Factors Affecting Ocean-Going Cargo Ship Speed and Arrival Time

Erwin Filtz[(✉)], Emanuel Sanchez de la Cerda,
Mathias Weber, and David Zirkovits

Vienna University of Economics and Business, Vienna, Austria
erwin.filtz@gmail.com, mathias_weber@gmx.at,
{h0850863,david.zirkovits}@wu.ac.at

Abstract. Due to the high density of ocean traffic and the influence of marine weather on the route planning of vessels, as well as berth allocation in harbors, it is important to be able to predict arrival times as precise as possible. This paper shows the influence of marine weather on ship speed by analyzing publicly available ship traffic and weather data from different sources. A linear regression model is created to explain recorded ship speed in terms of certain ship properties and marine weather. The model has an adjusted R^2 value of 83.98% with a significant correlation of many weather related data such as wind direction (0.211), significant wave height (0.195), peak wave period (0.133), as well as ship-related data including ship type, dead weight tonnage, and gross register tonnage. Given the variables in the model the speed of the ship could be estimated fairly well. These variables along with other factors are tested regarding their usefulness for the prediction of arrival times.

Keywords: Cargo ship speed · Marine vessel speed prediction · Marine weather effects · Predicting arrival times

1 Introduction

In 2012 about 9,2 billion tons of goods destined for seaborne trade were loaded in ports worldwide. With a steady growth rate, maritime transport has more than doubled since 1980 and can be considered as one of the most important transport modes in today's global economy. Between 2012 and 2013 the number of seagoing merchant vessels of 100 GT (gross tonnage) and above grew by 6% to a total of 1,628,783 [1]. With this ever increasing number of ships and freight capacity the market is highly competitive and declining freight rates reduce earnings, requiring operators to increase efficiency and to cut costs.

One significant factor for shipping operators is costs caused by vessel delays. This paper aims to determine factors that allow for a better prediction of ship arrival times and therefore enable involved parties in the shipping process to better deal with

---

possible delays. However, it is not only shipping operators but also other businesses holding a stake in ships being on time. In today's logistics where companies usually source from multiple suppliers and production gets ever more time critical it is advantageous to have accurate estimates on the arrival of goods allowing adjusting production and procurement accordingly. By enhancing process management with further information, businesses will be able to increase performance and optimize their processes.

Reasons for cargo vessels not arriving on time are numerous. A classification found in marine delay insurances divide them into shore side incidents and ship related incidents [2, 3]. The former include dock worker strikes, fire, lawful closures, and physical obstructions while the latter comprise crew strikes, collisions, strandings, crew illness, quarantine, and piracy. Another factor affecting vessel speed and therefore arrival time often mentioned is the weather along the shipping route [4], which includes precipitation, water levels, wave height, swell, wind speed and direction and a number of other factors.

Even though all of these can be causes for serious delays, most of them are either hard to predict or not publicly available and therefore not suitable factors for arrival prediction models. For the scope of this research external factors were limited to the effect of marine weather conditions on cargo ship speed.

The main purpose of this research is to identify internal (ship related, e.g. ship type and size, year of build) and external (non-ship related, e.g. weather, waves) variables that affect ship speed. The aim is to create a model with these variables explaining the actual recorded speed of a given ship. It shall be shown which variables and to which extent are useful to this model. To identify the importance of influencing variables a multiple linear regression model is used.
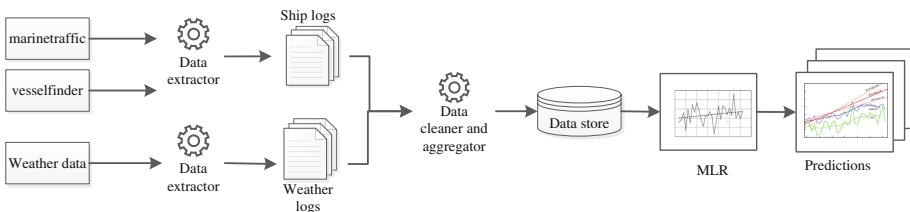

Fig. 1. Conceptual framework

Figure 1 shows the conceptual framework of this research work. Ship data is acquired from two different sources and saved in log files. Parallel to the ship data, weather data is also acquired from the weather source and saved in log files. Afterwards, the log files are combined to get a full log of ship data enhanced with weather data. This data is the basis for analysis via a multiple linear regression and on the predictability of arrival times.

