University of Cagliari

University of Antwerp

Joint Ph. D in

LAND ENGINEERING AND APPLIED ECONOMICS

MANAGING VESSEL ARRIVAL UNCERTAINTY IN CONTAINER TERMINALS: A MACHINE LEARNING APPROACH

ICAR/05

Author:
Claudia Pani

Supervisors:
Prof. Dr. Gianfranco Fancello
Prof. Dr. Thierry Vanelslander

Co-Supervisor:
Prof. Dr. Paolo Fadda

Ph.D Coordinator:
Prof. Dr. Roberto Deidda

CYCLE XXVI

Final defence academic year 2012-2013

MANAGING VESSEL ARRIVAL UNCERTAINTY IN CONTAINER TERMINALS: A MACHINE LEARNING APPROACH

DICAAR

Department of Civil-Environmental Engineering and Architecture
University of Cagliari, Italy

TPR

Department of Transport an Regional Economics
University of Antwerp, Belgium

To my grandfather.

Acknowledgements

Completion of this doctoral dissertation was possible thanks to several people who supported me during these past three years. I would like to express my sincere gratitude to all of them.

A heartfelt thank you goes to my supervisor and my co-supervisor in Cagliari, Prof. Gianfranco Fancello and Prof. Paolo Fadda, who gave me the opportunity to study a very interesting topic and who always believed in my abilities as a researcher and as an engineer. Thank you, also, for sharing your valuable network of contacts at the port of Cagliari with me.

I am totally indebted to my supervisor in Antwerp, Prof. Thierry Vanelslander, for introducing me to a very inspiring and stimulating work system, for his valuable support and suggestions and for always having appreciated my research. His hard work and dedication are an example for me.

The thesis would not have come to successful completion without the great help of Prof. Francesco Mola and his statistics team. In particular, thank you very much to Luca Frigau and Massimo Cannas who gave me a great deal of advice within the complex word of Data Mining. Thanks also to the colleagues at the department of Economics Antonio, Vincenzo, Claudio, Elisabetta and Maria Bonaria for their kindness and moral support.

I am grateful to the following organizations for the help they gave me in collecting such a large amount of data: CICT (Cagliari International Container Terminal), ISPRA (Institute for Protection and Environmental Research) and the port authority of Antwerp. In particular thanks to Daniele Cuzzocrea and Roberto Inghilesi for always being willing.

I greatly benefitted from all those Professors who have approached me over these years with questions and ideas and who have contributed to increasing the value of my research. I would like to thank my referees Prof. Michele Acciaro and Prof. Yves

Crozet for providing me with useful comments on my thesis and for giving me important advice for the future research. Special thanks go to Professor Antoch, that I met in Prague, and to Professor Heaver, that I met in Antwerp, with whom I had insightful discussions. Finally, I would like to thank the members of my dissertation committee that gave me the opportunity to start the Ph.D, Professors Giovanni Sechi, Michele Campagna and Maria Grazia Badas.

A very important role in these years has been covered by my engineering colleagues and friends, Patrizia, Michele and Daniela. Special attention owed to Patty, who shared with me the best and the worst moments of our doctoral period, for our wonderful trips and for our long Skype conversations between Antwerp and Hamburg.

It has been a great privilege to spend part of my PhD research time in the Department of Transport and Regional Economics in Antwerp. I am very grateful to Professors Ann Verhetsel, Eddy Van De Voorde, Hilde Meersman, Christa Sys and, of course, Thierry for their highly professional qualities and for always welcoming me to the department. Special thanks go to Christa for her enthusiasm and for the help she gave me to better structure my work. I am thankful to the wonderful group of colleagues who made me feel at home from the beginning and who created a truly supportive atmosphere for research and for writing the thesis. I wish to thank, in particular, Anne and Katja for their sweetness and for the attention that they paid to me in this last year (Anne I will always remember the chocolate bomb that you bought for me the day of the thesis deadline!). Thank you very much to Flo, the free-mind of the group, for our wonderful days in the *International-Office*, for our long discussions and for our great evenings in Antwerp. My warmest thanks go to Kat, my Belgian sister, for the significant support she gave me from the beginning and for her beautiful surprises that always made me cry. Finally, thanks to Yasy, my colleague, neighbor and special friend, for her concrete and continuous help and for her great strength. I learned a lot from our different cultures and personalities.

Thanks to all the members of my warm family and my friends who supported me in every situation. A special thought in this moment goes to Sandry, Marta, Giaky, Giuly,

Carlo, Fede, Francy, Sivy, Alby and Lau, for always being present in my life, especially in these last weeks.

There are no words that can express my gratitude towards my sister Carla and towards my parents. Thanks Mom and Dad for always allowing me to realise myself.

And last but not least the greatest thanks go to Marco for always believing in me and understanding my choices, for his patience, for his complete support and for his unconditional love.

Thanks also to all the others that contributed to make this challenging period a very beautiful and important chapter of my life.

Summary

A container terminal is a complex system where a broad range of operations are carried out involving a wide array of resources that need to interact over a 24 hour operating cycle. Since the various activities are mutually related to each other, there is a need not only to maximise the efficiency of each one, but also to ensure proper coordination, hence to solve integrated decision-making problems. Several factors can affect the quality of the services provided and the overall efficiency. Vessel arrival uncertainty further complicates the task of the planners and, as a result, of the effectiveness of the planning itself, in particular at the operational level. Each arrival produces high peak loads for other terminal activities, as well as for the supporting arrival activities (pilotage, towage, etc.) and hinterland transportation (waiting, congestion etc.). Deviating arrivals only worsen this peak load.

On a daily level, the actual time of arrival of the vessels often deviates from the scheduled time. Despite contractual obligations to notify the Estimated Time of Arrival (ETA) at least 24 hours before the arrival, ship operators often have to adapt and update the latest ETA due to unexpected circumstances. This aspect results in a last-minute change of plans in terminal operations resulting in higher costs. In fact, the ability to predict the actual time of a vessel's arrival in a port 24 hours in advance is fundamental for the related planning activities for which the decision-making processes need to be constantly adapted and updated. Moreover, disruptions in container flows and operations caused by vessel arrival uncertainty can have cascade effects within the overall supply chain and network within which the port is part.

Although vessel arrival uncertainty in ports is a well-known problem for the scientific community, the literature review highlights that in the maritime sector the specific instruments for dealing with this problem are extremely limited.

The absence of a reference model that specifies the relationship between vessel arrival uncertainty and the involved variables resulted in the application of a specific machine learning approach within the Knowledge Discovery in Database process. This

approach, that abandons all prior assumptions about data distribution shape, is based on the self-learning concept according to which the relation between an outcome variable Y and the set of predictors X is directly identified from the historical collected data.

The approach has been validated thanks to two different case studies: the container terminal of Cagliari, located in the Mediterranean basin, and one of the main container terminals of Antwerp, located at the North Sea.

Depending on the framework and planning purposes several estimates can provide useful information on vessel arrivals. Sometimes, it can be useful for planners to infer a quantitative estimate of the delay/advance in minutes, sometimes it may be useful to have a qualitative estimate, even only knowing whether or not an incoming vessel is likely to arrive before or after the scheduled ETA. For this reason a two-step instrument is proposed is made up of two different modules.

The fitted algorithmic models used to obtain predictions are Logistic Regression, CART (Classification and Regression Trees) and Random Forest. All the proposed models are able to learn from experience, following the well-known Data Mining paradigm “learning from data”.

From a practical point of view, the probability, associated to the continuous estimation, of specifically identifying the work-shift of the incoming vessel is calculated. In all predictions Random Forest algorithms still show the best performance. This aspect can help planners, in the daily strategy decision making process, in order to improve the use of the human, mechanical and spatial resources required for handling operations. This could maximise terminal efficiency and minimise terminal costs, hence improving terminal competitiveness.

Moreover, the interpretation of the discovered knowledge, made it possible to evaluate the most discriminating variables of the analysis, even thanks to graphical visualisation of the Importance-plots.

Sommario

Il terminal marittimo è un sistema complesso, al suo interno si svolgono una molteplicità di operazioni che coinvolgono una varietà di risorse che devono interagire in un ciclo operativo che abbraccia le intere 24 ore. Il terminalista mira, nella gestione, a massimizzare l'efficienza interna del terminal, ovvero a garantire il maggior numero di movimentazioni al minor costo possibile. Le attività operative che si svolgono all'interno di un terminal sono caratterizzate da un elevato numero di variabili e vincoli che contribuiscono ad accresce il livello di complessità: numerosi elementi, non sempre facilmente controllabili, possono influenzare la qualità del servizio offerto e l'efficienza complessiva del terminal.

A rendere ulteriormente complesso il lavoro dei Planners e, conseguentemente l'efficacia stessa della pianificazione, è il problema dell'incertezza dell'orario di arrivo delle navi in porto. Gli operatori di linea sono obbligati da vincoli contrattuali ad inviare periodicamente l'ETA (Estimated Time of Arrival), secondo cadenze temporali predefinite. Anche l'ultimo ETA inviato è però spesso soggetto ad aggiornamenti e modifiche successive, a causa del verificarsi di eventi imprevisti quali condizioni meteo-marine avverse, ritardi nelle lavorazioni ai porti precedenti, etc. A livello giornaliero permane, dunque, l'incertezza sull'orario di arrivo delle navi in porto.

Considerata la forte dipendenza dei processi di pianificazione dal flusso informativo in ingresso, una migliore gestione dei ritardi e degli anticipi risulta fondamentale per una pianificazione più efficiente delle risorse del terminal (umane, spaziali e meccaniche) necessarie per le operazioni di movimentazione, in particolare con riferimento al breve periodo. Attualmente infatti, in uno scenario di pianificazione giornaliero, le risorse sono spesso sovra o sottodimensionate per sovvenire alle caratteristiche d'incertezza che caratterizzano gli arrivi.

L'approccio al problema appare complesso considerato l'elevato numero di variabili e vincoli che influenzano il processo, che riguardano principalmente:

- Struttura del naviglio (lunghezza, pescaggio, stazza lorda, capacità,..);

- Servizio effettuato (sailing direction, rotazione dei porti,..);
- Piano di carico e tipologia di containers;
- Organizzazione/Disponibilità del porto precedente;
- Fattori esterni (condizioni meteo-marine, scioperi,..).

I processi decisionali dedicati a tali funzioni sono, di solito, talmente complessi da risultare ingestibili senza il supporto di adeguati strumenti metodologici. L'obiettivo generale della presente tesi risiede nello sviluppo di uno strumento di previsione dell'orario di arrivo delle navi in un Terminal Container di transhipment, nel breve periodo.

Lo studio dello stato dell'arte e il confronto con gli operatori del settore hanno evidenziato, che nonostante gli sviluppi senza precedenti dell'innovazione tecnologica, l'incertezza degli arrivi rimane ancora una sfida per i gestori dei porti. Le specifiche applicazioni nel settore appaiono fortemente limitate. Per questo motivo l'approccio metodologico utilizzato ricade nel campo del Data Mining e più specificamente del Machine Learning. L'osservazione del reale andrà a formare la base di conoscenza fondata sull'apprendimento dal passato.

Gli algoritmi impiegati al fine di ottenere una previsione discreta e continua del ritardo sono: regressione logistica, CART (Classification and Regression Tree) e Random Forest.

L'approccio utilizzato è stato validato grazie alla sperimentazione condotta su due porti di transhipment Europei: il terminal container di Cagliari, situato al centro del Mar Mediterraneo, e uno tra gli otto terminal container principali del porto di Anversa, situato nel Mare del Nord.

Lo strumento di previsione proposto si configura come uno strumento di supporto alle decisioni degli operatori portuali in grado di fornire risposte analitiche all'incertezza delle informazioni sui flussi in ingresso al Terminal.

Da un punto di vista operativo, conoscere, con almeno 24 ore di anticipo, l'orario effettivo di arrivo delle navi in porto permetterebbe una pianificazione più efficiente

delle attività giornaliere del terminal. Questo consentirebbe una notevole riduzione dei costi e un miglioramento dell'efficienza interna rendendo, quindi, il terminal maggiormente competitivo. Più generale, uno strumento di questo tipo consentirebbe un miglioramento delle performance dell'intera catena logistica o rete all'interno della quale il porto stesso si inserisce.

Samenvatting

Een containerterminal is een complex system waarin een brede waaier aan operaties wordt uitgevoerd. Dit houdt een hele lijst van middelen in die met elkaar in wisselwerking zijn binnen een werkingscyclus van 24 uur. Aangezien de verschillende activiteiten onderling met elkaar gerelateerd zijn, is het nodig om niet enkel de efficiëntie van elke activiteit te maximaliseren, maar ook de geschikte coördinatie te verzekeren en dus geïntegreerde beslissingsproblemen op te lossen. Meerdere factoren kunnen de kwaliteit van de aangeboden diensten en globale efficiëntie beïnvloeden. De aankomstonzekerheid van schepen maakt de taak van de planners nog moeilijker en dus ook de effectiviteit van de planning zelf, in het bijzonder op operationeel niveau. Elke aankomst produceert sterke piekbelastingen voor andere terminalactiviteiten, alsook voor de ondersteunende aankomstactiviteiten (loods- en sleepdiensten, etc.) en het hinterlandtransport (wachttijden, congestie, etc.). Afwijkende aankomsten maken deze piekbelasting alleen nog maar erger.

Op dagelijks niveau wijkt de werkelijke aankomsttijd van de schepen vaak af van de geplande tijd. Ondanks contractuele verplichtingen om de verwachte aankomsttijd (ETA) tenminste 24 uur voor de aankomst te melden, moeten scheepsoperatoren de laatste ETA vaak aanpassen en updaten omwille van onverwachte omstandigheden. Dit aspect resulteert in een verandering van de plannen voor de terminaloperaties op het laatste moment, wat resulteert in hogere kosten. Het vermogen om de werkelijke aankomsttijd van een schip in de haven 24 uur op voorhand te voorspellen, is in feite fundamenteel voor de gerelateerde planningsactiviteiten waarvoor de beslissingsprocessen constant moeten worden aangepast en geüpdatet.

Hoewel de aankomstonzekerheid van een schip in de haven een bekend probleem is voor de wetenschappelijke gemeenschap, benadrukt het literatuuronderzoek dat de specifieke instrumenten om dit probleem aan te pakken extreem beperkt zijn in de maritieme sector.

De afwezigheid van een referentiemodel dat de relatie tussen de aankomstonzekerheid van een schip en de betrokken variabelen specificeert, resulteerde in de toepassing van een specifieke machine learning benadering in het Knowledge Discovery in Database proces. Deze benadering, die afstand doet van alle voorgaande assumpties over de verdelingsvorm van de data, is gebaseerd op het self-learning concept volgens hetwelke de relatie tussen een afhankelijke variabele Y en de set onafhankelijke variabelen X direct geïdentificeerd wordt uit de historisch verzamelde data.

De benadering werd gevalideerd door twee verschillende gevalstudies: de containerterminal van Cagliari, gelegen in het Middellandse Zeebekken, en een van de belangrijkste containerterminals van Antwerpen, gelegen aan de Noordzee.

Het voorgestelde twee-stappen instrument is opgesteld door twee verschillende gerelateerde modules:

- de discrete-schatting module die een kwalitatieve evaluatie van de aankomstonzekerheid van de schepen aanlevert;
- de continue-schatting module die een kwantitatieve evaluatie van de vertraging/ het te vroeg aankomen aanlevert in minuten.

De gepaste algoritmische modellen gebruikt om voorspellingen te verkrijgen, zijn logistische regressie, CART (classificatie en regressiebomen) en Random Forest. Alle voorgestelde modellen kunnen leren door ervaring, volgens het welbekende Data Mining paradigma “learning from data”.

Het voorgestelde instrument kan planners helpen in hun dagelijks strategisch beslissingsproces en zo het gebruik van menselijke, mechanische en ruimtelijke middelen nodig voor de afhandelingsoperaties verbeteren. Dit kan de terminalefficiëntie maximaliseren en de terminalkosten minimaliseren, waardoor de concurrentiekracht van de terminal gemaximaliseerd wordt.

Table of Contents

Acknowledgements I
Summary IV
Sommario VII
Samenvatting IX
List of Figures XV
List of Tables XVII

CHAPTER 1: INTRODUCTION 1
1.1 Research area 1
1.2 Rationale of the study 3
1.2.1 The problem setting 3
1.2.2 Research objectives 5
1.3 Methodological approach 6
1.4 Research structure 8

CHAPTER 2: CONTAINER TERMINAL SYSTEM 11
2.1 Introduction 11
2.2 General layout and Technical Equipment 12
2.2.1 System using gantry cranes for container storage 14
2.2.2 Pure straddle carrier system 16
2.3 Port performance 17
2.4 Main planning activities 19
2.4.1 Berth planning 20
2.4.2 Loading/unloading operations 21
2.4.3 Transport of containers 21
2.4.4 Yard stacking 22
2.5 Operations centre 22
2.6 Standard gang composition 24
2.7 Importance of reliable vessel arrival time 25

CHAPTER 3: THE VESSEL ARRIVAL UNCERTAINTY PROBLEM 27
3.1 Introduction 27
3.2 Related planning activities 30
3.2.1 Berth scheduling process 31
3.2.2 Human resources and equipment allocation 32
3.2.3 Yard planning process 33
3.3 Arrival uncertainty in the air transport sector 34
3.4 Required approach 35
3.4.1 Classification methods in supervised learning 37
3.4.2 Regression approach in classification problems 38

CHAPTER 4: METHODOLOGICAL FRAMEWORK 40
4.1 Knowledge discovery in databases approach 40
4.2 Algorithms 41
4.2.1 Logistic Regression 41
4.2.2 Classification and Regression Trees (CART) 42
4.2.2.1 Regression Trees 44
4.2.2.2 Classification Trees 46
4.2.3 Random Forest 47
4.3 Cross validation 48
4.4 Performance metrics 50
4.4.1 Mean absolute prediction error 50
4.4.2 Kappa statistic and percentage of misclassified instances 50

CHAPTER 5: THE PORT OF CAGLIARI CASE STUDY 53
5.1 Introduction 53
5.2 KDD process 56
5.2.1 Understanding the application domain 56
5.2.2 Data selection 57
5.2.2.1 Vessel-related variables 57
5.2.2.2 Weather-related variables 57
5.2.2.3 The outcome variable 61

5.2.2.4 Database structure 61
5.2.3 Data preparation 62
5.2.3.1 Data cleaning 63
5.2.3.2 Creation of new variables 63
5.2.3.3 Relationships among variables 63
5.2.3.4 Exploratory analysis 63
5.2.3.4 Ranking the delay severity at the daily level 63
5.2.4 Data mining 73
5.2.5 Interpretation of results 76
5.2.6 Consolidation of the discovered knowledge 80

CHAPTER 6: THE PORT OF ANTWERP CASE STUDY 82
6.1 Introduction 82
6.2 KDD process 84
6.2.1 Understanding the application domain 84
6.2.2 Data selection 86
6.2.2.1 Vessel-related variables 86
6.2.2.2 Weather-related variables 87
6.2.2.3 The outcome variable 89
6.2.2.4 Database structure 89
6.2.3 Data preparation 89
6.2.3.1 Data cleaning 89
6.2.3.2 Creation of new variables 90
6.2.3.3 Relationships among variables 90
6.2.3.4 Exploratory analysis 91
6.2.4 Data mining 94
6.2.5 Interpretation of results 97
6.2.6 Consolidation of the discovery knowledge 100

CHAPTER 7: CONCLUSIONS 105
7.1 The contribution of the research 105
7.2 Practical implications 107

7.3 Suggestions for future research 108

BIBLIOGRAPHY 110

APPENDIX 1 124

APPENDIX 2 130

List of Figures

Figure 1.1: Global container trade, 1996- 2013 2
Figure 1.2: EU27 freight container transport performance from 2007 to 2012 2
Figure 1.3: Main steps constituting the KDD process 7
Figure 1.4: Research structure 10

Figure 2.1: General layout of a container terminal 13
Figure 2.2: Quay Crane 14
Figure 2.3: Track-Trailer 14
Figure 2.4: Yard Crane 15
Figure 2.5: System using yard cranes for container storage 15
Figure 2.6: Straddle Carrier 16
Figure 2.7: Pure straddle carrier system 17
Figure 2.8: The key factors in port performance 18
Figure 2.9: Operations Center structure observed at the Cagliari container terminal .. 24

Figure 3.1: Major logistic processes in container terminals 27
Figure 3.2: Different steps in the KDD approach 36

Figure 4.1: A two-dimensional feature space partitioned by recursive binary splitting 43
Figure 4.2: Binary tree corresponding to the partitioned feature space 43
Figure 4.3: γ -fold cross validation 49

Figure 5.1: Location of the Cagliari port 53
Figure 5.2: The structure of the “porto canale” 54
Figure 5.3: Selected points in the Mediterranean Sea 61
Figure 5.4: First database structure 62
Figure 5.5: Second database structure. 62

Figure 5.6: Time series of arrivals in the Cagliari container terminal 67
Figure 5.7: Delay distribution in the Cagliari container terminal 68
Figure 5.8. Distribution of vessel arrivals during the day, the week and the year 68
Figure 5.9: Importance of predictors for the Random Forest algorithm (continuous model)..... 78
Figure 5.10: Uncertainty range of vessel arrival at Cagliari container terminal 81

Figure 6.1: Location of the Antwerp port 83
Figure 6.2: Port of Antwerp container terminals..... 85
Figure 6.3: Selected points in the North Sea..... 88
Figure 6.4: Delay distribution at Antwerp port, by terminal..... 92
Figure 6.5: Importance of predictors for the Random Forest algorithm (discrete model - port level) 98
Figure 6.6: Importance of predictors for the Random Forest algorithm (discrete model - terminal level) 99
Figure 6.7: Frequency distributions of the delays (a) and advances (b) at Terminal 7102
Figure 6.8: Uncertainty range of vessel arrival at the Antwerp container terminal ... 104

List of Tables

Table 2.1: The main decision-making problems in Container Terminals.....20

Table 4.1: Value of Kappa and agreement level.....52

Table 5.1: Main characteristics of the Cagliari Container Terminal.....54
Table 5.2: Cagliari Port statistics (2011-2012).....55
Table 5.3: Summary statistics of the continuous variables.....59
Table 5.4: Longitude and latitude of the points selected in the Mediterranean Sea 60
Table 5.5: $\chi 2$ and p-value values for the categorical variables 65
Table 5.6: Pearson coefficient and p-value values for the continuous variables 66
Table 5.7: Delay summary statistics (in minutes).....67
Table 5.8: Values of cluster validation indices 72
Table 5.9: Cluster means and standard deviations.....73
Table 5.10: Predictive performance for the discrete outcome 75
Table 5.11 Predictive performance for the continuous outcome 76

Table 6.1: Main characteristics of the Antwerp container terminals 85
Table 6.2: Summary statistics of the continuous variables.....87
Table 6.3: Longitude and latitude of the points chosen in the North Sea 88
Table 6.4: Pearson coefficient and p-value values for the variables related to the vessel features 91
Table 6.5: Delay summary statistics by terminal (in minutes).....93
Table 6.6: Predictive performance for the discrete outcome (port level) 95
Table 6.7: Predictive performance for the discrete outcome (terminal level).....95
Table 6.8: Predictive performance for the continuous outcome (port level) 96
Table 6.9: Predictive performance for the continuous outcome (terminal level) 96
Table 6.10: Delay and advance summary statistics. 101
Table 6.11: Predictive performance for the continuous outcome (terminal level-delay) 103

Table 6.12: Predictive performance for the continuous outcome (terminal level-advance) 103

CHAPTER 1: Introduction

The general research topic of this study is the assessment of the competitive conditions of the container liner shipping industry. This introductory chapter explains the purpose and the methodology of this research. It is divided into four parts. Subsequently, the container liner shipping industry, some forms of cooperation, two anti-trust regulations as well as the conceptual framework of this study will be presented.

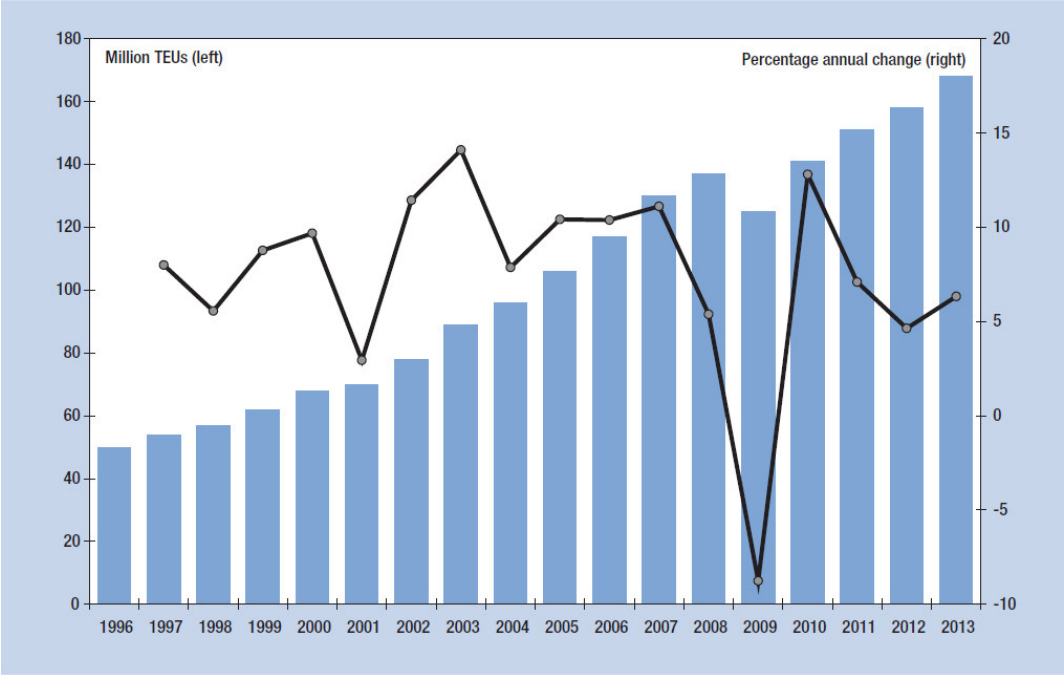
1.1 Research area

In the last few decades, the expansion of trade areas, resulting from the process of globalisation and from the reinforcement of the world economy, has heightened the need to move freight and information faster and more economically. Maritime transport with the unitisation of cargo is an effective means of meeting these needs. The introduction of the container and the consequent development of intermodality have allowed the considerable reduction in transportation costs and the increase of transcontinental cargo flows (Woodburn et al., 2008, Kumar and Hoffmann, 2002).

Today, around 80% of global trade by volume and over 70% by value is carried by sea and is handled by ports worldwide (Thana, 2013). In particular, international sea-freight container transportation has grown exponentially, and nowadays container terminals represent a key player in the global shipping network.

