EJTIR

Issue 15(4), 2015
pp. 536-550
ISSN: 1567-7141
tlo.tbm.tudelft.nl/ejtir

Prediction of late/early arrivals in container terminals – A qualitative approach

Claudia Pani[1]
Department of Civil and Environmental Engineering and Architecture, University of Cagliari, Italy.

Thierry Vanelslander[2]
Department of Transport and Regional Economics, University of Antwerp, Belgium.

Gianfranco Fancello[3]
Department of Civil and Environmental Engineering and Architecture, University of Cagliari, Italy.

Massimo Cannas[4]
Department of Business and Economic Science, University of Cagliari, Italy.

This document describes the formatting guidelines for the EJTIR journal. Please save this file to your computer and use it as the basis for formatting your own document.
All styles are included and can be found under "Apply Styles". The styles are described in more detail in the remaining of this document. Vessel arrival uncertainty in ports has become a very common problem worldwide. Although ship operators have to notify the Estimated Time of Arrival (ETA) at predetermined time intervals, they frequently have to update the latest ETA due to unforeseen circumstances. This causes a series of inconveniences that often impact on the efficiency of terminal operations, especially in the daily planning scenario. Thus, for our study we adopted a machine learning approach in order to provide a qualitative estimate of the vessel delay/advance and to help mitigate the consequences of late/early arrivals in port. Using data on delays/advances at the individual vessel level, a comparative study between two transshipment container terminals is presented and the performance of three algorithmic models is evaluated. Results of the research indicate that when the distribution of the outcome is bimodal the performance of the discrete models is highly relevant for acquiring data characteristics. Therefore, the models are not flexible in representing data when the outcome distribution exhibits unimodal behavior. Moreover, graphical visualisation of the importance-plots made it possible to underline the most significant variables which might explain vessel arrival uncertainty at the two European ports.

Keywords: *classification tree, container terminal, data mining, late/early arrivals, random forest.*

1. Introduction

The efficiency of container handling operations can significantly affect terminal competitiveness (Tongzon and Heng, 2005; Vanelslander, 2005) and the competitiveness of the entire container supply chain or network that the port is part of (Sciomachen et al., 2009; Notteboom and Rodrigue, 2008). In addition, port technology, geographical position and terminal structure are

[1] A: Via Marengo 2, 09123 Cagliari, Italy T: +39 070 6755267 F: +39 070 6753209 E: claudia.pani@unica.it
[2] A: Prinsstraat 13, 2000 Antwerp, Belgium T: +32 3 2654034 F: +32 32654799 E: thierry.vanelslander@uantwerp.be
[3] A: Via Marengo 2, 09123 Cagliari, Italy T: +39 070 6755274 F: +39 070 6753209 E: fancello@unica.it
[4] A: Viale Sant'Ignazio 83, 09123 Cagliari, Italy T: +39 070 6753410 E: massimo.cannas@unica.it

the result of strategic decisions and hence cannot be altered in the short to medium term. At the tactical and operational levels however, it is possible to adopt methodologies for the optimal management of the terminal's resources and the logistics processes involved.

This study is a step towards better understanding the needs of terminal operators in a daily planning scenario and it proposes a specific instrument that is able to support planners in the short-medium term planning of activities. The latest ETA (Estimated Time of Arrival), sent at least 24 hours prior to the expected arrival time of the vessel, often has to be updated due to unexpected events, and the actual time of vessel arrival remains uncertain. This results in serious consequences directly associated with the related planning processes. A review of the literature highlighted that punctuality of the vessel's arrival commonly affects:

- Berth scheduling (Hendriks et al., 2010; Han et al., 2010; Moorthy and Teo; 2006, Du et al., 2010; Zhen et al., 2011; Salido et al., 2011; Ambrosino and Tanfani, 2012);
- Human resources and equipment allocation (Di Francesco et al., 2014; Gambardella et al., 1998; Fancello et al., 2011; Legato and Monaco, 2004);
- Yard planning (Bruggeling et al., 2011; Ku at al., 2012).

Although vessel arrival uncertainty in ports is a well-known problem for the scientific community, the literature review highlighted that in the maritime sector the specific instruments for dealing with this problem are extremely limited and vessel arrival uncertainty still remain a challenge for port operators. The problem was raised by Fancello et al. (2011) and Pani et al. (2014) who used a neural network algorithm and a regression tree algorithm, respectively, to deal with the problem of late arrivals in a Mediterranean port.

Furthemore, arrival uncertainty has also been the topic of several studies in the air transport sector. Flight delays at airports have become a very common problem. In particular, a number of empirical studies on this topic were carried out by several authors that used past data in order to identify the causes behind flight delays (Xu, 2007; Zonglei et al., 2008).

In this work, two different case studies are considered: the port of Cagliari and the port of Antwerp, located in the Mediterranean basin and in the North Sea respectively. The two different scenarios were crucial in order to better understand the specific characteristics of the problem being analysed before broadening and generalising the conclusions. In the first stage of the study, all the variables that may potentially influence late/early arrivals in port were collected, after which an analysis was conducted in order to extract useful information on the delay/advance of future arrivals using historical data on previous arrivals.

The remainder of the paper is set up as follows; Section 2 summarizes the methodological approach and the various algorithms that were employed as classifiers, Section 3 introduces the collected data, Sections 4 and 5 describe the port of Cagliari case-study and the port of Antwerp case-study, respectively, Section 6 concludes and proposes future developments.

2. Methodological framework

Overall, the literature showed that there are two different approaches towards the use of statistical modelling to reach conclusions from data: one approach assumes that the data are generated by a given stochastic data model, while the other treats the data mechanism as unknown and uses algorithmic models (Breiman, 2001). The approach we used falls within the latter one, in particular it focuses on the machine learning discipline based on methodologies for exploring and understanding historical arrivals. This approach is especially appropriate in this specific instance where there are currently no reference models that are able to specify the functional form between the outcome (vessel delay/advance) and the potential predictors (Breiman, 2001). The classification and regression algorithms used in machine learning share the

idea of understanding the specific link between the outcome and the predictors directly from the data. The real differences between the most recently notified Estimated Time of Arrival and the recorded Actual Time of Arrival will go on to form an historical knowledge base upon which the models are built. Many classifications of learning algorithms exist based on the underlying learning strategy. The literature highlights three different approaches that can be taken when dealing with classification problems: the discriminative approach (neural networks, support vector machines), the regression approach (logistic regression, decision trees, random forest), or the class-conditional approach (Bayesian classifiers). There is no general rule regarding which approach works best, it is mainly related to the researcher's goal and to data characteristics. In this specific application a regression approach is taken. First of all because, as compared to the discriminative approach models and class-conditional approach models, the regression models can be explained and interpreted more intuitively. Moreover, from a statistical point of view, the literature showed that Decision Trees and Random Forest outperform Neural Networks (NNs) for this specific case (Pani et al., 2014).

