Early estimates of maritime traffic using innovative data sources

Nikolaos Roubanis, Eurostat, nikolaos.roubanis@ec.europa.eu

Boryana Milusheva, Eurostat, boryana.milusheva@ec.europa.eu

ABSTRACT

In 2023, Eurostat and the European Maritime Safety Agency established a cooperation agreement to develop methods and produce early estimates of European ports vessel traffic, exploring the use of Automatic Identification System (AIS) and other administrative, and commercial data available in EMSA. As a first step, EMSA data on vessel traffic were aggregated according to the Eurostat vessel type classification and compared to data for selected high-traffic ports submitted to Eurostat under the Directive 2009/42/EC on statistical returns in respect of carriage of goods and passengers by sea. In a second step, quarterly trends in vessel traffic at port and country level for the years 2015 to 2019 were calculated for each dataset. The comparative data analysis indicated a good match, particularly of the trends at EU level. A method was then developed to estimate vessel traffic (port calls) for the most recent quarter with a model integrating Eurostat data of the previous years and the most recent EMSA data. The method was tested on quarterly data of the past 5 years, showing results with a deviation at EU level for all reporting ports of 2-4% from the statistical data Eurostat received a year later. Further work to reduce identified differences includes better understanding of how countries classify certain vessel types reported to Eurostat and improving the aggregation of EMSA data to better match Eurostat vessel classification. It will allow for more accurate estimates and more granular short-term traffic statistics at port and vessel type level across the EU. Furthermore, the project will improve the quality of maritime statistics by improving timeliness and accuracy in the classification of data by vessel type.

Keywords: Transport, Experimental statistics, Maritime traffic, Innovative data sources

1. INTRODUCTION

Quarterly data on the dataset F2 of the Directive 2009/42/EC of the European Parliament and of the Council on statistical returns in respect of carriage of goods and passengers by sea are reported to Eurostat, once per year in August, covering the previous year reporting period. To make available more timely maritime traffic data, Eurostat in cooperation with the European Maritime Safety Agency (EMSA) launched a new cooperation project in February 2023.

A detailed analysis of Eurostat and EMSA data was made aiming to match these datasets at the lowest possible disaggregation level and compare aggregates by type of vessel at port, country and EU level. The comparison exercise aimed also at identifying potential methodological reasons which would explain eventual differences.

Estimations methods were then tested to produce the best estimates of the statistics provided in dataset F2, using the EMSA data.

The steps followed, together with the project results, are described in this methodological report.

2. DESCRIPTION OF THE DATA SOURCES

Four datasets were considered for the analysis: Eurostat F2 dataset and three datasets from EMSA (SafeSeaNet (SSN), MARINFO and ‘Detected port calls’).

2.1. Eurostat

Eurostat dataset F2 covers statistics on vessel traffic in European ports (vessels arriving at ports). This dataset provides quarterly information on two variables: the number of vessels and the gross tonnage of vessels.

To ensure the quality of the data transmitted to Eurostat by reporting countries, several validation checks are performed such as:

- Intra-dataset checks: time series checks (outliers), distribution by type of vessel and port, share of the category 'Unknown' to the total;

- Inter-dataset checks: average tonnes per vessel, average number of passengers embarked/disembarked per vessel (for the categories ‘General cargo, non-specialised’ (33), ‘Passenger’ (35) and ‘Cruise passenger’ (36)).

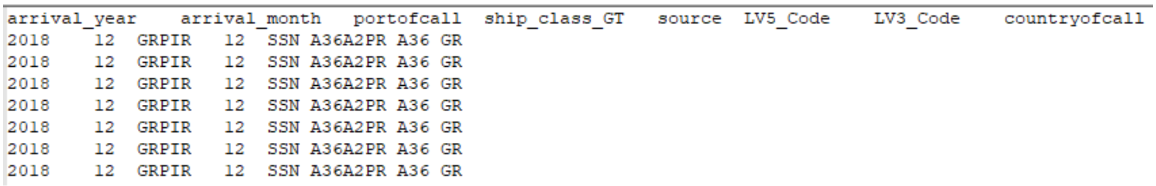
2.2. EMSA

The three EMSA datasets analysed were the SafeSeaNet dataset (administrative data on notifications provided by the Member States), and MARINFO and the ‘Detected port calls’ (DPC), both based on Automatic Identification System (AIS) signals.

2.2.1. SafeSeaNet

SafeSeaNet (SSN) is a vessel traffic monitoring and information system. It was set up as a network for maritime data exchange, linking together maritime authorities across Europe. It enables European Union Member States, Norway, and Iceland, to provide and receive information on vessels, vessel movements, and hazardous cargoes.

The dataset is referred as ‘EMSA-SSN’ in this report. The structure of the dataset provided to Eurostat is as follow:



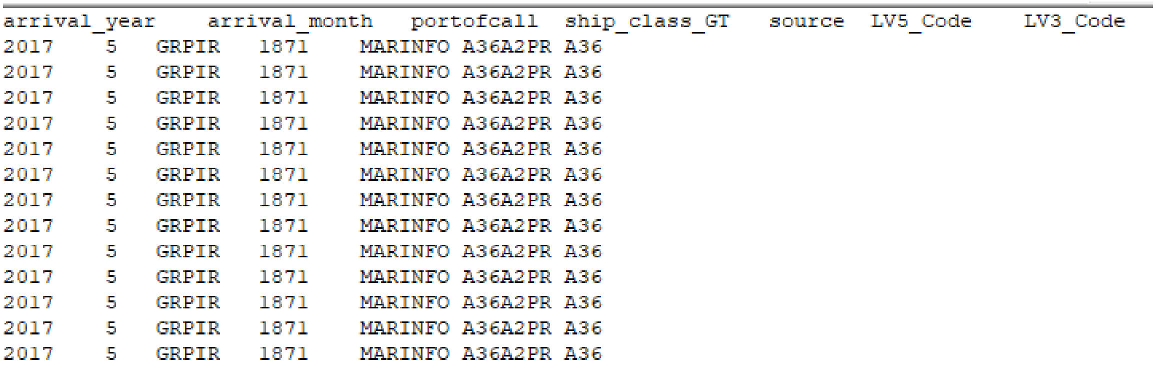
```
arrival_year    arrival_month    portofcall   ship_class_GT    source   LV5_Code     LV3_Code     countryofcall
2018     12   GRPIR    12   SSN A36A2PR A36 GR
2018     12   GRPIR    12   SSN A36A2PR A36 GR
2018     12   GRPIR    12   SSN A36A2PR A36 GR
2018     12   GRPIR    12   SSN A36A2PR A36 GR
2018     12   GRPIR    12   SSN A36A2PR A36 GR
2018     12   GRPIR    12   SSN A36A2PR A36 GR
2018     12   GRPIR    12   SSN A36A2PR A36 GR
```

The microdata is collected and processed at national level according to an agreed data quality criteria. In addition, upon submission, EMSA performs independent data quality validation checks to reduce missing reports from Member States. The average number of missing reporting is below 1 %.

2.2.2. *MARINFO*

MARINFO is built on AIS signals from commercial data providers.

The dataset is referred as ‘EMSA-MARINFO’ in this report. The structure of the dataset provided to Eurostat is as follow:



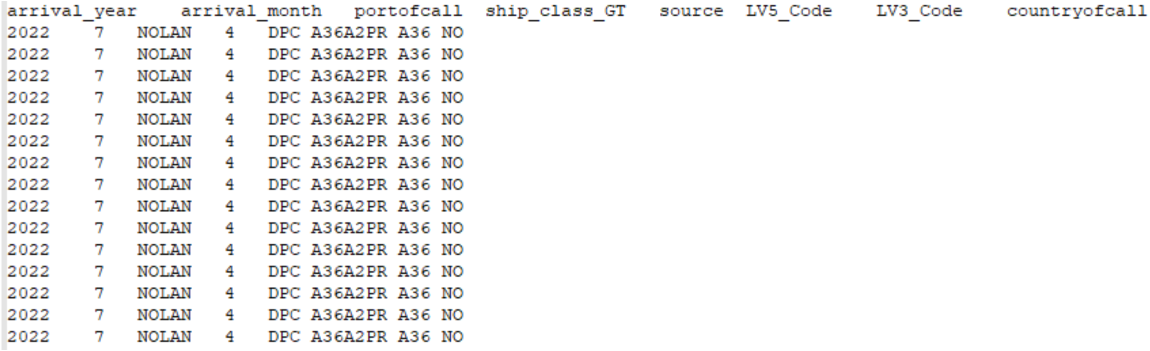
```
arrival_year    arrival_month    portofcall   ship_class_GT    source   LV5_Code     LV3_Code
2017     5    GRPIR    1871     MARINFO A36A2PR A36
2017     5    GRPIR    1871     MARINFO A36A2PR A36
2017     5    GRPIR    1871     MARINFO A36A2PR A36
2017     5    GRPIR    1871     MARINFO A36A2PR A36
2017     5    GRPIR    1871     MARINFO A36A2PR A36
2017     5    GRPIR    1871     MARINFO A36A2PR A36
2017     5    GRPIR    1871     MARINFO A36A2PR A36
2017     5    GRPIR    1871     MARINFO A36A2PR A36
2017     5    GRPIR    1871     MARINFO A36A2PR A36
2017     5    GRPIR    1871     MARINFO A36A2PR A36
2017     5    GRPIR    1871     MARINFO A36A2PR A36
2017     5    GRPIR    1871     MARINFO A36A2PR A36
2017     5    GRPIR    1871     MARINFO A36A2PR A36
```

The data are collected and processed by the commercial data provider. The data provider reviews all vessels every day and do not carry out selective updating, thus following a process of continuous updating.

2.2.3. *Detected port calls (DPC)*

This dataset consists in data from AIS signals, and it is a newly established data sources of EMSA and for this reason its time-coverage is shorter than the other sources (time series starting in 2018).

The dataset is referred as ‘EMSA-DPC’ in this report. The structure of the dataset provided to Eurostat is as follow:



```
arrival_year    arrival_month    portofcall   ship_class_GT    source   LV5_Code    LV3_Code    countryofcall
2022    7    NOLAN    4    DPC A36A2PR A36 NO
2022    7    NOLAN    4    DPC A36A2PR A36 NO
2022    7    NOLAN    4    DPC A36A2PR A36 NO
2022    7    NOLAN    4    DPC A36A2PR A36 NO
2022    7    NOLAN    4    DPC A36A2PR A36 NO
2022    7    NOLAN    4    DPC A36A2PR A36 NO
2022    7    NOLAN    4    DPC A36A2PR A36 NO
2022    7    NOLAN    4    DPC A36A2PR A36 NO
2022    7    NOLAN    4    DPC A36A2PR A36 NO
2022    7    NOLAN    4    DPC A36A2PR A36 NO
2022    7    NOLAN    4    DPC A36A2PR A36 NO
2022    7    NOLAN    4    DPC A36A2PR A36 NO
2022    7    NOLAN    4    DPC A36A2PR A36 NO
2022    7    NOLAN    4    DPC A36A2PR A36 NO
2022    7    NOLAN    4    DPC A36A2PR A36 NO
```

Being a recent data collection, the data still require additional data quality improvements by EMSA. At this stage, the validation of port calls is focused in excluding duplicate records. The validation of the port polygons was ongoing at the time of the project.