Figure 1.1 shows the global container trade from 1996 to 2013. The trend underlines a decrease in 2008 and 2009 due to the crisis that has affected the world's largest economies with major repercussions on the shipping sector. In 2011, world container port throughput increased by an estimated 5.9% to 572.8 million 20-foot equivalent units (TEUs), its highest level ever (UNCDAD 2012).

Figure 1.1: Global container trade, 1996- 2013



Source: UNCDAD 2012

Since 2011, the European market was one of the largest in terms of handled containers, amounting to 741,28 million tonnes in 2011 (own calculations based on data from the European Commission). The EU27 freight container transport performance in 2011 reached a level which is more than 7% greater than in 2007, i.e., before the global economic crisis (Figure 1.2). This huge amount of traffic means that port infrastructures need to be continually upgraded to ensure efficient transfer of containers within the world global transportation network.

Figure 1.2: EU27 freight container transport performance from 2007 to 2012

Source: European Commission, 2012

Furthermore, the trend towards gigantism and the prevalent hub & spoke policy are bringing about changes in the international container market (Coulter, 2002). Traffic is being concentrated into hub ports of appropriate size and infrastructure that are able to accommodate mega containerships and terminal managers are having to deal with increasing competitiveness among terminals (Tongzon and Heng, 2005; Heaver et al., 2001). Hub ports are changing their role not only as individual places that handle vessels but also within supply chains and networks therefore they aim to maximise network functioning and effects.

For the reasons mentioned above, competition among ports has become more and more complex and dynamic (Notteboom, 2012). In particular, with reference to the European context, Mediterranean and North European hub ports compete to increase their traffic travelling from the Far East and the Eastern shore of the Pacific Ocean to Europe.

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 (Sciomachen et al., 2009 Notteboom and Rodrigue, 2008). In addition, port technology, geographical position and terminal structure are the result of strategic decisions and hence cannot be altered in the short-medium term. At the tactical and operational levels however, it is possible to adopt methodologies for the optimal management of a terminal's resources and the logistics processes involved.

Therefore, this research is a step towards better understanding the needs of terminal operators in a daily planning scenario and to develop specific instruments able to support planners in the short-term planning activities.

1.2 Rationale of the study

1.2.1 The problem setting

A container terminal is a complex system where a broad range of operations are carried out involving a wide array of resources that need to interact over a 24 hour

operating cycle. Since the various activities are mutually related to each other, there is a need not only to maximise the efficiency of each one, but also to ensure proper coordination, hence to solve integrated decision-making problems (Salido et al., 2011, Won and Kim, 2009, Murty et al., 2005). Several factors, that are not always easy to control, can affect the quality of the services provided and the overall efficiency.

Vessel arrival uncertainty strongly complicates the task of the planners and, as a result, of the effectiveness of the planning itself, in particular at the operational level. On a daily level, the actual time of arrival of the vessels often deviates from the scheduled time. Despite contractual obligations to notify the Estimated Time of Arrival (ETA) at least 24 hours before the arrival, ship operators often have to update the latest ETA due to unexpected circumstances. This aspect requires several last-minute change of plans in terminal operations, for which the decision-making processes need to be constantly adapted, resulting in higher costs.

Although vessel arrival uncertainty in port is a well known problem for the scientific community, the literature review highlights that in the maritime sector the specific instruments for dealing with this problem are extremely limited.

Therefore, the purpose of this research is to identify a specific approach that can provide, at least 24 hours in advance, reliable estimates on vessel arrivals in order to support port managers in short-term planning activities. At the terminal level, information on vessel arrival time would facilitate the allocation of the human, mechanical and spatial resources required for handling operations. To date, this task has been delegated to the planners, i.e., professionals who operate mainly on the basis of hands-on experience. The decision-making processes that are involved are often so complex as to be unmanageable without the support of adequate methodological instruments. The problem solution approach appear extremely complex, considering the large number of variables and constraints involved. They are related in particular to:

- Vessel structure;
- Vessel service;

- Vessel owner;
- Organisation/availability of previous port;
- Loading plan and type of containers to be loaded/unloaded;
- Weather/sea conditions;
- External factors (strikes, breakdowns, etc.);
- Human resources management (contractual obligations, labor regulations, etc.);
- Equipment management (repairs, out of service for maintenance, etc.);
- Space management (berth space and relative distance from the stacked containers).

1.2.2 Research objectives

The general objective of this study is to develop a forecasting instrument that can provide reliable information of vessel arrival times in Container Terminals in order to support port managers in the daily strategy decision process.

This general objective leads to a series of specific and operational objectives.

A first objective of the study is to identify the methodological approaches that can best provide analytical and objective answers to the vessel arrival uncertainty problem.

After defining the approach, a second important goal of the research is to determine the algorithm(s) to be used in order to handle late/early arrival times. This may help planners to allocate, with greater certainty, all the human, mechanical and spatial resources required for handling operations, that are often under/overestimated at the planning stage.

Furthermore, a third purpose is to build a general instrument that can be used in various contexts by planners, on the basis of their experience.

Therefore, this specific research uses two case studies: the terminal container of Cagliari and the terminal container of Antwerp, located in the Mediterranean basin and in the North Sea respectively. The different characteristics of the phenomenon under

study in the two European terminals were crucial in order to expand and generalise the conclusions.

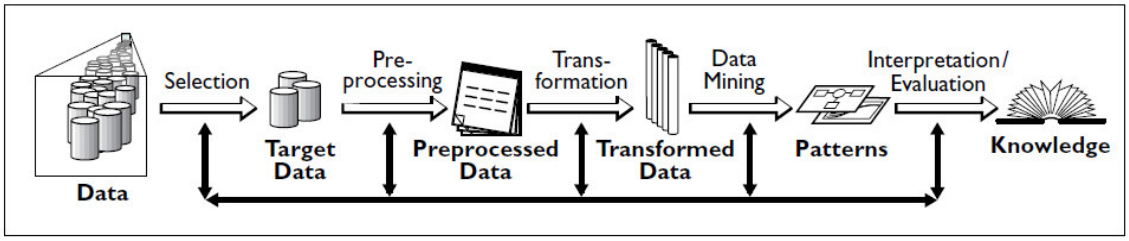
From an operational point of view the main objective of this work is to increase the probability of specifically identifying the work-shift the incoming vessel will fall within. This will reduce terminal operating costs while improving terminal efficiency and hence increasing terminal competitiveness.

1.3 Methodological approach

The basic idea of the study is to find a statistical approach that allows to make predictions of vessel arrivals in container terminals using a set of collected predictors that may affect vessel delay or advance. The study of the scientific literature showed that there is no reference model specifying the relationship between the delay/advance of vessels and the variables that are involved. For this reason, the basic idea of the research is to explore the phenomenon by collecting all the variables that may potentially influence late/early arrivals in port and then by assessing their predictive power. This last aspect required defining the methodological algorithms that are able to extract information on the delay/advance of future arrivals using historical data on previous arrivals.

Thus, the methodological approach taken falls into the Knowledge Discovery and Data Mining interdisciplinary area that focuses upon methodologies for extracting useful knowledge from data. In particular, Data Mining is one step within the Knowledge Discovery in Database (KDD) process (Figure 1.3) that allows to extract not known information from a data set and transform it into an understandable structure for further application.

Figure 1.3: Main steps constituting the KDD process



Source: Fayyad et al., 1996.

Moreover, machine learning is the name of a broad collection of computational and statistical techniques used in data mining. The classification and regression techniques used in machine learning share the idea of learning the relationship between an outcome variable Y and a set of predictors X directly from the data. This leads to a modelling approach which avoids the prior specification of the functional form between the outcome and the predictors in favor of flexible models that locally fit the data. This approach was required in this specific context where the study focuses on the prediction of future outcomes and where it is very difficult to specify the functional form linking the predictors.

The research concerns a purely operative setting where planners have to make predictions about future arrivals in order to have useful information that can be used in the daily planning decision process. Depending on the framework and planning purposes several estimates can provide useful information on vessel arrivals. Sometimes, it can be useful for planners to infer a quantitative estimate of the delay/advance in minutes, sometimes it may be useful to have a qualitative estimate, even only knowing whether or not an incoming vessel is likely to arrive before or after the scheduled ETA.

Therefore, the research focuses on learning techniques that are used for qualitative and quantitative estimates. In particular, the algorithms that are used for predictive purposes include:

- Logistic Regression;
- Classification and Regression Trees (CART);

- Random Forest.

The algorithmic models were built thanks to two different case studies. Considering the different characteristics of the outcome distribution in the two cases, the goodness of fit of the models was tested and the main variables influencing the process were identified. Final models were evaluated considering both goodness of fit and interpretability of the results.

1.4 Research structure

The structure of this dissertation is shown in Figure 1.4.

Chapter 1 introduces some general remarks as well as the research setting, the purpose of the research, the main objectives and the methodological approach that was taken.

The main characteristics of a container terminal system are briefly described in **Chapter 2** in an effort to highlight the complexity of the processes and operations involved therein on an operational level. The first part focuses on the general layout of a terminal container and the equipment that is employed. The second part briefly describes the main scheduling decisions that planners have to make for each incoming vessel.

Chapter 3 describes the problem of vessel arrival uncertainty in container terminals. The literature review focuses on three main goals: the first one is to define the problem and to understand how the scientific community addresses it. The second one is to highlight how the problem affects all terminal decisions associated with both the main planning activities and with the allocation of available resources. The third one is to identify a specific methodological approach in order to obtain accurate forecasts.

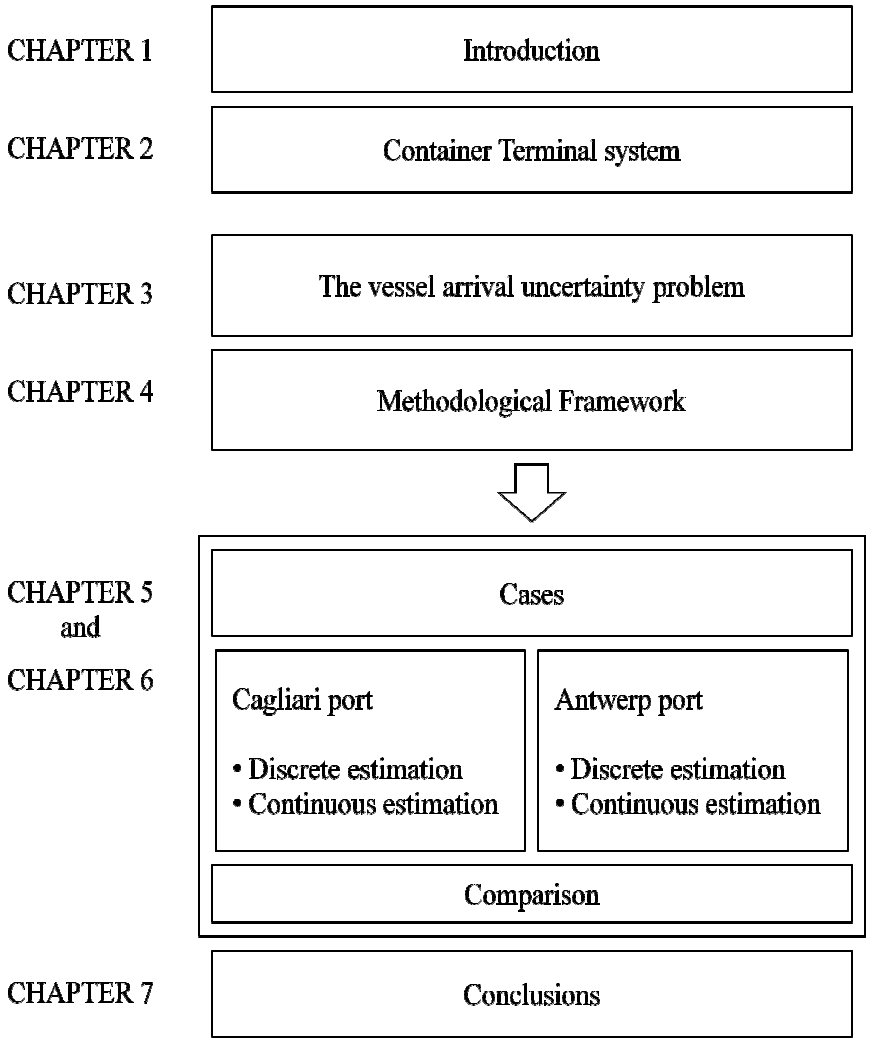
Chapter 4 provides a theoretical overview of the machine learning algorithms adopted in the KDD approach for estimating, in qualitative and quantitative terms, the late/early arrival of ships in the container terminals. Special attention is paid to the

logistic regression algorithm as well as to the CART (Classification and Regression Trees) algorithm and the Random Forest algorithm.

Chapter 5 and **Chapter 6** analyse the case studies in detail. Chapter 5 describes the case of Cagliari, a Mediterranean transhipment container terminal, while Chapter 6 concerns a transhipment container terminal in the North Sea (Antwerp). In both cases, the results regarding the discrete and the continuous estimate of vessel arrival are presented. The two sections are presented in sub-sections according to the main steps within the KDD approach that is used. Firstly, the chapters describe the examined ports, the collected data and the main characteristics of the available variables. Then the various phases involved in model application are addressed and the main results are discussed. Finally, the two applications are compared.

The Conclusions are presented in **Chapter 7,** which is the final chapter. Contributions and practical implications of the research are highlighted and future developments are proposed.

Figure 1.4: Research structure



CHAPTER 2: Container terminal system

The main aim of this descriptive chapter is to underline the main characteristics of a container terminal system and to emphasise the decisive role that vessel arrival time plays in terminal organisation, in particular in a daily planning scenario.

2.1 Introduction

A container terminal can be described in general terms as a dynamic system where a variety of handling operations is carried out for moving containers arriving and leaving the terminal by different transport modes, e.g. by ship, truck or train.

Container terminals can be classified as regional terminals or transhipment terminals, depending on the main type of service they provide. Regional terminals are terminals where freight arrives and is sorted according to mode of inland transport system. Thus, they are usually situated in strategic positions with respect to the important consumption areas. On the other hand, the main function of a hub terminal is to accommodate the larger container vessels that normally provide transoceanic services. The cargo is then loaded onto feeder vessels that are employed for short-sea shipping for delivery to the local market. An important aspect to be considered is that different vessel sizes imply different container-handling costs and organisation in the terminals.

In the last few years the share of transhipments with regard to the totality of maritime containerised traffic has grown significantly. By using an intermediate hub terminal in conjunction with short sea shipping services, it is possible to reduce the number of port calls and increase the throughput of the port calls left. Maritime shipping companies also elect for transhipment as a way to use their networks more rationally.
One important factor for determining the type of service generally provided by a container terminal is its geographical position, which makes it more or less strategically attractive for a specific function.

Summing up, both regional and hub terminals handle containers:

- arriving by ship that continue their journey inland;
- arriving overland that continue their journey by ship;
- arriving by ship that use the terminal as a sorting point for onward transportation by other ships.

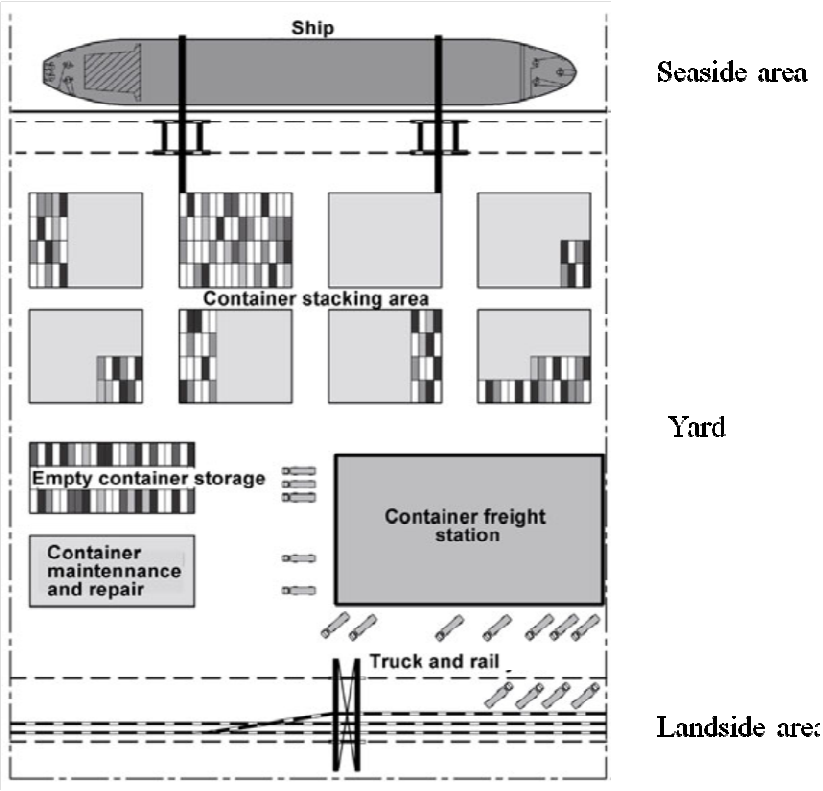
2.2 General layout and Technical Equipment

The layout of a container terminal consists of four main sub-systems (Figure 2.1) (Brinkmann, 2011, Gunter and Kim, 2006, Steenken et al., 2004):

1. the seaside area where containers are loaded and unloaded onto/from vessels;
2. the yard area (or stacking area) for storing the containers;
3. the landside area where containers are loaded and unloaded onto/off the trucks and trains;
4. the cargo handling system, that connects seaside and landside areas to the yard. The handling system consists of a variety of equipment for moving containers inside the terminal.

The yard area is usually divided into different blocks which are differentiated into rows, bays and tiers. Some yard areas are given over to special types of containers such as reefer containers, that require power supply, containers carrying dangerous goods or containers of non-standard size that cannot be stacked in the normal way. The terminal can also include other areas like parking, office buildings, customs facilities, container freight station with an area for stuffing and stripping containers, empty container storage, container maintenance and repair areas, etc.

Figure 2.1: General layout of a container terminal



Source: Brinkmann, 2011

Various pieces of technical equipment can be used for the main handling operations for: loading/unloading vessels, transporting containers to and from the different areas and stacking containers in the yard. The array of equipment used in a container terminal is commonly referred to as “Operation System”.

Depending on the handling equipment, available container terminals can be broadly divided into two main types: systems using gantry cranes for container storage and pure straddle carrier systems (Brinkmann, 2011, Steenken et al., 2004).

The choice of technology to be adopted depends on a number of factors, though primarily on the size of the yard area. However, not only space constraints but also financial and historical reasons play an important role.

2.2.1 System using gantry cranes for container storage

A container vessel that arrives in port is assigned to a berth equipped with specialised Quay Cranes (QC) to load and unload containers (Figure 2.2). These cranes are mounted on rail tracks alongside the quay. They travel back and forth along the quayside. They are equipped with a trolley that travels along the arm of the crane. Special devices attached to the trolley, called spreaders, are provided with guide brackets at the corners, flippers, for picking up containers that are then moved from ship-to-shore or vice versa. The technical performance of Quay Cranes can change between 22-30 moves per hour.

Figure 2.2: Quay Crane

The QC lowers the containers directly onto horizontal transport vehicles waiting on the quayside. A variety of vehicles can be used for this purpose. This specific operations system employs horizontal vehicles such as trucks with trailers (TT), that can transfer the containers to the storage yard but are not able to lift the containers by themselves (Figure 2.3).

Figure 2.3: Track-Trailer

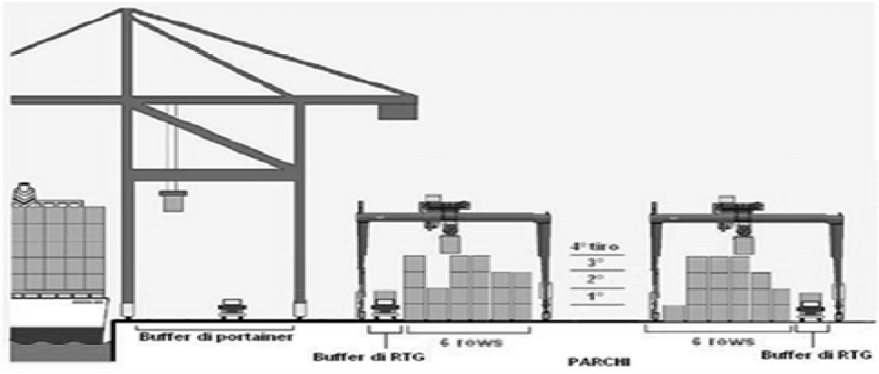
Once the container arrives in the stacking area a second type of crane, Yard Crane (YC), attached to stacks places it into position (Figure 2.4). The position of each

container in the yard is therefore identified by a block (and area if necessary), and by the row-bay-tier combination. The yard cranes are composed of a solid load-bearing structure and a mechanism comprising the cabin, a hoist and a spreader for stacking the containers in the yard. There are three different types of cranes: RMG (rail mounted gantry cranes), RTG (rubber tired gantries) and OBC (over-head bridge cranes). Each type has different characteristics: in general, RMGs are more stable and RTGs are more flexible in their operation. Yard crane efficiency is normally around 20 moves per hour.

Figure 2.4: Yard Crane

Thus, container handling is performed in a three level system: QC-TT-YC (Figure 2.5). This system is generally adopted in ports where space is limited, and containers have to be stacked in four or more tiers.

Figure 2.5: System using yard cranes for container storage



Source: Pisano, 2008

This type of system requires careful synchronisation of the horizontal transport vehicles with the quay cranes on the seaside and with yard cranes on the landside in

order to ensure that operations are executed rapidly. For example, each time the gantry crane proceeds to unload a container, the transport vehicle must be close by, along the quay. If no truck is available, then the crane remains inactive.

2.2.2 Pure straddle carrier system

The alternative Operation System comprises Quay Cranes (QC) and Straddle Carriers (SC). Thus, handling operations are performed in a two-level system (QC-SC) known as the pure straddle carrier system. With this system the containers are unloaded from the ship by the QC and are positioned onto the quay.

Straddle carriers are horizontal transport vehicles that are able not only to transport containers but also to stack them in the yard up to 3 or 4 tiers (Figure 2.6). They are very flexible and dynamic systems that have free access to containers regardless of their position in the yard.

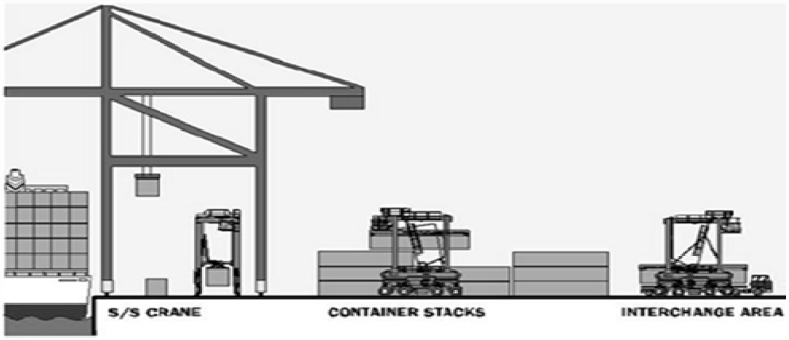
Figure 2.6: Straddle Carrier

This handling system is usually used in more spacious terminals. In fact, the yards where straddle carriers are employed are set up such as to leave enough space between two rows of containers for the vehicles to operate easily. Furthermore, the blocks are spaced in such a way as to enable two straddle carriers to travel and maneuver safely at the same time.

In addition to the straddle carriers, other horizontal transport vehicles, such as forklifts and reach stackers are used for handling the containers. They are generally used to move and stack light containers, in particular empty ones. The disadvantage of these vehicles is that they operate at low speed.

A diagram of the second handling system is shown in Figure 2.7.

Figure 2.7: Pure straddle carrier system



Source: Pisano, 2008

As is clear from the above described general layout, a container terminal is a complex structure. Terminal operations need to be efficiently executed and properly coordinated since any loss of efficiency in one operation can have deleterious repercussions on all the others, thus drastically reducing terminal productivity.

2.3 Port performance

Productivity is one of the basic elements for assessing terminal performance. In particular, productivity is affected by a number of factors (Figure 2.8). The interconnection among all factors, along with numerous other exogenous factors, makes performance assessment a highly complex issue.

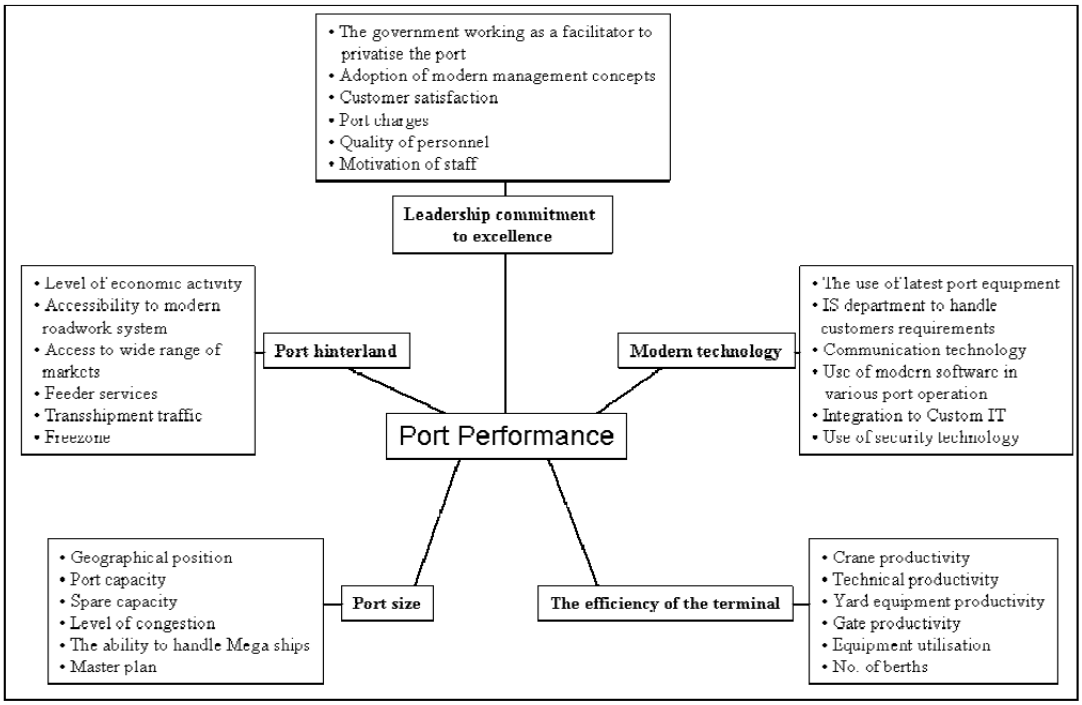
Shipping companies, which are the main clients of container terminals, usually perceive the quality of the services that are provided by considering:

- the ability to offer a high level of service, which translates primarily into the total duration of the vessel time in port;
- the total amount of handling charges;
- the location of the terminal with respect to shipping routes;
- the connection of the terminal to intermodal networks.

