TRANSPORT
ISSN 1648-4142 / eISSN 1648-3480
2014 Volume 29(2): 175–184
doi:10.3846/16484142.2014.930714

A DATA MINING APPROACH TO FORECAST LATE ARRIVALS IN A TRANSHIPMENT CONTAINER TERMINAL

Claudia Pani[1], Paolo Fadda[1], Gianfranco Fancello[1], Luca Frigau[2], Francesco Mola[3]

[1]*Dept of Civil, Environmental and Architectural Engineering, University of Cagliari, Italy*
[2]*Dept of Life and Environmental Sciences, University of Cagliari, Italy*
[3]*Dept of Economics and Business Sciences, University of Cagliari, Italy*

Submitted 5 April 2013; resubmitted 14 May 2013; accepted 20 August 2013

Abstract. One of the most important issues in Transhipment Container Terminal (TCT) management is to have fairly reliable and affordable predictions about vessel arrival. Terminal operators need to estimate the actual time of arrival in port in order to determine the daily demand for each work shift with greater accuracy. In this way, the resources required (human resources, equipment as well as spatial resources) can be allocated more efficiently. Despite contractual obligations to notify the Estimated Time of Arrival (ETA) 24 hours before arrival, ship operators often have to revise it due to unexpected events like weather conditions, delay in a previous port and so on. For planners the decision-making processes related to this topic can sometimes be so complex without the support of suitable methodological tools. Specific models should be adopted, in a daily planning scenario, to provide a useful support tool in TCTs. In this study, we discuss an exploratory analysis of the data affecting delays registered at a Mediterranean TCT. We present some preliminary results obtained using data mining techniques and propose a Classification and Regression Trees (CART) model to reduce the range of uncertainty of ship arrivals in port. This approach is compulsory to manage vast amounts of unstructured data involved in estimating of vessel arrivals.

Keywords: container terminal; vessels delay; data mining; CART; decision trees.

Reference to this paper should be made as follows: Pani, C.; Fadda, P.; Fancello, G.; Frigau, L.; Mola, F. 2014. A data mining approach to forecast late arrivals in a transhipment container terminal, *Transport* 29(2): 175–184. http://dx.doi.org/10.3846/16484142.2014.930714

Introduction

The vessel arrival uncertainty problem, characteristic of all logistics systems, is a constraint to planning effectiveness in container terminals where the decision-making processes need to be constantly adapted and updated. To estimate with good approximation, the effective time of ship arrival in port is decisive in improving terminal organization, especially in the short term.

On a daily basis, the information regarding ships' loads to be handled is often known while the actual time of arrival remains uncertain. Despite the contractual obligations to send the Estimated Time of Arrival (ETA) 24 hours in advance the arrival, ports are often forced to make last minute changes due to unexpected events (weather and sea conditions, delays in previous ports, etc.).

Reliable estimation of delay would make it possible to determine the actual time of arrival of ships in port with greater accuracy, hence the workload for each work shift. This would facilitate allocation of the resources (human, mechanical and spatial) required to satisfy the expected demand, which are often overestimated at the planning stage. Up to now, this task has been delegated to the planners, professionals who operate chiefly on the basis of hands-on experience. The decision-making processes involved are often so complex as to be unmanageable without the support of adequate methodological tools. The problem solution approach is extremely complex, considering the large number of variables and constraints influencing the process, in particular:

- vessel structure (length, draft, gross tonnage, capacity, etc.);
- service (sailing direction, port rotation, etc.);
- loading plan and type of containers;
- organization/availability of previous port;
- external factors (weather/sea conditions, strikes, etc.);
- human resources management (contractual obligations, labour regulations, etc.);

Corresponding author: Claudia Pani
E-mail: *claudia.pani@unica.it*

Taylor & Francis
Taylor & Francis Group

– equipment management (repairs, out of service for maintenance, etc.);
– space management (berth space and relative distance from container stacked).

These key issues form the basis of the present research line. The main objective being to identify those scientific approaches that best lend themselves to addressing the problem, along with suitable tools for providing an analytical answer to the problem.

The need to handle the large number of unstructured data involved in estimating ship arrival times calls for the application of specific data mining techniques. In particular, in this paper we propose a Classification and Regression Trees (CART) model. As known, this is a non-parametric method, that abandoning any assumption about data distribution shape, is based on the self-learning concept, such that experience forms the knowledge base for generating and calibrating predictions.

Using this methodology it has been possible to develop and implement a specific algorithm for forecasting ship arrival times providing an important decision support tool for port operators.

The model has been tested in the Transhipment Container Terminal (TCT) of Cagliari (we are grateful to the people for their availability and valuable research support). This kind of tool is essential for enabling a rapid re-planning of the resources required to meet the expected demand, following the occurrence of unexpected events.

1. Literature Review

The analysis of scientific literature confirmed that numerous complex problems of different nature coexist in a container terminal system, most of which need integrated solutions. Thus, the solution to one problem often becomes decisive for the other problems related thereto (Murty *et al.* 2005; Salido *et al.* 2012; Won, Kim 2009).

Vis and De Koster (2003), Stahlbock and Voß (2008) provide an interesting overview of the classification of decision problems in a container terminal on the basis of the five main logistic processes: arrival of the vessel, loading/unloading, moving the containers from quayside to yard and vice versa, stacking containers in the yard and transport of containers outside the terminal with other vehicles. Therefore, the whole process starts with the vessel arrival. Providing an analytical solution to the uncertainty of ship arrivals is thus essential for improving availability and functionality of the handling systems as a whole. In the event of a ship's delay, its berth space has to be re-allocated and the containers, that are already stacked in the yard on the basis of the original space assigned, have to be re-located as quickly as possible to minimize the berthing time (Berth Allocation Problem) (Zhen *et al.* 2011; Salido *et al.* 2012). Moreover, to optimize resources (personnel and equipment) management for handling operations and for establishing maintenance schedules it is important to know the effective arrival time of vessels (Fancello *et al.* 2011; Gambardella *et al.* 1998).