3. Comparability of the data sources

Before defining and testing estimation methods, the comparability between Eurostat F2 dataset and each of EMSA's dataset was analysed. To perform this comparison, EMSA provided the EMSA-SSN, EMSA-MARINFO and EMSA-DPC datasets for the reference period 2015-2019. To compare data by type of vessel, a correspondence table was created to categorize the vessel types from EMSA datasets according to Eurostat's classification, before aggregation.

3.1. Eurostat data treatment

To compare Eurostat data with EMSA data, two Eurostat sources were used: first, the dataset F2 on European port vessel traffic, by port, type and size of vessels (loading or unloading cargo, embarking or disembarking passenger), and second, the list of ports reporting to Eurostat (maintained by Eurostat on the basis of countries regular feedback). The port information from dataset F2 was matched with the corresponding port from Eurostat's list to ensured that the port codes are standardised across all records and no errors in matching with EMSA port codes is encountered.

3.2. EMSA data treatment

For each of EMSA data, three sets were needed to aggregate the data: first, the data received from EMSA; second, the corresponding table for the type of vessels; and third, Eurostat's list of ports.

The data were also aggregated by quarter (to match the Eurostat data availability), port and type of vessel. The gross tonnage (GT) was also calculated, when available.

3.3. Comparison exercise

In this section, the conclusions drawn for the different comparisons performed are presented.

3.3.1. Comparison at port level

As a first comparison exercise, EMSA provided data for two ports, Dublin and Rotterdam, for the reference year 2019. These data were compared with Eurostat F2 dataset in number of vessels and GT.

Eurostat data for the port of Dublin showed low deviation on the number of vessels with EMSA-MARINFO, leading to the conclusion that the use of EMSA-MARINFO data is more adequate than the other EMSA datasets.

The only significant deviation in comparing number of vessels of Eurostat and EMSA-MARINFO data for the port of Dublin for 2019 concerned the ‘Containers’ (31) vessel type. A difference of 20% was observed for 2019 that requires further analysis assisted by the reporting countries.

Rotterdam is one of the biggest EU ports, with wide range of shipping activities and the large total volume of EU maritime transport of goods. The differences between Eurostat F2 dataset and EMSA-MARINFO and EMSA-SSN were lower than with EMSA-DPC and would therefore be more suited candidates for estimating Eurostat F2 dataset.

In addition, it appeared that the configuration of ‘statistical port of Rotterdam’ as defined by Eurostat does not correspond to ‘the port of Rotterdam’ in EMSA datasets (the regrouping of sub-ports or terminals for data provision is different) and would need further clarification of the port terminals recorded under ‘port of Rotterdam’ in order to better analyse data.

This analysis showed some of the challenges that might be faced if estimates are done at lower level of aggregation such as at port level. Differences in comparability of EUROSTAT and EMSA data may exist for a specific data subset (e.g., only container vessels, for a particular year). Further cooperation with the countries might be needed to run in-depth analyses of the data reporting at port level and vessel level.

3.3.2. Comparison at country level for all reporting countries

The comparisons at country level were performed for the reference years 2018 and 2019, also at a more detailed level considering vessel type and for the reference period 2015-2019.

When focusing on the reference years 2018 and 2019, all three EMSA data sources - EMSA-SSN, EMSA-MARINFO and EMSA-DPC - showed matching parts but also several deviations. Deviations varied across countries, as well as over time. The differences however between Eurostat and EMSA data were lower without taking into account the vessel type than when considering it. No clear pattern could be seen between ports depending on the volume of the maritime traffic. Based on the analysis by vessel type, countries were classified in four groups according to the level of the observed differences in the number of vessels. First group of three countries with major differences (Denmark, Greece, Croatia), second group with six countries having differences that may differ depending to the EMSA source used (Malta, Italy, Latvia, Estonia, Cyprus and Romania), third group of other six countries where the differences are less important (Germany, Belgium, the Netherlands, Portugal, Sweden and France), and lastly a group of seven countries with insignificant differences, especially with EMSA-SSN data source (Ireland, Bulgaria, Slovenia, Lithuania, Poland, Spain and Finland).

When analysing the trends of Eurostat F2 dataset and EMSA datasets over the reference period 2015-2019, a similar seasonality could be seen in the data from Eurostat F2 dataset, EMSA-SSN and EMSA-MARINFO datasets. The same pattern could not be observed in EMSA-DPC data compared with other data sources; explained mainly by the fact that this data source is a recent one with shorter timespan.

The analysis at country level showed variable comparability of Eurostat-EMSA data at vessel type level. Producing reliable short-term estimates at country level based on the absolute number of EMSA port calls require improved match of vessel types. Similarities however in trends of Eurostat and EMSA data give good assurance in using EMSA data for the estimation of Eurostat F2 statistics.