To achieve a high level of service, terminal operators constantly seek out different strategies that are able to enhance port productivity (Soberon, 2012), particularly in terms of:

- capacity, both physical (availability of space, length of berth,..) and operational (congestion);
- efficiency of equipment and operators' skills;
- level of technology adopted in handling equipment and information management.

Figure 2.8: The key factors in port performance



Source: Ferraro, 2006

Over the last twenty years, in response to the increasingly frequent requests by terminal operators having to continually cope with new and complex problems, a variety of statistical, simulation and information tools have been devised and tested. These have been successfully implemented in operations planning and management in numerous terminals, thereby increasing productivity significantly (Steenken et al., 2004).

Efficient scheduling of activities plays a decisive role in enhancing efficiency within a port terminal. The support of advanced instruments is now becoming indispensable for providing high quality services that allow terminals to reduce their operating costs and to maintain or enhance their productivity targets (Steenken et al., 2004; Vis and de Koster, 2003).

2.4 Main planning activities

Three different levels can be identified within the planning process of a container terminal: strategic, tactical and operational levels. These differ both as to the type of decision to be made and to the time horizon that is covered (Ghiano e Musmanno, 2000).

The strategic level refers to long-term decisions. Decision making at the strategic level mainly concerns the infrastructural (layout, handling equipment, berth and yard capacity,..) and economic (contracts with shipping lines,..) aspects of the terminal. The time horizon for decisions that are made at the strategic level may involve several years. These decisions lead to the definition of a set of constraints under which the decisions at the tactical and operational levels have to be made.

The tactical level refers to medium-term decisions and involves berth and yard operations planning. The time horizon in this case usually covers a one-month period.

The operational level refers to short-term decisions and involves the decisions pertaining to quayside and landside operations, on the basis of choices made at the tactical level. The operational planning involves more detailed manpower and equipment allocation with a view to maximising productivity while minimising costs. The time horizon is around 24 hours.

The research that was carried out refers to the operational level, where the decisions that are made have a very short term impact. The major decision-making problems involved for each incoming vessel are summarised as follows (Table 2.1):

Table 2.1: The main decision-making problems in Container Terminals

	General Layout			
Strategic level	Berth dimensioning	Quay crane selection	Handling equipment selection	Yard crane selection
Tactical level	Berth allocation and scheduling	Crane assignment	Transport vehicles assignment	Storage and stacking policies
Operational level		Stowage planning Crane scheduling	Routing and scheduling of vehicles	Block scheduling
	Berth allocation	Loading/Unloading	Internal Transport	Yard stacking

Comprehensive overviews of the main decision making issues at container terminals are described by Bierwirth and Meisel (2010), Stahlbock and Voß (2008), Vacca et al. (2007), Steenken et al. (2004), Vis and de Koster (2003).

This descriptive chapter has been developed in order to emphasise and underline the complexity of the main processes directly related to each arrival. Moreover, each arrival produces high peak loads for other terminal activities as well as for support activities upon arrival and hinterland transportation, in particular at the operational level.

2.4.1 Berth planning

The berth planning process is the process whereby berths and processing times are assigned to vessels arriving in port to be loaded/unloaded.

The purpose of the berth allocation is to optimise the berth utilisation and to minimise the total vessel turnaround time, reducing the distance of each container from its origin/destination parks in the yard as much as possible. Since the arrival of vessels in a container terminal follows a predetermined cycle, it is possible to plan berth allocation over the medium-term period.

2.4.2 Loading/unloading operations

The ship loading/unloading process can essentially be considered as being made up of various sub-processes: stowage planning, quay crane assignment and quay crane scheduling.

When the vessel is berthed, at the tactical level the number of QCs required to simultaneously load/unload the ship has to be determined. While, at the operational level, once the stowage plan has been drawn up, the next step is to efficiently coordinate the QC operations and to distribute the workload among them in order to reduce the amount of time required to complete handling operations. Despite being an operational issue, the stowage plan is drawn up by both the shipping company and the vessel planner. From the terminal operator's point of view, the objectives to be optimised differ significantly from those of the shipping company. In fact, while the aim of the shipping company is to maximise capacity utilisation while minimising the number of movements required for loading/unloading the containers in ports, the aim of the terminal operator is to complete the loading/unloading process as quickly as possible. A number of operational constraints have to be observed during this process.

2.4.3 Transport of containers

Organising the internal transport of containers is a rather complex process. The horizontal transport vehicles have to be synchronised with the cranes. The type of vehicle to be used is a strategic decision and depends on yard configuration. At the tactical level, the decision to be made concerns the number of vehicles that are required. At the operational the vehicle routing and scheduling needs to be decided.

Horizontal transport vehicles can be allocated in different ways. One solution consists in assigning to each quay crane a predetermined number of vehicles, that operate synchronously with the crane. However, this solution presents two major disadvantages. The first is that crane and vehicle productivity are closely related: any pauses in crane operation will necessarily slow down the vehicles assigned thereto and vice-versa. The second concerns the travel time of each vehicle. Every time a vehicle moves with a full load, it has to return empty in the opposite direction. An

alternative solution consists in allocating a group of vehicles to two or more adjacent cranes, performing both loading and unloading operations, thereby substantially reducing the number of empty journeys.

2.4.4 Yard stacking

The yard may reasonably be considered the heart of the container terminal. The problems involved in yard management differ depending on the type of container traffic and on how the yard is set up. Depending on operating requirements, the yard is usually divided up into different areas. Each area is split up into blocks and each block is arranged by bay, row and tier.

Storage and stacking policies are determined at the tactical level, while at the operational level the decisions involve determining the specific place in the storage area for each container and the specific number of Yard Cranes that are needed in order to ensure an efficient storage process.

Before loading the containers on board the vessels, they can be transferred to the quayside, in accordance with the stowage plan. Repositioning containers in different yard areas is known as *housekeeping* operations. Although these operations are not carried out during peak workloads, they are nonetheless time consuming and labor intensive.

At the tactical level, yard areas can be set up in two different ways: either by assigning a specific area to a specific ship or alternatively by assigning each berth its own yard area.

2.5 Operations centre

All the operations conducted within a container terminal are set up and controlled in real time by the O*perations Centre*, under the supervision and coordination of the Planning Managers. The work conducted in each office of the Operations Center is closely related and the offices exchange information on the basis of variable schemes.

In order to schedule operations, the shipping company must provide information for each vessel arriving in port. A documentation office exists for maintaining customer contacts and its work consists in gathering the necessary information and passing it on to the competent offices in the Operations Center to ensure the functional organisation of the seaside, yard and landside areas.

In particular, the most important information that has to be received is the vessel Estimated Time of Arrival (ETA) in port within the port rotation schedule. This information is updated several times before the vessel arrival date and the reliability of the information improves as the arrival date approaches.

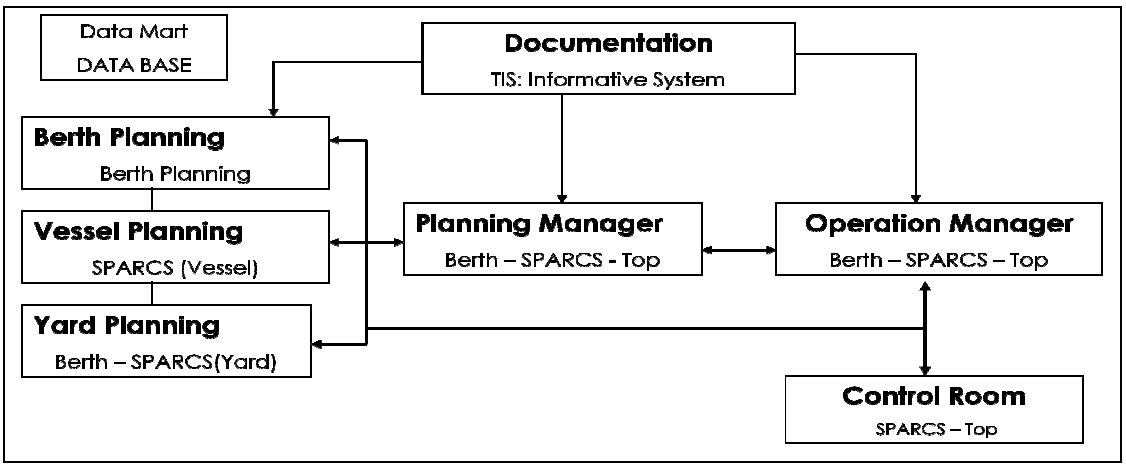
The other main information that is acquired consists in:

- the BAPLIE, which provides the characteristics of the containers on-board, indicating the ones that need to be unloaded, the ones that are to remain on board and the ones that have to be restowed;
- the inbound COPRAR, which contains the BAPLIE as well as the final destination of the containers to be unloaded the next port of call and the ship they are to be loaded onto;
- the outbound COPRAR, which contains the container load list for outbound vessels;
- the MOVINS that contains the loading instructions.

The submission times of the above documentation differ depending on the services that are requested. Once this information has been received, the relevant planning decisions can be made by the competent offices.

The following graph summarises the offices that are responsible for planning (Figure 2.9).

Figure 2.9: Operations Center structure observed at Cagliari container terminal



2.6 Standard gang composition

Putting into practice any decisions made by the planners requires a sufficient numbers of workers and equipment. A standard gang is defined as a team of human and associated handling equipment generally composed of:

- concerning the workers:
  - one quay crane driver;
  - one deck man;
  - two twist handlers;
  - one checker;
  - three track drivers;
  - one to three yard crane drivers;
  - three to eight lashing unlashing operators.

- concerning the equipment:
  - one quay crane;
  - three track trailers;
  - one to three yard cranes.

Each gang has to be assigned to one or more working shifts. In order to offer a good level of service, the number and composition of the gangs must be appropriate at all

times. Because of the uncertainty of arrivals in the medium period, the schedule for each day assigns “fixed” gangs to one specific shift and “flexible” gangs to a shift to be decided during daily scheduling once the actual time of arrival has been determined with greater certainty. Contractual terms for flexible shifts specify that a flexible worker will be assigned to a shift with only 24 hours notice. In the event demand requires additional manpower, then external workers can be hired. The cost of the gang differs depending on the specific working shift, for example shifts at night and on Sundays are more expensive. Since the cost of hiring external manpower is high, it should be avoided as much as possible. The operation managers as well as the control room operators may decide to increase those resources when necessary.

2.7 Importance of reliable vessel arrival time

As can be seen from the above description, the various levels of planning suffer from the same critical issues:

- temporal fragmentation and uncertainty of received information
  Information is received at different, undefined times and has to be continuously updated; thus the information is uncertain.

- complexity of the planning processes
  A multitude of decisions have to be made, with different though closely interconnected characteristics. This means that a large number of variables and constraints have to be taken into account.

Vessel arrival uncertainty plays a decisive role in terminal organisation. To know in advance the effective time of vessel arrival in port with greater precision would allow terminal operators to assign all the resources required for handling operations more accurately, avoiding under/over manning at the planning stage. This is especially important in short term forecasting, the most crucial time period for successful management in terminal operations (Fadda et. al., 2014, Fancello et al., 2011).

Moreover, disruptions in container flows and operations caused by vessel arrival uncertainty can have cascade effects within the overall supply chain within which the

port is part. The unreliability of vessel arrival time can affect both hinterland transport and logistics costs (Chung and Chiang, 2011) due to the high correlation among the various segments of the chain (Sciomachen et al., 2009, Wang and Cullinane, 2006). This aspect can incur additional operating costs for the shipping lines related to unproductive vessel time and the rescheduling of vessel operations (Vernimmen, 2007). Furthermore, additional logistics costs to the customers can be mainly caused due to additional inventory and production costs resulting from, for exemple, a late delivery of materials (Notteboom and Rodrigue, 2008, Notteboom, 2006).

From the above considerations, knowing the actual time of vessel arrival in port in advance would substantially reduce operating costs while enhance the efficiency of the services that are provided. Lower operating costs combined with increased productivity would enhance the competitiveness of the terminal and the supply chain as a whole.

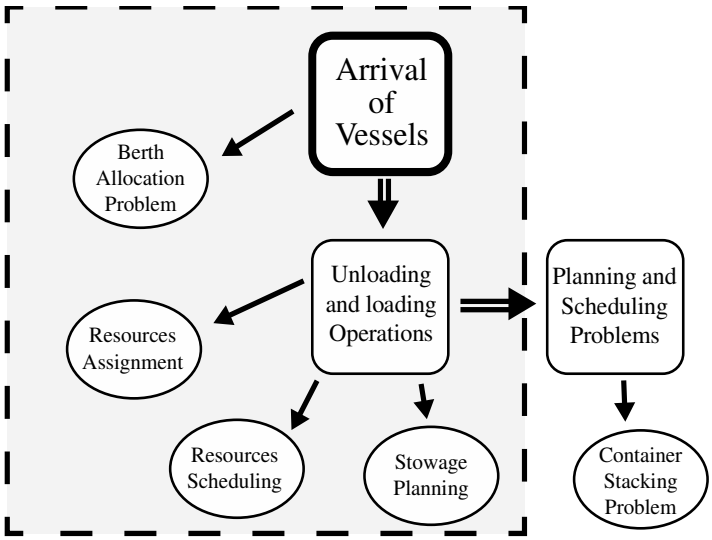
CHAPTER 3: The vessel arrival uncertainty problem

This Chapter describes how the scientific community addresses the problem of vessel arrival uncertainty in container terminals. Although it is a recognised problem in the maritime sector, there are no reference models that provide planners with objective and analytical answers. Therefore, one of the specific aims of the study is to find a methodological approach that may help fill this gap in the literature.

3.1 Introduction

The analysis of the scientific literature confirmed that a container terminal is a complex system in which a variety of different complicated, inter-related problems coexist, most of which require integrated solutions. Thus, the solution to one problem often becomes decisive for the other problems related thereto (Salido et al., 2011, Won and Kim, 2009, Murty et al., 2005). The issues discussed in the previous chapters show that the main planning processes depend strongly on vessel arrival, the first of the five major logistic processes in container terminals (Stahlbock and Vob, 2008, Vis and De Koster, 2003) (Figure 3.1).

Figure 3.1: Major logistic processes container terminals



Source: based upon Salido et al., 2011, p. 5.

On a daily basis the actual time of arrival remains uncertain. Shipping lines are contractually bound to notify their Estimated time of Arrival (ETA) at predetermined time intervals. The operator, agent or master of a ship shall notify to the port authority the last ETA at the port of destination or pilot station[1]:

- at least twenty-four hours in advance or;
- at latest, at the time the ship leaves the previous port, if the voyage time is less than twenty-four hours;
- if the port of call is not known or it is changed during the voyage, as soon as this information is available.

Despite contractual obligations, ship operators often cannot comply with the declared last ETA and they have to update it due to unforeseen circumstances. Vessels that serve a specific service may arrive in port at different times from week to week due to unexpected events like weather or sea conditions, waiting times or delays that may have occurred in the previous port, and so on. Thus, in the short-term planning scenario, the uncertainty surrounding vessel arrival time in port persists. The actual times of vessel arrival in port are extremely uncertain: only half of the vessels arrive on time and many vessels do not arrive on the scheduled day (Drewry Shipping Consultants, 2008). This aspect strongly affects the number of ships that may have to be handled/operated concurrently. Moreover, each arrival produces high peak loads for other terminal activities as well as for support activities upon arrival (pilotage, towage, etc.) and hinterland transportation (ensuing waiting time , congestion etc.). Deviating arrivals can only exacerbate this peak load.

However, some ports are equipped with information systems that indicate the position of the incoming vessel in real time. It can be useful for port operators to have instruments that provide reliable information about vessel arrival some time in advance. Thus, the instruments can support planners in a daily planning scenario by allowing them flexibility when making decisions regarding terminal operations management.

[1] article 4 of the Directive 2002/59/EC of the European Parliament and of the Council of 27 June 2002.

To date, this task has been assigned to the planners that operate mainly on the basis of their personal experience. However, for planners it is difficult to manage the vessel arrival uncertainty problem, in particular because many variables and constraints can affect the process. The main factors which may affect arrival time can be summarised as follows:

- Vessel features: length, draft, gross tonnage, capacity, etc.;
- Vessel service: sailing direction, port rotation, etc.;
- Organisation of previous port: congestion, distance from the port of interest;
- Vessel owner;
- Type of containers to be loaded/unloaded;
- Weather conditions;
- External factors like strikes, breakdowns, etc.

Thus, although vessel arrival uncertainty in port is a well known problem for port operators and for the scientific community, the specific instruments for dealing with this problem are highly limited and there is no standard definition or way to measure vessel delay/advance in ports.

This problem was raised by Fancello et al. (2011). They presented a decision support system for supporting port operators in day to day management that was made up of two different modules: a forecasting module and a human resources optimisation module. Using a neural network algorithm, the first module faces the problem of handle late arrivals in a Mediterranean port.

The nearest contributions in the maritime sector concern the estimates of container throughput in a daily time horizon in order to provide reliable input data for correctly scheduling handling operations. Gambardella et al. (1996) proposed a forecasting module for estimating the daily container flow in and out of a terminal container, which combines two different estimators. The first predicts the number of expected 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 expected to enter the terminal by truck, as a function of time till the ship's ETA. Sideris et al. (2002)

developed a tool for predicting daily demand variations in terms of the number of containers that are moved through a terminal. In particular, they developed a static modelling approach in order to estimate the daily percentages of container movements. They built two different models in order to calculate the percentage of the total import/export container arrivals/departures on a specific day. With the first model they considered historical data based on the previous arrivals/departures, while with the second model they readjusted the estimate, using on-line information regarding the progress of the arrival/departure process in the terminal. Chou et al. (2008) proposed a regression model for forecasting volumes of Taiwan's import containers, and Chen and Chen (2010) attempted to forecast container throughputs at Taiwan's major ports using genetic programming.

3.2 Related planning activities

Vessel arrival uncertainty may, for instance, require additional handling operations within the terminal. Therefore, the problem is commonly studied in conjunction with the related processes. Knowing the possible deviation from the scheduled arrival time in advance can 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.

A review of the literature highlighted that punctuality of the vessel's arrival commonly effects:

- Berth scheduling;
- Human resources and equipment allocation;
- Yard planning.

3.2.1 Berth scheduling process

In general, berth spaces are allocated so as to reduce vessel loading/unloading times and the distance from the origin/destination container areas in the yard. Often, in the event of ship delay, as the containers to be loaded onto the vessel have already been moved to the stacking yard, a remarshaling plan is needed to minimise berthing time (Zhen et al., 2011, Salido et al., 2011).

Moorthy and Teo (2006) published one of the earliest studies on berth plan template associated with vessel arrival uncertainty. They propose a sequence pair based simulated annealing algorithm to solve the problem. The results they obtained show that the proposed methodology is able to construct an efficient template for transshipment hub operations. Du et al. (2010) extended Moorthy and Teo's solution method and introduced a feedback procedure to the berth allocation problem with stochastic vessel delays. In this procedure earlier iterations generate feedback to the model to adjust the time buffers for the future iterations.

Hendriks et al. (2010) studied the berth plan template problem jointly with the quay crane reservation problem. They assume that the number of reserved quay cranes for a vessel is a function of the punctuality of the vessel's arrival, and further assume that the number of reserved quay cranes is proportional to the cost of operating the berth. They propose a mixed integer linear program in order to identify a robust berth plan that minimises the crane reservation. Han et al. (2010) considered vessel uncertainty to berth assignment and quay crane sequencing problems simultaneously, and developed a stochastic mixed integer program. They then solved it with a genetic algorithm.

Zhen et al. (2011) developed a two stage model to solve the berth allocation problem under uncertain arrival time. The aim of the study was to calculate the cost associated with the initial schedule and the expected costs of deviation from the initial schedule due to late arrivals. Moreover, depending on the contract with the liner, the yard may start receiving export containers for a vessel more than one week before the vessel arrival. One of the main goals in berth planning is to ensure that the location of containers in the yard is as close as possible to the berth location. When vessels that were scheduled to use the same berth have an overlap (because one of them or both

did not arrive at the expected time) there is a berth conflict: the berth location becomes dependent on the arrival time (Bruggeling et al., 2011)

Ambrosino and Tanfani (2012) used an integrated simulation-optimisation approach to solve the Quay Crane Assignment problem and the Quay Crane Scheduling problem on the operational level. They use an optimisation model to assign the number of quay cranes to the various bays of each ship served by the terminal for each work-shift. The results of this first model are used as input for a simulation model that is able to reproduce the system's behavior. The developed model is able to consider some causes of variability in the system including unplanned delays.

3.2.2 Human resources and equipment allocation

It is also important to manage vessel arrival uncertainty in order to determine the number of workers required for handling operations (Di Francesco et al., 2013, Gambardella et al., 1998). This is a major issue in a port system where the cost of manpower is high. In general, mid-term scheduling ensures a gang is available for each working day, in conformity with shift arrangements as well as contractual obligations and labor regulations. Because of the uncertainty of arrivals in that period, the schedule for each day assigns "fixed" workers to one specific shift and "flexible" workers to a shift to be decided during the short-term scheduling i.e., once the arrivals time are known with greater certainty. Contractual terms for flexible shifts specify that a flexible worker will be assigned to a shift with only 24 hours advance notice. Sometimes it is also possible to activate additional contracts with external workers or to assign a worker two consecutive shifts within the same day. A "double" shift can be assigned once or at most twice a week. The additional shift is assigned within the daily scheduling (Fancello et al., 2011, Legato and Monaco, 2004). Thus, the authors underline that the main difficulty in the daily planning process comes from the inherent and unavoidable uncertainty of workforce demand due to uncertainty in arrival times.

Moreover, it is important to know the effective arrival time of vessels in order to optimise equipment management for handling operations and for establishing

maintenance schedules. This aspect is essential for ensuring the availability and functionality of the handling systems that are used (Fancello et al., 2010).

3.2.3 Yard planning process

The uncertainty of the vessel arrival schedule is one of the main factors that may impact the efficiency of yard operations. This is why it is also studied in conjunction with the yard planning process. One problem that may arise when a vessel is late is that a transshipment link can broken: containers scheduled to be unloaded and then loaded onto another vessel may be stocked in the yard because the other vessel has already departed. Thus, the yard space that was supposed to be free will be occupied for a longer time. Moreover, when the vessels do not arrive at the expected time, there is a chance that two vessels will be close to each other, and they have to handle the containers simultaneously (Bruggeling et al., 2011). Terminal operators can then choose to de-conflict by moving some of the containers to another location before the arrival of the vessels, but that will mean incurring extra costs. However, if they are not moved, it will result in a concentration of activities and likely contention for yard cranes.

The first paper that directly addresses the yard planning problem with respect to vessel arrival uncertainty was written by Ku at al. (2012). They found a yard template that can be modified in the event of changes in service arrival schedule. Many container terminals adopt the yard planning strategy whereby containers to be loaded onto the same vessels are stacked in groups. Due to this consolidated strategy, a change in vessel arrival schedule may cause congestion of trucks at yard locations where groups of containers in the vicinity are loading simultaneously.

In conclusion, the state-of-the-art and the discussions with planners showed that although it is a recognised problem in the maritime sector, vessel arrival uncertainty still remain a challenge for port operators.

However, in recent years, arrival uncertainty has been the topic of several studies in the air transport sector.

3.3 Arrival uncertainty 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 in order to identify the causes behind flight delays in the U.S. National Airspace System (NAS). Aircraft delay is defined as the difference between the Estimated and the actual time of an aircraft's departure or arrival.

Mayer and Sinai (2003) examined two potential factors that might explain the extent of air traffic delays in the United States: network benefits from hubbing and congestion externalities. They used data at the individual flight level on all domestic flights by major US carriers and collected data on more than 66 million previous flights.

Wesonga et al. (2012) used a logistic model in order to determine the daily probability of aircraft departure and arrival delays. They examined the causes of flight delays and cancellations in the American National Airspace System. All flights that arrived or departed earlier than expected were considered as being on time. The study analysed ground delays and air holding delays at Entebbe International Airport during 1,827 days of activity between 2004 and 2008. The overall average probability for departure and arrival delay over this period was estimated by including the meteorological and aviation parameters while computing the exact probability of delay.

Tu et al. (2008) presented a model for estimating flight departure delay distributions. The study focused on data collected at Denver International Airport during a two-year period. The purpose was to identify the main factors influencing flight departure delays, and to develop a strategic departure delay prediction model.

Some other machine learning models have been used to describe flight delays, such as Ning Xu's Bayesian network model (Xu, 2007) and Zonglei Lu's decision tree model (Zonglei et al., 2008). To date, there is no existing model that is able to accurately predict flight delays. The existing models only provide some reference of the prediction.

Zonglei et al. (2008) used a machine learning approach in order to predict the daily level of delay in the hub-airport of China. The first step involved cluster analysis to define five levels of daily alarm related to flight delays at their study airport using data recorded in 2006. Then they compared three techniques (Naive Bayes, Decision Trees and Neural Network) to obtain an estimate of the alarm level for each day of the airport activity.

3.4 Required approach

In the maritime sector there are no reference models that specified the relationship between vessel arrival uncertainty and the variables that can affect the process. Therefore, one of the aims of the study is to find a statistical approach that may help fill this gap in the literature.

Overall, the literature review shows that there are two 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). Algorithmic modelling, both in theory and practice, has developed rapidly in many fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modelling on smaller data sets.

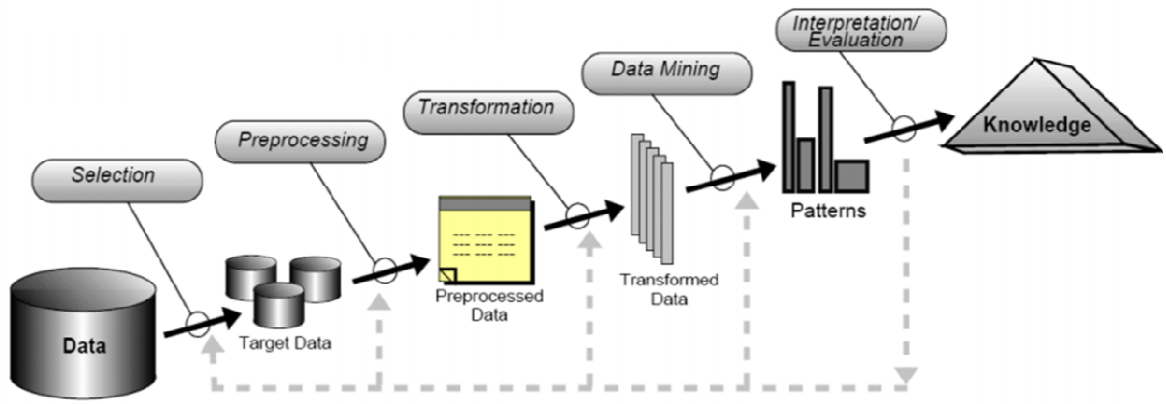
Moreover, when major irregularities are present, or for discontinuous and irregular time series, the interesting contributions to the literature by numerous authors have demonstrated that traditional approaches perform poorly. Traditional statistical models do not perform well when data have to be continuously updated. The fitting procedures are moreover difficult to implement by non-experts. Concerning dynamic learning models (Carbonneau et al., 2007, Potvin and Smith, 2001), the forecast update procedure is more robust and can be carried out in a more flexible manner.