These aspects are essential for ensuring the availability and perfect functionality of the handling systems used and for avoiding under-manning or equipment being out of service.

The diagram (Fig. 1) shows the main planning and scheduling problems in container terminals (Salido *et al.* 2012). It highlights the central role of the arrival of vessels.

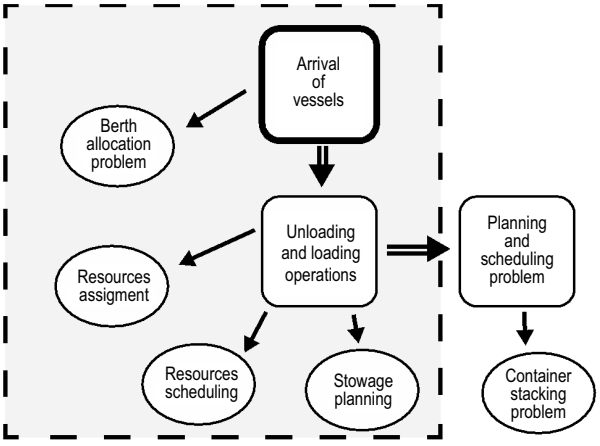

Fig. 1. Main planning and scheduling problems in container terminals

The state-of-the-art study showed that notwithstanding unprecedented technological innovation, the uncertainty and variation in daily demand forecasting still remain a challenge for port operators. Moreover, the specific applications are strongly limited. The most appropriate contributions concern container flow prediction in and out of container terminals over a daily time horizon. Here again, the problem is addressed along with the other spheres of the terminal's activities.

The Hong Kong container terminal is equipped with information systems that indicate in real time container flows and the resources necessary for their handling. This information, in addition to being essential for the safety of the terminal, provides reliable input data for correctly scheduling handling operations (Murty *et al.* 2005; Fung 2002). Sideris *et al.* (2002) have developed a tool, using on-line data, for predicting daily demand variations in terms of the number of containers moved through a terminal. The purpose of being able to predict these variations is to improve allocation of the handling equipment and manpower required, so as to improve work scheduling. Lastly, Gambardella *et al.* (1996) proposed a forecasting module for estimating the daily container flow in and out of a terminal, combining two different estimators. The first predicts the number of containers to be loaded onto a ship due to arrive in port, based on past data. The second calculates the percentage of the total number of containers that should be transported by truck to the terminal, as a function of the ship's ETA. In particular, the only model capable of predicting ship arrival times has been calibrated by Fancello *et al.* (2011). The decision support system presented here reduces the interval of uncertainty on ar-

rival time in port to around 6 hours, employing a neural network model. By so doing, the terminal in question is able to plan resources around just two work shifts, instead of 3 or 4.

From the literature review conducted, it has also emerged that developing advanced vessel arrival time prediction tools for a transhipment container terminals is closely related with the problem of managing and elaborating large amount of data.

The increasing complexity of the data to be processed and the need to conduct ever more sophisticated analyses has made essential to develop specific data mining techniques (Dunham 2002). On the other hand, a blind application of data mining would be detrimental as it could lead to discovering meaningless patterns (Fayyad *et al.* 1996). To extract useful information therefore it requires referring to data mining within the Knowledge Discovery in Databases (KDD) process, defined as the extraction of useful and not known information from data (Frawley *et al.* 1992; Fayyad *et al.* 1996). The other steps in the KDD process, like data preparation, data selection, data cleaning, and correct interpretation of the results, are mandatory to be able to extract information from data (Fayyad *et al.* 1996).

Numerous data mining techniques can be used for predictive purposes.

Analysis of the scientific literature showed CART decision trees to outperform Neural Networks (NNs) for this specific case. CART patterns can be applied to larger data problems and are able to handle smaller data sets than NNs (Markham *et al.* 2000). Moreover, CART performs better than NNs models when data sets are smaller with large numbers of irrelevant attributes (Brown *et al.* 1993). Decision trees are used either as prediction tools or as exploratory tools. They aim to detect to which class of a response variable belong data records, knowing the values or the categories of one or more explanatory variables. The recursive algorithm splits data applying a depth-first approach (Hunt *et al.* 1966) or a breadth-first approach (Shafer *et al.* 1996) until all records are classified. At each step data are split using impurity measures (Quinlan 1992). The decision tree structure consists of a root, no terminal and terminal nodes (leaves). The model obtained enables one to classify new unknown records. The decision tree algorithm consists of two main tasks: tree growing and tree pruning. Tree growing follows a top-down approach. Here the data set is recursively partitioned until all records belong to the same class label (Hunt *et al.* 1966). On the other hand, tree pruning follows a bottom-up approach. In this phase the algorithm minimizes over-fitting improving prediction accuracy (Mehta *et al.* 1996).

A multitude of decision tree models have been developed since the 1960s. The first to appear was the Automatic Interaction Detection (AID), in which the outcome variable is quantitative (Morgan, Sonquist 1963). Several other algorithms followed such as Exploration of Links and Interaction through Segmentation of an Experimental Ensemble (ELISEE) (Cellard *et al.* 1967) and THeta AID (THAID) (Morgan, Messenger 1973) for categorical response variable, and MAID (Gillo 1972) for quantitative response variable. Numerous algorithms were later developed such as CHi-square Automatic Interaction Detection (CHAID) (Kass 1980), CART (Breiman *et al.* 1984), ID3 (Quinlan 1986) and C4.5 (Quinlan 1992). Some authors have proposed variations to the CART method that develop non-binary trees (Loh, Vanichsetakul 1988) or that reduce computation time (Mola, Siciliano 1997). Of these decision trees the most relevant statistical contribution was provided by the CART method, because it distinguishes between a classification tree in which the response variable is categorical and regression tree in which the response variable is quantitative.

