different supervised machine learning methods for regression, time series analysis and classification.
Sharing of results	Results initially available to project partners. External use to be decided.
License and terms of use	https://www.navcen.uscg.gov/?pageName=NAISdisclaimer
DPO assessment (XLAB)	By law, only large vessels (ships of more than 300 gross tonnage, passenger ships, and fishing vessels with a length of above 15 metres) are required to carry an AIS transponder. These vessels are owned by legal entities and not natural persons. **Hence, data from these large vessels do not contain any personal information.** Other, smaller vessels (including those, which could be owned by natural persons) are not required to be fitted with AIS. If they are fitted with AIS, they provide data to the maritime authorities on voluntary basis considering and accepting their terms of use and privacy policies (which explain that the collected data can be shared with other organisations). Nevertheless, since directly no personal data relating to natural persons are included in the AIS calls (only vessel ID), re-identification of natural persons would require significant (financial, technological, timewise) resources. It can be assumed that, currently, such means would not reasonably likely to be used. According to the Recital 26 of the GDPR, it can be concluded that, **even in the case of smaller vessels, this dataset is anonymized and the GDPR does not apply.** The dataset can thus be processed as indicated above without any limitation if the processing is aligned with the licence of the data provider.

Dataset name	AISHub data
Data Source	http://www.aishub.net/

Description	AIS data-sharing platform. To receive data, one needs to connect own raw AIS data feed. Mostly amateur AIS receiving stations are present, all over the world. The data is provided through an API (decoded, processed AIS data) and as raw NMEA (National Marine Electronics Association) data through a dedicated TCP (Transmission Control Protocol) port. Information that is captured is the same as provided with historical datasets already presented. This data will be used to support tasks regarding the use of AIS data, mainly in the port area and for short-term ETA prediction.
Usage in PIXEL	This data will be used to support tasks regarding the use of AIS data, mainly in the port area and for short-term ETA prediction.
Algorithms	Different supervised machine learning methods for regression, time series analysis and classification.
Sharing of results	Results initially available to project partners. External use to be decided.
License and terms of use	http://www.aishub.net/
DPO assessment (XLAB)	By law, only large vessels (ships of more than 300 gross tonnage, passenger ships, and fishing vessels with a length of above 15 metres) are required to carry an AIS transponder. These vessels are owned by legal entities and not natural persons. **Hence, data from these large vessels do not contain any personal information.** Other, smaller vessels (including those, which could be owned by natural persons) are not required to be fitted with AIS. If they are fitted with AIS, they provide data to the maritime authorities on voluntary basis considering and accepting their terms of use and privacy policies (which explain that the collected data can be shared with other entities). Nevertheless, since directly no personal data relating to natural persons are included in the AIS calls (only vessel ID), re-identification of natural persons would require significant (financial, technological, timewise) resources. It can be assumed that, currently, such means would not reasonably likely to be used. According to the Recital 26 of the GDPR, it can be concluded that, **even in the case of smaller vessels, this dataset is anonymized and the GDPR does not apply.** The dataset can thus be processed as indicated above without any limitation if the processing is aligned with the licence of the data provider.

Dataset name	Thessaloniki port traffic data
Data Source	ThPA (data source). Sample data for 1 day was sent and are prepared to share more historical data. However, real-time data collection is more of interest.
Description	List of vehicles entering and leaving the port, including gate number and timestamps (so as dwell time). Registration plates, RFID (Radio-frequency identification) and company have been excluded.
Usage in PIXEL	Correlation of timing operations to detect congestion at the gates
Algorithms	Both real-time estimation and predictive algorithms
Sharing of results	Results initially available to project partners. Any publishing must be reviewed with the data owner.
License and terms of use	Usage allowed for the PIXEL consortium for research purposes as stated in the Grant Agreement.
DPO assessment (XLAB)	This dataset specifies the following parameters: vehicle type, marker, model, colour, entry time, entry gate, exit time, exit gate, dwell time

	According to the Recital 26 of the GDPR, we can conclude that the dataset is anonymized (after removing registration plates and RFIDs, it is unreasonably likely that any individual could be directly or indirectly identified) and the GDPR does not apply. The dataset can be freely processed without any limitation, given that the processing is in line with the license of the dataset owner/provider.