The algorithms, which are briefly described below, made it possible to have a qualitative estimate of the delay/advance by determining whether or not an incoming vessel is likely to arrive before or after its scheduled ETA. This section also describes the performance metrics used for evaluating the predictive power of the models.

2.1 Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable that can assume only two values, zero or one. The conditional probability of Y_i being one can be modeled as:

$$\Pr(Y_j = 1 | X) = \frac{\exp(\sum_i \beta_i X_i)}{1 + \exp(\sum_i \beta_i X_i)} \quad (1)$$

Where:

- Y is the outcome, coded as zero if a given vessel arrived earlier than the expected ETA, and one if it was delayed.
- X denotes the vector of input variables: X=(X1, X2,…,Xk) that can be numerical or categorical.
- The beta coefficients are usually unknown and must be estimated from the data.

2.2 CART

CART models (Breiman et al., 1984) can be considered local models in the sense that they indirectly specify different conditional distributions of $Y \mid X$, depending on the region of the covariate space where unit *i* lies. This is in contrast with the global relation imposed by classical modelling strategies and allows for greater flexibility. On the other hand, this localization makes it more difficult to assess the overall explanatory power of the predictors. In order to partition the covariate space, CART uses a binary algorithm, graphically visualized as a binary tree, which subsequently splits the observations into subsets where the distribution of Y becomes more and more homogeneous. The splitting procedure is defined in each node on the basis of covariate values: for a quantitative predictor the split value assigns the ith observation to the right or to the left subnode depending on whether $x_i \leq s \geq$ or $x_i > s$, while for a qualitative predictor the splitting rule depends on whether $x_i \in M$ or not, where M is a subset of the categories of the qualitative predictor. The best splitting variable and splitting point at each node are determined using a greedy algorithm that evaluates the homogeneity of the outcome variable in the resulting nodes using a homogeneity measure and stops the splitting process when homogeneity is not significantly improved. If the outcome variable is nominal, the model is called a classification tree

and the Gini impurity index is used as a homogeneity measure. If the outcome variable is continuous, the term regression tree is employed and variance or entropy are the common measures of homogeneity.

2.3 Random Forest

A Random Forest (Breiman, 2001) is a multitude of correlated trees that can be used for classification or regression purposes. In particular, a prediction for a continuous outcome can be obtained by averaging single-tree predictions, while a prediction for a categorical outcome can be obtained by majority voting. The trees of the forest are correlated via random selection: in particular in the R implementation of the random forest used here (Liaw and Wiener, 2002), (i) about two thirds of the data are randomly re-sampled to grow each tree and (ii) at each node the best splitting variable is selected from among a randomly chosen subset of all predictors. The random selection process is meant to improve the stability of predictions by differentiating the trees and then averaging the results. Moreover, the left out observations (Out Of Bag, OOB) are used to build an estimator of the prediction error (in a similar way to cross-validation) and to rank the relative importance of the variables in the prediction task. A natural measure of performance for a classifier is the difference between the proportion of votes for the correct class and the maximum proportion for other classes. This difference is calculated using the OOB data before and after a permutation of the values of the variable. If the variable is not important for a good classification, then the difference should be small and we can define an importance measure by averaging this difference over all OOB and trees of the forest.

2.4 Performance metrics

Two common performance metrics for evaluating the predictive power of a classifier are employed: the percentage of misclassified instances and the kappa statistic, which show how accurate the prediction is for each algorithmic model.

Percentage of misclassified instances

The percentage of misclassified instances is simply the percentage of incorrectly classified delay levels.

Kappa statistic

For a prediction problem involving a dichotomous variable, a binary classifier can classify an individual instance into the following four categories: false positive (FP), true positive (TP), false negative (FN) and true negative (TN). Various performance measures can be derived after recording the frequencies of each category on test data. The total prediction accuracy (ACC) and Cohen's Kappa coefficient for assessing the prediction accuracy are given by the following formulae:

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} = \frac{\mathrm{TP}+\mathrm{TN}}{N} \tag{2}$$

$$\mathrm{Kappa} = \frac{\frac{TP+TN}{N} - \frac{TP+FP}{N}\times\frac{TN+FN}{N}}{1-\frac{TP+FP}{N}\times\frac{TN+FN}{N}} \tag{3}$$

The ACC is simply the proportion of correctly classified instances and it can be misleading when the proportion of positive and negative outcomes are very different. The kappa statistic (Cohen, 1960) takes into account the agreement occurring by chance and so it can be regarded as a more reliable indicator of good prediction performance. It ranges from zero (no better prediction than that occurring by chance) to one (perfect prediction).

2.5 Cross-Validation

In practice, one option is to evaluate the generalization performance of the method or, in other words, its ability to generalize on new samples. In this case the performance measures are calculated on independent test data via cross-validation. A k-fold cross validation (usually k=10) requires that the dataset first be partitioned into k non overlapping subsets of approximately the same size. Then for i=1,…,k the model is fitted after removing the *ith* subset, which is left out in order to evaluate the error on independent test data. The final performance measure can be obtained by averaging the errors in the k test data set.

The algorithmic models in this paper were built and evaluated thanks to two different case studies: the port of Cagliari, located in the Mediterranean basin, and the port of Antwerp, located in the North Sea. Data collection for both ports is described in detail in the next section.

3. Data

A preliminary stage of the analysis was necessary in order to study the port structures and to interview planners about the problems they perceived. Subsequently, the most promising variables for predicting vessel delay were identified and information about all the mother and feeder vessels arriving at the ports over a fixed observation period was collected. The available collected variables are summarised in Table 1.