Over the last few years decision trees algorithms have been improved and new models developed embodying this approach. Many hybrid approaches have also been developed. Conversano (2002) proposed the Generalized Additive Multi-Mixture Models (GAMMM) using the decision trees approach for regression smoothing. Other authors have pursued the same path, for example Chan and Loh (2004), Su *et al.* (2004), Choi *et al.* (2005) and Hothorn *et al.* (2006). To improve the accuracy of traditional decision tree methods, they have been combined to produce, for example, the tree averaging approach. Another approach is the Ensemble methods: Freund and Schapire (1996) introduced an Ensemble method called Adaptive Boosting, while Breiman (1996) developed the Bootstrap Aggregating, and Random Forest (Breiman 2001).

2. How TCT of Cagliari Works

This investigation has concerned the TCT of Cagliari. A six-month period of observation at the terminal was required for investigating the phenomenon and for building the database, essential aspects for the choice of methodology adopted. During this period terminal operations were closely observed, making it possible to:

- observe the main causes of delays in ship arrivals in real conditions;
- analyse the main critical aspects and frequent operational issues in a container terminal associated with late arrivals;
- analyse impacts of ship delays on the other terminal areas (space, human resources and equipment planning, maintenance scheduling);
- analyse the dynamic response of the terminal, in terms of supply of port services, in reacting to the delay.

The work day in the terminal is divided up into four 6-hour shifts. Resources scheduling is performed at two main sublevels that differ in the type of decision and time horizon. The purpose of the first level is to plan handling operations and activities in the different sectors (ship, quayside and yard) over an annual time horizon. The second level, object of the present paper, addresses the specific allocation of resources (personnel, equipment and space) for maximizing productivity and minimizing costs. In this case the time horizon is about 24 hours.

Both levels are characterized by fragmentation of time and information uncertainty: the information arrive at different and undefined times and are repeatedly updated, hence their content is uncertain.

Furthermore, the planning processes were observed to depend strongly on the flow of incoming information, especially over short time horizons, hence the dynamicity of the process.

The database includes all arrivals at the container terminal over a period of 12 months (year 2010). Considering the 779 statistical units collected (corresponding to arrivals of both mother and feeder ships) only 29% of ships actually arrived at the time indicated, the remaining 71% were delayed or arrived early. The time series of arrivals, divided into mother and feeder ships, is shown below (Fig. 2).

As can be seen in Fig. 2, the time series of arrivals at the TCT of Cagliari is complex and irregular, especially for the feeder vessels, which appear to be more prone to early or late arrivals.

From a preliminary descriptive analysis other important aspects came out about the terminal operations over the period examined. Fig. 3 shows the distribution of vessel arrivals, respectively, during the hours of the day, the days of the week and the months of 2010. It is possible to make interesting considerations (Fig. 3):

– 64% of vessel arrive during the first two work shifts (from 1:00 to 13:00);
– 37% of arrivals are at weekends;
– ship arrivals are regularly distributed over the 12 months.

The data analysis also gave other important information, in particular the average vessel turnaround time around 21 hour. Moreover:

– the average waiting time for a berth is around 2 hours;
– the average waiting time in berth before loading/unloading operations is around 2 hours and 30 minutes;
– once loading/unloading has been completed, the average time before unberthing is 1 hour and 50 minutes.

It is an essential aim for the terminal to minimize mentioned times.

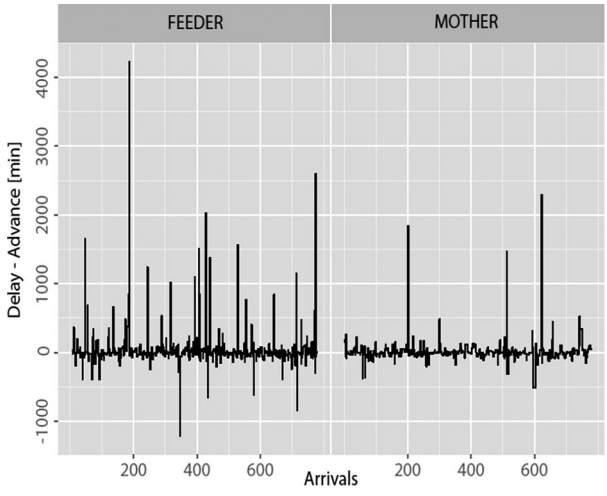

Fig. 2. Time series of arrivals at the TCT of Cagliary

3. Methodology

3.1. Decision Trees Overview

Decision trees are considered as powerful tools for extracting meaningful patterns from data sets with records characterized by a dependent variable and a set of explanatory variables (Hastie *et al.* 2013). These trees aim to classify unknown records using the pattern obtained. The algorithm is very simple. It recursively splits the feature space (usually binary splits) into several regions using explanatory variables and split-points to obtain the best fit, until a stopping rule terminates the process. Suppose, for graphical reasons, that we have just two explanatory variables, X_1 and X_2 as in Fig. 4. The first step consists in splitting the feature space at $X_1 = a_1$. Then the

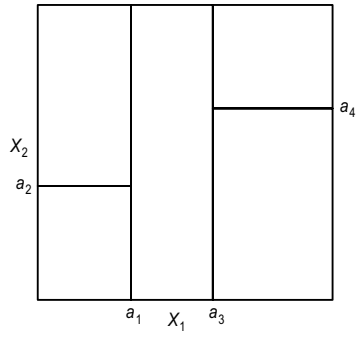

Fig. 4. A two-dimensional feature space partitioned by recursive binary splitting