This framework does not represent a monitoring or management system on its own. This work aims to investigate the effects of specific factors on the speed and arrival times of vessels. The findings are ought to be incorporated into models that deal with

the prediction of delays of vessels that support business process management systems in the fields of berth allocation and ship operating.

The remainder of this paper is structured as follows: Section 2 describes the data acquisition and preparation process for both the ship and weather data. Section 3 outlines the correlations between variables and the multiple linear regression is explained in Sect. 4. Section 5 provides a closer look into the arrival deviation at the port of Rotterdam. It is followed by Sect. 6 where related work is mentioned. Finally, Sect. 7 concludes the paper.

2   Data

The data used within this research originates from two different sources. Ship movement data is broadcasted by marine vessels worldwide via the Automatic Identification System (AIS) and made available by AIS service providers on the Internet. Weather data required for the model is made available by the Environmental Research Division's Data Access Program (ERDDAP). Both datasets are then combined and used to create the prediction model.

2.1   Vessel Data

As there is no ready-to-use ship movement dataset available free of charge, it is necessary to aggregate this data with specialized scripts.

For the vessel data it was decided to follow two different approaches with two different data sources, namely marinetraffic.com and vesselfinder.com. This approach enables us to try out different ways of data collection and provides us, in case one approach turns out to be a dead end, with data to work with from the other data source.

Based on the data provided by vesselfinder.com an area around the harbor of Rotterdam with a $\sim$ 200 mile radius is selected where all vessels sailing through are recorded at a 15 min interval. While there were millions of data points recorded over the course of four weeks, they are not used for further analysis as it proved too difficult to filter and link them with accurate weather data.

Parallel to the vesselfinder data acquisition, data is also collected from marinetraffic, in which all vessels sailing to Rotterdam are recorded worldwide. This method provides several advantages over the other approach including longer observation periods for each ship, obstacle free tracking and more information provided by the website (e.g. estimated arrival time, vessel status).

The following list describes the data we retrieved from both approaches whereas the source is indicated in brackets after the variables. "m" indicates data retrieved from marinetraffic, "v" data from vesselfinder: timestamps (m, v) of the query and currentness of data, International Marine Organization Number IMO (m), Maritime Mobile Service Identity MMSI (m,v), name (m), call-sign (m), flag (m), ship type (m,v), gross weight and deadweight tonnage (m), length and width (m), built (m), status (m), area (m), latitude and longitude (m,v), activity (m), speed (m,v), course (m,v), draught (m), estimated time of arrival (m), wind speed (m), wind direction as classification (m), wind

direction in degrees (m), air temperature in °C (m), departure time at previous port (m), name of previous port (m) and the destination (m).

Not all variables are used in the final prediction model as they are not all contributing to the intended purpose. The finally used variables for the prediction model are described below.

2.2   Weather Data

As our data sources for AIS data do not provide any marine weather information, but wind information only, it is required to get marine weather information.

The marine weather information is provided by ERDDAP which allows downloading marine weather information based on time, latitude, longitude and selected variables as gridded data, which means that it contains the selected variables for a chosen area in a 0.5 degree grid.

The weather information is based on the third-generation wind-wave model WAVEWATCH III developed by the Marine Modeling and Analysis Branch (MMAB) of the Environmental Modeling Center (EMC) of the National Centers for Environmental Protection (NCEP). The third-generation model differs from its predecessors in major points, for instance physical approaches [5].

We do not investigate different possible routes from a vessels origin to a specific destination based on historic ship and current marine weather data as the acquired dataset only includes the current positions of the vessels.

Marine weather information that is available and part of our model includes the peak wave direction in degrees, peak wave period in seconds, significant wave height in meters, swell peak wave direction in degrees, swell peak wave period in seconds, swell significant wave height in meters, wind peak wave direction in degrees, wind peak wave period in seconds and wind significant wave height in meters.

The significant wave height is defined as the average height (trough to crest) of the highest one-third of the waves [6]. The wave period in seconds describes the time between two peaks of a wave at the same point in space. The direction indicates where a wave is coming from.

The different weather information is related to the types of existing waves. Wind waves are generated through wind blowing over large area (called fetch). By contrast, swells are also called surface gravity waves and are caused not by local but distant weather systems.