Dataset name	TrafficThess data – road traffic information of Thessaloniki city
Data Source	Historical data: https://www.trafficthess.imet.gr/exporter.aspx Live feed: https://www.trafficthess.imet.gr/
Description	Free flow, average, minimum, maximum, median and geometric mean speed of the traffic in a street of the city of Thessaloniki in a certain period.
Usage in PIXEL	Correlation of road traffic (selection of the 5 surrounding streets of the port) of the city with the traffic at the gates of the Port of Thessaloniki.
Algorithms	Prophet model based on baseline gates congestion data adding road traffic as a new regressor.
Sharing of results	Results available to project partners. ThPA will use this incorporated as Dockerised model to run in its pilot in T7.4. External use to be decided.
License and terms of use	Same criteria apply as for the open data IMET source, a citation is required: http://opendata.imet.gr/about. At the bottom of the website, link to open data license is provided: https://opendefinition.org/od/2.1/en/
DPO assessment (UPV)	DPO checked the following datasets: ● **Historical data:** Summary of the different datasets above (in opendata.imet analysis) condensed and filtered per "road branch" in the whole Thessaloniki city. No personal data is contained. Considering the GDPR definition of personal data, these datasets only contain non-personal data. The datasets can be processed as indicated above without any limitation, considering that the processing is in line with the licence of the dataset owner/provider.

Dataset name	Stratus meteo of Greece
Data Source	Historical data: http://stratus.meteo.noa.gr/front Live feed: http://penteli.meteo.gr/stations/helexpo/
Description	Set of meteorological stations of Greece that are depicted in a Google maps Layout. Behind, the different station pinpoints are connected to data sources, providing weather information (temperature, dew point, humidity, pressure, wind speed and wind direction, rain…) in real-time and different historical options. The selected station redirects to a data provision service, consisting of a tool result of an InterReg Balkan-Mediterranean project: BeRTISS.
Usage in PIXEL	Correlation of weather data of the closest station to the Port of Thessaloniki (station id: LG9M) with the traffic at the gates of the Port of Thessaloniki. Inclusion of the meteorological data into the predictive model together with traffic gates to predict congestion.
Algorithms	Prophet model based on baseline gates congestion data, adding weather as a new regressor.
Sharing of results	Results initially available to project partners. External use to be decided.
License and terms of use	At the bottom of the data origin website, link terms of use are provided: https://www.meteo.gr/terms.cfm According to them: "The content and services of this node are the property of EAA (National Observatory of Athens) and are available to visitors of the node

	strictly for personal use. Their use is prohibited, in any other medium as well as for commercial purposes, without the written permission of the EAA.” UPV (Universitat Politècnica de València) asked for permission of use this data with academic purposes, specifying the processing to be applied, the Grant Agreement number of the project and the number of the sensor (node) of interest. Access was granted and permission of use in the context of PIXEL was allowed. Email exchange has been stored as formal proof and will be provided if needed. Citation to the network of meteorological sensors is required and provided below: http://onlinelibrary.wiley.com/doi/10.1002/gdj3.44/full
DPO assessment (UPV)	Dataset was checked. Timely registers from an installed meteorological station containing weather information (temperature, dew point, humidity, pressure, wind speed and wind direction, rain, among others of the same nature). No personal data is contained. Considering the GDPR definition of personal data, these datasets only contain non-personal data. The datasets can be processed as indicated above without any limitation, considering that consent has been provided and that all processing is in line with the license of the dataset owner/provider.

Dataset name	SILI system traffic data
Data Source	Data provided by INSIEL (Insiel SpA) - project partner, by authorization of ASPM and Central Directorate of Infrastructure and Territory of Friuli Venezia Giulia, as the data owners.
Description	Sample data was provided by INSIEL for one of the gates that are monitored by a traffic camera connected to the SILI system. Data consists of the following fields: gate name and direction of the traffic, lane identifier, exact date and time, license plate and nationality extracted from the license plate, vehicle type. Note that the license plate and nationality extracted from the license plate attributes were removed before sharing the dataset with the consortium.
Usage in PIXEL	This will be used for short-term traffic volume prediction and correlation with operations in the port.
Algorithms	Different supervised machine learning methods for regression, time series analysis and classification.
Sharing of results	Results initially available to project partners. Any publishing must be reviewed with the data owner.
License and terms of use	Usage allowed for the PIXEL consortium for research purposes as stated in the Grant Agreement.
DPO assessment (XLAB)	The dataset does not include any personal data (note that the dataset was anonymised by the data owner before sharing - licence plates and derived nationalities were removed). The dataset can be processed as indicated above without any limitation, considering that the processing is in line with the data-sharing agreement or licence of the dataset owner/provider.