3.3.3. Comparison at EU level

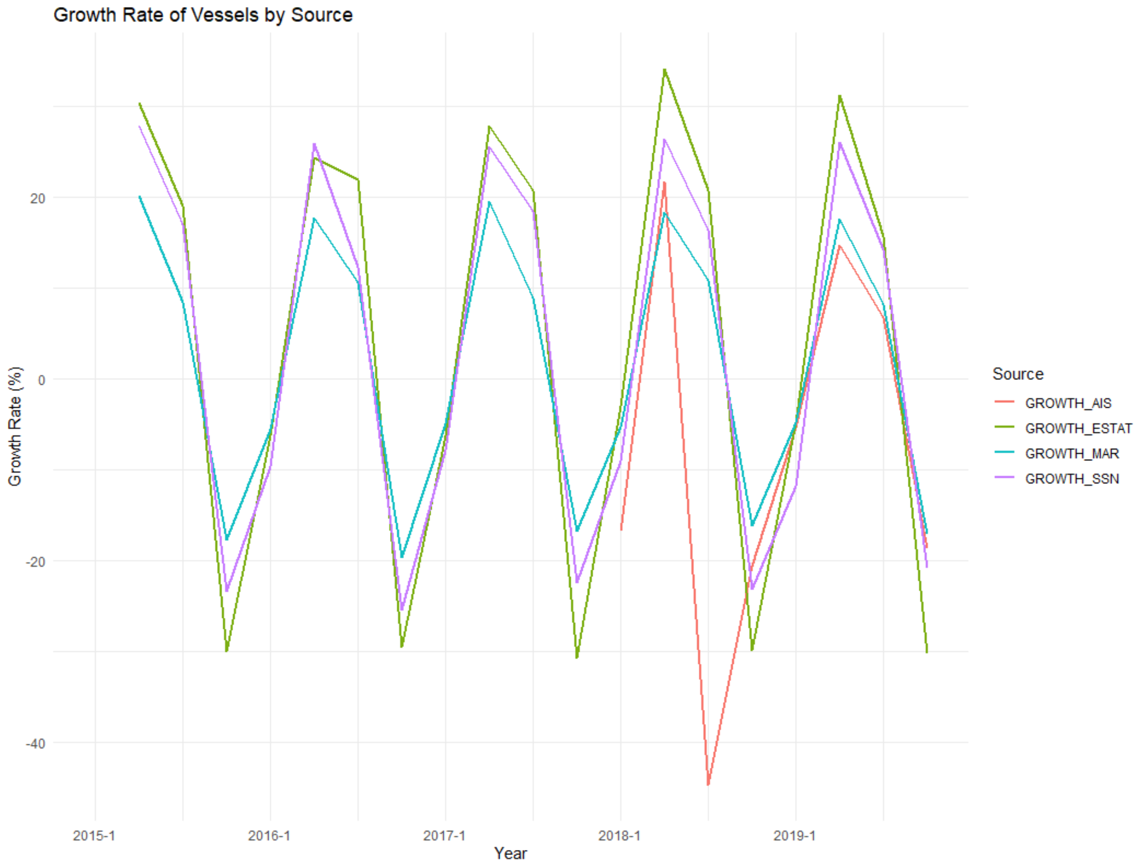
Further to the country level comparisons, number of vessels (vessel traffic) in Eurostat and EMSA data were also compared at EU level for the years 2018-2019. The analysis of the differences between Eurostat F2 dataset and EMSA-MARINFO, EMSA-SSN and EMSA-DPC data sources revealed varying levels of deviation for different vessel types.

For vessel types ‘General cargo, non-specialised’ (33) and ‘Passenger’ (35), which represent a high share of vessels, the EMSA-MARINFO data provided more consistent and comparable results. In some other cases, EMSA-SSN data seemed to be the best candidate.

This would suggest that EMSA-SSN and EMSA-MARINFO should be combined in order to get better estimations for Eurostat dataset F2.

For the reference period 2015-2019, it could be seen that the comparisons between Eurostat and EMSA three data sources vary depending on the category of vessels. To the performed comparison, it could be added that, in general, the differences between Eurostat and EMSA data tend to slightly decrease in the more recent years. In addition, the trends over the period 2015-2019 were similar for the data from Eurostat F2 dataset, EMSA-SSN and EMSA-MARINFO and only EMSA-DPC data showed significant deviations, particularly at the beginning of the series.

Figure 1



Note: GROWTH_AIS is EMSA-DPC; GROWTH_ESTAT is Dataset F2 of Eurostat; GROWTH_MAR is EMSA-MARINFO and GROWTH_SSN is EMSA-SSN.

The analysis at EU level shows a strong similarity of trends between the EMSA-SSN and EMSA-MARINFO data and Eurostat dataset that would allow proceeding with their use for estimations of Eurostat F2 data at EU level.

3.3.4. Potential reasons for differences

While performing comparisons between Eurostat F2 dataset and EMSA datasets, the following potential reasons for the differences were identified.

- Classification by type of vessel: classifying EMSA vessel information according to the Eurostat vessel category reporting is crucial for the comparability of the two sources. Furthermore, some records in the EMSA dataset do not contain information on the type of vessel. These vessels are consequently missing from their vessel category. Even if the share of missing information is low (less than 1% of the total number of vessels for each EMSA dataset at EU level), it could have some influence on improving accuracy.

- Scheduled traffic between two ports: it was noticed that several countries report similar values for some ports and vessel types to Eurostat, which could be explained by the scheduled traffic between pairs of ports. It concerns the categories of vessel

‘General cargo, non-specialised’ which includes ‘Ro-ro passenger’ vessels. It could also be observed that for most of these ports, there is no traffic reported in EMSA data sources.

