Predicting Anchorage Duration of Commercial Vessels

Milad Malekipirbazari, Vural Aksakalli, and Y. Volkan Aydogdu

***Abstract*—Seaborne transportation accounts for more than 90% of world's trade. Anchorages serve as a temporary waiting area for commercial vessels for reasons such as supply, waiting for passage, and as refuge from bad weather. In many anchorages, authorities do not pose any restrictions on the anchorage duration and incoming vessels are not obligated to disclose an estimated duration time, which makes it difficult to manage traffic flow inside the anchorage in an efficient manner. In this study, we first provide a brief statistical analysis on an anchorage dataset for Istanbul Strait anchorages between 2006 and 2014. Next, using this dataset, we propose a data mining framework for predicting anchorage duration for an incoming vessel, which is critical for efficient anchorage planning. Our results indicate that decision trees provide superior prediction performance compared to alternative methods and reason of anchorage has the highest association with anchorage duration.**

***Index Terms*—Anchorage, Istanbul Strait, decision tree.**

I. Introduction

The academic field of maritime traffic engineering was introduced by Toyoda and Yahei in 1971 with the aim of improving maritime traffic regulations and better utilization of maritime resources [1]. Recently, anchorage utilization optimization and anchorage planning have been the center of focus and debate. In this regard, Bijwaard and Knapp [2] performed a duration analysis and generated vessel life cycles for better assessment of vessel conditions and reducing the possibility of incidents. The capacity of multiple anchorages was evaluated by [3] using a simulation-based model and several methods regarding improvement of space utilization were proposed. Concerning optimal navigation of vessels in the presence of obstacles, a graph theoretical approach was presented and applied on an ice navigation case study [4]. Silveira *et al*. analyzed the risk of collision near the ports of Portugal [5]. There exists some research on analysis and improvement of maritime traffic specifically in the Istanbul Strait via different approaches such as offering a mathematical formulation of the current scheduling practices [6], proposing a specific navigation safety support model [7], suggesting local traffic separation schemes [8], evaluating the performance of an online precise point positioning service for vessel positioning in Halic Bay [9], and using generic fuzzy analytic hierarchy methods for risk evaluation [10].

Manuscript received August 30, 2015; revised January 11, 2016.

M. Malekipirbazari and V. Aksakalli are with the Department of Industrial Engineering, Istanbul Sehir University, Istanbul, 34662 Turkey (e-mail: miladmalekipirbazari@std.sehir.edu.tr, aksakalli@sehir.edu.tr).

Y. V. Aydogdu is with the Maritime Faculty, Istanbul Technical University, Istanbul, 34940 Turkey (e-mail: yvaydogu@itu.edu.tr).

There exist several studies using machine learning and data mining techniques in order to analyze maritime traffic and extract information for efficient management of maritime traffic flow. Among these, different classification techniques were applied on vessel arrivals at a port in order to predict their future locations [11]. Using clustering and statistical methods, a data mining approach was presented for prediction of maritime traffic flow patterns [12]. Tsou employed association rule discovery method for analysis of Keelung Harbor navigation conditions [13]. Clustering algorithms along with three neighborhood models were employed in order to detect vessel traffic areas in the Shanghai Strait [14]. While these studies proved that machine learning and data mining methods are capable of providing useful information concerning maritime traffic management, none of them considered the issue of anchorage duration. To our knowledge, our work is the first in the literature that tackles the problem of anchorage duration prediction.

The Istanbul Strait is one of the busiest waterways in the world and it is the only sea route between Mediterranean, Aegean, and the Black Sea. The Strait divides the City of Istanbul into European and Asian parts and makes the city a significant logistics node in the entire region [15]. Commercial vessels have the right to pass freely through the Istanbul and Dardanelles Straits and drop anchor at north and south sides of these straits in peacetime [16]. However, heavy local traffic causes complications for transient commercial vessels in the Straits as there are no alternative sea routes in the region. Considering rapid expansion of global shipping and limited anchorage capacity of Istanbul anchorages, a comprehensive analysis of anchorage traffic and prediction of the anchorage duration of vessels are in order.

This manuscript is concerned with developing a systematic approach for predicting vessel anchorage duration in Istanbul anchorages based on data recorded between 2006 and 2014 consisting of attributes including anchorage reason, vessel type, length, flags and, date and duration of anchor [17]. In this work, we first analyze the nature and relationships of these attributes via basic statistical analysis techniques. Afterwards, in order to identify the best prediction model, we perform a thorough investigation for eliminating unimportant attributes as well as performing attribute transformations with a proper discretization method. Finally, we evaluate the accuracy of several data mining methods on the pre-processed data and we present a performance comparison of these methods.

II. The Data

The available data was recorded by Turkish Directorate General of Coastal Safety and it includes historical records of 13 attributes related to the anchorage of vessels from 2006 to

doi: 10.18178/ijmlc.2016.6.1.573

2014 in the anchorages of Istanbul. There are 443339 observations in the dataset with both categorical data such as vessel type, flag, and anchorage reason, as well as numeric attributes such as anchorage duration, length, and gross tonnage of vessels. In order to gain a better understanding of this data, the type and number of levels of these attributes are reported in Table I.