Dataset name	HERE Traffic Data
Data Source	https://developer.here.com/documentation/traffic/dev_guide/topics/what-is.html

Description	Status of road traffic for different locations inside a bounding box provided. Apart from current speed, road descriptions, intersections and direction, the status of traffic flow is also added.
Usage in PIXEL	Exploratory data analysis and road traffic prediction.
Algorithms	Algorithms used in road traffic predictions like Prophet
Sharing of results	Results available for publication.
License and terms of use	https://developer.here.com/terms-and-conditions#terms_sec2
DPO assessment (Prodevelop)	The dataset can be processed as indicated in the terms without any limitation in PIXEL, as it was stated by direct permission, given by HERE Europe. This agreement is available to EC under request.

Dataset name	OpenWeather Data
Data Source	https://openweathermap.org/current
Description	Current weather data for any location. Different weather parameters like temperature, humidity, pressure, wind, rain and snow predictions and so on.
Usage in PIXEL	Impact of weather on road traffic.
Algorithms	Algorithms used in road traffic predictions, introduced as additional attributed in time series forecasting algorithm.
Sharing of results	Results available for publication.
License and terms of use	The OpenWeather API is free to use in the context described above.
DPO assessment (Prodevelop)	The dataset can be processed as indicated above without any limitation, considering that the processing is in line with the data-sharing agreement or licence of the dataset owner/provider.

Dataset name	ESA Sentinel satellite imagery (Sentinel Hub)
Data Source	ESA Sentinel imagery provided through external provider Sentinel Hub
Description	Satellite imagery provided by ESA and further distributed through external provider Sentinel Hub, which offers API for simplified access.
Usage in PIXEL	ESA Sentinel imagery will be used to develop methods for ship detection and classification from satellite imagery. Sentinel-1 (SAR) and mostly Sentinel-2 (optical) imagery will be used. Satellite imagery will also be used for data fusion with AIS data.
Algorithms	To train different CNN based object detection and classification methods.
Sharing of results	Results initially available to project partners. External use to be decided.
License and terms of use	ESA Sentinel data: https://sentinel.esa.int/documents/247904/690755/Sentinel_Data_Legal_Notice Sentinel Hub: https://sentinel-hub.com/tos
DPO assessment (XLAB)	This dataset does not include any personal information. The dataset can thus be processed as indicated above without any limitation if the processing is aligned with the licence of the data provider.

Dataset name	Planet Labs satellite imagery
Data Source	Openly available satellite imagery over California - OpenCalifornia. https://www.planet.com/products/open-california/
Description	Satellite imagery provided by PlanetScope (Dove) constellation of satellites with 3m resolution and daily revisit time (14-day delay in case of OpenCalifornia). Only optical imagery is available.
Usage in PIXEL	Planet Labs imagery will be used to develop methods for ship detection and classification from satellite imagery. Increased resolution (3m) and daily revisit time will offer additional capabilities especially for ship classification. Satellite imagery will also be used for data fusion with AIS data.
Algorithms	To train different CNN based object detection and classification methods.
Sharing of results	Results initially available to project partners. External use to be decided.
License and terms of use	https://creativecommons.org/licenses/by-sa/4.0/ https://www.planet.com/assets/pdfs/planet-open-ca-license-faqs.pdf
DPO assessment (XLAB)	This dataset does not include any personal information. The dataset can thus be processed as indicated above without any limitation if the processing is aligned with the licence of the data provider.

Dataset name	Kaggle Airbus ship detection
Data Source	https://www.kaggle.com/c/airbus-ship-detection
Description	Satellite imagery with annotated ships for ship detection from optical imagery.
Usage in PIXEL	To develop the methods for ship detection which results will be further used for traffic analysis and prediction in and around the port.
Algorithms	To train different CNN based object detection methods.
Sharing of results	Results initially available to project partners. External use to be decided.
License and terms of use	Section B, rule 7: https://www.kaggle.com/c/airbus-ship-detection/rules
DPO assessment (XLAB)	Considering the GDPR definition of personal data and considering that the dataset relates to cargo vessels (which are large ships, owned by legal entities and not individuals), this dataset only includes non-personal data: images and metadata (IDs, labels, coordinates, etc.). These parameters cannot be used to, by reasonable means, directly or indirectly identify any data subject. The dataset can be processed as indicated above without any limitation, considering that the processing is in line with the data-sharing agreement or licence of the dataset owner/provider.