Thus, the approach that is taken falls within the machine learning discipline that focuses upon methodologies for exploring and understanding historical arrivals. This approach seems to be particularly appropriate in this specific instance where the

functional form between the vessel arrival uncertainty and the potential predictors is not specified (Breiman, 2001). In fact, 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 ETA and the recorded ATA will go on to form an historical knowledge base upon which the model is built.

The machine learning theory and methods refer to the broader interdisciplinary area of Data Mining. 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). Extracting useful information from data requires referring to data mining within the KDD multi-step process (Figure 3.2), defined as the extraction of potentially useful information from data (Fayyad et al., 1996, Frawley et al., 1992). Data mining is the core of the knowledge discovery process and it refers to a specific step in the process. The other steps in the KDD process, such as data selection, data cleaning, data preparation, and correct interpretation of the results, are mandatory to be able to extract information from data (Fayyad et al., 1996).

Figure 3.2: Different steps in the KDD approach



Source: Fayyad et al., 1996

Although the other steps are also very important for the successful application of KDD, the data mining component has received the greatest amount of attention by researchers in the literature.

In general, this step consists of a mixture of three main components: the model, the preference criterion and the search algorithm. The literature of learning algorithms often does not focus on the description of these three aspects, but they are often included in a description of a particular algorithm.

A wide variety and number of data mining algorithms have been described in the field of machine learning.

3.4.1 Classification methods in supervised learning

Learning problems can be classified as supervised or unsupervised. Classification methods are used in the setting of supervised learning. The literature highlights three different approaches that can be taken in classification problems: the discriminative approach, the regression approach, or the class-conditional approach.

The discriminative approach (neural networks, support vector machines,..) attempts to directly map the explanatory variables $\mathbf{X_i}$ to one of the **k** possible target categories $y_1, y_2, \ldots y_k$. The input space is partitioned into different regions which have a unique class label assigned to them.

The regression approach (logistic regression, decision trees,.) calculates the posterior class distribution P(Y|x) for each case and chooses the class for which the maximum probability is reached.

The class-conditional approach (Bayesian classifiers) starts by explicitly specifying the class-conditional distributions $P(X|y_i,\theta_i)$. After estimating the marginal distribution P(Y), Bayes rule is used to derive the conditional distribution P(Y|x). Parametric, semi-parametric, and non-parametric methods can be used to estimate the class-conditional distribution.

There is no general rule regarding which approach works best, it is mainly related to the researcher's goal and to data characteristics.

3.4.2 Regression approach in classification problems

In this specific application a regression approach is taken. First of all because compared with 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 outperform Neural Networks (NNs) for this specific case. Decision tree patterns can be applied to larger data problems and are able to handle smaller data sets than NNs (Markham et al.,. 2000). Moreover, they perform better than NN models when data sets are smaller with large numbers of irrelevant attributes (Brown et al.,. 1993). Decision trees can be used either as prediction tools or as exploratory tools. They aim to identify which class of a response variable the data records belong to, knowing the values or the categories of one or more explanatory variables. The recursive algorithm splits data by applying a depth-first approach (Hunt et al., 1966) or a breadth-first approach (Shafer et al., 1996) until all records are classified. Data are split at each step using impurity measures (Quinlan, 1993). The decision tree structure consists of a root, a number of non-terminal nodes and terminal nodes (leaves). The obtained model 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 contrary, tree pruning follows a bottom-up approach. In this phase the algorithm minimises over-fitting, thus 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 (Morgan and Sonquist, 1963), in which the outcome variable is quantitative. Several other algorithms followed, including Exploration of Links and Interaction through Segmentation of an Experimental Ensemble, ELISEE (Cellard et al., 1967) and THeta AID, THAID (Morgan and Messenger, 1973) for categorical response variables, and MAID (Gillo, 1972) for quantitative response variables. Numerous algorithms were later developed such as CHi-square Automatic Interaction Detection, CHAID (Kass, 1980),

Classification And Regression Trees, CART (Breiman et al., 1984), ID3 (Quinlan, 1986) and C4.5 (Quinlan, 1993). Some authors have proposed variations to the CART method that develop non-binary trees (Loh and Vanichsetakul, 1988) or that reduce computation time (Mola and Siciliano, 1997). Among 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 a regression tree in which the response variable is quantitative. Therefore, it is especially indicated in this research where the aim is to have a discrete and continuous estimate of the vessel arrival uncertainty.

Over the last few years, decision tree algorithms have been improved and new models embodying this approach have been developed. Many hybrid approaches have also been developed. In 2002, Conversano proposed the Generalized Additive Multi-Mixture Models (GAM-MM) using the decision tree 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 Horton et al. (2006). In order to improve the accuracy of traditional decision tree methods, these algorithms 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 (2001).

CHAPTER 4: Methodological framework

This Chapter gives a theoretical overview of the learning techniques adopted within the KDD approach in order to obtain qualitative and quantitative estimates of late/early arrivals in container terminals. Moreover, the main statistical measures used for performance evaluation are described.

4.1 Knowledge discovery in databases approach

According to Fayyad (1996), these specific approach implemented, which is based on the KDD process, involved six main steps that can be summarised as:

1. Learning the application domain, a step that was needed in order to understand the context, the relevant prior knowledge and the goals of the application;
2. Data selection, in order to create the data set, or to focus the analysis on a limited subset of data that had to be explored;
3. Data preparation, that comprised basic operations in order to clean and prepare the data so as to be in the most suitable form for use;
4. Data mining, the core of the KDD process. In this step the purpose of the model derived by the machine learning algorithm(s) was decided. After defining the function of data mining, the specific algoritm(s) to be used to search for patterns within the data were chosen;
5. Evaluation and interpretation of the results included interpreting the discovered models and, if necessary, going back to any of the previous steps. In this step a graphical visualisation of the results was very helpful;
6. Using the discovered knowledge, consisted in incorporating and using the knowledge discovered in order to draw the conclusions.

4.2 Algorithms

This section presents the algorithms that are used to exploit information on past vessel arrivals in order to predict the future arrival status of a vessel in port.
In the following, it is assumed to have a dataset consisting of an $n \times p$ matrix where:

- n is the number of independent observations;
- p is the number of variables recorded for each observation.

In particular, Y denotes the target variable (the continuous or binary outcome measuring the delay/advance of the ship) and X denotes the vector of quantitative and qualitative predictors: $X=(X_1, X_2,\ldots,X_k)$.

The algorithms used to obtain predictions are explained below:

- Logistic Regression;
- CART;
- Random Forests.

4.2.1 Logistic Regression

Logistic regression is the standard way of modelling binary outcomes, i.e., outcomes that can assume only two values, zero or one. It assumes that the conditional probability of Y_i being one can be modeled as:

$$\Pr(Y_i = 1 \mid X) = \frac{\exp(\sum \beta_i X_i)}{1+\exp(\sum \beta_i X_i)} \quad (1)$$

Where:

- Y is the outcome. The dependent dummy variable (Y_i) is zero if a given vessel arrived earlier than the expected ETA and one if is delayed;
- X denotes the vector of input variables: $X=(X_1, X_2,\ldots,X_k)$;
- the beta coefficients are usually unknown and must be estimated from the data. When β_i is positive it implies an increasing rate. When β_i is negative it implies a

decreasing rate. When $\beta_i = 0$ it would mean that the delay/advance of a vessel is independent of X_i.

The fitted model can be used to obtain predictions for new cases. The role of a variable in explaining the outcome can be evaluated using classic statistical test theory, for example, testing the null hypothesis that the jth coefficient is zero provides an easy way to assess the strength of the link between Xj and the outcome variable.

4.2.2 Classification and Regression Trees (CART)

CART models (Breiman et al., 1984) can be considered local models in the sense that they indirectly specify different conditional distributions of $Y|X_i$, depending on the region of the covariate space where unit *i* lies. This is in contrast with the global relationship imposed by classic modelling strategies and allows for greater flexibility. On the other hand, this localisation makes it more difficult to assess the global explanatory power of the predictors.

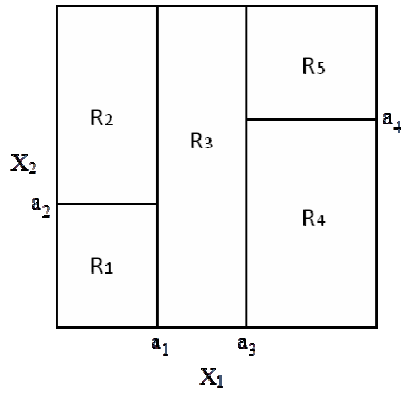
Decision trees are considered powerful tools for extracting meaningful patterns from data sets whose records are characterised by a dependent variable and a set of explanatory variables (Hastie et al., 2009). These trees attempt to classify unknown records using the obtained patterns. The algorithm recursively 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.

Assume, for graphical reasons, there are only two explanatory variables, X_1 and X_2 as in Figure 4.1. The first step consists in splitting the feature space at $X_1 = a_1$. Then the algorithm splits region $X_1 \leq a_1$ at $X_2 = a_2$ and region $X_1 > a_1$ at $X_1 = a_3$, and lastly, 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 corresponding regression model predicts Y with a constant value (c_k) in region R_k.

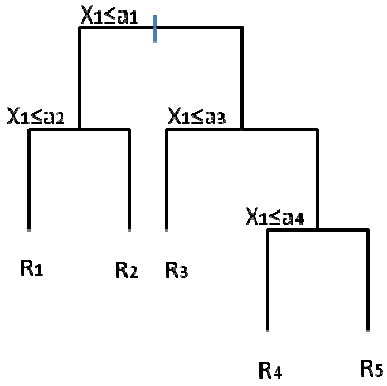
$$f(x) = \sum_{m=1}^{5} c_m \, I(X_1, X_2) \in R_m) \qquad (2)$$

Figure 4.1: A two-dimensional feature space partitioned by recursive binary splitting



The same model can be represented by a binary tree (Figure 4.2).

Figure 4.2: Binary tree corresponding to the partitioned feature space



The complete data set is located at the top of the tree. Records satisfying the condition at each subnode are assigned to the left branch and the others to the right one. The terminal nodes (leaves) correspond to regions R_1,R_2, R_3,R_4, R_5.

The algorithm works in the same way when there are more than two explanatory variables. Among all decision tree algorithms, CART, the Classification And Regression Trees (Breiman et al., 1984) is considered a landmark.

Using this method it is possible to distinguish between regression trees and classification trees i.e., they are Regression trees when the response variable is numerical, and classification trees when it is categorical.

4.2.2.1 Regression Trees

The regression tree algorithm involves two main 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 recognised patterns to other data, as large trees give unsatisfactory results when applied to new data. Moreover, an oversised tree contains a large number of terminal nodes, making its interpretation difficult and running the risk of over-fitting. Data are usually divided into two subsets: a training set and a test set . The training set is used for tree growing, while the test set is used for tree pruning in order to select the optimal tree.

Maximal tree construction. Assume there is a data set with p explanatory variables X, one dependent variable Y and N records.

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,p} \end{bmatrix} \boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad (3)$$

Considering the training set, the algorithm splits the feature space into several regions, for instance into M regions, by selecting the explanatory variables and split-points. In each region the algorithm models the response variable as a constant c_m:

$$f(x) = \sum_{m=1}^{M} c_m \, I(x \in R_m) \quad (4)$$

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 measurement. The most common measurements of homogeneity for Regression Trees are variance and entropy.

The best value of c_m, that minimises the sum of squared deviation between y_i and $f(x_i)$, is the average value of y_i in region R_m:

$$min_{f(x_i)} \sum \big(y_i - f(x_i)\big)^2 \Rightarrow \hat{c}_m = ave(y_i | x_i \in R_m) \quad (5)$$

Starting from a predictor variable j and split-point s, with the first split it is possible to define:

$$R_1(j,s) = \{X | X_j \leq s\} \quad (6)$$

and

$$R_2(j,s) = \{X | X_j > s\} \quad (7)$$

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

$$\min_{j,s}[\min_{c_1} \Sigma_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2} \Sigma_{x_i \in R_2(j,s)}(y_i - c_2)^2] \quad (8)$$

Where:

$$\hat{c}_1 = ave(y_i | x_i \in R_1(j,s)) \quad (9)$$

$$\hat{c}_2 = ave(y_i | x_i \in R_2(j,s)) \quad (10)$$

Once the best j variable and s split-point have been found, it is necessary to repeat the previous step by dividing each region into two sub-regions until a stopping rule terminates the process. Tree pruning is necessary in order to find a good trade-off between goodness of fit and the interpretation and generalisation to new data. Available data should drive the best choice of the tree size.

Tree Pruning for model selection. As a Tree T_0 has been built, it needs to be trimmed to improve the interpretability of the tree and to avoid over-fitting. It is necessary to define $T \subseteq T_0$ a subtree of T_0 obtained by pruning a number of subnodes.

The CART procedure uses a specific tree pruning method known as cost-complexity pruning. Let, $\alpha \, \varepsilon \, [0, \infty)$, called the complexity parameter, express the trade-off between tree size and goodness of fit.

Cost-complexity pruning is defined as:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T| \quad (11)$$

where:

- $|T|$ is the number of terminal nodes in T;
- N_m is the the number of records within the R_m region

$$N_m = \#\{x_i \in R_m\} \quad (12)$$

- C_m is the average of y_i in R_m

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i \quad (13)$$

- $Q_m(T)$ is the squared-error node impurity measure:

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2 \quad (14)$$

The idea is to find a $T_\alpha \subseteq T$ minimising $C_\alpha(T)$. Breiman et al. (1984) demonstrate that there is a unique sub-tree T_α. that minimises $C_\alpha(T)$. To find T_α, a *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.

4.2.2.2 Classification Trees

Classification Trees work like Regression Trees, the only difference is that they try to predict a nominal outcome rather than a continuous one. In order to partition the covariate space, they use a binary algorithm, graphically despicted as a binary tree, which subsequently splits the observations into subsets where the distribution of Y becomes more and more homogeneous. The algorithm starts from a single node which contains all records. The splitting procedure is defined in each node on the basis of covariate values: for a quantitative predictor the split value s assigns the observation to the right or to the left subnode depending on whether $x_i \leq s$ 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 is determined by using a greedy algorithm that evaluates

the homogeneity of the outcome variable in the resulting nodes using a homogeneity measurement, and stops the splitting process when the homogeneity does not significantly improve. When the outcome variable is nominal the Gini impurity index is used as the homogeneity measurement.

In a node m, the proportion of class k observations is defined as:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k) \quad (15)$$

where:

- k=1,2,…,K and $p_{m1}+p_{m2}+\ldots+p_{mK}=1$;
- m is the node representing the R_m region;
- N_m is the total number of records within the R_m region.

Then, the impurity of a node is maximum when all classes of the dependent variable are present in the same proportion. The node impurity is minimum when the node contains observations belonging to a single class.

If k is the majority class in node m, all observations in node m are classified, via majority rule, as class k observations:

$$k(m) = arg\ max_k \hat{p}_{mk} \quad (16)$$

The Gini index is defined as:

$$G = \sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \quad (17)$$

4.2.3 Random Forest

A Random Forest, introduced by L. Breiman (2001), is a multitude of correlated trees that can be used for both classification and regression purposes. When the algorithm is used for regression, the prediction for a continuous outcome can be obtained by averaging single-tree predictions. When the algorithm is used for classification, the 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 implementation of the random forest algorithm used in this research:

a. about two thirds of the data are randomly resampled to grow each tree;
b. at each node the best splitting variable is selected from among a randomly chosen subset of all predictors.

The idea in random forest algorithms is to improve the variance reduction of bagging by reducing the correlation between the trees without increasing the variance too much. This is achieved in the tree-growing process through random selection of the input variables. Specifically, before each split the algorithm selects a number $m \leq p$ of the input variables at random as candidates for splitting. For classification purposes, the value used for m is $\sqrt{3}$ and the minimum node size is one. For regression purposes, the value for m is p/3 and the minimum node size is five.

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 (similarly to the cross-validation process) and to rank the relative importance of the variables in the prediction task. A natural measurement of performance for a classifier is the difference between the proportion of votes for the correct class and the max 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 it is possible to define an importance measurement by averaging this difference over all OOBs and trees of the forest.

4.3 Cross validation

In practice, a straightforward option is to calculate the performance measurement on the data used to estimate the model. These data are usually referred to as the learning sample. In this case, the problem with evaluating such a model is that it may

demonstrate adequate prediction capability on the training data, but might fail to predict future unseen data.

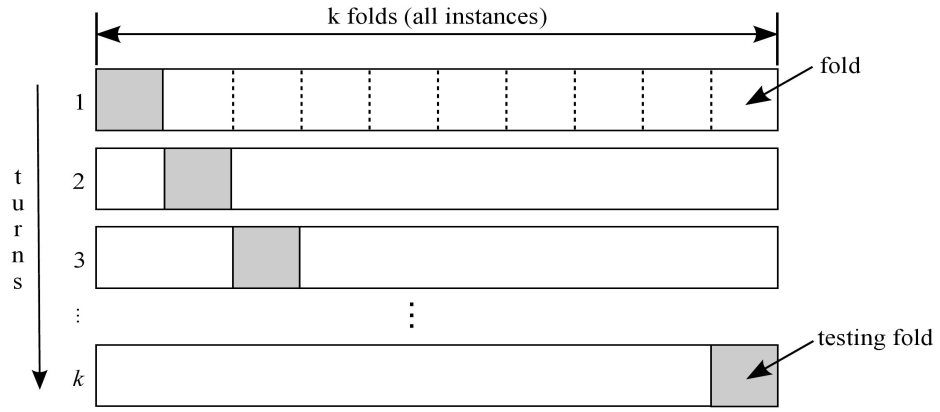
Another option is to evaluate the generalisation performance of the method or, in other words, its ability to generalize on new samples. In this case the performance measurements are calculated on independent test data, usually via a γ-fold cross-validation approach.

Currently, cross-validation is widely accepted in the data mining and machine learning community, and serves as a standard procedure for performance estimation and model selection.

Then for i=1,…,k the model is fitted after removing the ith subset, which is left out to evaluate the error on independent test data. The final k tests the data set.

A γ-fold cross-validation requires the previous random partition of the data set in γ non-overlapping subsets (or folds) of approximately the same size. The training set is split into γ parts, each of size $\frac{N}{\gamma}$. A tree is grown γ times, with each one having a different training set consisting of a γ combinations of $\gamma - 1$ original parts. The final performance measurement can be obtained by averaging the errors in the generated γ trees. In data mining and machine learning, 10-fold cross-validation ($\gamma = 10$) is the most common type (Figure 4.3).

Figure 4.3: γ -fold cross validation



This method is best for medium/small data sets, because it makes efficient use of limited amounts of data.

4.4 Performance metrics

This section discusses the main statistical measurements used for performance evaluation.

4.4.1 Mean absolute prediction error

When dealing with a prediction problem involving a continuous or interval scale variable, a natural performance measurement is the mean prediction error. It is defined as the mean absolute difference between the observed and the predicted value:

The mean absolute error is a common index for measuring the learning results:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | \bar{y} - y | \qquad (18)$$

Where: $\bar{y}$ is the predicted value of the delay and y is the real actual observed.

4.4.2 Kappa statistic and percentage of misclassified instances

For a prediction problem involving a dichotomous variable, a binary classifier can classify an individual instance into the following four categories:

- false positive (FP) = the instance is incorrectly identified;
- true positive (TP) = the instance is correctly identified;
- false negative (FN) = the instance is incorrectly rejected;
- true negative (TN) = the instance is correctly rejected.

		Observed Value	
		P	**N**
Predicted Value	**T**	True positive	False positive
	F	False negative	True negative

Various performance measurements can be derived after recording the frequency of each category on test data. The total prediction accuracy (ACC) and Cohen's Kappa coefficient for assessment of the prediction accuracy are given by the following formulas.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{N} \tag{19}$$

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.

$$Kappa = \frac{\frac{TP + TN}{N} - \frac{TP + FP}{N} \cdot \frac{TN + FN}{N}}{1 - \frac{TP + FP}{N} \cdot \frac{TN + FN}{N}} \tag{20}$$

Where:

- $\Pr(a) = \frac{TP + TN}{N}$ is the relative frequency of agreement between predicted and observed levels;
- $\Pr(e) = \frac{TP + FP}{N} \cdot \frac{TN + FN}{N}$ is the probability of agreement by chance.

The kappa statistic (Cohen, 1960) takes into account the agreement occurring by chance, thus it can be considered as a more reliable indicator of good prediction performance. It ranges from zero (no better prediction than what occurs by chance) to one (perfect prediction).

The kappa statistic is a common statistical measurement of inter-rater agreement for categorical items. According to the scale proposed by Landis and Koch, different values of Kappa could be associated with different agreement levels (Table 4.1).

Table 4.1: Value of Kappa and agreement level

value of Kappa	agreement level
<0	Less than chance agreement
0–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–0.99	Almost perfect

Table 4.1 may help the researcher to "visualise" the interpretation of kappa.

When interpreting kappa, it is also important to consider that the estimated kappa itself could be due to chance. Therefore, reporting a *p*-value of a kappa requires calculating the variance of kappa and deriving a z statistic. The *p* value tests whether the estimated kappa is not due to chance, but it does not test the strength of agreement. The *p* value is sensitive to sample size, and with a large enough sample size, any kappa above 0 will become statistically significant.

Even relatively low values of kappa can nonetheless be significantly different from zero but not of sufficient magnitude to extract useful conclusions. Thus, another performance metric for evaluating predictive power of a classifier is also used i.e., the percentage of misclassified instances. For each algorithm, it shows how accurate the prediction is considering the percentage of incorrectly classified delay levels.

CHAPTER 5: The port of Cagliari case study

This Chapter discusses the case study of Cagliari. According to the main steps of the KDD approach, the statistical results regarding the discrete and the continuous estimate of late/early arrivals are presented. Moreover, the interpretation of the discovered knowledge, made it possible to evaluate the most discriminating variables of the analysis, even thanks to graphical visualisation of the Importance-plot.

5.1 Introduction

Thanks to its position in the centre of the Mediterranean Sea (Figure 5.1), 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. The port handles conventional and bulk cargo, Ro-Ro, containers, as well as passenger ferries and cruise ships.

Figure 5.1: Location of the Cagliari port

Source: Port of Cagliari Authority

The port comprises two areas: the historic port and the industrial port, known as “porto canale”. The old port has an overall quay length of 5,800 meters and it serves commercial, Ro-Ro and passenger traffic. The industrial port extends over an area of

some 400,000 m^2, with a further 500,000 m^2 of potentially developable area to meet the growing traffic demand. The industrial port 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 (Port of Cagliari Authority) (Table 5.1; Figure 5.2).

Table 5.1: Main characteristics of the Cagliari container terminal

Quay length (m)	**Area** (ha)	**Quay cranes**	**MHC**	**Yard cranes RTG**	**Capacity** (1,000 TEU)
1,520	40	7	1	17	1.3

Figure 5.2: The structure of the "porto canale"

Source: Port of Cagliari Authority

In 2012, the port handled a total of 35,379,123 tonnes of cargo (liquid bulk, dry bulk, roll-on/roll-off, break-bulk and containers). Table 5.2 shows the port statistics for 2011 and 2012.

Between January and August 2013, the port of Cagliari handled 435,059 TEUs, showing an increase of 11.5% over the same period in 2012, without experiencing any significant congestion problems. The largest customer of the port is Hapag Lloyd (CONTSHIPITALIA, CICT, Port of Cagliari Authority).

Table 5.2: Cagliari Port statistics (2011-2012)

	2011			2012			Difference	
	In-bound	**Out-bound**	**Total**	**In-bound**	**Out-bound**	**total**	**Tot**	**%**
liquid bulk	15,336,619	11,952,296	27,288,915	13,938,519	11,304,626	25,243,145	-2,360,504	-8.65%
dry bulk	418,867	113,145	532,012	365,559	196,459	562,018	30,006	5.64%
Ro-Ro	1,456,985	1,356,266	2,813,251	1,349,043	1,249,536	2,598,579	-214,672	-7.63%
TEUs	307,630	305,559	613,189	314,518	313,091	627,609	14,420	2.35%
containers	201,458	200,145	401,603	205,910	204,421	410,331	8,728	2.17%
pax	90,331	95,100	417,549	73528	85234	239,317	-178,232	-42.69%

5.2 KDD process

The aim of this section is to present the results of the algorithmic models built on Cagliari data to solve the problem of vessel arrival uncertainty at the Mediterranean container terminal.

Two different estimates are proposed:

- a *discrete-estimate*, with which it is possible to know whether or not an incoming vessel is likely to arrive before or after the scheduled ETA.
  In this case, the output variable is a binary one, codify as 0-1.
- a *continuous-estimate*, which provides a quantitative evaluation of the difference between the scheduled ETA and the actual time of arrival in minutes.
  In this case the outcome is a continuous one.

This section, which is based on the KDD approach, consists of six main tasks.

5.2.1 Understanding the application domain

A specific theoretical study was carried out for the first task with the purpose to understand the context and to select the variables that can provide a theoretical explanation for the vessel arrival uncertainty. Therefore, a six-month period of observation was required at the Cagliari container terminal in order to define the database and to interview planners about the problems they actually encounter. Terminal operations were closely observed during this period, thus making it possible to:

- identify the main causes of delay/advance in ship arrivals in real conditions;
- analyse the critical aspects and most frequent operational issues associated with late/early arrivals in a container terminal;
- analyse how the terminal reacts to vessel arrival uncertainty, in terms of supplying of port services.

5.2.2 Data selection

Data were collected at the individual arrival level. The final database includes all arrivals at the container terminal for a period of 30 months, i.e., from January 2010 to June 2012.

The available variables that may potentially influence late/early arrivals in port can be divided into two main groups:

- *vessel-related* variables, collected thanks to the CICT (Cagliari International Container Terminal);
- *weather-related* variables, collected thanks to the ISPRA (Institute for Protection and Environmental Research).

5.2.2.1 Vessel-related variables

The input variables collected in port can easily be divided into five main classes:

1. Variables related to vessel features, :
    - *length* [m];
    - *gross tonnage* [tons];
    - *capacity* [TEUs];
    - *vector type* (mother or feeder).
2. Information on the ship owner:
    - *owner's name*;
    - *owner's nationality*.
3. Variables related to vessel service:
    - *port rotation*;

- *sailing direction*;

- *previous port*.

The shipping lines provide service on several routes. Sometimes the routes have one sailing direction (standard) or several directions (eastbound/westbound - Tyrrhenian bound/Levant bound).

4. Variables providing an indication of the number of containers to be handled:

    - number of *containers to be loaded*;

    - number of *containers to be unloaded*;

    - number of *containers to be restowed*.

    These variables are used for descriptive purposes, in particular for ranking the daily alarm level created by late arrivals at the port of Cagliari during the examined period (see section 5.2.3.5).