Table 1. Collected Variables

Variable type	Port of Cagliari	Port of Antwerp
Variables related to the physical structure of the vessel	Length *[m]* Gross tonnage *[tons]* Capacity *[TEU]* Vector Type *(mother/feeder)*	Length *[m]* Gross tonnage *[tons]* Capacity *[TEU]* -
Variables providing information about vessel owner	Owner's name Owner's frequency	Owner's name Owner's frequency
Variables providing information about vessel service*	Port rotation Sailing direction Previous port of call	- - Previous port of call
Variables that give an indication about vessel position**	Last Estimated Time of Arrival (ETA) *[dd/mm/yyyy]* Actual time of Arrival (ATA) at the pilot point *[dd/mm/yyyy]*	Last Estimated Time of Arrival (ETA) *[dd/mm/yyyy]* Actual Time of Arrival (ATA) at the Pilot Station, the Flushing, the Coordinatiepunt *[dd/mm/yyyy]*
Variables related to the specific terminal of arrival***	- -	Berth number Presence of a lock before reaching the terminal
Variables related to the weather conditions along the route	ug: geostrophic wind speed in the x (positive towards east) *[m/s]* vg: geostrophic wind speed in the y (positive towards north) *[m/s]* Hs: significant wave height m *[ft]* Tp: spectral peak wave period *[m]* θd: vector mean wave direction	ug: geostrophic wind speed in the x (positive towards east) *[m/s]* vg: geostrophic wind speed in the y (positive towards north) *[m/s]* Hs: significant wave height m *[ft]* Tp: spectral peak wave period *[m]* θd: vector mean wave direction

* Considering the large amount of data collected at the Antwerp port, these variables were not available at the port level but only at the terminal level.

** In the case of Antwerp, it was necessary to consider that before 09:59 a.m. on May 1st, 2012, the ETA refers to the moment the vessel passed Flushing, while after 09:59 on May 1st, 2012, the ETA refers to the moment the vessel passed the Pilot Station.

*** These variables were collected for the Antwerp port alone because the port of Cagliari is composed of only one container terminal.

The weather-related variables were available for four time intervals per day and were collected for some selected points in the Mediterranean basin and in the North Sea. Based on the previous port of call and on the ideal route travelled by each vessel, a match was created so that each arrival could be associated with the weather conditions that were observed in the points nearest to its route. The points were chosen in order to be representative of the weather conditions in the Mediterranean Sea (Figure1) and in North Sea (Figure 2). In both cases, the selected points are located at a distance corresponding to 24 hours and 12 hours sailing before arriving at the specific port.

Point	N	E
1	39.14	9.10
2	39.03	13.0
3	40.54	12.6
4	41.86	10.8
5	41.36	6.4
6	38.56	5.3
7	36.95	11.88
8	37.34	1.51
9	35.05	15.13

Figure 1. Selected points in the Mediterranean Sea

Point	N	E
1	51.44	3.34
2	53.27	2.96
3	50.53	0.63
4	49.47	-3.87
5	56.15	0.56
6	56.06	5.15
7	54.19	7.86

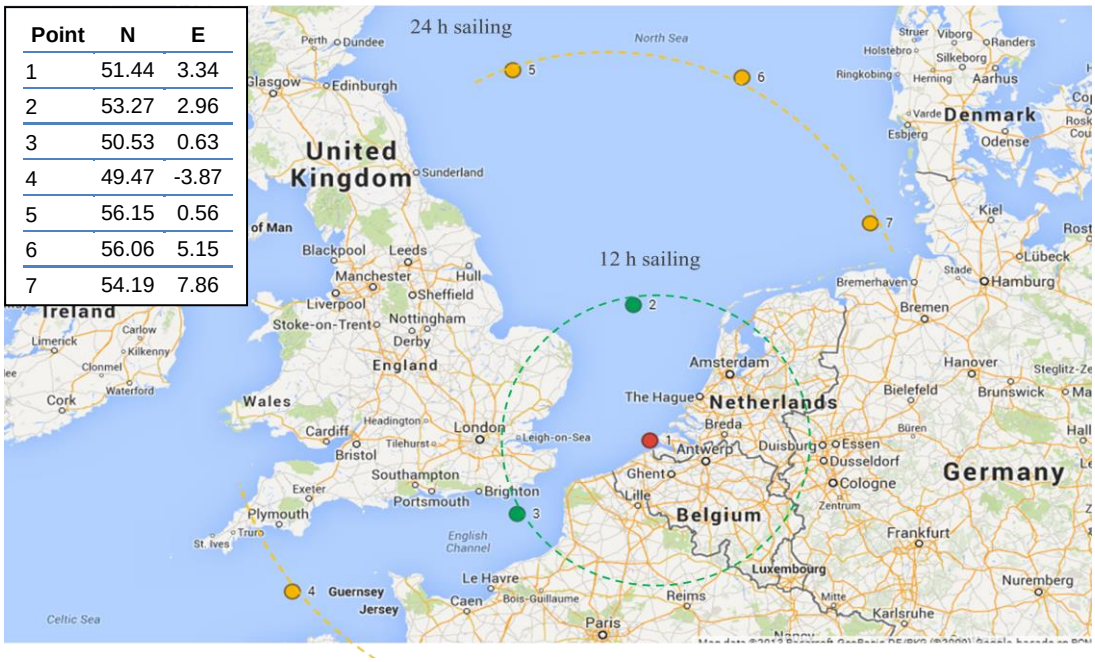

Figure 2. Selected points in the North Sea

4. The Cagliari case

Thanks to its position in the Mediterranean Sea (Figure 3), the port of Cagliari plays a major and strategic role as a trade hub. It lies just 11 miles from the ideal Gibraltar-Suez route and is thus one of the hubs for transhipment activities in the western Mediterranean.

Figure 3. Structure of the Cagliari port

The industrial port extends over an area of some 400,000 m², with a further 500,000 m² that could potentially be developed to meet growing traffic demand. It has an overall quay length of 1,500 m with five berths for container ships. Handling equipment includes: 7 quay cranes, 17 RTGs, 4 Reach Stackers, 8 Front Loaders, 28 Trucks and 26 Trailers.

From January to August 2013, the port of Cagliari handled 435,059 TEUs, an increase of 11.5% over the same period in 2012, without experiencing any significant congestion problems. The port's largest customer is Hapag Lloyd (Port of Cagliari Authority, 2014).

4.1 Outcome distribution

Considering the 1,969 statistical units that were collected, and setting a tolerance threshold of 15 minutes, only 30% of ships arrived at the expected time (i.e., within the interval (ETA-15, ETA+15)), the remaining 70% were delayed or arrived early. Table 2 shows the summary statistics of the outcome. The threshold is set at 15 minutes for operational reasons since a delay/advance of a quarter of an hour or less does not cause any disruptions in port.