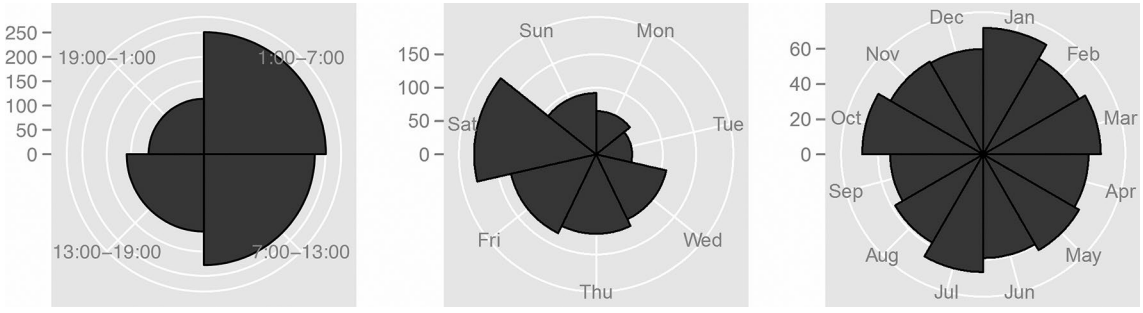

Fig. 3. Distribution of arrivals

algorithm splits region $X_1 \le a_1$ at $X_2 = a_2$ and $X_1 > a_1$ at $X_1 = a_3$. Finally, the region $X_1 > a_3$ at $X_2 = a_4$ until five regions are generated. The algorithm assigns a specific value or label to each region.

The algorithm works in the same way when there are more than two explanatory variables. Among all decision tree algorithms – CART (Breiman *et al.* 1984) is considered a landmark. Using this method it is possible to distinguish between regression trees and classification trees. Regression trees when the response variable is numerical, or classification trees when it is categorical. The algorithm is explained below. All formulas are from Hastie *et al.* (2013).

3.2. CART: Regression Trees

The regression tree algorithm involves two phases: tree growing and tree pruning. In the first phase a tree is built. The aim of the second phase is to reduce tree size in order to be able to apply the recognized patterns to other data, as large trees give unsatisfactory results when applied to new data. Moreover, an oversized tree contains a large number of terminal nodes, making its interpretation difficult and falling in risk of the over-fitting. Normally data are divided into two subsets: a training set (with two-thirds of the data) and a test set (with the other third). The training set is used for tree growing, while the test set is used for tree pruning to select the optimal tree. Nowadays, due to the fact that it is possible to use more powerful computers and, the computation time is strongly reduced, a γ-fold cross-validation approach is considered.

Tree Growing. Suppose that we have a data set with p explanatory variables X, one dependent variable Y and N records:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,p} \end{bmatrix};$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}. \quad (1)$$

Let us consider the training set. The algorithm splits the feature space into several regions, for instance into M regions, selecting the explanatory variables and split-points automatically. In each region the algorithm models the dependent variable as:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m). \quad (2)$$

To decide the value of c_m it minimizes the sum of squared deviation between y_i and $f(x_i)$, obtaining c_m equal to average of y_i in region m:

$$\min_{f(x_i)} \sum \left(y_i - f(x_i)\right)^2 \Rightarrow \hat{c}_m = \text{ave}\left(y_i | x_i \in R_m\right). \quad (3)$$

In tree growing the regression trees algorithm makes a series of local decisions about which predictor variable j and split-point s to use in binary recursive partitioning in order to create homogeneous regions. With the first split we have:

$$R_1(j,s) = \left\{X | X_j \le s\right\} \quad (4)$$

and

$$R_2(j,s) = \left\{X | X_j > s\right\} \quad (5)$$

to know the optimal j variable and s split-point it suffices to solve:

$$\min_{j,s}\left(\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2\right). \quad (6)$$

Once the algorithm has found the best j variable and s split-point, it repeats the previous step dividing each region into two sub-regions until a stopping rule terminates the process.

Tree Pruning. As a Tree T_0 has been built, it needs to be trimmed in order to improve its interpretability and in order to avoid over-fitting. The obtained classification rule can now be applied to new data. CART uses a specific tree pruning method known as cost-complexity pruning. Let, $\acute{a} \in [0,\infty)$, called complexity parameter, expressing trade-off between tree size and goodness of fit. Indicate T as a subtree of T_0 obtained by pruning, such that $T \subseteq T_0$ and $|T|$ is the number of terminal nodes in T. Letting:

$$N_m = \#\left\{x_i \in R_m\right\}; \quad (7)$$

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i; \quad (8)$$

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2. \quad (9)$$

Cost-complexity pruning is defined as follows:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|. \quad (10)$$

The idea is to find a $T_\alpha \subseteq T$ minimizing $C_\alpha(T)$. To find T_α the weakest link pruning approach is applied. This approach is developed by successively collapsing the internal node that produces the smallest per-node increase in $\sum_m N_m Q_m(T)$ until the single node (root) is obtained. In this way a finite sequence of subtrees has been generated containing the optimal subtree T_α (Breiman *et al.* 1984).

If the data only contain a few records, it is possible to use γ-fold cross-validation. The training set is split into γ parts each of size $\frac{N}{\gamma}$. A tree is grown γ times, each one with a different training set consisting of a γ combinations of γ−1 original parts. The generalization error is calculated averaging over each γ tree generated. Finally, α is chosen minimizing the cross-validated $C_\alpha(T)$.

4. Discussion

This study, based on the KDD approach (all statistical analyses were performed with the aid of the software *R*), consists of six main tasks (Fayyad 1996):

1) understanding the application domain;
2) data selection;
3) data preparation;
4) data mining;
5) interpretation of results;
6) consolidation of the discovered knowledge.

4.1. Understanding the Application Domain and Data Selection

A specific theoretical study was carried out for the first two tasks. In order to forecast container ship delays the variables that were able to provide a theoretical explanation for the delay were examined. Once these had been identified, then data collection could start. The database contains information about mother and feeder vessels arriving at the container terminal in 2010. It consisted of 779 records and 44 variables.

The variables were divided into two groups: the first composed of 11 variables having a direct influence on late/early arrivals, the second containing 33 variables that does not affect or does not have a direct influence on delay. These variables will be used for a post-analysis, an essential step for examining the economic/organizational fallout caused by late/early arrivals in the system as a whole.