5. Variables related to vessel position:

    - last *Estimated Time of Arrival* (ETA) at the pilot point [dd/mm/yyyy];

    - *Actual time of Arrival* (ATA) at the pilot point [dd/mm/yyyy];

    - *berthing time* [dd/mm/yyyy];

    - *unberthing time* [dd/mm/yyyy];

    - *start operations time* [dd/mm/yyyy];

    - *end operations time* [dd/mm/yyyy].

    These variables provide an indication about the position of the vessel from the time it arrives at the pilot point to the unberthing time. In particular, the ETA and the ATA are essential for calculating the output variable, while the others are useful for descriptive purposes.

Table 5.3 shows the summary statistics of the continuous potential predictors[2].

Table 5.3: Summary statistics of the continuous variables

Variable	Min	Q_1	Mean	Median	Q_3	Max	Standard deviation
length	99	149	184	198	264	338	60.9
gross tonnage	3,784	10,310	17,660	27,470	40,300	97,825	21,733.5
capacity	350	1,079	1,560	2,519	4,253	8,749	1,884.3
discharged containers	0	107	212	236	333	1,556	176
loaded containers	0	114	213	241	348	1,397	161
restows	0	2	6	18	16	492	42

5.2.2.2 Weather-related variables

Data concerning weather conditions were collected because they can intuitively strongly affect the uncertainty of vessel arrival. From the reference model that was taken into consideration[3], the following quantities are considered in various points of the Mediterranean Sea:

- u_g: geostrophic wind speed in the x (positive towards east) [m/s];

- v_g: geostrophic wind speed in the y (positive towards north) [m/s];

- H_s: significant wave height m [ft];

- T_p: spectral peak wave period [m];

- θ_d: vector mean wave direction;

The predictions are available for four time intervals per day:

- night (00-06);

[2] The summary statistics of the categorical predictors are shown in the Appendix 1.

[3] The version of the ECMWF (European Centre for Medium-Range Weather Forecasts.) wave forecasting system is based on WAM cy4 model as described by Komen et al. (1996).

- morning (06-12);
- afternoon (12-18);
- evening (18-24).

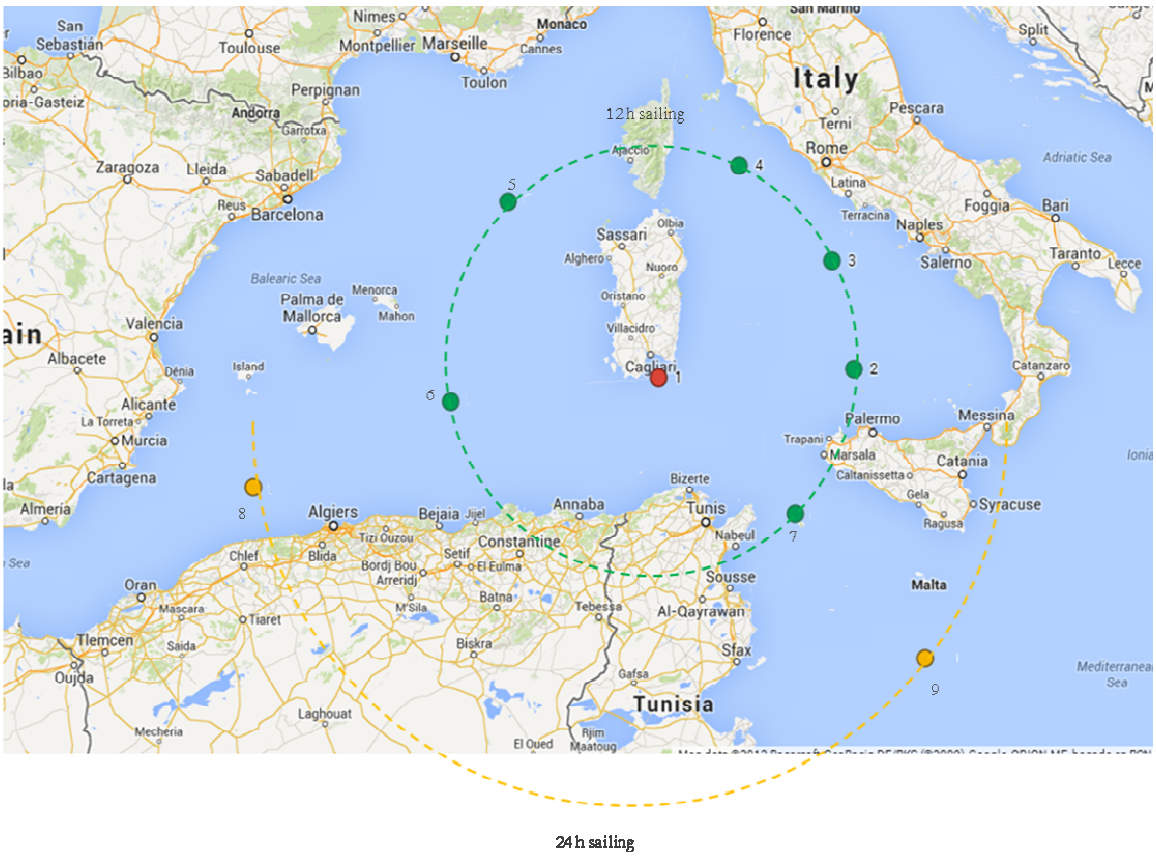
The points were selected so as to be representative of the weather conditions in the Mediterranean Sea area (Figure 5.3) for each time interval of a given day. Nine points were chosen on the basis of their longitude and latitude coordinates (Table 5.4):

- one is near Cagliari;
- six are located at a distance corresponding to 12 hours' sailing before arriving in the port;
- two at a distance corresponding to 24 hours' sailing before arrival.

Table 5.4: Longitude and latitude of the points selected in the Mediterranean Sea

Point	N	E	sailing distance from Cagliari
1	39.14	9.10	-
2	39.03	13.0	12 hours
3	40.54	12.6	12 hours
4	41.86	10.8	12 hours
5	41.36	6.4	12 hours
6	38.56	5.3	12 hours
7	36.95	11.88	12 hours
8	37.34	1.51	24 hours
9	35.05	15.13	24 hours

Figure 5.3: Selected points in the Mediterranean Sea



5.2.2.3 *The outcome variable*

The outcome variable of the study is the delay of the ship, in minutes. It is calculated as the difference between the last notified Estimated time of Arrival (ETA) and the recorded Actual Time of Arrival (ATA). The variable is expressed in negative values for vessels that arrive earlier than expected.

5.2.2.4 *Database structure*

The first database contains information about 913 mother and 1,056 feeder vessels arriving at the Cagliari container terminal over the examined period. It consists of 1,969 total arrivals and 18 vessel-related variables. Figure5.4 shows the structure of the first database.

Figure 5.4: First database structure

	Predictors										Outcome
2	univoyage	Vesselname	length	grt	teu	ETA	ATA	owner	PreviousPort	...	delay
3	5561	UASC SHUWAIKH	261	40300	4253	01/01/2010 05:00	01/01/2010 08:00	UASC	PORT SAID	...	180
4	5604	HILDE A	184	17665	1560	12/01/2010 12:00	12/01/2010 12:35	EMES	ISTANBUL	...	35
5	5742	BERLIN EXPRESS	320	88493	7506	02/01/2010 17:00	02/01/2010 16:00	HAPAG LLOYD	SOUTHAMPTON	...	-60
6	5757	RBD BOREA	129	7545	698	23/01/2010 18:00	23/01/2010 18:00	XPC	RAVENNA	...	0
7	5761	RBD BOREA	129	7545	698	07/02/2010 05:00	07/02/2010 06:30	XPC	RAVENNA	...	90
8	5769	CORELLI	169	15120	1119	21/01/2010 12:00	21/01/2010 11:30	EMES Ship&Transp	BARCELLONA	...	-30
9	5779	HILDE A	184	17665	1560	06/02/2010 05:00	06/02/2010 08:00	EMES	ISTANBUL	...	180
10	5781	GENOA EXPRESS	270	40435	3266	05/01/2010 05:30	05/01/2010 05:42	HAPAG LLOYD	NEW ORLEANS	...	12
11	...	...	...	...	...	...	...	...	...	...	...

The second database takes into consideration 915 days of activity, each of which is divided into four main intervals. Therefore, it is composed of 3,660 rows and five columns for each selected point in the Mediterranean Sea (Figure 5.5).

Figure 5.5: Second database structure

1	Date	Time interval	u_g	v_g	H_s	θ_d	T_p	point
2	01/01/2010	00:00:00	14,30	2,43	1,90	250,18	5,94	1
3	01/01/2010	06:00:00	16,08	4,56	3,14	244,94	8,24	1
4	01/01/2010	12:00:00	17,66	0,91	3,64	246,33	11,08	1
5	01/01/2010	18:00:00	15,81	-0,98	2,97	248,57	11,84	1
6	02/01/2010	00:00:00	14,32	-2,78	2,62	253,79	11,43	1
7	02/01/2010	06:00:00	12,26	-6,41	2,46	271,90	11,65	1
8	02/01/2010	12:00:00	9,56	-2,28	1,61	269,85	10,49	1
9	02/01/2010	18:00:00	7,77	-3,75	1,25	274,77	9,31	1
10	03/01/2010	00:00:00	7,86	-4,16	1,10	281,22	3,96	1

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. Thus, two or three different points can be associated with each arrival, depending on whether the vessel sailed from a port that is more or less than 24 navigation hours away from Cagliari. The number of associated weather variables can therefore range from 10 to 15 for each arrival.

5.2.3 Data preparation

Several sub-steps are needed to transform the collected data into the most suitable form for analysis. This aspect is crucial in order to improve the quality of data and of the data mining results.

5.2.3.1 *Data cleaning*

This phase refers to the correction of data problems, including missing values, extremely values, or values that are logically inconsistent in the dataset.

First of all, data were cross checked in order to evaluate logical correspondence among variables and, then, possible error types or illogical correspondence were corrected.

Furthermore, *missing values* and *outliers* were removed. Missing values were deleted because they make up less than 5% of the observations. Observations with extremely high or low values of delay were removed using the 1.5 rule (21). On the basis of this rule, the outliers were discarded prior to the analysis due to their extra-ordinary behaviour and their potentially misleading impact on performance assessment.

$$\text{Delay} < Q_1 - 1.5 \cdot |Q_3 - Q_1| \quad \text{or} \quad \text{Delay} > Q_3 + 1.5 \cdot |Q_3 - Q_1| \qquad (21)$$

One of the nine quantile algorithms discussed in Hyndman and Fan (1996).

After outliers and missing data were deleted, the final dataset includes 1,625 observations.

5.2.3.2 *Creation of new variables*

Creating new variables from one or more existing variables is a common procedure in data preparation. On the basis of the practical assistance of the experts, new variables have been created that can be useful for estimating vessel arrival uncertainty:

- ETA has been re-elaborated and broken down into three new variables i.e. *ETA-month*, *ETA-day* and *ETA-hour*. This variable could be used to evaluate whether the reliability of the ETA is different depending on the moment it is sent, for example during the night or during the day;
- *Freq_owner.* This numerical variable indicates the frequency with which a company serves a terminal that in general may influence the service provided by the terminal itself;

- *Previous port distance* [nautical miles]. The distance, in nautical miles, from the previous port to the Cagliari container terminal has been calculated.

- *Sailing*: divided into two categories: sailed and not sailed. This variable indicates whether the vessel notified its ETA once it left the previous port or while it was still in port. It is calculated as the ratio between the previous port distance and the vessel's average speed.

Table 5.4 shows the summary statistics of the new continuous potential predictors[4].

Table 5.3: Summary statistics of the new continuous variables

Variable	Min	Q_1	Mean	Median	Q_3	Max	Standard deviation
Previous Port distance	154	354	751	1,387	1,874	5,210	1,546.2
Freq_owner	2	31	134	152	168	616	149.6

5.2.3.3 Relationships among variables

An analysis of the correlations makes it possible to emphasise the linear relationships among the various groups of vessel-related variables and to identify the variables that are strongly related to others. A chi-square test of independence is used to demonstrate whether two categorical variables are related to each other, while the Pearson correlation coefficient is calculated among the continuous ones.

Chi square test

The chi square test is a test of independence between two variables:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (21)$$

where O_i is the observed frequency and E_i is the expected frequency.

[4] The summary statistics of the new categorical predictors are shown in the Appendix 1.

To test the null hypothesis that two variables are independent, the p-value[5] is compared to the significance level of the test, which was set at 1% in this specific study.

As expected, the p-values between the variables that provide information about vessel service are statistically highly significant ($p < 0.001$) (Table 5.5).
The results of the significance test were then analysed among the different groups of categorical variables. (Table 5.5).

Table 5.5: $\chi 2$ and p-value values for the categorical variables

	Service	**Previous Port**	**Sailing Direction**
Service	1	$1.9 \cdot 10^3$ (< 2.2e-16)	$3.7 \cdot 10^3$ (< 2.2e-16)
Previous Port	$1.9 \cdot 10^3$ (< 2.2e-16)	1	$4.8 \cdot 10^3$ (< 2.2e-16)
Sailing direction	$3.7 \cdot 10^3$ (< 2.2e-16)	$4.8 \cdot 10^3$ (< 2.2e-16)	1

Pearson's r correlation coefficient

The linear association between two variables using Pearson's r coefficient can change within the range [-1, +1], where the value r=0 means no correlation. The equation for Pearson's r coefficient is:

$$r = \frac{\sum xy - N\bar{x}\bar{y}}{\sqrt{(\sum x^2 - N\bar{x}^2)(\sum y^2 - N\bar{y}^2)}} \quad (22)$$

[5] The p-value is the probability that the statistics test exceeds the observed value so it tends to be small when the null hypothesis is true.

$$p = \frac{\bar{x} - \mu}{s/\sqrt{n}} < (n-1; \alpha)$$

Where:
t = the t statistic;
$\bar{x}$= the mean of the sample;
μ = the comparison mean;
s = the sample standard deviation;
n = the sample size.

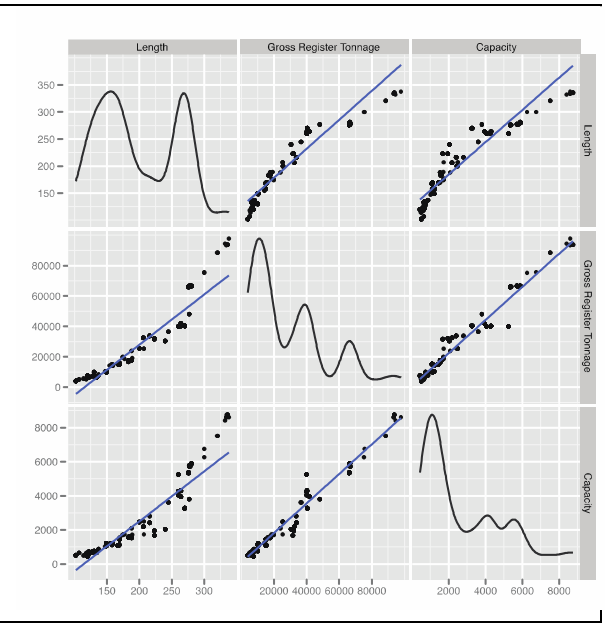
This equation requires to calculate the sum of the product of all data pairs, the means of both variables, and the sum of the squared values.

In order to determine whether the correlation is statistically significant or not, a test of *no* correlation based on Pearson's product moment correlation coefficient is used[6]. To test the null hypothesis that two variables are independent, the p-value is compared to the significance level of the test, which was set at 1% in this study.

The results of the significance test were then analysed among the various groups of continuous variables i. e. among variables related to vessel features (Table 5.6).

Table 5.6: Pearson coefficient and p-value values for the continuous variables

	Length	**GRT**	**Capacity**
Length	1	0.943 (<2.2e-16)	0.938 (<2.2e-16)
GRT	0.943 (<2.2e-16)	1	0.97
Capacity	0.938 (<2.2e-16)	0.975 (<2.2e-16)	1



Since the value of the coefficient is very close to 1 and the p-value<0,01, the variables related to the vessel features are positively and linearly strongly related. The strong relationship is also evident at a graphical level.

As a result of the previous steps, the dimensions of the database changed: the number of records decreased to 1,625 and the number of variables increased to 37.

[6] It is possible to refuse the null hypothesis if:

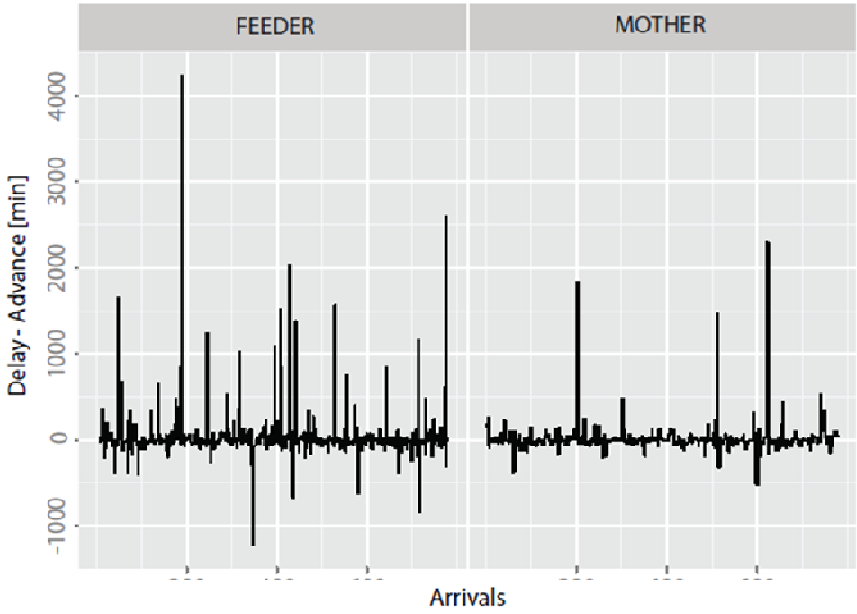
$$p = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} < (N-2, \alpha)$$

5.2.3.4 Exploratory analysis

A descriptive exploratory analysis was performed in order to identify and visualise the most important characteristics of the phenomenon.

As can be seen in Figure 5.6, the time series of arrivals in the Cagliari container terminal is complex and irregular, especially for the feeder vessels, which appear to be more prone to early or late arrivals.

Figure 5.6: Time series of arrivals in the Cagliari Container Terminal



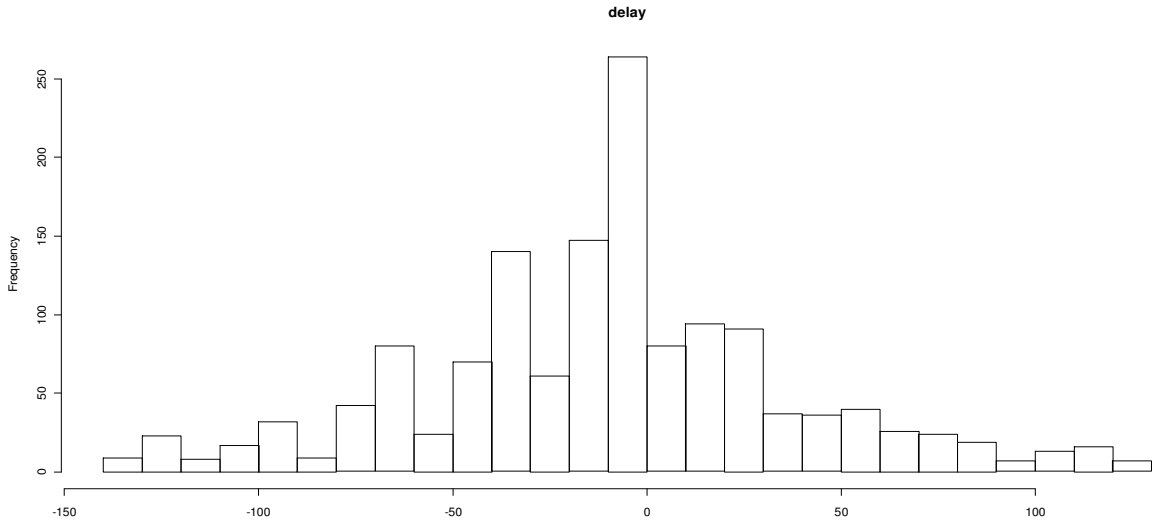
Setting a 15 minutes tolerance threshold reveals that only 30% of ships arrived at the expected time (i.e., within the interval (ETA-15, ETA+15)), the remaining 70% were delayed or arrived early. Table 5.7 shows the summary statistics of the outcome. The threshold is set at 15 minutes for operational reasons, since a delay/advance of a quarter of an hour does not cause any disruptions.

Table 5.7: Delay summary statistics (in minutes).

Sample	Min	Q1	Mean	Median	Q3	Max	Standard deviation
All vessels	-6420	-41	-3	19	30	8670	50.8

The histogram of delay distribution is shown in Figure 5.7 for the entire set of container vessel calls. Visual inspection of the histograms suggests that the frequency distribution is a unimodal distribution that exhibits only one peak.

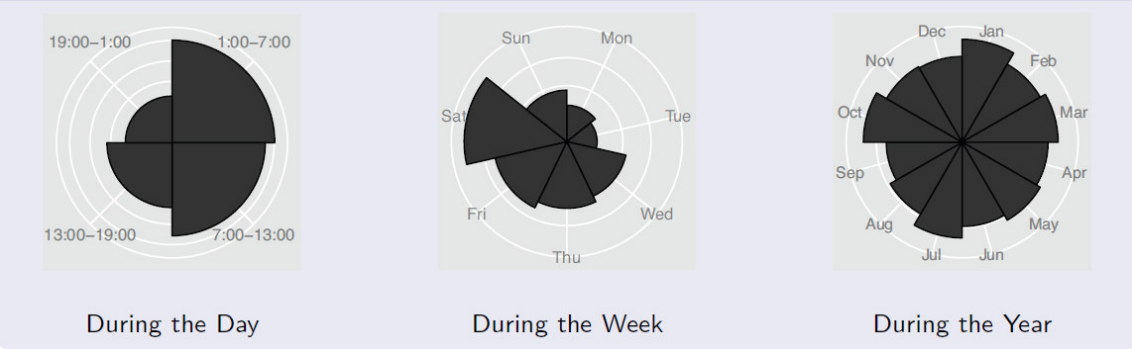
Figure 5.7: Delay distribution in the Cagliari Container Terminal



A descriptive analysis showed other important aspects regarding the terminal operations over the examined period. Considering that the working day in the Cagliari container terminal is divided into four work-shifts of six hours each:

- 64% of vessels arrive during the first two work shifts (between 1 a.m. and 1 p.m.);
- 37% of arrivals are at weekends;
- ship arrivals are regularly distributed over 12 months (Figure 5.8).

Figure 5.8. Distribution of vessel arrivals during the day, the week and the year



Data analysis also provided important information about the vessel waiting times. The average vessel turnaround time is found to be approximately 21 hours. Moreover:

- the average waiting time for a berth is approx. 2 hours;
- the average waiting time in berth before loading/unloading operations is approx. 2 hours and 30 minutes;
- once loading/unloading operations have been completed, the average amount of time before unberthing is more than 2 hours.

Minimising the above mentioned times is an important goal for the terminal to achieve.

5.2.3.5 Ranking the delay severity at the daily level

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 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. Nevertheless, it was decided to consider both late and early arrivals in the forecasting phase in order to obtain a more exhaustive analysis of the vessel arrival uncertainty. For descriptive purposes, on the other hand, the exploratory analysis of the phenomenon is completed by indicating the daily alarm rate at the port of Cagliari that is created by late arrivals during the examined period (see Pani et al., 2013).

There is no standard index for ranking delay severity in maritime transportation. Thus, the variables that are considered as the "dimensions" over which the severity of the late arrivals can be measured, were selected. In particular, the following variables were chosen for each day of activity:

1. proportion of delayed vessels;
2. total delay hours;
3. total length of the delayed vessels;
4. total number of delayed mother vessels;
5. total number of delayed containers (import, export and restows).

These are the main indices related to delay complexity management in ports, from the port operators' point of view.

The first variable represents the proportion of delayed ships on a given day with respect to the total number of vessels arriving that day. Ships arriving up to 15 minutes before and no later than 15 minutes after the notified ETA are considered on time. The second variable considers the total time of delayed arrivals on a given day, expressed in hours. The third variable indicates the total length of the vessels not arriving on time and provides important information concerning vessel size and berth occupancy. The fourth specifies the total number of mother ships arriving late on a given day, while the last variable indicates the total number of containers to be handled for each delayed arrival on a given day (import+export+restows). It is strongly related to vessel berthing time.

At the end of this process a data set was created composed of 765 rows containing information for 765 days of operations. The index ranking the delay severity was obtained by partitioning the records and ordering the final groups with respect to the chosen dimensions.

In particular, cluster analysis is used to partition the records into groups maximising some measurements of internal homogeneity and external heterogeneity, so that the profiles of the objects in the same cluster are very similar, whereas the profiles of objects in different clusters are quite distinct.

Hierarchical cluster analysis

Cluster analysis techniques can be classified as hierarchical or partitioning. In the hierarchical method the number of clusters is not established *a priori*. Instead, a series of partitions occur, which may run from *n* clusters containing a single object to a single cluster containing all objects. The process proceeds sequentially, starting from single original records and yielding a nested arrangement of records in groups. A common used approach in hierarchical clustering is Ward's method. In Ward's method (Ward, 1963), at each stage the algorithm merges the two clusters that result in the least increase in "information loss", usually measured by the within-cluster variance.

By contrast, partitional algorithms, such as the k-means algorithm (Hartigan and Wong, 1979), require the prior choice of the number of groups. The algorithm randomly chooses a set of initial centres, and then assigns each record to the group showing the least distance from its centre.

A number of graphical procedures and numerical indices have been developed in order to choose the "best" partition among the output of a hierarchical cluster analysis. In particular, several authors have proposed cluster validity indices, using different approaches. As reported by Theodoridis and Koutroubas (1999), these indices can be classified as external (based on previous knowledge about data), internal (based on the information intrinsic to the data alone) and relative (based on comparison of different clustering schema). It is possible to distinguish some specific indices (stability measures) that work very well when data are highly correlated (Brock et al., 2008), especially among internal validation indices. The stability measures are based on comparison between clusters achieved using all variables of data and clusters achieved removing the variables, one at a time.

In this case, both k-means and Ward's method are used to cluster the daily records. The two partition methods substantially overlap (only six records have different classification), so the results below refer to Ward's solution.

Cluster validity indices

Six internal indices were chosen from the literature review that was carried out, three of which are stability measurements, to identify which partition would be the optimal one.

The three stability measures used are: the Average Proportion of Non-overlap (APN), the Average Distance between Means (ADM) and the Figure Of Merit (FOM). The APN measures the average proportion of observations that are not included into same cluster, considering clusters achieved on all data and on data with a variable removed. The ADM measures the average distance between cluster centres, calculated as the mean of observations of the cluster, considering clusters achieved on all data and on data with a variable removed. The FOM measures the average of variance of the

observations in the removed variable. For all these three stability measures small values correspond better performances.