2.3   Data Preparation

Data preparation is necessary to join ship and weather data. The latitude and longitude of the ships and the weather data is indicated in degrees, but the weather data is provided in a 0.5 degree grid. In order to being able to match both datasets, the ship positions must be adjusted and rounded to next integer or half of an integer. Furthermore, the longitude of the weather data is not indicated in the range ± 180, but 0 to 359 degrees. Hence, the longitude of the ship position must be added to 180. The adjusted position data is then used to join both datasets.

2.4   Data Recorded in Numbers

The data collection phase is divided into two periods. From 14th December 2014 until 23rd December 2014 we collected 55,776 ship positions. The second collection period lasted from 2nd January 2015 till 12th January 2015 and involved 68.395 observations. Hence, in total 124,170 observations. It must be noted that this number refers to the already cleaned dataset as the AIS data might not be updated with every data query, e.g. the ship is not within the range of an AIS receiver. Therefore, duplicate entries have been removed from the dataset.

As the work focuses on vessel highways located in Europe, weather data was only downloaded for that area (02N25W to 72N35E) for both vessel data acquisition periods which resulted in total 3,170,387 observations. The ship observations are then enhanced with marine weather data and result in a new dataset with 54,554 entries. The deviation of the two datasets is caused by the area where weather information is downloaded.

The focus is on the analysis of cargo ships and tankers sailing long distances over the ocean, other special ship types, e.g. "Tug" or "Dredger" not showing the typical behavior of ocean ships are removed from the dataset.

Sailing vessels show the activity "Underway using Engine". For the analysis of the ship speed all other observations with not-sailing activities of the ships, e.g. "Stopped" or "At Anchor" were removed from the dataset to avoid distorting the evaluations. The recorded data also includes inland water, e.g. "Kiel Canal", "Elbe River" or "Europe, Inland", and inter-port traffic through canals, e.g. Rotterdam to Antwerp. These ships are not exposed to equal environmental conditions as ocean ships, e.g. wave height and currents, and therefore as well removed from the dataset.

3   Correlations

As discussed in the previous section, the dataset was cleaned and irrelevant observations removed from the data. Thereafter, a meaningful analysis can be applied on the dataset. The aim is to identify properties that affect the speed of the vessel. Generally, properties are divided into vessel-related and weather-related properties. Correlation tests are used to find out which weather conditions and vessel properties impact the speed of the ocean ships. For the weather related analysis 11 variables as shown in Table 3 were collected.

The absolute wind direction given in degrees (0-360°) was adjusted by each ship's sailing course resulting in a new variable that shows the relative wind direction *appwinddir* ranging from full head winds to full tail winds, but does not indicate port or starboard side. The following method is used to calculate the new variable:

$$AppWindDir = \begin{cases} 360 - |ShipCrs - WindDir|, & if \ |ShipCrs - WindDir| > 180 \\ |ShipCrs - WindDir|, & otherwise \end{cases}$$

The results of our correlation tests are shown in Table 1. The highest correlation is identified between wind direction (degrees) and speed of the vessel. Also the significant

wave height and the swell significant wave height show quite a moderate correlation to the speed of the vessel (18%-19%). Wind speed has just a low negative correlation with about -8%. The higher the wind speed, the slower the vessel. Peak wave direction and swell peak wave direction do not correlate with the speed of the vessel. The low P-value in most of the results can be explained by the high number of observations (23,810).

Table 1. Correlation results

Variable	Correlation	P-Value
winddirdeg (wind direction, degrees)	0.2113996	$< 2.2 * 10^{-16}$
windspd (wind speed)	-0.08453314	$< 2.2 * 10^{-16}$
wdir (peak wave direction, degrees)	0.0513367	$2.22 * 10^{-15}$
tper (peak wave period, seconds)	0.1325658	$< 2.2 * 10^{-16}$
thgt (significant wave height, meters)	0.1947815	$< 2.2 * 10^{-16}$
sdir (swell peak wave direction, degrees)	-0.00387547	0.5629
sper (swell peak wave period, seconds)	0.1088287	$< 2.2 * 10^{-16}$

4   Multiple Linear Regression

The previous section deals with the correlation of weather data on the ship data. This section focuses on applying multiple linear regressions [7] on the data in order to find out if speed of the vessel can be described by corresponding variables. The multiple linear regression consists of two parts, the dependent variable (regressand) and the independent variable (regressor). While the regressand represents the effect of the formula, regressors are input variables that constitute the causes that lead to the effect. In our context, the speed of the ship (effect) has to be predicted by corresponding variables (causes). A variable selection technique lets the algorithm decide which attribute is considered as relevant and which is omitted. Observations with missing values have to be removed, because this technique cannot deal with them.