The first group of variables comprises: ETA at pilot point (the pilot point is a conventional imaginary point into the port used to indicate when vessels arrived in port) – 24 hours, actual time of arrival at pilot point, length, gross tonnage, capacity, vessel type, previous port, shipping line, service, sailing direction, average speed. The second group includes variables such as characteristics and number of containers to be unloaded, characteristics and number of containers to be loaded, date and time of berthing, of start operations, of end operations, of unberthing. In the present study only the first group of variables was examined.

Furthermore, preliminary investigations and frequent consultations with professionals revealed that the inconvenience created by the uncertainty surrounding arrivals at the TCT of Cagliari is caused mainly 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 without creating major difficulties. For this reason, in this work, only delays were analysed.

4.2. Data Preparation

Data preparation consisted of several steps:

- missing values and outliers were deleted as standard statistical procedure;
- new variables were created:
  - *delay* [minutes]: calculated as the difference between the actual time of arrival at the pilot point and the ETA. As this variable was only intended to express delay, it was set at zero for early arrivals;
  - *previous port distance* [nautical miles]: calculated as the distance of the previous port from the TCT of Cagliari;

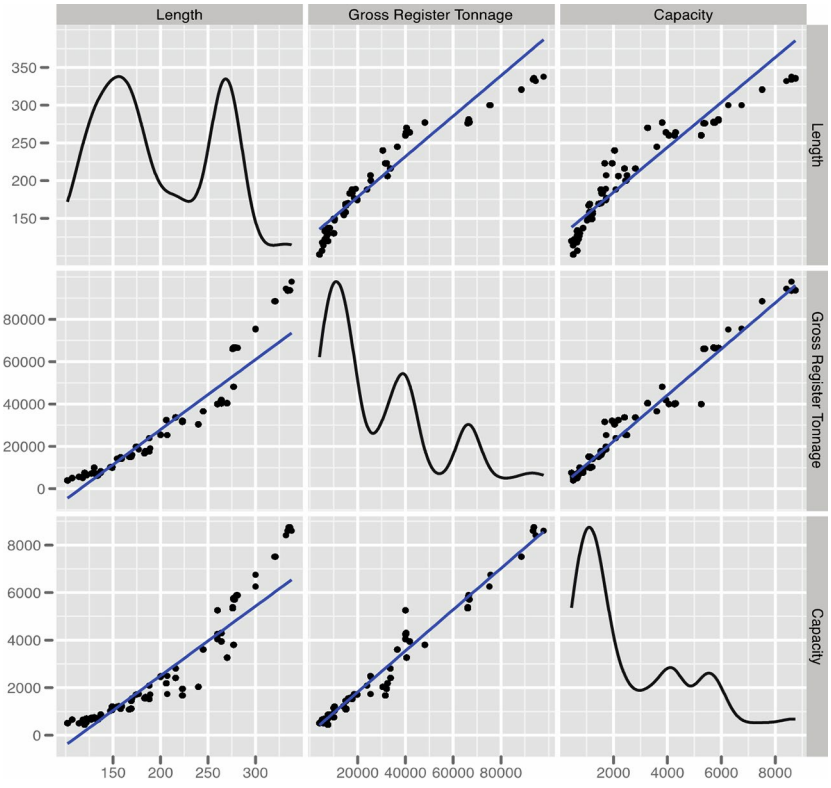

Fig. 5. Correlation matrix, Length–GRT–Capacity

◦ *sailing*: divided into two classes sailed and not sailed. It indicates if vessel notified the ETA once it had left the previous port or while it was still in port.

– the redundant variables, variables strongly correlated with others that do not provide any additional information on the phenomenon being studied, were eliminated. Length, Gross Register Tonnage and Capacity were found to be strongly correlated (Fig. 5). The correlation coefficients are 0.94 (GRT–Length), 0.93 (Capacity–Length) and 0.97 (Capacity–GRT). To remove redundant information it was necessary to include one variable only in the analysis. Length was chosen as it also expresses berth occupancy;
– a statistical exploratory analysis was performed to identify the most important characteristics of the data;
– the data were cross checked (e.g. previous Port with service).

As a result of the analysis, the dimensions of the database changed: records being reduced to 752 and variables of the first group increasing to 12.

4.3. Data Mining

The CART method has been considered. Several regression trees were built using different combinations of all twelve variables. The tree with a good trade-off between goodness of fit and its interpretation and generalization to new data was chosen (Fig. 6).

Length, ETA Day, ETA Shift, Vector Type and Sailing were used as predictor variables in the tree.

The response variable is Delay. A brief description of them is given below.

Delay [minutes] was calculated as the difference between the ETA notified and the actual time of arrival at the port.

Length [metres] is an indicator of the vessel's physical structure that directly affects both the speed of handling operations in previous ports (cranes on board, position of bridge) and navigation (speed, ability to withstand adverse weather and sea conditions). It also provides important information concerning berth occupancy.

ETAs, indicated by two variables ETA Day and ETA Shift, constitute the expected date of arrival. This information is updated several times as the date approaches, in general monthly, weekly, 48 hours and 24 hours in advance. This study concerns the last ETA 24 hours before the arrival.

Vector Type (mother or feeder) considers not only the objective physical/structural differences between the two types of vessels but also the different services provided by the container terminal to cope with the delay.

Sailing (Ship departed/not departed) provides important information about the position of the ship over the last 24 hours. It is calculated on the basis of transit time from the previous port. A transit time of more than 24 hours means that the ship notified the ETA once it had left the port. In this case any delay will most likely depend on weather/sea conditions alone.

For a transit time of less than 24 hours the ETA is notified while the ship is still in port. In this case, the delay could be caused not only by weather or sea conditions but also by the organization/ occupancy of the previous port. The transit time was calculated as the ratio between the distance in nautical miles of the last port of call from the port of Cagliari and the average speed, in knots, of the vessel.