The other three internal indices used are: the Connectivity, the Dunn index and the Silhouette width. The Connectivity measures for each observation the number of own nearest neighbours not belonging to same cluster. The Dunn index measures the ratio of the smallest distance between observations not in the same cluster and the maximum distance between observations in cluster. The Silhouette width is the average of each observation of the Silhouette value. The Silhouette value measures the normalised difference between two the average distances: the first one is the average distance between a single observation and all other observations in the same cluster, and the second one is the average distance between an observation and the observations in the nearest neighboring cluster. In order to correspond better performances, Silhouette width and Dunn index should be maximised, instead Connectivity minimised.

Table 5.8 shows the values of cluster validation indices for Ward's method and the optimal scores are highlighted in bold.

Table 5.8: Values of cluster validation indices

Indices	3 clusters	4 clusters	5 clusters	6 clusters	Best partition
APN	**0.0524**	0.0864	0.1080	0.1099	**3**
ADM	**0.1897**	0.2531	0.3332	0.3969	**3**
FOM	0.5952	0.5608	0.5459	**0.5416**	6
Connectivity	**6.9091**	24.9250	37.0095	37.1095	**3**
Dunn	**0.1575**	0.0759	0.0670	0.0670	**3**
Silhouette	0.6755	0.6884	**0.6895**	0.6829	5

As four indices out of six suggest, a three cluster solution is used.

Table 5.9 shows the mean values of each variable for the three cluster solutions using Ward's method.

Table 5.9: Cluster means and standard deviations

Variables	Cluster 1	Cluster 2	Cluster 3
Total number of mother ships delayed	0 ± 0	0.49 ± 0.50	1.53 ± 0.56
Proportion of delayed vessels	0.02 ± 0.08	0.69 ± 0.27	0.80 ± 0.19
Total length of delayed vessels	9.29 ± 34.22	231.0 ± 84	533.0 ± 123
Total number of delayed containers	15.13 ± 70.5	572.0 ± 277.6	1,421.3 ± 393.3
Total delay hours	0.03 ± 0.20	1.03 ± 0.98	3.04 ± 1.83

As expected, the clusters can be naturally sorted by increasing values of the five dimensions. The three resulting groups are classified as “low”, “medium” and “high” severity, respectively. Therefore, the third class is characterised by the days with:

- the highest proportion of delayed vessels;
- the highest number of delayed mother ships, which are the most difficult to process;
- the highest number of delayed longer ships, which are the most difficult to berth;
- the highest number of containers to process, which require longer berthing times.

Using this classification it is possible to rank the alarm level of each day in the data set and conclude that in the period that was taken into consideration at the Cagliari port:

- 32% of days were characterised by a “low” delay alarm level;
- 25% of days were characterised by a “medium” delay alarm level;
- 43% of days were characterised by a “high” alarm level due to late arrivals.

5.2.4 Data mining

In this section Classification and Regression model functions are used in order to obtain a discrete and a continuous estimates of late/early arrivals. In the first case the outcome is a binary one codified as 0-1 (0: delays, 1: advance), in the second case the output variable is expressed as the difference between ETA and ATA in minutes.

The prediction errors are calculated both on the learning sample and using 10-fold cross validation. Referring to the cross validation, part of the data set is used to estimate model parameters, and the other part to assess the predictive ability of the model. The training set is split into 10 parts, each one of size $\frac{N}{10}$. Each tree is grown 10 times, with each one having a different training set consisting of 10 different combinations of 10-1 original parts. The final performance measurement can be obtained by averaging the errors in the generated 10 trees. The training data sets were built based on the normalised data of each arrival.

Classification and regression models were built using R software[7] and were generated by considering the two main different steps (see section 4.2.2): tree growing and tree pruning. The first step consisted in the identification of the full regression tree using training data. The second one consisted in the definition of the pruned tree obtained using validation data by defining the complexity parameter associated with the smallest cross-validated error. α is chosen by minimising the cross-validated $C_\alpha(T)$ (11).

Random forests improve predictive accuracy by generating a large number of bootstrapped trees, classifying a case using each tree in the new forest, and deciding a final predicted outcome by combining the results across all of the trees (an average in regression, a majority vote in classification) (see section 4.2.3).

Several models were built in both cases (discrete and continuous) using different combinations of all input variables. The model with a good trade-off between goodness of fit and its interpretation and generalisation was chosen.

To evaluate the goodness of fit of the algorithms to the data, the performance metrics illustrated in section 4.4 are considered. The statistical results are a necessary condition in order to evaluate and compare the models. Moreover, the interpretation of the results is very important in order to understand the different discriminating power of the predictors.

[7] Version 2.15.1 GUI 1.52 on a Leopard OS build 32-bit.

Table 5.10 shows the kappa statistic and the percentage of misclassified instances (see section 4.4.2) for the algorithms used for the discrete estimate i.e., Logistic Regression, Classification Tree and Random Forest.

Table 5.10: Predictive performance for the discrete outcome

Algorithm	Sample	misclassified instances	Kappa Statistic	Observed agreement	Expected agreement
Logistic Regression	sample test	32.8%	0.12	67.2%	62.74%
	10-fold cross validation	32.4%	0.10	66.9%	62.12%
Classification Tree	sample test	30.4%	0.22	69.55%	63.74%
	10-fold cross validation	31.7%	0.20	68.35%	63.85%
Random Forest	sample test	30.3%	0.23	64.81%	60.21%
	10-fold cross validation	31.5%	0.21	65.87%	58.89%

According to the scale proposed by Landis and Koch (see Table 4.1) for the evaluation of the kappa statistics, the predictive performance for the discrete outcome ranges from slight (0.10 for logistic regression) to fair (0.21 for Random Forest). The tree methods substantially overlap and don't provide a good estimate of the binary outcome. As expected, the results obtained on the sample tests are slightly better than the results obtained by cross validation. Cross validation is a good estimate of generalisation performance in order to evaluate the goodness of fit of the models. Referring to the results that were obtained via cross validation, it is possible to test the models' ability by evaluating their performance on a set of data that was not used for training. This makes it possible to avoid the overfitting problem. Thus, the results are discussed with regard to the cross validation sample.

Table 5.11 shows the mean absolute error (see section 4.4.1) for the algorithms used to obtain the continuous estimate i.e., Regression Tree and Random Forest.

Table 5.11 Predictive performance for the continuous outcome

Algorithm	**Sample**	**MAE** (min)
Regression Tree	sample test	36.80
	10-fold cross validation	38.78
Random Forest	sample test	15.89
	10-fold cross validation	35.12

In the prediction of the continuous outcome, the Random Forest algorithm still shows the best performance (Table 5.11). With regard to the results obtained by cross validation, the mean prediction error is around 35 minutes. The absolute value of this result should be considered since it takes into account a prediction of 35.12 minutes late/early, giving an uncertainty range for incoming ships of approx 1 hour and 10 minutes.

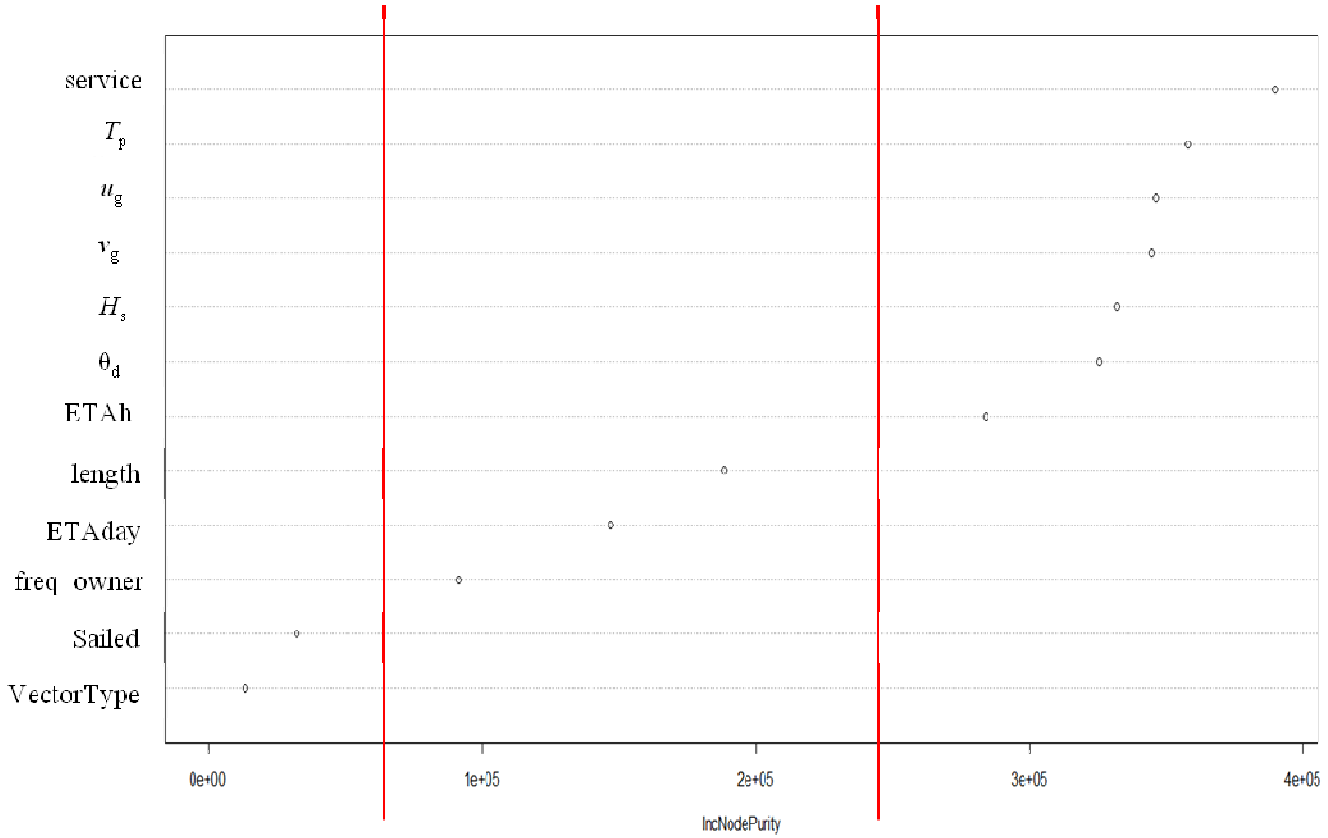
The result that was obtained is very encouraging from both the scientific and the operational point of view. In particular, if the performance is compared with a Neural Network model that was tested on the same port (presented in Fancello et al., 2011), the mean error in delay prediction changes from around 2 hours and 40 minutes to about 35 minutes. However, the results also shows that if both the service and the weather-related variables are considered predictors, the mean prediction error is reduced by approximately half (see Fadda et al., 2014).

5.2.5 Interpretation of results

In engineering applications, an understanding of the intuitive meaning of the models is needed in order to check their validity. The structure of the trees is not intuitive enough to be read, in particular with variables (such us service variables) that have a large number of modalities. Thus, the *importance-plot* of the continuous Random Forest model is depicted (Figure 5.9). The importance of the predictors is determined

by the Gini Index. Therefore, it is possible to observe the different predictive power of the most discriminating input variables and to make some assumptions.

Figure 5.9: Importance of predictors for the Random Forest algorithm (continuous model)



The variables used as predictors can be easily grouped into three categories going from the most significant to the least significant.

Service – This variable considers all the three variables related to service together i.e. port rotation, sailing direction and previous port. It appears the most discriminating variable on vessel arrival uncertainty probably because can provide important information about the service performance and the organisation/occupancy of the previous port.

Weather/sea conditions - the plots underline that the variables capturing weather conditions are 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 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.

Length - Another important discriminating variable is the vessel length. It has been chosen as an indicator of the vessel's features because as compared to the other variables in the same group, it also provides important information concerning berth occupancy. In general, longer ships are more difficult to process, in particular if they do not at the expected times.

ETA hour and ETA day - These variables underline the fact that the reliability of the ETA may depend on the moment in which it has been sent, and in particular, if the information has been notified while the vessel is still in port. In this case it can highlight variations in performance of the port operators (for example port operators working at night experience greater mental and physical fatigue).

Owner frequency - This variable indicates that the frequency with which a company serves a terminal can affect the service offered by the terminals itself.

Vector Type and *Sailed* - These type of variables substantiates the fact that once the ship has actually set sail for its destination port, then the information becomes more

reliable. Information that is 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 the ship has left the port, any uncertainty will most likely depend on weather/sea conditions alone. This can be also the reason because, as the time series of arrivals suggest (Figure 5.6), mother ships have a greater tendency to arrive on time than feeder vessel

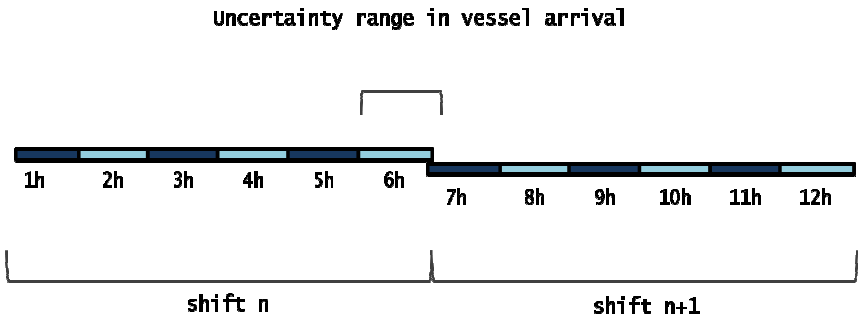
5.2.6 Consolidation of the discovered knowledge

The results obtained confirmed Breiman's approach whereby the model that is built using historical data is evaluated on the basis of the forecasting accuracy of the algorithms that are used and their ability to fit the data. Based on the tests conducted using the data regarding the Cagliari container terminal, the continuous approach proved to perform very well, whereas the discrete approach did not yield satisfactory estimates.

In operational terms, the result is nevertheless satisfactory as it enables planners to determine the shift of vessel arrival with a good degree of reliability simply by considering the continuous estimate. The best result that was obtained reduces the interval of uncertainty regarding arrival time in port to approx 1 hour and 10 minutes (± 35 minutes) (Figure 5.10). Two basic considerations emerge from this result:

- the vessel arrival uncertainty can fall within one or, at most, two work shifts. Thus, there is the certainty that the resources can be scheduled over 2 work shifts at the most;
- the probability of unequivocally identifying the work-shift of arrival is very high, i.e. around 90%.

Figure 5.10: Uncertainty range of vessel arrival at Cagliari container terminal



CHAPTER 6: The port of Antwerp case study

This Chapter analyse the case study of Antwerp. The goodness of fit of the continuous and discrete models are discussed and the most promising variables affecting vessel arrival uncertainty at Antwerp port are highlighted.

6.1 Introduction

Overall, the Port of Antwerp is Europe's second largest cargo port, after Rotterdam, with 184,134,000 tonnes being handled in 2012. That volume makes it the 17th port world-wide (Vlaamse Havencommissie). Furthermore, the Port of Antwerp is Europe's third largest port in terms of container traffic, after Rotterdam and Hamburg, thus making it the 15th port world-wide. In 2012, 14,593 vessels called at the port, the majority of which were are container vessels. In 2012, the port handled 8.64 million TEUs. In tonnage, containers represent the largest volume, but the port also handles substantial volumes of liquid bulk, dry bulk, roll-on/roll-off and breakbulk (Vlaamse Havencommissie, Containerisation International).

The Port of Antwerp generated a direct value added of € 9,765.3 mn in 2012, as well as 60,815 jobs. The level of investments in Antwerp was at € 2,339.3 mn in 2011. The vast majority of value added and jobs are related to the large firms in the port, even though the port hosts more SME's than large firms. As to value added, the non-maritime cluster represented twice the amount of the maritime cluster. All terminals are privately operated, under concession. The largest container terminal operators are PSA-Antwerp and DP World, the former retaining more than 80% of the market (PSA-Antwerp, DP World Belgium). The largest customer of the port is MSC.

The port covers more than 13,000 ha of land. It is located inland and is connected with the North Sea by the River Scheldt, which is a tidal river (Figure 6.1). The distance between the port and the North Sea is about 125 km (Vlaamse Havencommissie).

Figure 6.1: Location of the Antwerp port

The port can receive the largest of the currently existing container vessels, but not fully laden. This is not due to draught restrictions inside the port but rather to the access way through the river Scheldt, which features a number of sandbanks. Pilots and towage boats are needed for most of the vessels during the full stretch on the river Scheldt, as well as inside the port itself. Outside the port, these services are provided by government agencies, while inside the port they are provided by a private company. When entering the port through the river Scheldt, vessels have to take aboard a pilot at the roadstead at Flushing, who will take control of the vessel until it passes through the same point when leaving. Waiting can hence be involved at the roadstead.

The port is located at a crossroads of international motorways, highways and inland waterways and this is also reflected in the Port's mode split: 38% by road, 52% by inland waterway, and 10% by rail. In the container segment, the balance is slightly different, with 57% road, 34% inland waterway, and 9% rail (Vlaamse Havencommissie, Meersman et al., 2010-2011, Port of Antwerp Authority). Combined with other road traffic (passenger and freight), the substantial use of road as a port hinterland mode leads to strong congestion. Antwerp is the second most congested region in the world, with an average queuing load per working day of 221 km.hour in

2012. Furthermore, port traffic contributes to additional congestion, but can also lead to port-related road traffic being hampered by overall congestion.

6.2 KDD process

Even in this case the application is structured in six main tasks on the basis of the main steps within the KDD process.

6.2.1 Understanding the application domain

On account of its complex structure, the application to the Port of Antwerp is more involved. Before focusing the analysis on a single container terminal, the study must focus on the port level.

The port features eight main container terminals: six located on the oldest right bank of the river Scheldt, and two located on the newer left bank (Figure 6.2). Table 6.1 gives the details of the respective terminals. Two terminals i.e., the PSA-Churchill terminal and the DP World Delwaide dock terminal, are multi-purpose terminals, which implies that the container capacity cannot be specified only for containers in those cases (PSA-Antwerp; DP World Belgium). On the right bank, two of the main terminals are located in front of the locks. Locks do not imply waiting, as that would be too dangerous, but ships will adjust their speeds when approaching a closed lock. The same goes for the approach of an occupied berth.

Figure 6.2: Port of Antwerp container terminals

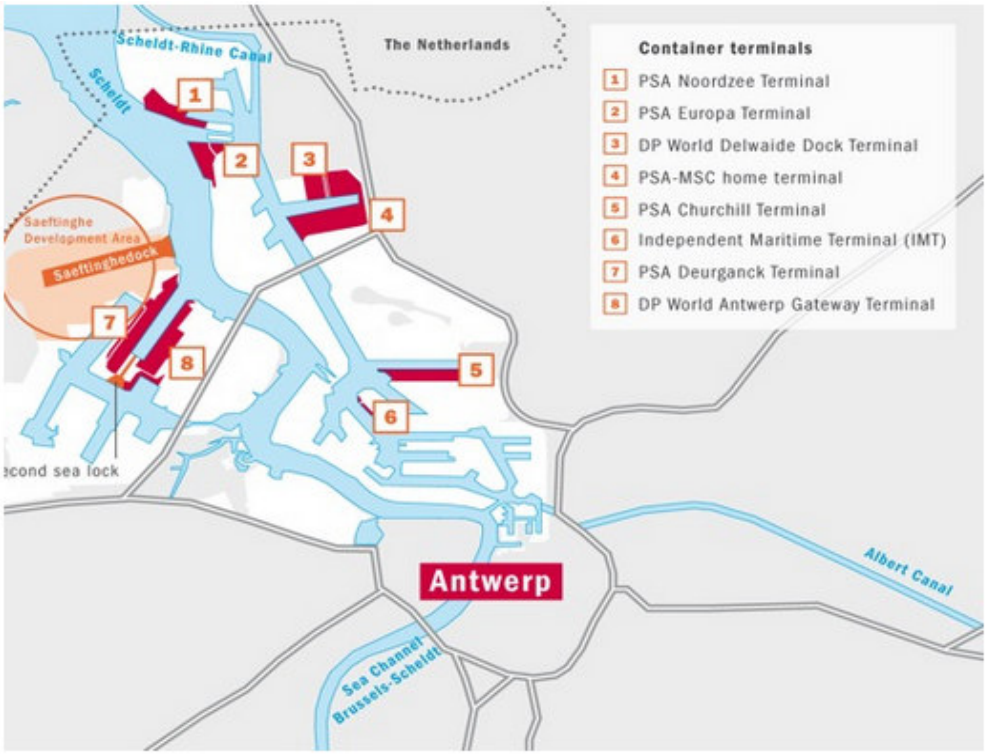


Table 6.1: Main characteristics of the Antwerp container terminals

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 container terminals are multi-user terminals, although at the PSA-MSC Home Terminal, with its 50% divided ownership between PSA-Antwerp and MSC, the latter shipping company is the main user. Furthermore, the DP World Antwerp Gateway is a joint venture, and its shareholders include DP World (42.5%), Zim Ports (20%), Cosco Pacific (20%), Terminal Link/CMA-CGM (10%) and Duisport (7.5%).

6.2.2 Data selection

A first stage of the analysis was necessary in order to study the port structure. Subsequently, all the vessels arriving at Port of Antwerp over a two year observation period were taken into consideration. On the basis of the practical assistance of the experts and on the data availability, the variables that may have influenced late/early arrivals were collected and divided into two main classes: vessel-related variables and weather-related variables. The final data made up of 10,611 rows concerning all container arrivals at the eight main CTs of this Northern Range port in 2011-2012.

6.2.2.1 Vessel-related variables

The vessel-related variables, collected thanks to the Port Authority of Antwerp, are divided in four main groups:

1. variables related to the physical structure of the vessel:
    - *length* [m];
    - *width* [m];
    - *gross tonnage* [tons];
    - *TEU's* carried.
2. Variables providing information about the vessel owner:
    - *owner's name*;
    - *owner's nationality*.
3. Variables related to the specific terminal of arrival:
    - the *berth number*;
    - the presence of a *lock* before reaching the terminal.
4. Variables that give indications about the vessel position:

- last *Estimated Time of Arrival* (ETA) [dd/mm/yyyy];
- the *Actual Time of Arrival* (ATA) in specific points located in the Scheldt river before the port entrance, in particular at the Pilot Station, the Flushing, the Coordinatiepunt [dd/mm/yyyy];
- *berthing time* [dd/mm/yyyy];
- *unberthing time* [dd/mm/yyyy].

The *Previous Port*, was available only at a terminal level and was collected in a second stage of the analysis thanks to the PSA-Antwerp terminal.

Table 6.2: Summary statistics of the continuous variables[8]

Variable	**Min**	**Q1**	**Mean**	**Median**	**Q3**	**Max**	**Standard deviation**
length	92.75	161.53	220.35	224.90	281.00	397.71	68.4
width	11.80	25.23	32.20	30.73	32.30	56.40	7.3
gross tonnage	2,906	15,933	42,533	37,518	55,994	166,085	302.8
TEU's	301	1,306	3,628.83	2,824	4,729	44,25	2,992.11

6.2.2.2 Weather-related variables

On the basis of the ECMWF model, the following data are available at all points selected in the North Sea and at each time interval that was considered:

- $u_{g:}$ geostrophic wind speed in the x (positive towards east) [m/s];
- $v_{g:}$ geostrophic wind speed in the y (positive towards north) [m/s];
- H_s: significant wave height m (ft)
- T_p: spectral peak wave period; m
- θ_d: vector mean wave direction;

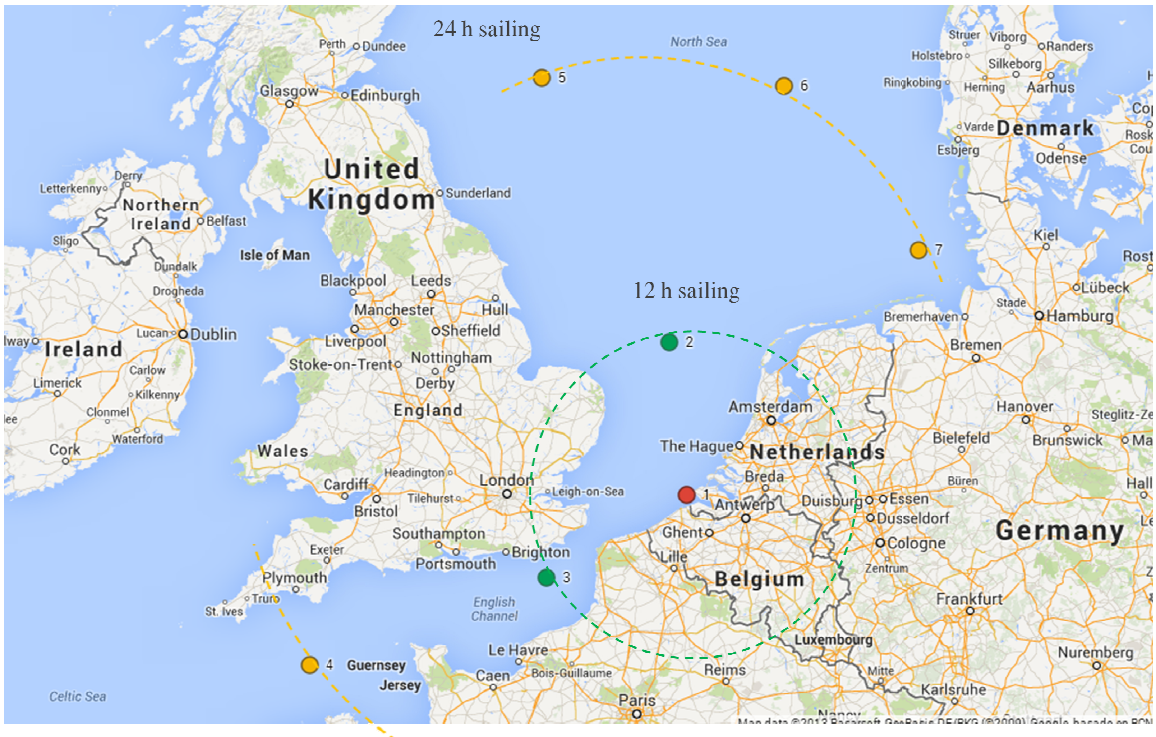
[8] The summary statistics of the categorical predictors are shown in the Appendix 2.

Even in this case, the points were chosen in order to be representative of the weather conditions in the North Sea (Figure 6.3) for each time interval and for each day of activity, considering the main vessel routes. Seven points were chosen in the North Sea: one near Antwerp, two are located at a sailing distance of 12 hours from Antwerp and four are located at a sailing distance corresponding to 24 hours (Table 6.3) from Antwerp.