In our context, not only the weather data is taken into account but also some other attributes that are related to the ships and to the current location of the vessels: deadweight tonnage, gross, built and area. For multiple linear regression analysis, this paper focuses on specific areas on the oceans. These areas have several characteristics in common: Vessels are travelling in a straight line, are not encumbered by other vessels and can get top-speed in these areas. Hence, these areas are named as "vessel highways" and separated based on longitude/latitude specifications. The three vessel highways are: The English Channel, the west coast of Portugal in the Atlantic Ocean and the area between Sicily and Africa in the Mediterranean Sea. After the data filtering, 1,305 observations remained which we used for our analysis (Table 2).

Seven variables were removed from the model after applying the variable selection (see Table 3) via a stepwise regression. All other variables shown in Table 3 were considered as important from the selection algorithm.

Table 2. Linear regression models

Initial model	Results
shiptype, crs, windspd, thgt, whgt, tdir, wdir, sdir, shgt, area, activity, tper, sper, wper, dwt, built, gross, relwinddir	R^2: 87.76% Adjusted R^2: 84.94% P-Value: $<2.2 * 10^{-16}$
Final model	Results
shiptype, windspd, whgt, area, activity, tper, wper, dwt, built, gross, relwinddir	R^2: 86.71% Adjusted R^2: 83.98% P-Value: $<2.2 * 10^{-16}$

Table 3. Eliminated/Important variables

Variable	Description	Eliminated variable	Important variable
wdir	wind peak wave direction, degrees	X	
crs	course of the ship, degrees	X	
shgt	swell significant wave height, meters	X	
tdir	peak wave direction, degrees	X	
thgt	significant wave height, meters	X	
sdir	swell peak wave direction, degrees	X	
sper	swell peak wave period, seconds	X	
shiptype	type of the vessel (cargo, tanker)		**X**
windspd	windspeed		**X**
whgt	wind significant wave height, meters		**X**
activity	activity of the vessel		**X**
tper	peak wave period, seconds		**X**
wper	wind peak wave period, seconds		**X**
dwt	deadweight Tonnage		**X**
built	year the ship was built		**X**
gross	gross register tonnage		**X**
area	area in which the ship is located		**X**

As a result 86.71 % of the speed values can be explained via corresponding variables. However, these results have to be considered with some caution since variable selection is a dubious method against overfitting. Variable selection is highly discussed in the literature. On the one hand, theory tends to decline this approach while on the other hand it is frequently used in practice. The fact is, that it does not cover all issues with overfitting and even can come with several new problems [8]. An evidence for legitimating the dubiety of this method could be the fact that although variables *shgt* (swell significant wave height) and *thgt* (significant wave height) correlate more with the speed of the vessel than variables *wper* (wind peak wave period) and *tper*

(peak wave period) they were removed from the model. What is more, for some variables it is easier to explain effects and correlations with the speed of the vessel than for others. Variable *built* for example, doesn't say anything about maintenance and modifications of a ship. A 10 year old ship could basically be as fast as a 2 year old ship. However, the algorithm decided to consider this variable as important.

5   Arrival Deviation - Harbor of Rotterdam

In order to give the findings of the prior sections a practical value, they are tested for their applicability for predicting the arrival of vessels at their destination. The harbor of Rotterdam is chosen as destination of interest, as the collected data contains information about vessels heading to this harbor.

At first the dataset for arrival deviation analysis is prepared. Then the influence of the destination within the harbor is investigated. After that the findings are adapted to be applicable for testing their correlations with the arrival deviation.

5.1   Dataset for Analyses

For the purpose of investigating the arrival deviations, data about each shipping is needed. Additionally to the variables of the prior sections, the variables *ATA*, *ATA_moored*, *ETA_12hours* and *delay* have to be derived.

The actual time of arrival (*ATA*) is not directly available in the dataset. At the harbor of Rotterdam the *ETA* has to refer to the Maas Center buoy (5200.9'N, 00348.8'E) which is positioned in front of the harbors' main entrance from the ocean [9]. The point in time the vessel passes that point is considered as *ATA*.