Table 2. Delay summary statistics (in minutes)

Sample	Min	Q1	Mean	Median	Q3	Max	Standard deviation	Total arrivals
All vessels	-6,420	-41	-3	19	30	8,670	50.8	1,969

The histogram of delay distribution is shown in Figure 4 for the entire set of container vessel calls. The frequency distribution is unimodal and exhibits only one peak.

Preliminary investigations and frequent consultations with professionals revealed that the inconvenience created by the uncertainty surrounding arrivals at the container terminal of Cagliari is caused primarily by delays. As container traffic is not particularly heavy and the container terminal does not experience any significant congestion, ships arriving early that cannot be handled straight away due to unavailability of resources can wait until their assigned berthing space frees up without creating major difficulties. Nevertheless, it was decided to consider both late and early arrivals in order to obtain a more exhaustive analysis and to enable a comparison with the Antwerp case.

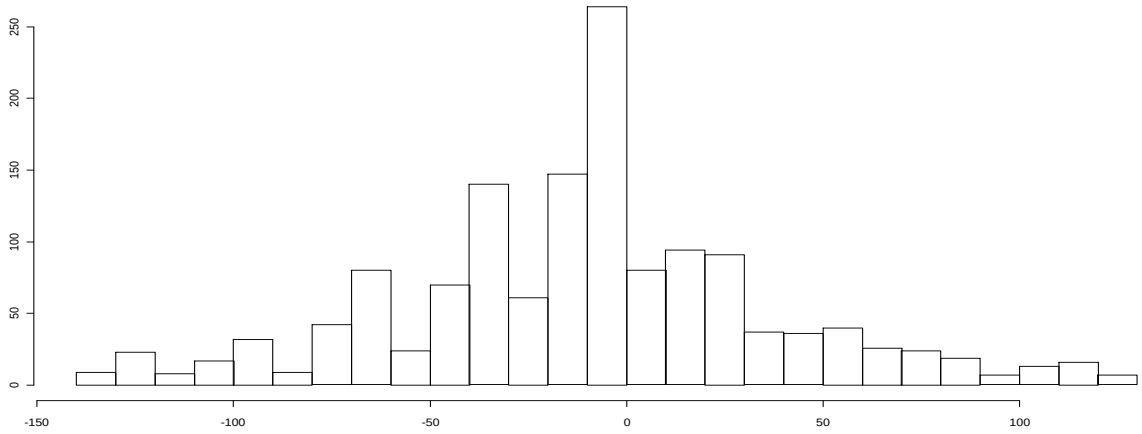

Figure 4. Delay distribution in the Cagliari Container Terminal

4.2 Data Mining

A first step of data pre-processing was required in order to identify the most important characteristics of the data and to transform data into the most appropriate form for use.

In particular:

- Missing values were deleted because less than 5% of the observations.
- Data were cross-checked in order to evaluate logical correspondence among variables.
- Eventual error types or illogical correspondence have been corrected.
- Outliers have been removed prior to the analysis due to their extra-ordinary behaviour and their potential misleading impact on performance assessment. Observations with extremely high or low values of delay were removed using the 1.5 rule:

$$\text{Delay} < Q_1 - 1.5 \times \left| Q_3 - Q_1 \right| \text{ or } \text{Delay} > Q_3 + 1.5 \left| Q_3 - Q_1 \right| \quad (4)$$

After outliers and missing data were removed, the final dataset included 1,625 observations.

The results of the machine learning models are shown in Table 3. The algorithmic models are described considering both the flexibility in representing the data and the interpretability of the results. Several models were built[5] using different subsets of all input variables. The prediction errors are calculated using 10-fold cross validation (k=10).

Considering the percentage of misclassified instances and the kappa statistic it is easy to conclude that the three methods substantially overlap and that they do not provide a good estimation of the binary outcome. There are more than 30% of misclassified instances and the evaluation of the kappa statistic ranges from 0.12 to 0.21. According to the scale proposed by Landis and Koch (1977), it ranges from zero (no better prediction than what occurs by chance) to one (perfect prediction). A value of 0.21 indicates a slight degree of agreement.

Table 3. Predictive performance for the discrete outcome

Algorithm	Misclassified instances	Kappa Statistic	Observed agreement	Expected agreement
Logistic Regression	32.4%	0.10	66.9%	62.12%
Classification Tree	31.7%	0.20	68.35%	63.85%
Random Forest	31.5%	0.21	65.87%	58.89%

[1] Models were run using R software (R Development Core Team) version 2.15.1 GUI 1.52 on a Leopard OS build 32-bit.

The *importance-plots* of the discrete Random Forest (Figure 5) model show the most discriminating variable on the y-axis, and their importance on the x-axis. The Gini coefficient is the measure of homogeneity that is used. The changes in Gini are summed for each variable and normalized at the end of the calculation. Variables that result in nodes with higher purity have a greater Gini coefficient decrease.

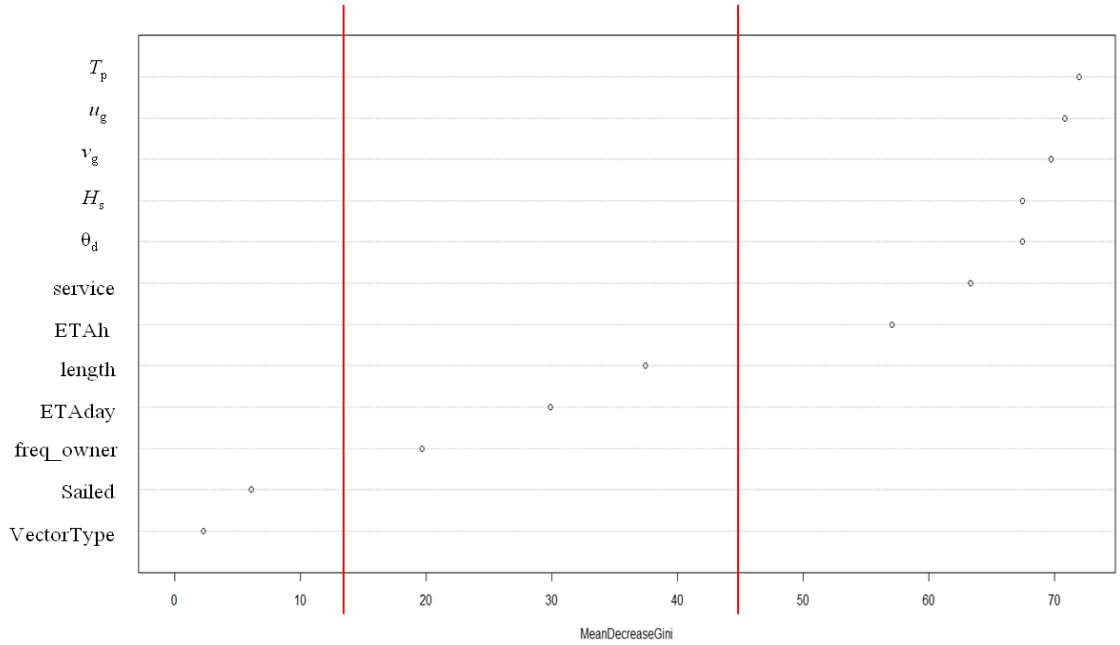

Figure 5. Importance of predictors for the Random Forest algorithm (Cagliari case)