Table 6.3: Longitude and latitude of the points chosen in the North Sea

Point	N	E	Navigation distance from Antwerp
1	51.44	3.34	-
2	53.27	2.96	12 hours
3	50.53	0.63	12 hours
4	49.47	-3.87	24 hours
5	56.15	0.56	24 hours
6	56.06	5.15	24 hours
7	54.19	7.86	24 hours

Figure 6.3: Selected points in the North Sea



6.2.2.3 The outcome variable

The outcome variable of the study is the delay of the ship. It measures the time difference between the ATA and the last notified ETA. Since the ETA point changed in the time window it was necessary to consider that before 09:59 of May 1st 2012, the ETA refers as the moment the vessel passed Flushing, while starting from 09:59 of May 1st 2012, the ETA refers to the moment the vessel passed the Pilot Station.

6.2.2.4 Database structures

The first database contains information about 10,611 vessels that arrived in port from January 2011 to December 2012. It is made up of 10,611 rows and 13 columns, which correspond to the vessel-related variables that were collected (see Figure 5.4).

The database take into consideration 915 days of activity for each point that was selected in the North Sea, each of which is divided into four main intervals, while the second database is composed of 3,660 rows and five columns for each selected point the North Sea (see Figure 5.5).

The two databases were merged to create the final one. At the port level, the match was created to be able to associate only the point near Antwerp to each arrival because the Previous port is unknown. At a terminal level, the associated weather variables for each arrival can therefore range from 10 to 15.

6.2.3 Data preparation

This step involved several tasks for manipulating and preparing the data for data mining in order to improve model accuracy.

6.2.3.1 Data cleaning

As per the application at Cagliari container Terminal:

- missing data were examined and then deleted since they accounted for less than 5% of the observations;

- data inconsistencies were verified thanks to frequency tables for categorical variables and histograms for continuous variables;

- outliers were removed. Observations with delay values that fall below $Q_1 - 1.5 \cdot |Q_3 - Q_1|$ or above $Q_3 + 1.5 \cdot |Q_3 - Q_1|$ are, on the basis of the 1.5 rule, identified as potential outliers and thus were deleted.

After removing outliers and missing data the final dataset at the port level comprises 9,857 observations.

6.2.3.2 Creation of new variables

Also in this case, thanks to the practical support of the experts, new variables have been created that can be useful for the analysis:

- ETA has been re-elaborated and broken down into three new variables: ETA day, ETA month and ETA hour;

- *Freq_owner.* In order to consider the frequency with which the vessel owner serves the port.

6.2.3.3 Relationships among variables

Relationships were identified among the continuous vessel-related variables by using the Pearson correlation coefficient. Also in this case, the significance level of the test is set at equal to 1%. (see section 5.2.3.3).

Table 6.4: Pearson's coefficient and p-value values for the variables related to the vessel features

	length	**gross tonnage**	**capacity**	**width**	**TEU's**
length	1	0,29 (<2.2e^{-16})	0.17 (<2.2e^{-16})	0.93 (<2.2e^{-16})	0.15 (2.75e-09)
gross tonnage	0,29 (<2.2e^{-16})	1	0,39 (<2.2e^{-16})	0,23 (<2.2e^{-16})	0.51 (<2.2e^{-16})
capacity	0.17 (<2.2e^{-16})	0,39 (<2.2e^{-16})	1	0,13 (<2.2e^{-16})	0,21 (<2.2e^{-16})
width	0.93 (<2.2e^{-16})	0,23 (<2.2e^{-16})	0,13 (<2.2e^{-16})	1	0.12 (1.33e-10)
TEU's	0.15 (2.75e-09)	0.51 (<2.2e^{-16})	0,21 (<2.2e^{-16})	0.12 (1.33e-10)	1

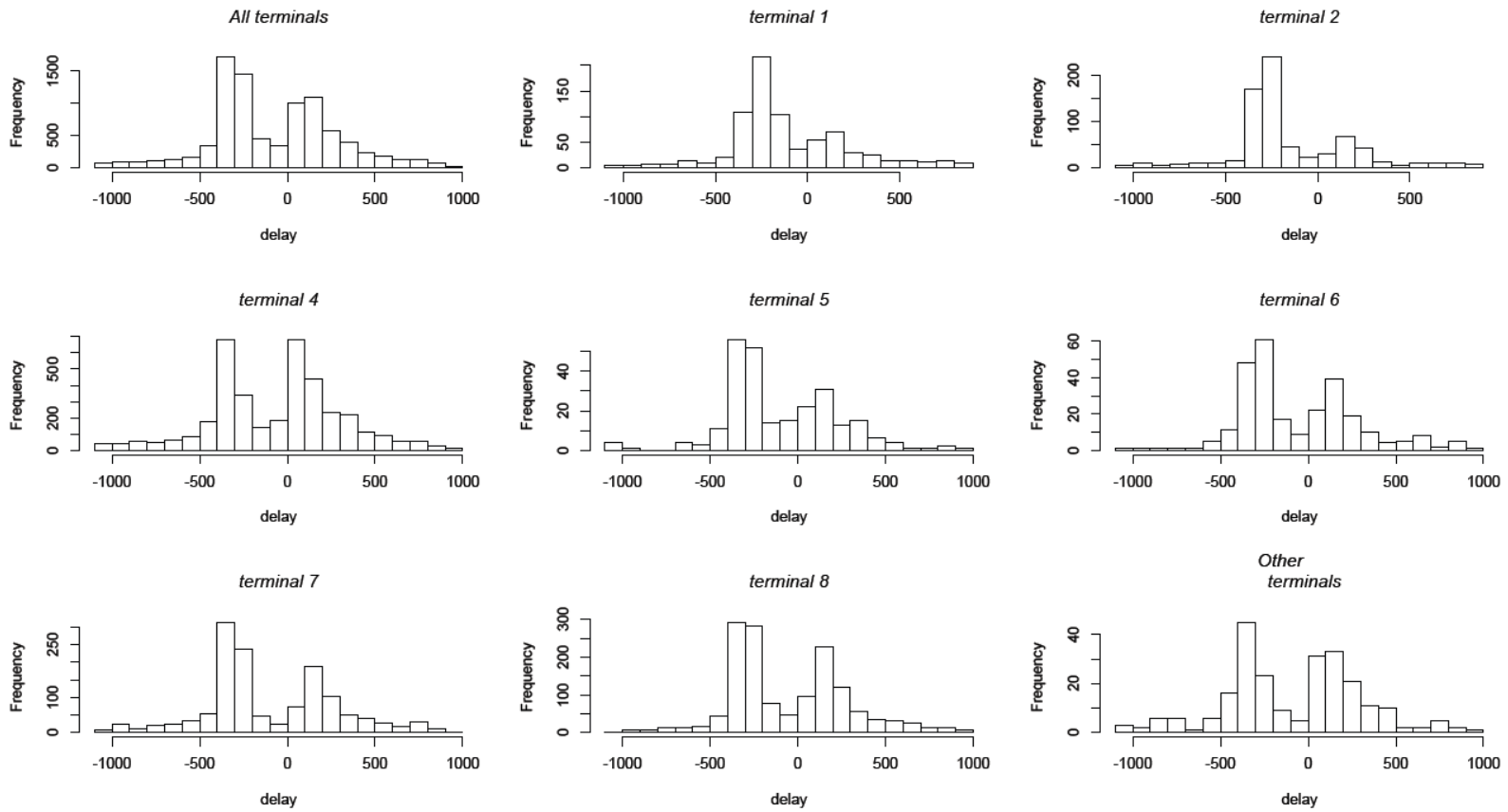
Table 6.4 shows that the p-values are statistically significant and the r coefficient values underline that some of the variables have a weak positive correlation.

6.2.3.4 Exploratory analysis

Histograms of delay distribution are shown in Figure 6.4 for the entire set of container vessel calls and by terminal (n.b.: 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). 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 vessels in advance and in delay differs across terminals (Table 6.5).

The two-peak distribution shape may probably be related to the sailing constraints due to tidal restrictions on port access. Tidal windows in maritime access channels complicate the service provided and could lead to changes in the order of port calls.

Figure 6.4: Delay distribution at Antwerp port, by terminal



At the port level, the average delay is approx minus 80 minutes; the median delay is even less. Most arrivals at Antwerp terminals are in advance but the proportion reverses in some of the terminals. The large value of the standard deviation implies that position measurements do not adequately summarise the delay values, as was evident from the bimodality. Various thresholds can be set based on the empirical distribution of delays. 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), 42.9% arrived later and the remaining 55.3% arrived earlier than expected.

Table 6.5: Delay summary statistics by terminal (in minutes)

Terminal	Min	Q1	Mean	Median	Q3	Max	Standard deviation	Total arrivals	Proportion of vessels on delay
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

The exploratory analysis of the data that was conducted at the port level with the assistance of the planners made it possible to choose Terminal number 7 for two main reasons:

- availability of data;
- database size: the number of observation that were collected is very similar to the number of observation 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. The data base contains information for each vessel and

includes 14 vessel-related variables and 15 weather-related variables. In fact, thanks to the Previous Port, weather conditions are considered in three points along the route for each arrival.

Also at the terminal level a preliminary step of data preparation was required. In this case it was possible to create as new variables also:

- *Previous port distance* [nautical miles]: this variable was created, only at a terminal level, and represents the distance in nautical miles between the previous port and the specific container terminal in Antwerp.

- *Sailing*. This variable indicates whether the vessel notified its ETA once it left the previous port or while it was still in port. It is calculated as the ratio between the previous port distance and the vessel's average speed.

6.2.4 Data mining

The results of the data mining application are shown in this section. The algorithmic models are described considering both the flexibility in representing the data and the interpretability of the results.

In this case study, several trees and forets were built using different subsets of all input variables and estimating the model parameters. The predictive performances are shown both on the learning sample and using 10-fold cross validation, but are evaluated with reference to the cross validation procedure. All models were built using R software (see section 5.2.4).

Table 6.6 and Table 6.7 show the results for the binary output variable at the port level and the terminal level, respectively.

Table 6.6: Predictive performance for the discrete outcome (port level)

Algorithm	Sample	misclassified instances	Kappa Statistics	Observed agreement	Expected agreement
Logistic Regression	Sample test	25%	0.46	73.64%	50.50%
	10-fold cross validation	28%	0.45	70.81%	50.04%
Classification Tree	Sample test	21%	0.58	79.89%	49.93%
	10-fold cross validation	26%	0.57	80.45%	51.06%
Random Forest	Sample test	15%	0.67	83.99%	50.14%
	10-fold cross validation	16%	0.72	84.93%	50.20%

Table 6.7: Predictive performance for the discrete outcome (terminal level)

Algorithm	Sample	misclassified instances	Kappa Statistics	Observed agreement	Expected agreement
Logistic Regression	Sample test	20%	0.59	79.26%	49.56%
	10-fold cross validation	22%	0.55	78.32%	49.50%
Classification Tree	Sample test	19%	0.61	79.78%	49.88%
	10-fold cross validation	22%	0.59	79.43%	49.76%
Random Forest	Sample test	20%	0.65	79.69%	50.19%
	10-fold cross validation	17%	0.63	80.20%	49.68%

As Table 6.6 and Table 6.7 clearly show, the discrete models perform well for the Antwerp container terminal data. This is demonstrated not only by the statistical kappa value and by the percentage of misclassified cases, but also by the percentages of the observed and expected agreement.

As the distributional form of the output variable suggests, with regard to the two distinct peaks, the algorithms fit the data very well in the discrete case, where the outcome is dichotomic.

Random forest showed the best performance in both cases. Based on the evaluation of the kappa statistics, the predictive performance for the discrete outcome ranges from moderate (0.45 for logistic regression) to substantial (0.72 for Random Forest). The percentage of misclassified instances is around 16% at the port level and 17% on terminal 7. In both cases the result is considerable from a statistical point of view. 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.

Table 6.8 and Table 6.9 show the results for the continuous output variable at the port level and the terminal level, respectively.

Table 6.8: Predictive performance for the continuous outcome (port level)

Algorithm	**Sample**	**MAE** (min)
Regression Tree	Sample test	217.44
	10-fold cross validation	218.89
Random Forest	Sample test	85.82
	10-fold cross validation	179.44

Table 6.9: Predictive performance for the continuous outcome (terminal level)

Algorithm	**Sample**	**MAE** (min)
Regression Tree	Sample test	209.32
	10-fold cross validation	218.24
Random Forest	Sample test	88.08
	10-fold cross validation	188.36

In the prediction of the continuous outcome, Random Forest algorithms still show the best performance even if the mean prediction error is quite high.

Though satisfactory in scientific terms, the results are not yet acceptable for the

specific operating context. In fact, in scientific terms it is easy to deduce that the mean prediction errors give an uncertainty regarding arrival time in port of around 5h 52' and 6h 16', respectively. Therefore, the probability of univocally identifying the work-shift of arrival is very low.

Regarding the two case studies that were examined, it has been shown that the models' forecasting accuracy is closely related to the distributional form of the output variable. The Antwerp data reveal that the continuous models are limited in their ability to capture bi-modality. Therefore, the continuous models do not perform as well as they do for the Cagliari container terminal where the output variable has unimodal distribution.

6.2.5 Interpretation of results

In this section some consideration are made concerning the predictive power of the predictors and their association with vessel arrival uncertainty. The importance plot shows each variable on the y-axis, and their importance on the x-axis. The Gini coefficient is the measurement of homogeneity that is used. The changes in Gini are listed for each variable and normalised at the end of the calculation. Variables that result in nodes with higher purity have a higher decrease in Gini coefficient.

The *importance-plots* of the discrete Random Forest models are compared in the two cases, at the port level and at the terminal level (Figure 6.5, Figure 6.6).

Figure 6.5: Importance of predictors for the Random Forest algorithm (discrete model - port level)

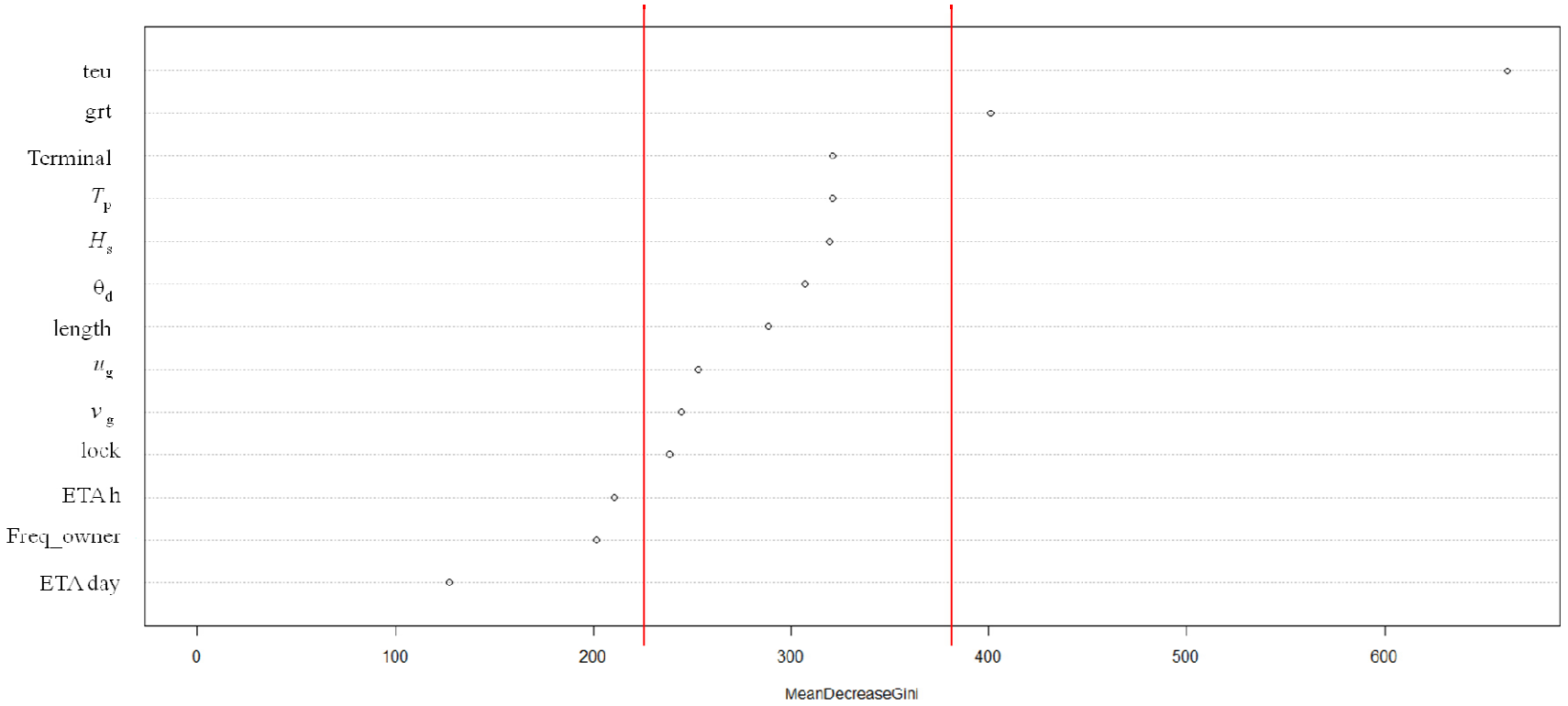
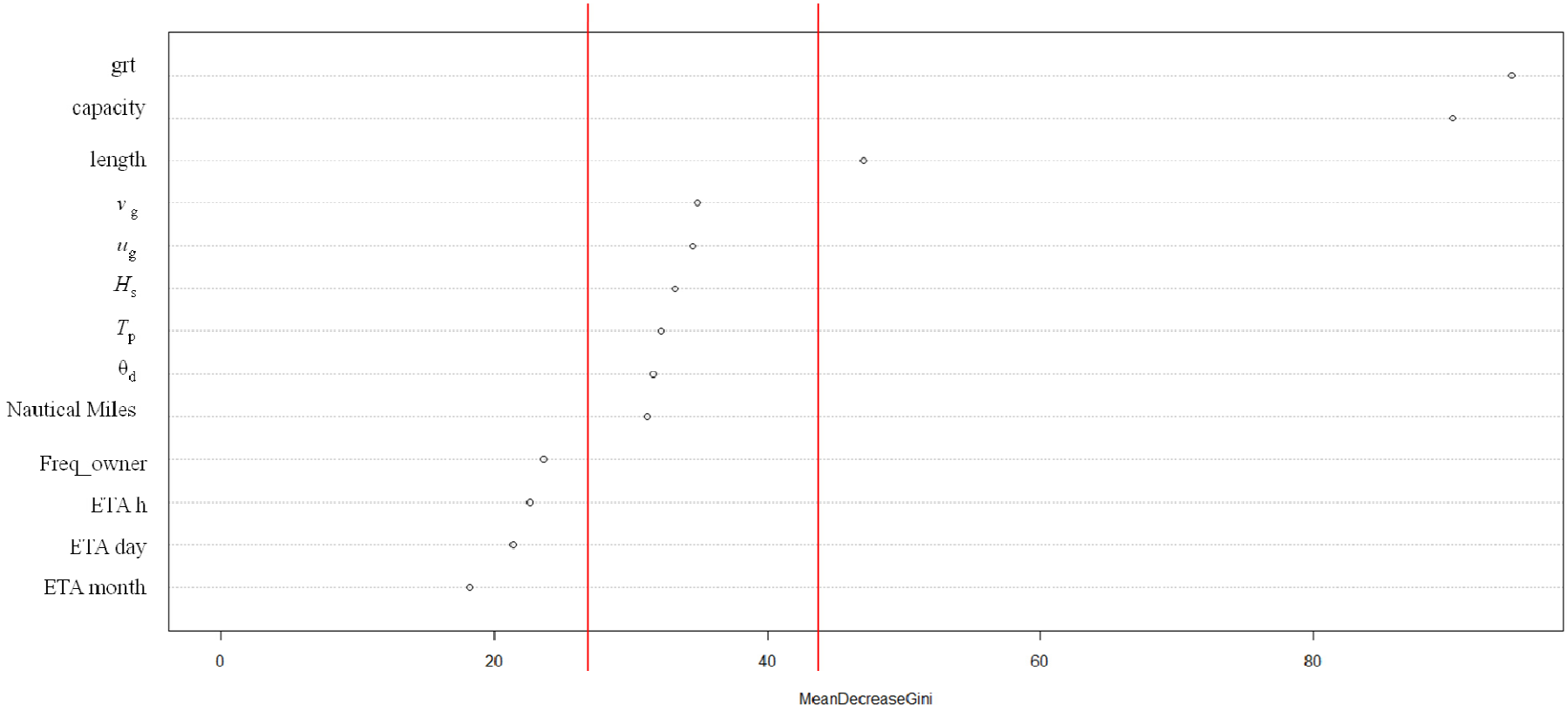


Figure 6.6: Importance of predictors for the Random Forest algorithm (discrete model - terminal level)



The plots clearly show the main variables impacting vessel arrival uncertainty.

In both cases, the variables capturing *vessel size* are important determinants of uncertainty. They can directly affect both navigation/sailing times and the possible effects on container handling productivity. This aspect can be important if the transit time from the previous port is less than 24 hours and the last ETA is notified before the vessel has already left the port.

At the port level, the variables characterising the specific *terminal*, the terminal location and the presence of the lock may reasonably impact on the early/late arrivals.

Another important aspect concerns the discriminating power of the *weather-related* variables. At the port level, these variables have less significance. The reason for this result can be explained by considering that although the point that was selected, in the North Sea close to Antwerp is indicative, it does not represent the weather conditions encountered during the route very well. The discriminating power of the weather-related variables increases at the terminal level since the point that was considered for the forecast is an intermediate point along the route.

Vessel *ownership* appears more discriminating at the terminal level, most likely because the frequency with which a company serves a terminal may affects the service offered by the terminal itself.

In conclusion, a possible time dependence in the reliability of the information based on ETA hour and ETA day emerges in the Antwerp case as well.

6.2.6 Consolidation of the discovery knowledge

The results for the Antwerp terminal confirm those obtained for Cagliari. From an interpretative point of view, the variables that most strongly influence arrival uncertainty remain substantially the same. Considering the goodness of fit of the algorithms, as far as the Antwerp container terminal is concerned the discrete models are very flexible in representing data when the distribution of the outcome is a double-peak distribution. However, the continuous algorithms show very relevant

performance in capturing data characteristics when the distribution exhibits a unimodal behavior.

The problem can therefore be solved by considering that in statistics a n-modal distribution is a continuous probability distribution with n different modes, that appear as distinct peaks in the probability density function. The bimodal distribution most commonly arises as a mixture of two different unimodal distributions. On the basis of this aspect, a test was conducted using the data regarding Terminal 7, examining late and early arrivals separately.

In both cases the frequency distribution is unimodal (Figure 6.7).

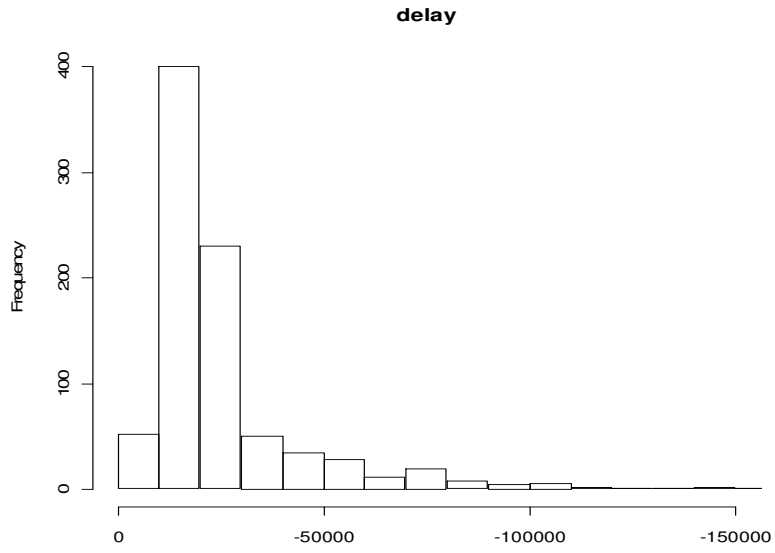
Table 6.10 shows the delay and advance summary statistics.

Table 6.10: Delay and advance summary statistics.

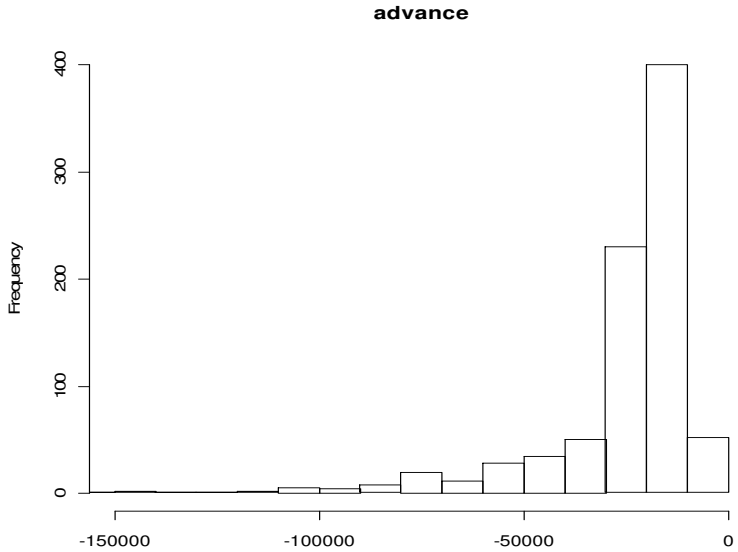
Dataset	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**	**Standard deviation**
Delay	15	130	207	281	376	929	208
Advance	-1,090	-386	-318	-364.5	-271	-16	188

Figure 6.7: Frequency distributions of the delays (a) and advances (b) at Terminal 7

(a)



(b)



a

The continuous estimates obtained for the two datasets are shown in Table 6.11 and in Table 6.12.

Table 6.11: Predictive performance for the continuous outcome (terminal level-delay)

Algorithm	**Sample**	**MAE** (min)
Regression Tree	Sample test	122
	10-fold cross validation	145
Random Forest	Sample test	65
	10-fold cross validation	119

Table 6.12: Predictive performance for the continuous outcome (terminal level-advance)

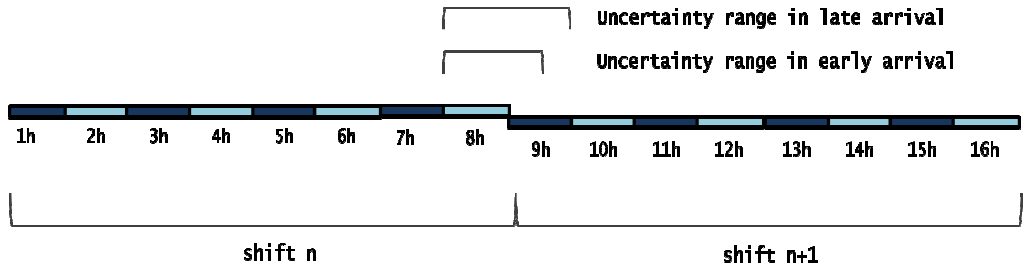
Algorithm	**Sample**	**MAE** (min)
Regression Tree	Sample test	112
	10-fold cross validation	123
Random Forest	Sample test	51
	10-fold cross validation	102

As expected, the results obtained with the continuous models are satisfactory. Moreover, in this case the prediction error that was obtained should not be evaluated in terms of absolute value but should be reduced by half, thus increasing the probability of unequivocally identifying the shift to be assigned to the vessel's arrival.