- Definition of the statistical ports: the definition of the ports included in EMSA datasets records are not always the same as the ‘statistical ports’ reported for dataset F2. This limits for the time being data comparability at port level in several ports.
- Activity of the vessel: in Eurostat dataset F2, a port call is recorded only if the vessel performed an activity in the port (loading/unloading goods or embarking/disembarking passengers). In each EMSA dataset, a port call is recorded regardless of if the vessel performed an activity in the port or not. This led to a small overestimation of the number of port calls in EMSA data comparing to Eurostat.
- Exemptions in reporting to EMSA: the exemption of reporting by some vessels in the EMSA-SSN dataset is another reason for differences. It was observed that in several of the cases, EMSA-SSN data were lower than those of Eurostat.

4. TESTED METHODS FOR THE ESTIMATES

The results of this comparison exercise showed the possibility of using EMSA data to estimate Eurostat F2 dataset on quarterly basis at EU level with good quality. Therefore, two estimations methods were tested and compared, in order to select the most reliable one. These methods were a multiple linear regression and the so-called Auto-Regressive Integrated Moving Average with Exogenous variables (ARIMAX) method.

4.1. Multiple linear regression

4.1.1. Description

During the comparison exercise between Eurostat and EMSA data, discrepancies were noted in vessel traffic reports for some ports. These were due to different reporting methods for scheduled traffic. Consequently, some ports reported identical vessel numbers to Eurostat but lower or zero values to EMSA. To maintain data accuracy and reduce error margins in the analysis, these ports were identified, and the relevant vessel types were excluded from the dataset used for linear regression.

A multiple linear regression model was used to evaluate the relationship between various factors. The following model was tested:

$$VESSELS_{ESTAT} \sim PERIOD + VESSELTYPE + VESSELS_{MARINFO} + (VESSELS_{MARINFO})^2 + VESSELS_{SSN} + (VESSELS_{SSN})^2$$

Where,

$VESSELS_{ESTAT}$ is the variable to be estimated (predicted)

PERIOD is the reference period (e.g. quarter)

VESSELTYPE is the vessel type according to the classification of Directive 2009/42/EC

$VESSELS_{MARINFO:}$ represent counts of vessels for specified periods and vessel types in EMSA-MARINFO (as shown in the specimen in section 2.2.2)

$VESSELS_{SSN:}$ represent counts of vessels for specified periods and vessel types in EMSA-SSN (as shown in the specimen in section 2.2.1).

To predict the number of Eurostat vessels, the model used the period (quarter), vessel type, the number of vessels from MARINFO and the number of vessels from SSN. The model also accounted for the number of vessels (squared) from EMSA data, to develop a more precise correlation.

For the development and testing of the model, the available data were separated into two sets: a 'training set' for model creation and a 'testing set' for validation, each chosen to be representative of the overall dataset.

To reinforce the model's reliability, a bootstrapping technique was applied, involving repeated resampling from the training set. This process that entailed 1,000 iterations, was essential in evaluating the model's consistency and strength.

4.1.2. Evaluation and results

4.1.2.1. Accounting for Excluded Data

To account for scheduled traffic between pairs of ports that was previously excluded, the EU average growth rate based on the previous period was applied. This approach enabled a comprehensive overview of the vessel statistics and their estimated values.

The following average growth rates, calculated between two consecutive quarters and rounded to the nearest whole number, were obtained for the specified period. For these records, Eurostat values from the preceding period were taken, and the corresponding average growth rate for each quarter was applied. This approach allowed the vessel statistics for previously excluded records to be estimated by leveraging the historical growth trends observed in the data.

4.1.2.2. Final Estimates

The objective of this step was to make estimates at EU level using the results from the linear regression analysis. Occasionally, for ports and vessel types with low vessel movement, predictions may result in values less than zero. Negative predictions were adjusted and set to zero, as having a negative number of vessels is not possible. Following these adjustments, the final estimates were produced.

The estimates of the scheduled traffic between pairs of ports were added to the estimations done on the basis of the multiple regression method.

The data were then aggregated by year, quarter, and vessel type, enabling to compare the calculation of the sum of the actual values (in Eurostat statistics) and the estimates for each category of vessels (see Table 1). The percentage difference between the model’s estimates and Eurostat’s data is shown at annual level (i.e. sum of the four quarters) presenting some significant deviations. Quarterly estimates are also indicated.

In conclusion, using this method results at EU level for 2015 to 2019 showed a difference in annual totals between -7.4% and 0.5%, and much larger differences at a more disaggregated level, by type of vessel.

Table 1: Comparison of dataset F2 and estimated EU vessel data by vessel type by quarter *(number of vessels for Q1/2015-Q4/2019)*