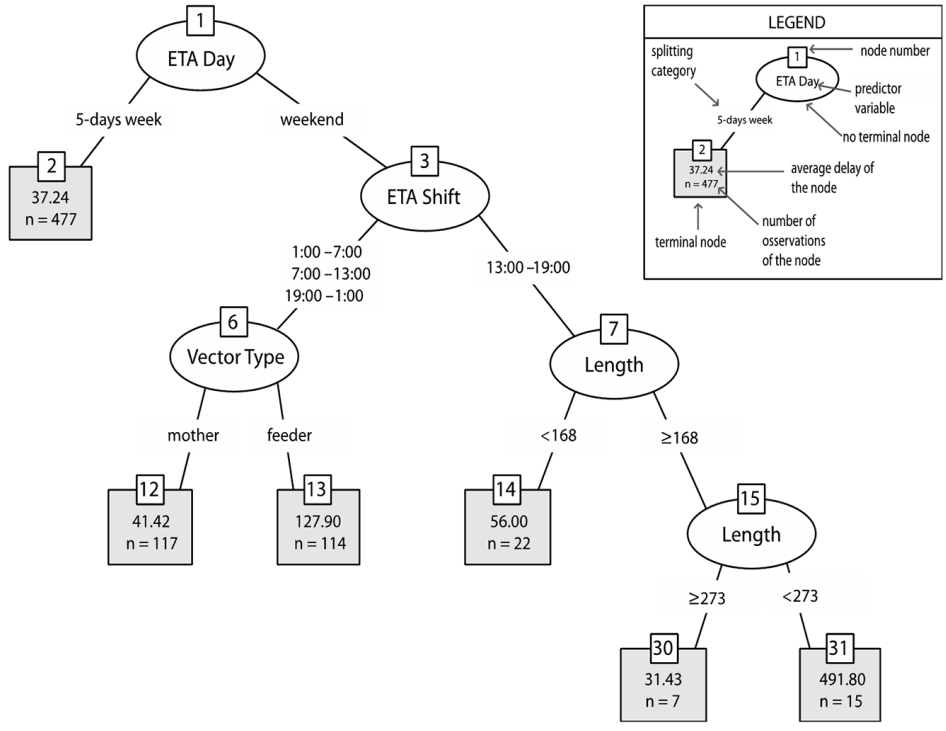

Fig. 6. Regression tree

4.4. Interpretation of Results

The regression tree is composed of 11 nodes, six of which are terminal nodes (Table). These are described below. Node 2, with an average delay of 37.24 minutes, comprises 477 observations: 57% concerns the feeder vessels, the remaining 43% the mother vessels. The discriminating variable is the ETA Day: all vessels arrive in the first 5 days of the week.

A clear distinction emerges between ships arriving at the weekend, more likely to be delayed, and those arriving on weekdays, more likely to arrive on time. This result can be interpreted considering that most ships that estimate to arrive in the port of Cagliari on Saturday or Sunday depart from the previous port on Friday, Saturday or Sunday when, as a general rule, there is often lower functionality.

Node 12, with an average delay of 41.42 minutes comprises 117 arrivals of mother ships. These vessels arrive at TCT of Cagliari at weekends and, the most, from ports at more than 24 hours sailing time away. Because they are already en route when the last ETA is notified, eventual delays cannot be attributed to inefficiencies in the previous port.

Node 13, with an average delay of 127.90 minutes, comprises 114 feeder vessels arriving in port at weekends. The majority of these were still in port when the ETA was notified.

Terminal nodes 12 and 13 are created by partitioning node 6 on the basis of vector type. The average delays for the two nodes suggest that mother ships have a greater tendency to arrive on time than feeder vessels. This result is confirmed by the fact that, in practice, the service contract terms for the two types of vessels differ. As the cost of their stay in port is higher, mother ships usually have priority over feeder vessels.

Moreover, the analysis of the two terminal nodes revealed that though the variable sailed/not sailed is not one of the discriminating variables, it contribute to differentiate between them. This type of variable substantiates the fact that information notified prior to sailing from the previous port is not very reliable because the extent of the delay also includes any efficiencies of the port.

Node 14, which has an average delay of 56 minutes, includes 22 feeder vessels arriving in the port at the weekend. Apart from the ETA Day and Vector Type, the other two discriminating variables are Length and ETA Shift. The ships belonging to this node in fact are over 135 metres long, and therefore tend to be shorter than the feeder ships belonging to node 13, which have an average length of 150 metres. For the feeder ships, which generally notify the ETA before sailing from the previous port, delays increase with length. This can be explained by the fact that longer ships are more difficult to process in terms of availability of port services (number of cranes, personnel, etc.) and, also, they are more difficult to berth as they occupy more space.

Moreover, all the ships belonging to node 14 arrive during the third work shift (13:00–19:00). Considering that the ports of origin (Malta, Barcelona, Naples, Catania and Genoa) are on average 17 hours away, the vessels leave the last port of call at around 23:00 and sail overnight, meaning that the ship will be processed in daylight. Therefore this variable provides an indication of processing times in the previous port, highlighting variations in performance for the different shifts: generally port operators working at night shifts experience greater mental and physical fatigue.

Node 30 is characterized by an average delay of 31.43 minutes. Similarly to node 12 it includes mother ships arriving at TCT of Cagliari at weekends with a transit time of more than 24 hours from the port of origin. Thus, as these vessels had already set sail when the ETA was notified, eventual delays cannot be attributed to inefficiencies at the previous port. The only difference between the two nodes lies in ship length. In this case however, as mother ships take priority over feeder ships this variable is not particularly discriminating. Vessels belonging to nodes 12 and 30 tend to arrive on time, differences in delays being insignificant.

Comparison of the terminal nodes 14 and 30 that originate from the same node 7, also confirms that mother ships are more likely to arrive on time than feeder ships.