The second actual time of arrival refers to the point in time when the vessel finished the mooring operation at the berth (*ATA_moored*).

The derived variable *ETA_12hours* contains the ETA that the vessel was communicating approximately 12 h before *ATA*. The decision of using this 12 hour difference was made, because the maritime weather data near the harbor will be used to investigate its effects on the arrival deviation. The time period 12 to 24 h prior to arrival also has a high impact on the vessel management at the harbor [10].

The *delay* is the difference between *ATA* and *ETA_12hours* in hours. Negative values indicate that the vessel was arriving earlier than expected.

The dataset made for analysis of the arrival deviation consists of one record for each observed shipping. These records contain the derived variables mentioned above and ship related variables necessary for analyses. Additionally each record contains weather related information which is averaged throughout the time period between *ATA* and *ETA_12hours*, in order to observe the influence on the arrival deviation near the harbor.

This dataset is cleaned by the following criteria. Only records with a *delay* ranging from -12 h to +12 h are kept in the dataset. Other delays are not considered as being realistic or considered as errors in the collected data. Furthermore if one of the derived variables is not computable for a record, the whole record is left out. The reasons for

that could be incomplete or erroneous information in the collected data or shipping that do not follow the required behavior for this analysis.

5.2 Destination Within the Harbor

As the harbor of Rotterdam consists of many terminals and extends over a wide area it is assumable that the destination of a vessel within the harbor is influencing the vessels behavior. For investigation the area of the harbor is geographically divided into three sectors. Sector 1 is nearest to and sector 3 is farthest away from the ocean entrance to the harbor.

The first assumption is that the further the destination terminal is away from the ocean, the more time it takes from the arrival at the harbor entrance to being moored at the final position. This assumption is approved as the first column of Table 4 shows. The time consumed is continuously increasing along with the distance from the ocean entrance.

The second assumption is that the distance of the destination terminal from the ocean is influencing the delay of the vessel at the harbor entrance. This assumption is partly approved. On the one hand the second column in Table 4 shows that there is no significant difference between sector 1 and sector 2. On the other hand sector 3 shows a significantly lower average delay than the other sections. It can be said that vessels that are heading to terminals in sector 3 tend to be on time as their average delay is close to zero.

Table 4. Delay by sector

	from *ATA* to *ATA_moored*	average *delay*
Sector 1	137 min	141 min
Sector 2	154 min	154 min
Sector 3	196 min	20 min

5.3 Applicability of Findings

In this section the findings of the previous sections are proved regarding their applicability on a prediction of delay. In order to do that the related variables are prepared in a suitable form and then their correlation with the delay is tested.

At first the variables that are found to be useful for predicting the vessels speed are tested. The variables *windspd*, *relwinddir*, *whgt*, *tper* and *wper* are averaged over the 12 h timeframe prior to arrival at the harbor and added to the data records. The variables *shiptype*, *dwt*, *built* and *gross* are not changing over time during one shipping, so they can be added as static values to the data records. The *activity* and the *area* are not considered in this test, as they are used to separate the individual shipping from each other and therefore not giving differences between them. The results of the tests can be seen in Table 5. As a result of these tests, the variables that could be useful for such prediction are only the ones that relate to weather, namely the peak wave period,

Table 5 Correlations between variables and delay

Variable	Correlation	P-Value
tper	-0.271860309754	0.0218218460564
wper	-0.27024008102	0.0226520841171
whgt	-0.229939767856	0.0537246472085

the wind peak wave period and the wind significant wave height. These results emphasize the assumption that the weather conditions mainly influence the delay.

Secondly the findings about the destination terminal within the harbor are tested. To this extent, the ship movements are classified according to the section they are heading to. The result is a correlation coefficient of -0.221585715593 with a p-value of 0.00368431138331. It can be assumed that the destination terminal within the harbor is of importance to the prediction of delay. Future research will investigate how to better define the individual sectors and what could be the reason for the differences.