	Dataset F2					ESTIMATE					
	Q01	Q02	Q03	Q04	Total	Q01	Q02	Q03	Q04	Total	Deviation of annual data (%)
2015	**398 415**	**519 823**	**619 101**	**433 469**	**1 970 808**	**432 981**	**576 093**	**651 607**	**461 468**	**2 122 149**	**-7.4**
10	19 642	20 552	21 668	20 008	81 870	22 019	22 876	34 206	22 248	101 349	-21.3
20	10 222	10 319	9 842	9 951	40 334	10 102	9 648	13 106	10 490	43 347	-7.2
31	17 109	18 093	18 210	18 078	71 490	22 265	25 971	27 143	22 468	97 847	-31.1
32	3 881	4 283	4 611	4 223	16 998	4 210	5 129	4 615	4 762	18 716	-9.6
33	308 011	373 328	426 837	321 647	1 429 823	319 073	413 598	447 590	331 436	1 511 697	-5.6
35	38 597	87 936	131 592	56 242	314 367	55 311	84 615	103 821	67 193	310 941	1.1
36	953	5 312	6 341	3 320	15 926	-	14 257	21 125	2 870	38 252	-82.4
2016	**406 320**	**505 464**	**616 473**	**434 519**	**1 962 776**	**429 439**	**580 408**	**648 310**	**420 469**	**2 078 625**	**-5.7**
10	19 665	20 590	21 683	19 727	81 665	20 205	22 062	29 947	18 769	90 983	-10.8
20	9 857	10 147	9 892	10 022	39 918	10 892	6 452	9 300	6 260	32 904	19.3
31	18 311	19 183	19 156	17 948	74 598	23 389	23 993	24 214	16 923	88 519	-17.1
32	4 073	4 452	4 400	4 248	17 173	5 996	5 250	5 048	5 367	21 660	-23.1
33	306 796	354 631	419 990	321 133	1 402 550	317 055	419 959	444 283	305 179	1 486 476	-5.8
35	46 619	90 071	133 398	58 100	328 188	51 901	87 909	111 932	64 700	316 442	3.6
36	999	6 390	7 954	3 341	18 684	-	14 785	23 585	3 272	41 641	-76.1
2017	**407 708**	**521 495**	**629 958**	**436 885**	**1 996 046**	**404 272**	**565 179**	**651 224**	**424 282**	**2 044 957**	**-2.4**
10	19 188	20 534	22 894	20 514	83 130	15 437	22 590	34 530	25 548	98 105	-16.5
20	10 333	10 392	10 237	10 842	41 804	8 945	6 612	10 557	10 272	36 386	13.9
31	17 832	18 613	19 139	19 283	74 867	16 900	19 218	25 692	22 959	84 769	-12.4
32	4 613	5 135	4 928	4 591	19 267	3 980	6 468	7 101	5 359	22 908	-17.3
33	312 299	367 981	436 846	323 194	1 440 320	305 050	403 322	433 510	289 847	1 431 730	0.6

35	42 352	93 292	129 167	54 693	319 504	53 960	92 254	118 145	66 075	330 434	-3.4
36	1 091	5 548	6 747	3 768	17 154	-	14 714	21 689	4 222	40 624	-81.2
2018	**424 022**	**568 903**	**687 320**	**482 214**	**2 162 459**	**429 672**	**594 405**	**691 793**	**466 992**	**2 182 863**	**-0.9**
10	19 462	20 880	22 634	20 721	83 697	21 609	16 989	31 631	18 580	88 809	-5.9
20	10 165	10 833	10 440	10 903	42 341	10 953	9 314	11 291	11 152	42 710	-0.9
31	18 549	19 237	19 051	18 008	74 845	22 285	22 793	24 338	20 286	89 701	-18.1
32	4 656	5 102	4 780	4 188	18 726	8 853	5 030	4 943	6 787	25 613	-31.1
33	329 200	410 093	478 910	364 892	1 583 095	315 951	425 288	478 109	339 315	1 558 663	1.6
35	40 588	96 624	143 897	59 482	340 591	50 021	97 665	118 882	66 044	332 612	2.4
36	1 402	6 134	7 608	4 020	19 164	-	17 328	22 599	4 828	44 754	-80.1
2019	**459 137**	**602 794**	**697 565**	**487 129**	**2 246 625**	**445 752**	**620 107**	**699 973**	**469 008**	**2 234 840**	**0.5**
10	20 346	22 410	22 823	21 094	86 673	22 900	26 699	34 101	26 866	110 566	-24.2
20	10 250	10 188	9 730	10 080	40 248	11 810	10 575	10 898	9 739	43 022	-6.7
31	17 483	18 088	17 957	17 434	70 962	18 540	18 254	20 227	17 559	74 581	-5.0
32	4 128	4 459	4 473	4 255	17 315	7 440	4 009	4 627	3 178	19 254	-10.6
33	360 034	434 793	482 219	372 346	1 649 392	337 965	442 464	476 274	345 896	1 602 598	2.9
35	45 561	106 692	153 456	57 557	363 266	47 097	99 668	129 065	59 301	335 131	8.1
36	1 335	6 164	6 907	4 363	18 769	-	18 438	24 782	6 468	49 688	-90.3

4.1.3. Conclusion

The approach using multiple linear regression models, enhanced by bootstrapping, effectively estimates vessel statistics. It provides a comprehensive analysis by accounting for scheduled traffic between port pairs. The method's robustness and the ability to calculate reliable confidence intervals for coefficients are notable strengths.

However, the model has limitations. It assumes linearity and may not capture all interactions between predictors, which could lead to inaccuracies. Challenges include managing scheduled traffic and accurately taking into account all reporting ports. Higher uncertainties at country level are also limiting its use for more granular estimates.

In addition, the accuracy of the model may decrease if data trends change. It might not adapt to new patterns or significant changes in maritime traffic. This emphasizes the importance of regularly updating the model to keep it relevant and accurate.

Given these limitations, an alternative approach using the ARIMAX model was explored.

4.2. Auto-Regressive Integrated Moving Average with Exogenous variables (ARIMAX)

4.2.1. Description

ARIMAX is a statistical method used for forecasting time-series data. In simpler terms, it predicts results, such as the number of vessels, based on past data and additional influencing factors. In this analysis, the ARIMAX model used historical data from the Eurostat F2 dataset and EMSA sources (as exogenous variables).