Lastly, node 31 comprises 15 mother and feeder vessels arriving at the terminal at the weekend. Average delay in this case is 491.8 minutes. Almost all the vessels in question have already set sail when the ETA was notified. This substantial delay is not completely explained by the model and it is most probably due to exogenous variables (weather variables, breakdown or navigation problems).

Table. Final nodes characteristics

Nodes	Delay [minutes]	Length [metres]	Vector type		ETA day		ETA shift				Sailing	
			feeder	mother	5-days week	weekend	1:00–7:00	7:00–13:00	13:00–19:00	19:00–1:00	yes	no
Node 2	37.24	196.9	57%	43%	100%	0%	33%	26%	24%	17%	61%	39%
Node 12	41.42	261.1	0%	100%	0%	100%	43%	49%	0%	8%	93%	7%
Node 13	127.90	148.2	100%	0%	0%	100%	38%	40%	0%%	22%	32%	68%
Node 14	56.00	135.0	100%	0%	0%	100%	0%	0%	100%	0%	14%	86%
Node 30	31.43	292.3	0%	100%	0%	100%	0%	0%	100%	0%	100%	0%
Node 31	481.80	209.0	47%	53%	0%	100%	0%	0%	100%	0%	80%	20%

4.5. Consolidation of Knowledge Found

CART methodology has been used to classify new data through obtained tree. We have considered all mother and feeder arrivals at TCT of Cagliari during the first 6 months of 2011, collecting for every vessel all information about their predictor variables. Each arrival trickles down the tree and is assigned to a terminal node. We then compared actual and estimated delays. The mean prediction error obtained, based on 339 arrivals, was 89.80 minutes.

The results obtained with CART methodology are very encouraging. In particular if we compare it with the Neural Networks model calibrated at the University of Cagliari built to forecast the delay in the same port (Fancello *et al.* 2011). Mean error on delay prediction was significantly lower, passing from around 2 hours 40 minutes to about 1 hour and 30 min.

Three basic considerations emerge from these results:

- the possibility that the vessel's predicted arrival time falls within 3 or 4 work shifts is entirely ruled out;
- there is the certainty that resources can be scheduled over 2 work shifts at the most;
- the probability of specifically determining the work shift of arrival is around the 75%.

Conclusions

The major issue for enhancing planning efficiency in a container terminal is the prediction of ship arrival times. Greater certainty on demand data would in fact improve port operations management. Furthermore, considering the strong dependence of planning processes on incoming information flow, a reliable estimation of the delay would reduce uncertainty in scheduling the resources (space, equipment and personnel), required to satisfy the predicted demand.

This paper concerns the development of a tool for predicting arrivals in a Mediterranean transhipment container terminal, for a short time horizon. Based on data collected and the available tools and after an analysis of the scientific literature it was decided to adopt the CART methodology.

In particular, referring to the TCT of Cagliari container terminal, where the working day is divided into four 6-hour shifts, a mean error of around 1 hour 30 minutes was obtained for delay prediction. In this way, the probability of the uncertainty interval falling entirely within a single work shift increases. In practical terms, this means that resources can be scheduled over most at the 2 work shifts. The results obtained are therefore very favourable, viewed from both the scientific standpoint and in the operational context.

They also provide basis for furthering the research work, which will focus on refining the model using new variables and observations. This will be followed by a careful economic analysis for examining the second group of variables and determining the repercussions of delays on the entire economic/organizational/management system.

In a broader perspective, the aim is to create a Decision Support System (DSS) for port operators that will assist planners and ultimately contribute to maximizing terminal efficiency and hence competitiveness, combining three main, closely interrelated modules, forecasting, resources optimization and equipment maintenance.

References

Breiman, L. 2001. Random forests, *Machine Learning* 45(1): 5–32. http://dx.doi.org/10.1023/A:1010933404324

Breiman, L. 1996. Bagging predictors, *Machine Learning* 24(2): 123–140. http://dx.doi.org/10.1023/A:1018054314350

Breiman, L.; Friedman, J.; Olshen, R. A.; Stone, C. J. 1984. *Classification and Regression Trees*. Chapman and Hall/CRC. 368 p.

Brown, D. E.; Corruble, V.; Pittard, C. L. 1993. A comparison of decision tree classifiers with backpropagation neural networks for multimodal classification problems, *Pattern Recognition* 26(6): 953–961. http://dx.doi.org/10.1016/0031-3203(93)90060-A

Cellard, J. C.; Labbé, B.; Savitsky, G. 1967. Le programme ELISEE, presentation et application, *Metra* 3(6): 511–519 (in French).

Chan, K.-Y.; Loh, W.-Y. 2004. LOTUS: an algorithm for building accurate and comprehensible logistic regression trees, *Journal of Computational and Graphical Statistics* 13(4): 826–852. http://dx.doi.org/10.1198/106186004X13064

Choi, Y.; Ahn, H.; Chen, J. J. 2005. Regression trees for analysis of count data with extra Poisson variation, *Computational Statistics & Data Analysis* 49(3): 893–915. http://dx.doi.org/10.1016/j.csda.2004.06.011

Conversano, C. 2002. Bagged mixtures of classifiers using model scoring criteria, *Pattern Analysis & Applications* 5(4): 351–362. http://dx.doi.org/10.1007/s100440200031

Dunham, M. H. 2002. *Data Mining: Introductory and Advanced Topics*. Prentice Hall. 315 p.

Fancello, G.; Pani, C.; Pisano, M.; Serra, P.; Zuddas, P.; Fadda, P. 2011. Prediction of arrival times and human resources allocation for container terminal, *Maritime Economics & Logistics* 13(2): 142–173. http://dx.doi.org/10.1057/mel.2011.3

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. 1996. Knowledge discovery and data mining: towards a unifying framework, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 2–4 August, 1996, Portland, Oregon, 82–88.