Dataset name	PVGIS data

Data Source	Dataset and project page: http://re.jrc.ec.europa.eu/pvg_download/data_download.html Publications: http://re.jrc.ec.europa.eu/pvg_static/Publications_in_proc.html Methods: http://re.jrc.ec.europa.eu/pvg_static/methods.html
Description	Web applications to browse and query GIS databases of solar radiation and other climatic parameters. With this data, it is possible to estimate PV electricity generation at any location in Europe, Africa, most of Asia, North America and most of South America
Usage in PIXEL	Based on historical irradiance data and associated weather conditions, obtained either by measurement or by satellite-based tools (PVGIS), a full methodology will be proposed to predict one-point irradiance for a time horizon from a day to a year. This prediction will reflect typical day, week, month, year based on past data.
Algorithms	To interact with web-services like PVGIS to obtain historical data and extract a typical irradiance.
Sharing of results	Results initially available to project partners. External use to be decided.
License and terms of use	The solar radiation data which are made available are long-term averages for each month and the year, based on data with hourly time resolution from a satellite. In all cases, the original data are freely available from the organizations that have produced the data sets. The use of these data is authorised if the source is acknowledged.
DPO assessment (CATIE)	According to the data described above and the one provided on the PVGIS website (http://re.jrc.ec.europa.eu/pvg_download/data_download.html), this dataset only includes non-personal data: solar radiation, geographical data, temperature, PV technical specification. The dataset can be processed as indicated above without any limitation, considering that the processing is in line with the data-sharing agreement or licence of the dataset owner/provider.

Dataset name	PVoutput data set
Data Source	https://pvoutput.org./list.jsp?id=15556&sid=13412
Description	PVOutput is a free online service for sharing and comparing photovoltaic solar panel output data. It provides both manual and automatic data uploading facilities. Output data can be graphed, analysed and compared with other PV output contributors over various periods. While PV output is primarily focused on monitoring energy generation, it also provides equally capable facilities to upload and monitor energy consumption data from various energy monitoring devices.
Usage in PIXEL	Based on historical production data, predictive algorithms for photovoltaic production are implemented based on past data of real production and associated weather conditions.
Algorithms	To train different machine learning models to predict solar energy production.
Sharing of results	Results initially available to project partners. External use to be decided.

License and terms of use	https://pvoutput.org./terms.html CATIE has acquired the donation status to be able to query historical data through the PV output API. Raw data have been downloaded, stored and analysed by CATIE.
DPO assessment (CATIE)	According to the data described above and the one provided on the PVoutput website (https://pvoutput.org), this dataset only includes non-personal data: solar energy production, energy consumption, geographical data, weather conditions, temperature, PV technical specification. The dataset can be processed as indicated above without any limitation, considering that the processing is in line with the data-sharing agreement or licence of the dataset owner/provider.

Appendix 2: External resources (“OTHER”)

Predicting vessel calls data from FAL forms and other sources

Git Repository: https://github.com/pixel-ports/gpmb_vessel_calls_eda

In this notebook, EDA work on vessel calls data from GPMB is presented. 8 years of vessel calls data was acquired, from the beginning of 2010 to the end of 2017. The data contains information about almost 4500 arrivals of cargo vessels and tankers. Even though GPMB has 7 terminals, all data is from one of them, Bassens. It is the largest terminal and capable of processing (loading and unloading) different types of cargo, bulk goods, cereals, containers, forestry products and heavy lift cargo. In the notebook, visualizations are presented, that provide insight in yearly trends of the number of arriving vessels and amount of processed cargo, seasonality of cargo types, regular vessels, turnaround time distributions and factors influencing on it (for example, tides).

Use of AIS data

Git Repository: https://github.com/pixel-ports/pixel_ais

Data for three PIXEL ports have been collected from the data-sharing portal AISHub: GPMB, ASPM and PPA. In the AIS notebook, work on analytics of AIS is presented: heatmaps of locations, most common navigation statuses in different areas in the ROIs, most common vessel types in different parts of the ports and neighbour areas, errors in AIS data, comparison of reported and fixed navigational statuses, voyages from entering the ROI, anchoring, mooring and leaving the region.