Historical Eurostat and EMSA data were grouped by vessel type and period at EU level to create time-series for prediction. The suitability of data for the ARIMAX modelling required at least two years of quarterly data.

The model fitting involved repetitive analysis for each vessel type with EU-level data. This process involved checking whether the data were appropriate for the model, creating time series and conducting statistical tests to detect and account for any seasonal variations in the model.

The weights assigned to the exogenous variables VESSELS_SSN and VESSELS_MARINFO are determined for each type of vessel to ensure that the specific impacts of external factors on forecast outcomes are accurately recognized. For each vessel category, distinct models are developed, reflecting the unique characteristics and trends of each type in the forecasted results.

The use of EMSA data in this method in addition to Eurostat previous quarters statistics, give assurance on capturing changes in vessel traffic (port calls) timely that may arise due to punctual events, such as closure of Suez Canal, pandemics, taxation, etc and will only appear in statistics later.

4.2.2. Evaluation and results

The model was run over the period 2015 Q1 to 2019 Q4 to simulate a nowcasting scenario of 2019 Q4, which involved making immediate short-term forecasts. This approach allowed for a direct comparison between the model's predictions and the actual observed values. By measuring the discrepancies between the predicted and real values, the quality and reliability of the ARIMAX model were assessed. This was essential for evaluating how effective the model was in generating early estimates.

Lastly, the totals of the actual values (Eurostat data) were compared to the totals of the estimated values (ARIMAX) by vessel type (see Table 2).

The results at annual level showed a difference of 5.3% in 2015, the first reference year in EMSA data. In the following more recent years the difference was significantly lower, at 0.5% in 2016, -1.3% in 2017, -1.8% in 2018, and 0.5% in 2019. The estimates by type of vessels also have lower differences comparing to the multiple regression model. These results indicate that this method is more suitable for producing estimates also by type of vessel at EU level.

The ARIMAX model produced more accurate estimates of the Eurostat data than the linear regression model at the EU level it is therefore better fitted for estimating the number of vessels' arrivals in EU ports.

Table 2: Comparison of dataset F2 and estimated EU vessel data by vessel type by quarter (number of vessels for Q1/2015-Q4/2019)

	Qtr1			Qtr2			Qtr3			Qtr4					
Vessel types	**ARIMAX**	**Dataset F2**	**deviation Q1**	**ARIMAX**	**Dataset F2**	**dev Q2**	**ARIMAX**	**Dataset F2**	**dev Q3**	**ARIMAX**	**Dataset F2**	**dev Q4**	**Total YEAR ARIMAX**	**Total YEAR dataset F2**	**Deviation year**
2015	**446 377**	**398 415**	**11%**	**528 321**	**519 823**	**2%**	**625 353**	**619 101**	**1%**	**477 055**	**433 469**	**10%**	**2 077 106**	**1 970 808**	**5.3%**
10	22 065	19 642	12%	21 999	20 552	7%	20 552	21 668	-5%	22 123	20 008	10%	86 739	81 870	5.8%
20	10 159	10 222	-1%	10 164	10 319	-2%	9 924	9 842	1%	10 153	9 951	2%	40 400	40 334	0.2%
31	16 731	17 109	-2%	17 109	18 093	-6%	17 199	18 210	-6%	17 522	18 078	-3%	68 561	71 490	-4.2%
32	3 916	3 881	1%	3 881	4 283	-10%	4 283	4 611	-7%	4 446	4 223	5%	16 526	16 998	-2.8%
33	343 777	308 011	11%	394 207	373 328	5%	444 617	426 837	4%	349 053	321 647	8%	1 531 654	1 429 823	6.9%
35	45 347	38 597	16%	79 097	87 936	-11%	124 575	131 592	-5%	69 491	56 242	21%	318 510	314 367	1.3%
36	4 382	953	129%	1 864	5 312	-96%	4 203	6 341	-41%	4 267	3 320	25%	14 716	15 926	-7.9%
2016	**384 404**	**406 320**	**-6%**	**529 229**	**505 464**	**5%**	**617 904**	**616 473**	**0%**	**441 558**	**434 519**	**2%**	**1 973 095**	**1 962 776**	**0.5%**
10	19 012	19 665	-3%	19 665	20 590	-5%	20 590	21 683	-5%	19 718	19 727	0%	78 985	81 665	-3.3%
20	10 222	9 857	4%	10 124	10 147	0%	10 125	9 892	2%	10 113	10 022	1%	40 584	39 918	1.7%
31	17 732	18 311	-3%	17 940	19 183	-7%	19 183	19 156	0%	18 804	17 948	5%	73 659	74 598	-1.3%
32	4 221	4 073	4%	4 138	4 452	-7%	4 341	4 400	-1%	4 320	4 248	2%	17 020	17 173	-0.9%
33	289 014	306 796	-6%	379 189	354 631	7%	425 722	419 990	1%	325 502	321 133	1%	1 419 427	1 402 550	1.2%
35	39 921	46 619	-15%	95 958	90 071	6%	133 727	133 398	0%	58 048	58 100	0%	327 654	328 188	-0.2%
36	4 282	999	124%	2 215	6 390	-97%	4 216	7 954	-61%	5 053	3 341	41%	15 766	18 684	-16.9%
2017	**395 780**	**407 708**	**-3%**	**514 345**	**521 495**	**-1%**	**616 046**	**629 958**	**-2%**	**444 910**	**436 885**	**2%**	**1 971 081**	**1 996 046**	**-1.3%**
10	19 882	19 188	4%	20 067	20 534	-2%	21 364	22 894	-7%	22 018	20 514	7%	83 331	83 130	0.2%
20	9 991	10 333	-3%	9 988	10 392	-4%	10 039	10 237	-2%	10 060	10 842	-7%	40 078	41 804	-4.2%
31	16 901	17 832	-5%	18 212	18 613	-2%	17 957	19 139	-6%	19 162	19 283	-1%	72 232	74 867	-3.6%
32	4 440	4 613	-4%	4 572	5 135	-12%	4 378	4 928	-12%	4 861	4 591	6%	18 251	19 267	-5.4%
33	301 889	312 299	-3%	364 798	367 981	-1%	413 913	436 846	-5%	329 016	323 194	2%	1 409 616	1 440 320	-2.2%
35	41 589	42 352	-2%	90 247	93 292	-3%	139 849	129 167	8%	56 105	54 693	3%	327 790	319 504	2.6%