The variables that are used as predictors can easily be grouped into different categories sorted from the most significant to the least significant. The relationship between the predictors and the delay/advance is interpreted below.

Weather/sea conditions - Figure 5 highlights how the variables representing weather conditions are the most important determinants of vessel arrival uncertainty. This result is extremely intuitive, in fact it is clear that the weather/sea conditions can strongly affect navigation times and hence, the arrival times. The best results are obtained by considering the weather-related variables at a distance of 12 hours from the port of Cagliari, most likely because this point, that lies in the middle of the route, is quite representative of the weather conditions along the whole route.

Service – The three variables that are related to the port service (port rotation, sailing direction and previous port) are all taken into consideration together in this variable. It can provide important information about the service performance and the organisation/occupancy at the previous port.

Length – Vessel length has been chosen as an indicator of the vessel's features because as compared to the other variables of the same group, it also provides important information concerning berth occupancy.

ETA hour and ETA day - This type of variable underlines the fact that the reliability of the ETA may depend on the moment in which it was sent.

Owner frequency - This variable indicates that the frequency with which a company serves a ports' rotation can affect the service offered by the terminals along the route.

Vector Type – This variable considers the different service contract terms for the mother and feeder vessels. As the cost of their stay in port is higher, mother ships usually have priority over feeder vessels, and therefore mother ships tend to arrive on time more often than feeder vessels.

Sailed - This type of variable substantiates the fact that once the ship has actually set sail for its port of destination, then the information becomes more reliable. Information notified prior to sailing from the previous port is less reliable because the extent of the delay may also include any inefficiencies of the previous port. On the contrary, if the information is sent after having left the port, any uncertainty will most likely depend on weather/sea conditions alone.

5. The Antwerp case

The port of Antwerp covers more than 13,000 ha of land, is located inland, and is connected to the North Sea by the River Scheldt, which is a tidal river.

The port is composed of eight main container terminals: six on the older right bank of the river Scheldt, and two on the newer left bank (Figure 6, Table 4).

Table 4. Main characteristics of the Antwerp container terminals

Terminal	Quay length (m)	Area (ha)	Quay cranes	Rail cranes	Barge cranes	Capacity (1,000 TEU)
PSA Deurganck	1,780	102	11	2		2,600
PSA Noordzee	1,125	79	8	1	1	2,000
PSA Europe	1,180	72	7	1	1	1,700
PSA-MSC Home	2,900	167	24	2	3	5,400
PSA-Churchill	2,260	84	3	3	-	-
DP World Antwerp Gateway	2,470	120	9	15	-	1,800

All the container terminals are multi-user terminals, although at the PSA-MSC Home Terminal, whose ownership is shared equally between PSA-Antwerp and MSC, the latter shipping company is the major user. Furthermore, the DP World Antwerp Gateway is also a joint venture whose shareholders include DP World (42.5%), Zim Ports (20%), Cosco Pacific (20%), Terminal Link/CMA-CGM (10%) and Duisport (7.5%).

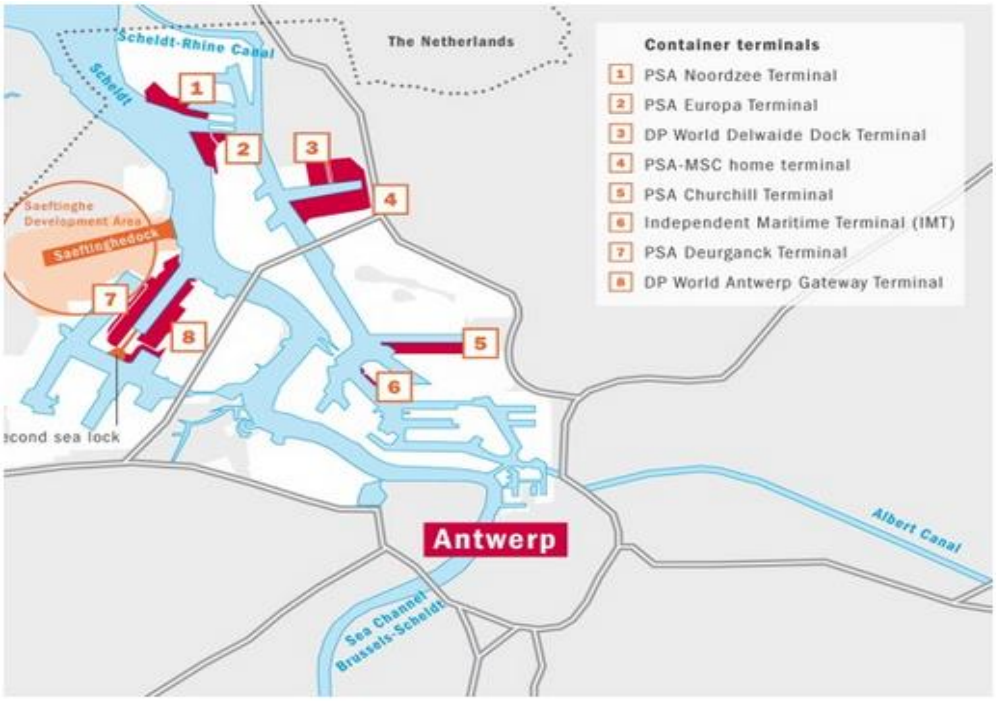

Figure 6. Structure of the Antwerp port

5.1 Outcome distribution

Histograms of delay distribution are shown in Figure 7 for the entire set of container vessel calls and by terminal[6]. The graphical visualisation of the histograms suggests that the delay distribution is bimodal both at the port and at the terminal level, but the proportion of "in advance" and "delayed" vessels differs across terminals (Table 5).