Frawley, W. J.; Piatetsky-Shapiro, G.; Matheus, C. J. 1992. Knowledge discovery in databases: an overview, *AI Magazine* 13(3): 57–70. http://dx.doi.org/10.1609/aimag.v13i3.1011

Freund, Y.; Schapire, R. 1996. Experiments with a new boosting algorithm, in *Machine Learning: Proceedings of the Thirteenth International Conference (ICML'96)*, July 3–6, 1996, Bari, Italy. 148–156.

Fung, M. K. 2002. Forecasting Hong Kong's container throughput: an error-correction model, *Journal of Forecasting* 21(1): 69–80. http://dx.doi.org/10.1002/for.818

Gambardella, L. M.; Bontempi, G.; Taillard, E.; Romanengo, D.; Raso, G.; Piermari, P. 1996. Simulation and forecasting in intermodal container terminal, in *Proceedings ESS'96: 8th European Simulation Symposium*, October 24–26, 1996, Genoa, Italy. 5 p. (CD).

Gambardella, L. M.; Rizzoli, A. E.; Zaffalon, M. 1998. Simulation and planning of an intermodal container terminal, *Simulation* 71(2): 107–116. http://dx.doi.org/10.1177/003754979807100205

Gillo, M. W. 1972. MAID: a honeywell 600 program for an automatized survey analysis, *Behavioral Science* 17(2): 251–252.

Hastie, T.; Tibshirani, R.; Friedman, J. 2013. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 7th edition. Springer. 745 p.

Hothorn, T.; Hornik, K.; Zeileis, A. 2006. Unbiased recursive partitioning: a conditional inference framework, *Journal of Computational and Graphical Statistics* 15(3): 651–674. http://dx.doi.org/10.1198/106186006X133933

Hunt, E. B.; Marin, J.; Stone, P. J. 1966. *Experiments in Induction*. Academic Press. 247 p.

Kass, G. V. 1980. An exploratory technique for investigating large quantities of categorical data, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 29(2): 119–127. http://dx.doi.org/10.2307/2986296

Loh, W.-Y.; Vanichsetakul, N. 1988. Tree-structured classification via generalized discriminant analysis, *Journal of the American Statistical Association* 83(403): 715–725. http://dx.doi.org/10.1080/01621459.1988.10478652

Markham, I. S.; Mathieu, R. G.; Wray, B. A. 2000. Kanban setting through artificial intelligence: a comparative study of artificial neural networks and decision threes, *Integrated Manufacturing Systems* 11(4): 239–246. http://dx.doi.org/10.1108/09576060010326230

Mehta, M.; Agrawal, R.; Rissanen, J. 1996. SLIQ: A fast scalable classifier for data mining, *Lecture Notes in Computer Science* 1057: 18–32. http://dx.doi.org/10.1007/BFb0014141

Mola, F.; Siciliano, R. 1997. A fast splitting procedure for classification trees, *Statistics and Computing* 7(3): 209–216. http://dx.doi.org/10.1023/A:1018590219790

Morgan, J. N.; Messenger, R. C. 1973. *THAID a Sequential Analysis Program for the Analysis of Nominal Scale Dependent Variables*. Institute for Social Research, University of Michigan. 92 p.

Morgan, J. N.; Sonquist, J. A. 1963. Problems in the analysis of survey data, and a proposal, *Journal of the American Statistical Association* 58(302): 415–434. http://dx.doi.org/10.1080/01621459.1963.10500855

Murty, K. G.; Liu, J.; Wan, Y.-W.; Linn, R. 2005. A decision support system for operations in a container terminal, *Decision Support Systems* 39(3): 309–332. http://dx.doi.org/10.1016/j.dss.2003.11.002

Quinlan, J. R. 1986. Induction of decision trees, *Machine Leaning* 1(1): 81–106. http://dx.doi.org/10.1007/BF00116251

Quinlan, J. R. 1992. *C4.5: Programs for Machine Learning*. Morgan Kaufmann. 302 p.

Salido, M. A.; Rodriguez-Molins, M.; Barber, F. 2012. A decision support system for managing combinatorial problems in container terminals, *Knowledge-Based Systems* 29: 63–74. http://dx.doi.org/10.1016/j.knosys.2011.06.021

Shafer, J.; Agrawal, R.; Mehta, M. 1996. SPRINT: A scalable parallel classifier for data mining, in *VLDB'96: Proceedings of 22th International Conference on Very Large Data Bases*, September 3–6, 1996, Mumbai (Bombay), India, 544–555.

Sideris, A. C.; Boilé, M. P.; Spasovic, L. N. 2002. Using on-line information to estimate container movements for day-to-day marine terminal operations, in *The 81st Annual Meeting of Transportation Research Board Compendium of Papers DVD*, 13–17 January, 2002, Washington, US. (DVD).

Stahlbock, R.; Voß, S. 2008. Operations research at container terminals: a literature update, *OR Spectrum* 30(1): 1–52. http://dx.doi.org/10.1007/s00291-007-0100-9

Su, X.; Wang, M.; Fan, J. 2004. Maximum likelihood regression trees, *Journal of Computational and Graphical Statistics* 13(3): 586–598. http://dx.doi.org/10.1198/106186004X2165

Vis, I. F. A.; De Koster, R. 2003. Transshipment of containers at a container terminal: an overview, *European Journal of Operational Research* 147(1): 1–16. http://dx.doi.org/10.1016/S0377-2217(02)00293-X

Won, S. H.; Kim, K. H. 2009. An integrated framework for various operation plans in container terminals, *Polish Maritime Research* 16(3): 51–61. http://dx.doi.org/10.2478/v10012-008-0034-4

Zhen, L.; Lee, L. H.; Chew, E. P. 2011. A decision model for berth allocation under uncertainty, *European Journal of Operational Research* 212(1): 54–68. http://dx.doi.org/10.1016/j.ejor.2011.01.021