Considering that the working day in Antwerp is set up in 3 eight-hours working shifts, it is easy to deduce that in case of late arrival the possibility of univocally determining the demand for each shift is 80%, while in case of early arrival it is around 86% (Figure 6.8).

Even in this case, vessel arrival uncertainty can fall within one or, at most, two work shifts, thus providing certainty that the resources can be scheduled over 2 work shifts at the most.

Figure 6.8: Uncertainty range of vessel arrival at the Antwerp Container Terminal



CHAPTER 7: Conclusions

This chapter discusses how this research contributes to the existing knowledge and the implications of the research itself. The results are discussed at the scientific and practical levels.

7.1 The contribution of the research

This research was set up to search for a solution to the vessel arrival uncertainty problem in container terminals.

The main issue for enhancing planning efficiency in container terminals is the prediction of ship arrival times. Furthermore, considering the strong dependence of planning processes on incoming information flow, a reliable estimate of the actual time of arrival 24 hours in advance could help planners in a daily planning scenario. In particular, this would facilitate the terminal operations management with regard to the allocation of the human, mechanical and spatial resources that are required for handling operations, which are often under/overestimated at the planning stage.

The literature review described in Chapter 3 reveals that very few studies deal explicitly with this problem. The absence of a reference model that specifies the relationship between vessel arrival uncertainty and the involved variables resulted in the application of a specific machine learning approach. This approach, that abandons all prior assumptions about data distribution shape, is based on the self-learning concept according to which the relation between an outcome variable Y and the set of predictors X is directly identified by the previously collected data.

The methodological approach has been validated thanks to two case studies: the transhipment container terminal of Cagliari and the transhipment container terminal of Antwerp.

A period of observation was conducted in both terminals in order to analyse the context, interview experts and collect all the variables that may potentially influence late/early arrivals in port.

On the basis of the output variable distribution shape, different estimates are required:

- a discrete-estimate that provides a qualitative evaluation of the delay/advance;
- a continuous-estimate that provides a quantitative evaluation of the vessel arrival uncertainty.

In the first case, the output variable is a binary one codified as 0-1, in the second case it is a continuous variable that is expressed in minutes. The fitted algorithmic models used to obtain predictions include Logistic Regression, Classification and Regression Tree and Random Forest. All the proposed models are able to learn from experience following the well-known Data Mining paradigm "learning from data".

Investigation of the case studies validates the methodology and provides important findings regarding the goodness of fit of the models on the data and the main variables affecting the process.

In particular, it was seen that the reliability of the prediction changes on the basis of the outcome distribution shape. The results obtained in the two terminals that were studied can be analysed considering the different frequency distribution of the delay: the Cagliari container terminal shows a unimodal distribution composed of a distinct peak, while the terminal container in Antwerp underlines a bimodal distribution composed of two different peaks. The applications highlight that, due to the strong bimodality, the discrete algorithms are very flexible in representing data when the distribution of the outcome shows two distinct modes. However, the continuous algorithms have a highly relevant performance in capturing data characteristics when the distribution exhibits unimodal behavior.

As expected, Random Forest algorithms still show the best performance in all predictions. Moreover, the evaluation of the discovered knowledge made it possible to highlight the most discriminating variables of the analysis, even thanks to the graphical visualisation of the Importance-plots.

7.2 Practical implications

The practical aspects of the research results make them relevant for application in container terminals.

The research concerns a purely operative setting where planners need to make predictions about future arrivals in order to have useful information that can be used in a daily strategy decision making process. In order to do that, it is crucial for planners to know whether the vessel will fall within the scheduled work shift or in a different one. Therefore, the practical usefulness of this research lies in identifying the probability, associated with the continuous estimate, of specifically identifying the work-shift the incoming vessel will fall within.

CAGLIARI CASE

If the delay has a unimodal distribution, the continuous model suffices to obtain a substantial range of uncertainty.

In the Cagliari case, in fact, the mean prediction error in minutes makes it possible to determine the uncertainty range for each ship arrival with a high degree of reliability: the probability of unequivocally identifying the work-shift of arrival is very high, i.e., around 90%.

ANTWERP CASE

If the delay distribution is not unimodal, the problem can be solved by considering that a n-modal distribution arises as a mixture of n different unimodal distributions.

In the Antwerp case the outcome distribution is a continuous probability distribution with two different modes, that appear as distinct peaks in the probability density function. This aspect suggested to use a two step instrument. The first step uses a discrete model to determine whether an incoming vessel is likely to arrive late or in advance. This information will allow us to consider the unimodal distribution of the delays or the advances separately in order to obtain a reliable forecast in minutes.

In this case the probability of unequivocally identifying the work-shift of arrival is approx 81% and 86%, respectively, for delays or advances.

The slightly higher mean prediction error in this case can be attributed to the higher

standard deviation value of the outcome.

The practical results of this research will contribute to the development of more efficient solutions for the management problems that terminal operators are called upon to solve. The instrument that is proposed can help planners, together with their experience, to have reliable indications on vessel arrival times. Moreover, once a machine learning algorithm has been successfully trained and tested, the planner can use it to obtain a prediction simply by substituting the input values for the vessel for which the forecast is required.

In summary, to know with greater certainty vessel arrival times can improve the use of the available resources (human resources and equipment as well as spatial resources) required for handling operations and for support activities. This could maximise terminal efficiency and minimise terminal costs, hence improving terminal competitiveness.

7.3 Suggestions for future research

This research also provides basis for further studies.

One of the main lines of research that results directly from this study concerns the introduction into the model of new variables. In particular:

- external factors like strikes, mechanical problems or breakdowns that, during discussions with planners, were mentioned as potential influential variables in vessel arrival uncertainty;
- tidal flow data, in order to investigate if the bimodality of the distribution in the Northern range port might be related to sailing constraint due to tidal restrictions on port access.

Another important future research area concerns the evaluation of management fallout caused by late/early arrivals in the system as a whole. In particular, it would be important to quantify the actual costs for a terminal and to analyse the economic and organisational benefits that might derive from the use of the proposed instrument.

With regard to future applications of the methodology, it must be taken into account that the outcomes of the analysis can be susceptible to the availability and the quality of the input data. It could be interesting to refer to a broader historical period of observation and evaluate any improvements in forecasting.

Lastly, it might be interesting to calculate the daily alarm rate at the port of Antwerp generated by late/early arrivals which, due to lack of data, is not included in the preliminary exploratory analysis.

BIBLIOGRAPHY

Ambrosino, D., & Tanfani, E. (2012). An Integrated simulation and optimization approach for seaside terminal operations Proceedings 26th European Conference on Modelling and Simulation.

Ambrosino, D., & Tanfani, E. (2011). Optimization and simulation for terminal container seaside. Presented at 21st Annual Conference of the Italian Operational Research Society (AIRO).

Beškovnik, B. (2008). Measuring and increasing the productivity model on maritime container terminals. *Pomorstvo*, *22*(2), 171-183.

Bierwirth, C., & Meisel, F. (2010). A survey of berth allocation and quay crane scheduling problems in container terminals. *European Journal of Operational Research*, *202*(3), 615-627.

Brachman, R. J., & Anand, T. (1996). The process of knowledge discovery in databases. In *Advances in knowledge discovery and data mining* (pp. 37-57). American Association for Artificial Intelligence.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

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

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Wadsworth Statistics/Probability Ser. Wadsworth Int. *Group, Belmont, CA*.

Brinkmann, B. (2011). Operations systems of container terminals: a compendious overview. In *Handbook of terminal planning* (pp. 25-39). Springer New York.

Bruggeling, M (2011). Abandoning the spherical container terminal: The support of container terminal berth planning by the integration and visualization of terminal information. Master's thesis, Delft University of Technology.

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.

Brueckner, J., 2002a. Airport congestion when carriers have market power. *American Economic Review,* 92 (5), 1357–1375.

Brueckner, J. K. (2002). Internalization of airport congestion. *Journal of Air Transport Management*, *8*(3), 141-147.

Bruggeling, M.,Verbraeck, A., Honig, H. J. (2011). Decision support for container terminal berth planning: integration and visualization of terminal information. Proceedings van de Vervoerslogistieke.

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, *184*(3), 1140-1154.

Cellard, J. C., Labbe, B., Savitsky, G. (1967). Le programme ELISEE, presentation et application. Metra 3 (6), 511-519.

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.

Chen, S. H., & Chen, J. N. (2010). Forecasting container throughputs at ports using genetic programming. *Expert Systems with Applications*, *37*(3), 2054-2058.

Choi, Y., Ahn, H., Chen, J. (2005). Regression trees for analysis of count data with extra poisson variation. *Computational Statistics and Data Analysis*, 49 (3), 893-915.

CICT Cagliari International Container Terminal http://www.cict.it. Accessed September 22, 2013.

Chou, C. C., Chu, C. W., & Liang, G. S. (2008). A modified regression model for forecasting the volumes of Taiwan's import containers.*Mathematical and Computer Modelling*, *47*(9), 797-807.

Chung, C. C., & Chiang, C. H. The Critical Factors: An Evaluation of Schedule Reliability in Liner Shipping.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

CONTSHIPITALIA, Port of Cagliari. http://www.contshipitalia.com. Accessed September 22, 2013.

Conversano, C. (2002). Bagged mixtures of classifiers using model scoring criteria. *Pattern Analysis & Applications*, *5*(4), 351-362.

Containerisation International. Top 100 Container ports 2012. http://europe.nxtbook.com/nxteu/informa/ci_top100ports2012/#/0 . Accessed July 30, 2013.

Coulter, D. Y. (2002). Globalization of maritime commerce: the rise of hub ports. *Globalization and Maritime Power*, 133-142.

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

DP World Belgium. Maritime Terminals. http://www.dpworld.be. Accessed July 2, 2013.

Du, Y., Xu, Y., & Chen, Q. (2010). A feedback procedure for robust berth allocation with stochastic vessel delays. In *Intelligent Control and Automation (WCICA), 2010 8th World Congress on* (pp. 2210-2215). IEEE.

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

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework, In Proceedings of KDD-96, Menlo Park, CA: AAAI Press, pp. 82–88.

Han, J., Kamber, M., & Pei, J. (2006). *Data mining: concepts and techniques*. Morgan kaufmann.

Han, X. L., Lu, Z. Q., & 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.

European Commission (2012a) Eurostat statistics Database, Maritime port freight and passenger statistics http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Maritime_ports_freight_and_passenger_statistics. Accessed September 23, 2013.

Fadda, P., Fancello, G., Frigau, L., Mola, F., & Pani, C. (2013). A data mining approach to forecast vessel arrival in a transhipment container terminal. *Transport*, forthcoming.

Fancello, G., Naseddu, M., Uccheddu, B. & Fadda, P. (2010). Modello di simulazione delle attività di manutenzione in un terminal container. Scenari di riferimento per i porti container italiani nel sistema euro-mediterraneo, Franco Angeli Editore.

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.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In Proceedings of KDD-96, Menlo Park, CA: AAAI Press, pp. 82–88.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI magazine*, *13*(3), 57.

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In *ICML* (Vol. 96, pp. 148-156).

Fung, M. K. (2002). Forecasting Hong Kong's container throughput: an error-correction model. *Journal of Forecasting*, 21(1), 69-80.

Gambardella, L. M., Bontempi, G., Taillard, E., Romanengo, D., Raso, G., & Piermari, P. (1996). Simulation and forecasting in intermodal container terminal. In *Proceedings of the 8th European Simulation Symposium*, 626-630. SCS International.

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

Ghiani, G., & Musmanno, R. (2000). *Modelli e metodi per l'organizzazione dei sistemi logistici*. Pitagora.

Gillo, M. W. (1972). Maid, a honeywell 600 program for an automatized survey analysis. Behaviorial Science 17 (2), 251-252.

Günther, H. O., & Kim, K. H. (2006). Container terminals and terminal operations. *OR Spectrum*, *28*(4), 437-445.

Han, X. L., Lu, Z. Q., & 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.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2009). The elements of statistical learning: data mining, inference and prediction. , 2nd edn. Springer-Verlag, New York.

Heaver, T., Meersman, H., & Van de Voorde, E. (2001). Co-operation and competition in international container transport: strategies for ports. *Maritime Policy & Management*, *28*(3), 293-305.

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

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3).

Hunt, E. B., Marin, J., & Stone, P. J. (1966). Experiments in induction. Academic Press, New York.

Hui, E. C., Seabrooke, W., & Wong, G. K. (2004). Forecasting cargo throughput for the port of Hong Kong: error correction model approach. *Journal of urban planning and development*, *130*(4), 195-203.

Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages.*The American Statistician*, *50*(4), 361-365.

Janssen, P., Bidlot, J. R., Abdalla, S., & Hersbach, H. (2005). Progress in ocean wave forecasting at {ECMWF}.

Janssen, P. A. (2008). Progress in ocean wave forecasting. *Journal of Computational Physics*, *227*(7), 3572-3594.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, 119-127.

Komen, G. J., Hasselmann, K., & Hasselmann, K. (1984). On the existence of a fully developed wind-sea spectrum. *Journal of Physical Oceanography*, *14*(8), 1271-1285.

Komen, G. J., Cavaleri, L., Donelan, M., Hasselmann, K., Hasselmann, S., & Janssen, P. A. E. M. (Eds.). (1996). *Dynamics and modelling of ocean waves*. Cambridge University Press.

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

Kumar, S., & Hoffmann, J. (2002). Globalisation: the maritime nexus. *The handbook of maritime economics and business*, 35-62.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 (1), 159-174.

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

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R news*, *2*(3), 18-22.

Loh, W. Y., & Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis. *Journal of the American Statistical Association*, *83*(403), 715-725.

Markham, I. S., Mathieu, R. G., & Wray, B. A. (2000). Setting through artificial intelligence: a comparative study of artificial neural networks and decision trees. *Integrated Manufacturing Systems*, *11*(4), 239-246.

Mathys, C. Economic Importance of the Belgian Ports: Flemish maritime ports, Liège port complex and the port of Brussels – Report 2011, National Bank of Belgium. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2293030. Accessed June 20, 2013.

Mayer, C., & Sinai, T. (2002). *Network effects, congestion externalities, and air traffic delays: Or why all delays are not evil*. The American Economic Review 93 (4), 1194–1215.

Mehta, M., Agrawal, R., & Rissanen, J. (1996). SLIQ: A fast scalable classifier for data mining. In *Advances in Database Technology—EDBT'96* (pp. 18-32). Springer Berlin Heidelberg.

Meersman, H., Sys, C., Van de Voorde, E. and T. Vanelslander. *Indicatorenboek Duurzaam Goederenvervoer 2010-2011*. http://www.ua.ac.be/download.aspx?c=.GOEDERENENPERSONENVERVOER&n=107948&ct=107948&e=309649. Accessed June 20, 2013.

Mola, F., & Siciliano, R. (1997). A fast splitting procedure for classification trees. *Statistics and Computing*, *7*(3), 209-216.

Monaco, M. F., Moccia, L., & Sammarra, M. (2009). Operations Research for the management of a transhipment container terminal: The Gioia Tauro case.*Maritime Economics & Logistics*, *11*(1), 7-35.

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

Morgan, J. N., & Messenger, R. C. (1973). THAID: a sequential search program for the analysis of nominal scale dependent variables. *Survey Research Center, Institute for Social Research, University of Michigan.*

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.

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.

Notteboom, T. (2012). Dynamics in port competition in Europe: implications for North Italian ports. In *Workshop 'I porti del Nord'–Milano*.

Notteboom, T. E. (2006). The time factor in liner shipping services. Maritime Economics & Logistics, 8(1), 19-39.

Notteboom, T., & 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.

Notteboom, T., & Rodrigue, J. P. (2012). The corporate geography of global container terminal operators. Maritime Policy & Management, 39(3), 249-279.

Pani, C., Cannas, M., Fadda, P., Fancello, G., Frigau, L., & Mola, F. (2013). A comparison of Machine Learning methods for delay level prediction in Transhipment Container Terminals. *Proceedings of the* the World Conference on Transport Research, forthcoming.

Pisano M., (2008), Un sistema di supporto alle decisioni per la pianificazione delle operazioni in un terminal container con funzione di transhipment., PhD Thesis, Facoltà di Ingegneria. Università di Cagliari.

Persson, A., & Grazzini, F. (2005). User Guide to ECMWF forecast products.*Meteorological Bulletin*, *3*, 2.

Port of Antwerp Authority. Port of Antwerp presentation. Antwerp, 2013.

Port of Cagliari Authority. http://www.porto.cagliari.it/index.php?option=com_content&view=article&id=79&Itemid=82&lang=it

Potvin, J. Y., & Smith, K. (2003). Artificial neural networks for combinatorial optimization. *Handbook of Metaheuristics*, 429-455.

PSA-Antwerp, Terminals. http://www.psa-antwerp.be/. Accessed July 2, 2013.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning* (Vol. 1). Morgan kaufmann.

R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. http://www.R-project.org.

Royston, P. (1995). Remark AS R94: A remark on algorithm AS 181: The W-test for normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *44*(4), 547-551.

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.

Santos, G., & Robin, M. (2010). Determinants of delays at European airports. *Transportation Research Part B: Methodological*, *44*(3), 392-403.

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

Shafer, J., Agrawal, R., Mehta, M. (1996). Sprint: A scalable parallel classifier for data mining. In: Proceedings of the 22nd international conference on very large data base. Mumbai, India.

Sideris, A., Boilé, M. P., & Spasovic, L. N. (2002). Using On-Line Information To Estimate Container Movements For Day-To-Day Marine Terminal Operations. 81st Annual Meeting of the Transportation Research Board, Washington.

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

Steenken, D., Voß, S., & Stahlbock, R. (2004). Container terminal operation and operations research-a classification and literature review. *OR spectrum*,*26*(1), 3-49.

Su, X., Wang, M., & Fan, J. (2004). Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, *13*(3), 586-598.

Thana, E. (2013). MARITIME TRANSPORT AND TOURISM SUSTAINABLE DEVELOPMENT. *European Scientific Journal*, *9*(19).

Tongzon, J., & 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.

UNCDAD 2012. *Review of Maritime Transport* 2012. Geneva: United Nations.

Tu, Y., Ball, M. O., & Jank, W. S. (2008). Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern. *Journal of the American Statistical Association*, *103*(481), 112-125.

Vacca, I., Bierlaire, M., & Salani, M. (2007). Optimization at container terminals: status, trends and perspectives. In *Proc. of Swiss Transport Research Conference*.

Vanelslander, T. (2005). The economics behind cooperation and competition in sea-port container handling. PhD thesis, Faculty of Applied Economics, University of Antwerp.

Vernimmen, B., Dullaert, W., & Engelen, S. (2007). Schedule unreliability in liner shipping: Origins and consequences for the hinterland supply chain. Maritime Economics & Logistics, 9(3), 193-213.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam Med*, *37*(5), 360-363.

Vis, I. F., & de Koster, R. (2003). Transshipment of containers at a container terminal: An overview. *European Journal of operational research*, *147*(1), 1-16.

Vlaamse Havencommissie. Ladingen & lossingen Vlaamse havens sinds 1980: overzichtnaar verschijningsvorm http://www.vlaamsehavencommissie.be/vhc/page/ladingen-lossingen-overzicht-naar-verschijningsvorm-vlaamse-havens-1980-2011. Accessed June 23, 2013

Wang, T. F., & Cullinane, K. (2006). The efficiency of European container terminals and implications for supply chain management. *Maritime Economics & Logistics*, *8*(1), 82-99.

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

Won, S., & Kim, K. (2009). An integrated framework for various operation plans in container terminals. *Polish Maritime Research* 3 (61) 16, 51–61.

Woodburn, A., Allen, J., Browne, M., & Leonardi, J. (2008). The Impacts of Globalization on International Road and Rail Freight Transport Activity–Past Trends and Future Perspectives. *Transport Studies Department, University of Westminster*, London, UK.

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.

Zonglei, L., Jiandong, W., & Guansheng, Z. (2008). A new method to alarm large scale of flights delay based on machine learning. In *Knowledge Acquisition and Modeling, 2008. KAM'08. International Symposium on* (pp. 589-592). IEEE.

Appendix 1

Summary statistics of the categorical predictors at the Cagliari CT

This section shows the summary statistics of the continuous potential predictors collected at the Cagliari container terminal.

VESSEL TYPE	**Frequency**
MOTHER	912
FEEDER	1057

VESSEL OWNER	**Frequency**
ARKAS CONTAINER TRANSPORT S.A.	268
CMA CGM	31
EMES SHIPPING & TRANSPORT	136
HAMBURG SUD	99
HANJIN SHIPPING	42
HAPAG-LLOYD	616
METZ CONTAINER LINE	15
NIPPON YUSEN KAISHA	5
ORIENT OVERSEAS CONTAINER LINE	158
SEA STAR LINE	4
UNITED ARAB SHIPPING COMPANY	7
UNITED FEEDER SERVICES	361
XPC	231

VESSEL SERVICE	Frequency
AEX	260
ASA	191
CBS	85
CMS	8
CSA	64
EU2	15
IAS	37
IAX	29
IOS	260
LES	1
LTX	68
LVA	75
MCA	134
MGX	131
MINA	8
MPS	103
MSX	4
NAX	18
SAX	53
SPOT	75
STX	79
TLB	207
TYR	56
WBS	8

The vessel service considers all the three variables related to service together i.e. port rotation, sailing direction and previous port. Some examples of the most frequent vessel service are illustrated as follow.

SERVICE	PORT 1	PORT 2	PORT 3	PORT 4	PORT 5	PORT 6	PORT 7	PORT 8	PORT 9	PORT 10	PORT 11	PORT 12	PORT 13	PORT 14	PORT 15
AEX	**Cagliari** ITALY	**Halifax** CANADA	**New York** USA	**Savannah** USA	**Norfolk** USA	**New York** USA	**Halifax** CANADA	**Cagliari** ITALY	**Jeddah** SAUDI ARABIA	**Colombo** SRI LANKA	**Singapore** SINGAPORE	**Vungtao** VIETNAM	**Laem Chabang** THAILAND	**Singapore** SINGAPORE	**Colombo** SRI LANKA

SERVICE	PORT 1	PORT 2	PORT 3	PORT 4	PORT 5	PORT 6	PORT 7	PORT 8	PORT 9	PORT 10
IOS	**Cagliari** ITALY	**Hamburg** GERMANY	**Tilbury** UNITED KINGDOM	**Antwerp** BELGIUM	**Tanger** MOROCCO	**Cagliari** ITALY	**Jebel ali** UNITED A. EMIRATES	**Karachi** PAKISTAN	**Nhava Sheva** INDIA	**Mundra** INDIA

SERVICE	PORT 1	PORT 2	PORT 3	PORT 4	PORT 5	PORT 6	PORT 7	PORT 8	PORT 9	PORT 10	PORT 11	PORT 12	PORT 13
TLB	**Cagliari** ITALY	**Genoa** ITALY	**La Spezia** ITALY	**Leghorn** ITALY	**Naples** ITALY	**Cagliari** ITALY	**Mersin** TURKEY	**Haifa** ISRAEL	**Ashdod** ISRAEL	**Alexandria** EGYPT	**Salerno** ITALY	**Naples** ITALY	**Trapani** ITALY

SAILING DIRECTION	Frequency
Eastbound	309
Westbound	297
Tyrrhenian bound	98
Levant bound	109
Standard	1156

PREVIOUS PORT	Country	Nautical Miles Distance	Frequency
Alexandria	Egypt	1,130	7
Algeciras	Spain	714	5
Algiers	Algeria	329	13
Ancona	Italy	846	88
Annaba	Algeria	154	2
Antwerp	Belgium	2,070	1
Ashdod	Israel	1,363	3
Barcelona	Spain	371	109
Castellon	Spain	448	9
Catania	Italy	371	7
Civitavecchia	Italy	234	1
Colombo	Sri Lanka	4,719	105
Damietta	Egypt	1,260	2
Fos Sur Mer	France	365	1
Genoa	Italy	354	16
Gioia Tauro	Italy	325	13
Halifax	UK	3,390	119
Haydarpasa	Turkey	1,119	3
Istanbul			59
Izmir	Turkey	961	71
Jeddah	Saudi Arabia	1,874	44
La Spezia	Italy	326	3
Livorno	Italy	301	42
Malta	Malta	329	84
Marseille	France	350	3
Mersin	Turkey	1,346	10
Messina	Italy	351	1
Montreal	Canada	3,889	4
Mundra	India	4,171	60
Naples	Italy	267	109
New Orleans	USA	5,210	43
Palermo	Italy	217	13
Piraeus	Greece	884	9
Port Everglades	Florida	4,530	8
Port Said	Egypt	1,241	9
Ravenna	Italy	920	3
Rijeka	Croatia	437	2
Salerno	Italy	284	32
Savannah	Georgia	4,235	14
Southampton	England	1,875	16

PREVIOUS PORT	Country	Nautical Miles Distance	Frequency
Tanger	Morocco	751	130
Tarragona	Spain	406	4
Thessaloniki	Greece	1,015	91
Trapani	Italy	177	3
Tunis	Tunisia	163	63
Vado Ligure	Italy	355	7
Valencia	Spain	456	228

Appendix 2

Summary statistics of the categorical predictors at the Antwerp CT

PREVIOUS PORT	Country	Nautical Miles Distance	Frequency
Aarhus	Denmark	511	13
Abidjan	Cote D'Ivoire	3,725	3
Algeciras	Spain	1,357	45
Boston	USA	3,091	1
Bremerhaven	Germany	319	443
Buenos aires	Argentina	6,361.81	1
Charleston	South Carolina	3,771	112
Copenhagen	Denmark	536	4
Dunkerque	France	98	20
Felixstowe	England	137	208
Gavle	Sweden	965	8
Gotthenburg	Sweden	560	31
Hamburg	Germany	370	209
Helsingborg	Sweden	549	4
Helsinki	Finland	998	4
Jeddah	Saudi Arabia	4,000	1
Kiel	Germany	387	1
Kotka	Finland	1,057	6
Liverpool	England	668	1
Port Kelang	Malesia	8,126	3
Port said	Egypt	3,280	2
Rauma	Finland	986	29
Rotterdam	Netherlands	126	167
Santa Marta	Colombia	4,510	2
Shanghai	China	10,463	1
Singapore	Singapore	8,300	1
St Petersburg	Russia	1,146	9
Tanger	Morocco	1,336	22
Tanjung Pelepas	Malaysia	8,288	2
Valencia	Spain	1,740	2
Yantian	Cina	11,206	1
Zeebrugge	Belgium	62	5