Use of AIS data – Event Detection

Git Repository: https://github.com/pixel-ports/AIS_Event

This repository contains, in a Jupyter Notebook, all the code developed for the AIS event detection part. In the first part, different heat maps to show areas of events impact and scatter plots of the physical characteristics of boats are developed. Also, all the strategies of feature extraction and selection are presented which take place around the KNN and Random Forest classification algorithms. Some of these techniques are the principal component analysis and linear discriminant analysis, among others. Finally, in the last part, the entire part of the neural network with LSTM cells for event detection given a sequence of AIS messages is developed, included the entire part of data processing architecture to mount such sequences.

Use of satellite imagery

Git Repository: https://github.com/pixel-ports/AIS_satellite_imagery_merge

In this repository, the code needed to merge AIS data from the U.S. Coast Guard with satellite imagery is provided, which presents a novel procedure for building large-scale weakly annotated ship detection or classification datasets[74]. The code is provided for both, open-source ESA Sentinel, as well as commercial Planet Labs satellite imagery.

Analysis and prediction of road traffic conditions for ASPM/SDAG

Git Repository: https://github.com/pixel-ports/aspm_traffic_prediction

The git repository contains Jupyter Notebooks used for short term traffic volume prediction on the regional roads around the Port of Monfalcone. The source of traffic data is the SILI (Sistema Informativo Logistico Integrato) system. The data was acquired from 11 locations in the Friuli Venezia Giulia region from March 2015 to August 2019. In total it consists of 95 million records of vehicles passing the gates. The repository is assembled from three parts. The first part is used for data pre-processing, the second part for exploratory data analysis and the third part is used for building the models and their evaluations. We used and evaluated two different machine learning algorithms. The first one is Facebook Prophet. It is a python library used

[74] Štepec, Dejan, Tomaž Martinčič, and Danijel Skočaj. "Automated System for Ship Detection from Medium Resolution Satellite Optical Imagery." OCEANS 2019 MTS/IEEE SEATTLE. IEEE, 2019.

for time series machine learning tasks. The second one is XGBoost. It is also a python library, which utilizes gradient boosting techniques. For this algorithm, we had to transform our problem from the time series into relational.

Piraeus Road Traffic Prediction

Git Repository: https://github.com/pixel-ports/PPA_Traffic_Prediction

This repository provides the results of the use case of road traffic prediction around the port of Piraeus. The first part includes the entire part of data processing and manipulation, as well as exploratory analysis. In this, utilizing different graphs such as heat maps or line graphs, an attempt is made to understand some of the underlying patterns of seasonality in the data, such as work migrations. In the second and third part, the prediction part is developed with the use of the Prophet time series algorithm, from an approach with only the base information to the inclusion of additional attributes, such as time and port activity, i.e. arrival-related data of passenger ships.

Traffic Prediction at the gates of the Port of Thessaloniki

Git Repository: https://github.com/pixel-ports/thpa_traffic_prediction

The git repository contains the Jupyter Notebooks used to explore (EDA) the data, to be later used for training and validating the predictive model, as explained in the section 6.2.3 - Prediction of traffic at gates of ThPA. The repository also includes the pre-processing of the data (Python scripts) and the different processing applied, to establish a common format for the predictive model training. The Facebook Prophet forecast model has been used, drawing from a congestion baseline data and adding individual regressors of weather in the area, traffic in the city surroundings and amount of vessels berthed at the port during the period April-2018 to February-2020. The scripts and notebooks work with CSVs of data owned by different entities, which have not been included due to copyright constraints.

Prediction of renewable energy production

Git Repository: https://github.com/pixel-ports/PV_prod_predic/tree/master/DELIVERABLE
The git repository contains the Jupyter Notebook used to benchmark production prediction models as explained in section *6. Prediction of renewable energy production*. The following forecast models have been included: SARIMAX, Holt-Winter, FBProphet, LSTM, and LSTM with weather data.

This notebook works with daily aggregated data from PVOutput ('aggreg_PRODUCTION_data.csv') and should be run with Jupyter Notebook for Python v.3.6+.