36	1 088	1 091	0%	6 461	5 548	15%	8 546	6 747	24%	3 688	3 768	-2%	19 783	17 154	14.2%
2018	**408 908**	**424 022**	**-4%**	**536 552**	**568 903**	**-6%**	**673 033**	**687 320**	**-2%**	**505 283**	**482 214**	**5%**	**2 123 776**	**2 162 459**	**-1.8%**
10	19 578	19 462	1%	20 582	20 880	-1%	21 320	22 634	-6%	21 211	20 721	2%	82 691	83 697	-1.2%
20	10 210	10 165	0%	10 224	10 833	-6%	10 292	10 440	-1%	10 292	10 903	-6%	41 018	42 341	-3.2%
31	18 715	18 549	1%	18 635	19 237	-3%	19 339	19 051	2%	18 988	18 008	5%	75 677	74 845	1.1%
32	4 543	4 656	-2%	4 901	5 102	-4%	4 417	4 780	-8%	4 701	4 188	12%	18 562	18 726	-0.9%
33	312 607	329 200	-5%	376 867	410 093	-8%	458 932	478 910	-4%	385 543	364 892	6%	1 533 949	1 583 095	-3.2%
35	41 894	40 588	3%	98 898	96 624	2%	151 380	143 897	5%	60 690	59 482	2%	352 862	340 591	3.5%
36	1 361	1 402	-3%	6 445	6 134	5%	7 353	7 608	-3%	3 858	4 020	-4%	19 017	19 164	-0.8%
2019	**463 903**	**459 137**	**1%**	**596 279**	**602 794**	**-1%**	**693 693**	**697 565**	**-1%**	**503 403**	**487 129**	**3%**	**2 257 278**	**2 246 625**	**0.5%**
10	19 793	20 346	-3%	21 189	22 410	-6%	22 082	22 823	-3%	21 667	21 094	3%	84 731	86 673	-2.3%
20	10 693	10 250	4%	10 488	10 188	3%	9 814	9 730	1%	10 491	10 080	4%	41 486	40 248	3.0%
31	16 914	17 483	-3%	17 805	18 088	-2%	18 049	17 957	1%	17 657	17 434	1%	70 425	70 962	-0.8%
32	4 331	4 128	5%	4 523	4 459	1%	4 076	4 473	-9%	4 384	4 255	3%	17 314	17 315	0.0%
33	364 976	360 034	1%	428 104	434 793	-2%	481 625	482 219	0%	381 609	372 346	2%	1 656 314	1 649 392	0.4%
35	45 948	45 561	1%	107 592	106 692	1%	150 130	153 456	-2%	63 519	57 557	10%	367 189	363 266	1.1%
36	1 248	1 335	-7%	6 578	6 164	6%	7 917	6 907	14%	4 076	4 363	-7%	19 819	18 769	5.4%

4.2.3. Conclusion

The ARIMAX method is a robust predictive model for estimating vessel traffic data, leveraging historical information from both Eurostat's F2 dataset and EMSA sources. The model undergoes a rigorous fitting process, which includes seasonal adjustments and ensures the data's suitability for time series analysis. Its effectiveness was evaluated during the periods from 2015 Q1 to 2019 Q4, providing a comprehensive view of the model's accuracy of estimates at EU level. Overall, the ARIMAX model offers a reliable method for capturing underlying patterns and trends in vessel traffic, contributing to more accurate and insightful analyses.

5. CONCLUSIONS

After testing and analysing the results of both estimation methods, the use of the ARIMAX model is considered more suitable for the early estimations of vessel traffic, statistics that is provided by the reporting countries several months later. The ARIMAX model is particularly performing also at country level, as it provides for more accuracy. The method shows better results also by vessel type.

The primary focus of this analysis aims at presenting estimates of maritime vessel traffic at EU level. Preliminary findings however already indicate the potential for reliable estimates also by type of vessel.

Future work is needed to produce accurate estimates, at the country level or for selected main ports, e.g., top 20 EU ports for which statistics are currently published by Eurostat. To address the fundamental factor affecting the quality and granularity of estimates, further work is needed, in cooperation with reporting countries to improve the attribution of vessels to the Eurostat reporting categories in dataset F2. This improvement will also allow a better match of EMSA records to these categories. More research into how vessel activity is attributed to ports for reporting statistics can also provide for better matching with the relevant EMSA data. These efforts can improve the models' accuracy and allow for improved comparisons of the basic data.