Considering the 10,611 statistical units that were collected, and setting a tolerance threshold of 15 minutes, the average delay is minus 78 minutes and the median delay is even less. However, the high standard deviation (Table 5) implies that position measures do not properly summarise the delay values.

By setting a tolerance threshold of 15 minutes, only 1.8% of ships arrived "on time" (i.e., within the (ETA-15, ETA+15) interval), while 42.9% arrived later than expected and the remaining 55.3% arrived earlier than expected.

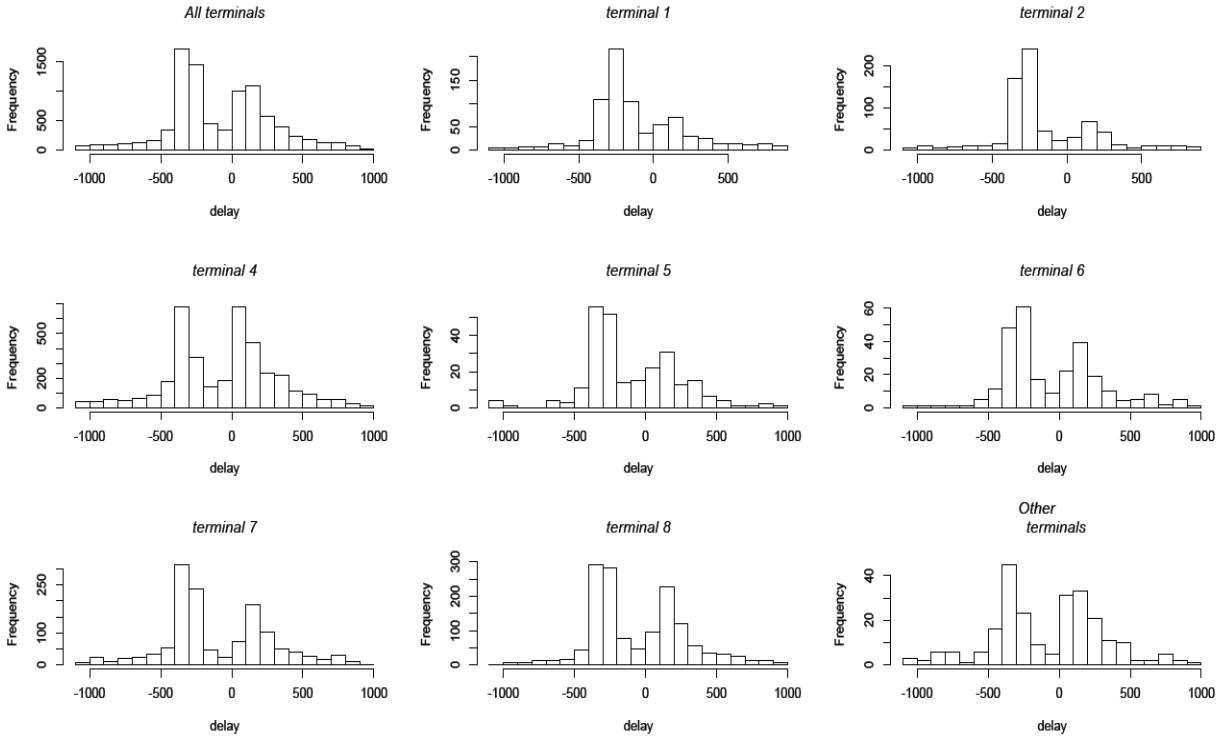

Figure 7. Delay distribution by terminal

Although the instrument is designed at the terminal level, a preliminary exploratory analysis at the port level was required. In a second stage, Terminal number 7 was chosen for two main reasons:

- Availability of data.
- Database size: the number of observations that were collected is very similar to the number of observations that were collected in the Cagliari container terminal.

Here again, it was possible to build a new database made up of 1,361 arrivals that was specific for Terminal 7.

[6] Terminal 3 handles a small fraction of total container ships and so it has been added to "other terminals", which comprises all the other terminals that are not specifically container terminals.

Table 5. Delay summary statistics by terminal (in minutes)

Terminal	Min	Q1	Mean	Median	Q3	Max	Standard deviation	Total arrivals	Proportion of delayed vessels
All Terminals	-1,045	-320	-78	-147	157	887	345.3	10,611	0.43
Terminal 1	-1,082	-292	-203	-110	85	871	323.2	772	0.31
Terminal 2	-1,051	-313	-267	-154	56	872	317.6	743	0.27
Terminal 4	-1,097	-330	20	-57	167	938	370.1	3813	0.51
Terminal 5	-1,081	-315	-211	-101	136	909	323.4	260	0.37
Terminal 6	-1,091	-300	-183	-56	169	919	339.6	273	0.42
Terminal 7	-1,090	-337	-243	-102	176	933	372.7	1,361	0.40
Terminal 8	-1,054	-313	-149	-59	180	924	338.6	1,442	0.44

5.2 Data mining

Even in this case, data pre-processing was conducted first in order to i) evaluate coherence of the information via cross checks ii) correct data problems iii) quantify missing values iv) identify outliers via the 1.5 rule. After outliers and missing data were removed, the final dataset included 9,857 observations at the port level and 1,298 at the terminal level. The results of the machine learning models are shown in Table 6 together with the corresponding estimate of the prediction error. The prediction error was calculated using 10-fold cross validation. All models were run using R software. The models with a good trade-off between goodness of fit and its interpretation and generalisation were chosen.

Table 6. Predictive performance for the discrete outcome

Algorithm	Level	Misclassified instances	Kappa Statistics	Observed agreement	Expected agreement
Logistic Regression	Terminal 7	22%	0.55	78.32%	49.50%
	Port	28%	0.45	70.81%	50.04%
Classification Tree	Terminal 7	22%	0.59	79.43%	49.76%
	Port	26%	0.57	80.45%	51.06%
Random Forest	Terminal 7	17%	0.63	80.20%	49.68%
	Port	16%	0.72	84.93%	50.20%

As Table 6 clearly shows, the discrete models perform very well for the Antwerp data. This is demonstrated not only by the value of the kappa statistic and by the percentage of misclassified cases, but also by the percentages of the observed and expected agreement. Random forest showed the best performance in both cases. Based on the evaluation of the kappa statistic, the predictive performance for the discrete outcome ranges from moderate (for logistic regression) to substantial (for Random Forest). The percentage of misclassified instances is around 16% at the port level and 17% at the terminal level. From a statistical point of view the result is noteworthy. In general, it is easy to see that all models generally performed better on the whole dataset than on the smallest subset of Terminal 7. This is because less information is available due to the limited size of the dataset.