5.4   Limitations

There are some limitations regarding the results of the arrival deviation analysis. Due to a very short time frame for data collection, the number of usable movements is quite small. Combining these movements with available weather data, the number of useful and complete records decreases even further. It would be of interest to have more data available for analysis and also to be able to investigate seasonal effects on the delay. Another limitation of the analysis is that they rely on the assumption that the delay is caused in the 12 h timeframe prior to the arrival. For a more precise analysis it would be necessary to know how much of the delay has already occurred prior to the investigated time frame. The decision of only collecting data of vessels heading to Rotterdam turned out to have the downside of not knowing about the complete traffic at the harbor. This made the analysis regarding the traffic not feasible with this dataset, although this could be of interest regarding a prediction of delay. It is in the plans to circumvent such limitation in future research.

6   Related Work

There are researches that mainly focus on the methods of berth allocation planning. [11] is dealing with robust berth scheduling and [12] is dealing with dynamic approaches for container handling at the berth. Our work is focusing on making the arrival times of vessels predictable which would be contributing to a more dynamic berth allocation planning. [13] provides approaches for the extraction of vessel routes and anomaly detection in movement patterns for decision support systems. On the other hand [14] is investigating the impacts of tides on seaside operations in container ports. Both are focusing on specific factors that influence the arrival times. Our work is investigating in particular the impacts of weather conditions on the arrival times. [10]

follows a similar approach to our work. A machine learning approach is applied using ship and weather related data for the prediction of arrival times. A main difference is that our work is using publicly available data whereas [10] is using data reported directly by the harbors of Antwerp and Cagliari.

7 Conclusions

There is data publicly available that can be used for estimating ship speed and arrival time. This data includes ship related and non-ship related information. A first challenge was to collect the data from different sources. Then they had to be combined and cleaned resulting in one common data store. After that the data store was used for making predictions on the ship speed at a certain point in time using multiple linear regression. It turned out that weather related data has a strong influence on the ship speed, but also some ship related variables are of importance. Regarding the prediction of arrival times this work can only give suggestions on input variables that could be used for a prediction model, because of the mentioned limitations. What is found to have a significant influence on the arrival time are weather related variables and the geographically location of the destination terminal.

This work serves as a basis for further researches on the prediction of arrival times of ships. We only had the possibility to investigate the effects of certain factors. The influence of the ongoing traffic in the harbor or in front of it is still an open question. The data could also be collected over a longer period of time to investigate seasonal effects. In the end the findings of this work along with other factors could be applied in a prediction model for estimating delays.

References

1. United Nations Conference on Trade and Development, Review of Marine Transport 2013. http://unctad.org/en/PublicationsLibrary/rmt2013_en.pdf
2. RaetsMarine Insurance B.V. http://www.raetsmarine.com/news/marine-delay-insurance
3. Nordic Marine Insurance, General information on fixed premium delay insurance. http://www.nmip.se/dwnds/General_Information_on_Delay_Insurance_Sept_2013.pdf
4. National Imagery and Mapping Agency, Principles of weather routing, http://msi.nga.mil/MSISiteContent/StaticFiles/NAV_PUBS/APN/Chapt-38.pdf
5. Tolman, H.L.: User manual and system documentation of WAVEWATCH III. Environ. Model. Center MMAB **276**, 194 (2009)
6. National Oceanic and Atmospheric Administration. http://www.srh.noaa.gov/key/?n=marine_sigwave
7. Brant, R.: Multiple Linear Regression (2007)
8. Babyak, M.A.: What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. Psychosom Med. **66**(3), 411–421 (2004)
9. Port of Rotterdam, Port Information Guide November 2014. http://www.portofrotterdam.com/en/Shipping/sea-shipping/port-information/Documents/port_information_guide.pdf

10. Pani, C.: Managing vessel arrival uncertainty in container terminals: a machine learning approach, Ph.D.-Thesis (2013)
11. Xu, Y., Chen, Q., Quan, X.: Robust berth scheduling with uncertain vessel delay and handling time. Ann. Oper. Res. **192**(1), 123–140 (2012)
12. Ku, L.P., Chew, E.P., Lee, L.H., Tan, K.C.: A novel approach to yard planning under vessel arrival uncertainty. Flex. Serv. and Manuf. Journal **24**(3), 274–293 (2012)
13. Pallotta, G., Vespe, M., Bryan, K.: Vessel pattern knowledge discovery from ais data: a framework for anomaly detection and route prediction. Entropy **15**(6), 2218–2245 (2013)
14. Du, Y., Chen, Q., Lam, J.S.L., Xu, Y., Cao, J.X.: Modeling the Impacts of Tides and the Virtual Arrival Policy in Berth Allocation. Transportation Science (2015)