The *importance-plots* of the Random Forest algorithm (Figure 8, Figure 9) show the different predictive power of the input variables. The plots clearly show that the variables do not have the same predictive power: in particular it seems that the variables capturing vessel "size" (TEU, grt, capacity and length) are very important determinants of delay/advance. These variables are indicators of the vessel's physical structure that can directly affect both the speed of handling operations in previous ports (cranes on board, position of bridge, crane intensity) and navigation (ability to withstand adverse weather and sea conditions). Other important discriminating variables at the port level are the ones that characterise the specific terminal. This result confirms the planner's opinion, according to which the terminal location and the presence of the lock strongly impact late/early arrivals. Even in the Antwerp case, the weather conditions are

important predictors. To conclude, the frequency with which a company serves a terminal affects the process, and the ETA hour and ETA day take into account possible time dependence in the reliability of the information.

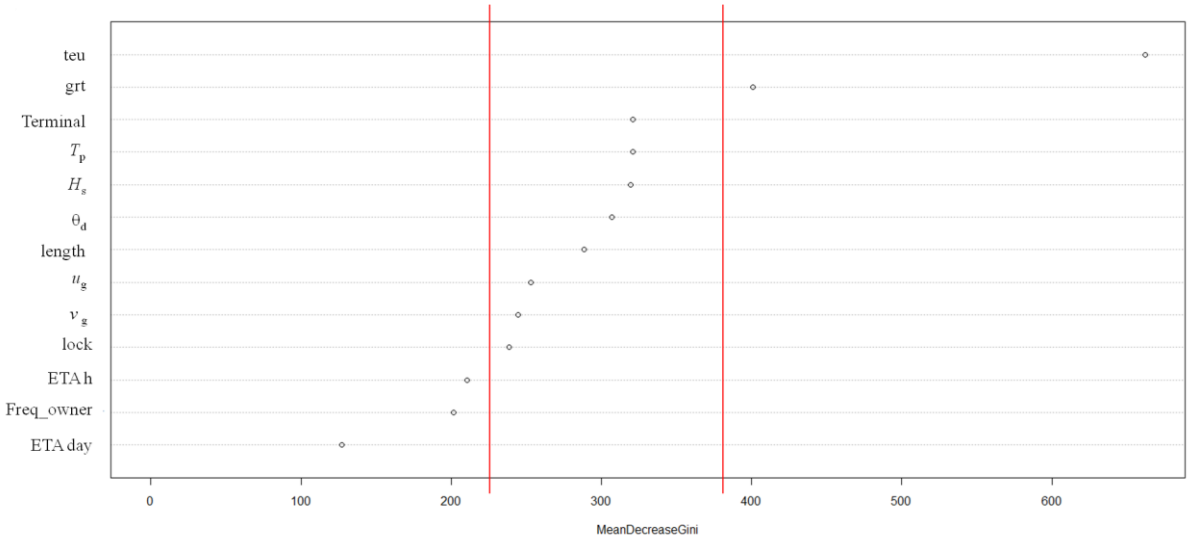

Figure 8. Importance of predictors for the Random Forest algorithm at the port level

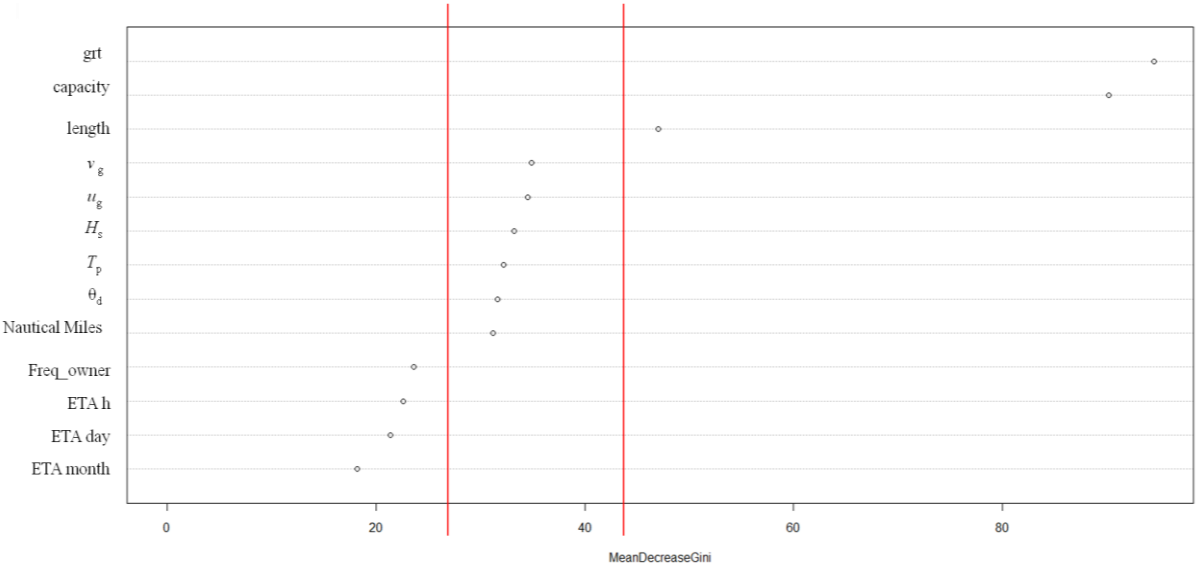

Figure 9. Importance of predictors for the Random Forest algorithm at the terminal level

6. Conclusions and future developments

The difficulties inherent to managing port operations due to vessel arrival uncertainty and to the complexity of the planning processes means that planners need to be assisted at each stage by tools that are able to support the decision-making process. Generally speaking, the last ETA is not respected by vessels due to unexpected events. In the short-term planning scenario this causes a series of inconveniences at the terminal level that are directly associated with the scheduling of the other terminal activities and with the resource allocation. Schedule unreliability also incurs additional inconveniences for the shipping lines and for the whole supply chain.

The lack of a reference model that describes the relationship between vessel arrival uncertainty and the involved variables led to the application of a specific machine learning approach based

on the concept of learning from historical data. In this study, a preliminary strategy is presented in order to help mitigate the consequences of late/early arrivals in port. The employed algorithms allow us to acquire a qualitative estimate of the delay/advance by knowing whether or not an incoming vessel is likely to arrive before or after the scheduled ETA. The case studies that were examined revealed that the ability of the discrete algorithms (which have a binary outcome) to capture bi-modality is noteworthy. Therefore, the forecasting accuracy of the models is lower when the distributional form of the delay shows only one peak. Moreover, thanks to graphical visualisation of the importance-plots, the most discriminating variables of the analysis have been highlighted.

The results that were obtained provide the basis for further research. In particular, from a research/policy perspective, this work falls within the framework of a broader project aimed at developing an instrument for terminal management that will allow policy makers to forecast the arrival time of each vessel through a continuous estimate in minutes. Knowing the possible deviation from the scheduled arrival time in advance would be important for planners in order to more efficiently allocate the manpower, equipment and spatial resources required to carry out handling operations. The main risk for planners is underestimating the resources. However, over-estimation within any given working period is also to be avoided since it would result in higher costs for the terminal. Thus, having this information could reduce operating costs, maximise terminal efficiency and hence competitiveness.

Acknowledgements

The authors would like to thank the following organizations for supporting the research by providing such a large amount of data: CICT (Cagliari International Container Terminal), ISPRA (Institute for Protection and Environmental Research), the Antwerp Port Authority and the PSA-Antwerp terminal.

References

Ambrosino, D. and Tanfani, E. (2012). An Integrated simulation and optimization approach for seaside terminal operations. Paper presented at the 26th*European Conference on Modelling and Simulation,* DOI: 10.7148/2012-0602-0609.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199-231.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Breiman, L., Friedman, J., Olshen and R., Stone, C. (1984). *Classification and Regression Trees.* Wadsworth Belmont, California.

Di Francesco, M., Fancello, G., Serra, P. and Zuddas, P. (2014). Optimal management of human resources in transshipment container ports, *Maritime Policy & Management*, DOI: 10.1080/03088839.2013.870355.

DP World Belgium. Maritime Terminals: http://www.dpworld.be. Accessed January 20, 2014.

Du, Y., Xu, Y., and Chen, Q. (2010).A feedback procedure for robust berth allocation with stochastic vessel delays. Paper presented at:*8th World Congress on Intelligent Control and Automation*, Jinan, China.

Dunham, M.H. (2003). *Data Mining. Introductory and Advanced Topics*. Prentice Hall.

Fancello, G., Pani, C., Pisano, M., Serra, P., Zuddas, P. and Fadda, P. (2011) Prediction of arrival times and human resources allocation for container terminal.*Maritime Economics & Logistics*, 13, 142-173.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework, *Proceedings of the 2nd International Conference in Knowledge Discovery and Data Mining (KDD-96)*, Portland, Oregon, August 1996, AAAI Press.

Gambardella, L. M., Rizzoli and A. E., Zaffalon, M. (1998). Simulation and planning of an intermodal container terminal.*Simulation*, *71*(2), 107-116.

Han, X. L., Lu, Z. Q. and Xi, L. F. (2010). A proactive approach for simultaneous berth and quay crane scheduling problem with stochastic arrival and handling time.*European Journal of Operational Research*, 207(3), 1327-1340.

Hand, D.J. and Yu K. (2001).Idiot's Bayes - not so stupid after all?*International Statistical Review,* 69, 385-399.

Hendriks, M., Laumanns, M., Lefeber, E. and Udding, J. T. (2010).Robust cyclic berth planning of container vessels.*OR spectrum*, *32*(3), 501-517.

Ku, L. P., Chew, E. P., Lee, L. H. and Tan, K. C. (2012). A novel approach to yard planning under vessel arrival uncertainty.*Flexible Services and Manufacturing Journal*, *24*(3), 274-293.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data.*Biometrics*, 33, 159-174.

Legato, P. and Monaco, M. F. (2004).Human resources management at a marine container terminal.*European Journal of Operational Research*, 156(3), 769-781.

Moorthy, R. and Teo, C. P. (2006). Berth management in container terminal: the template design problem. *OR spectrum*, 28(4), 495-518.

Notteboom, Theo E.(2006). The time factor in liner shipping services. *Maritime Economics & Logistics*, 8, 19-39.

Notteboom, T. and Rodrigue, J. P. (2008). Containerisation, box logistics and global supply chains: The integration of ports and liner shipping networks. *Maritime Economics & Logistics*,10(1), 152-174.

Pani, C, Fadda, P., Fancello, G., Frigau, L. and Mola, F. (2014). A data mining approach to forecast late Arrivals in a transhipment container terminal, *Transport*, 29, 175-184.

Port Authority of Cagliari. http://www.porto.cagliari.it. Accessed January 19, 2014.

PSA-Antwerp, Terminals. http://www.psa-antwerp.be. Accessed January 20, 2014.

Salido, M.A., Molins, M.R. and Barber, F. (2012). A decision support system for managing combinatorial problems in container terminals.*Knowledge-Based Systems*, 29, 63-74.

Sciomachen, A., Acciaro, M. and Liu, M. (2009).Operations research methods in maritime transport and freight logistics.*Maritime Economics & Logistics*, 11(1), 1-6.

Stahlbock, R. and Voß, S. (2008). Operations research at container terminals: a literature update. *OR Spectrum*, 30, 1-52.

Tongzon, J. and Heng, W. (2005). Port privatization, efficiency and competitiveness: Some empirical evidence from container ports (terminals). *Transportation Research Part A: Policy and Practice*, *39*(5), 405-424.

Vanelslander, T. (2005). The economics behind cooperation and competition in sea-portcontainer handling. *PhDthesis*, Faculty of Applied Economics, University of Antwerp.

Xu, N., Laskey, K. B., Chen, C. H., Williams, S. C. and Sherry, L. (2007). Bayesian network analysis of flight delays. Paper presented at the *Transportation Research Board 86th Annual Meeting,* Washington, DC.

Zhen, L., Lee, L.H. and Chew, E.P. (2011). A decision model for berth allocation under uncertainty.*European Journal of Operational Research*, 212, 54–68.

Zonglei, L., Jiandong, W. and Guansheng, Z. (2008).A new method to alarm large scale of flights delay based on machine learning. Paper presented at *International Symposium on Knowledge Acquisition and Modeling, 2008*.