TABLE I: DESCRIPTION OF THE ATTRIBUTES

Attribute	Type	Number of levels
Reason	Nominal	5
Zone	Nominal	3
Year	Ordinal	8
Month	Ordinal	12
Vessel Type	Nominal	73
Flag	Nominal	126
Arrival Country	Nominal	165
Departure Country	Nominal	200
Arrival Port	Nominal	1778
Departure Port	Nominal	1397
Length	Numeric	N/A
Gross Tonnage	Numeric	N/A
Duration	Numeric	N/A

Each year, more than 50,000 vessels anchor in Istanbul anchorages for various reasons including planning, supply, port, and as refuge from adverse weather conditions. These anchorages are comprised of three zones: Southern, Northern, and Eastern. Each vessel is distinguished with a flag demonstrating its country of registration and the vessel is required to follow the rules of its flag country.

A vessel's intended duration of stay in an anchorage is a critical parameter for efficient anchorage management [18]. Fig. 1 shows the log of anchorage duration histogram as well as its boxplot. We observe that commercial vessels anchor in Istanbul anchorages with a mean duration of about 12 hours.

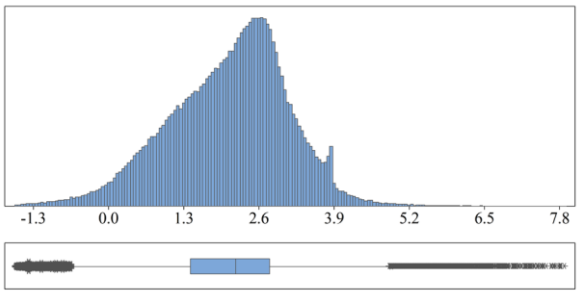

Fig. 1. Histogram and box plot of log vessel anchorage duration.

Mean anchorage durations of vessels with respect to the attributes of reason, zone, and year are shown in Fig. 2 and Fig. 3. Reasons of planning, port, and supply have a relatively stable duration mean over time. On the other hand, anchorage reasons of weather and other have relatively irregular trends. The variation in duration mean for rough weather conditions can be explained by atmospheric conditions, yet irregular alterations of other causes cannot be clarified due to the fact that the exact reason was not recorded. Unlike Southern zone, Northern and Eastern zones have very volatile duration means, which along with the unknown causes of anchorages makes duration prediction even more challenging.

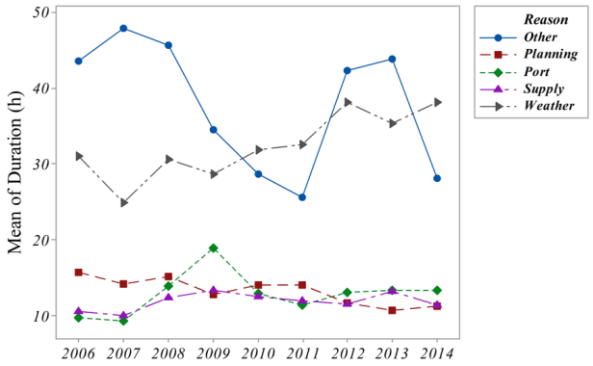

Fig. 2. Duration vs. year and reason.

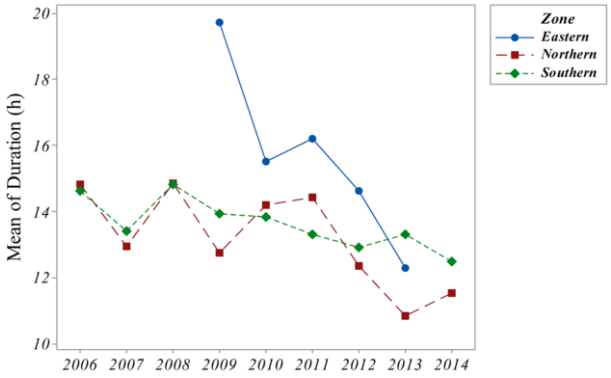

Fig. 3. Duration vs. year and zone.

III. PREDICTION OF ANCHORAGE DURATION

Data mining is the process of extracting information from a set of data. This computational process consists of discovering previously unknown patterns from typically large amounts of data using mathematical, statistical, and computer science tools. Besides numerous data mining software like R, Weka, and Rapid Miner, there are various data mining techniques each of which have their own advantages and disadvantages and should be selected based on the nature of parameters and constraints in the dataset. Most commonly used data mining techniques are artificial neural networks, support vector machines, decision trees, naïve Bayes, and nearest-neighbor classifiers.

As mentioned before, our main goal is to estimate the duration of vessel anchor by identifying the attributes responsible for the variation in the response variable. This section briefly reviews the concepts and techniques of data mining employed in this study.

A. Data Preprocessing

Prior to deployment of any data mining algorithms, the original data usually requires quality improvement, which is called data preprocessing. This step consists of data cleaning, transformation, normalization, and feature extraction and selection procedures.

1) Data cleaning

Data cleaning is the process of taking appropriate actions for inaccurate, missing, or corrupt observations. The anchorage dataset does not have any missing values and a preliminary analysis indicated that there are no outliers in the

data that requires special attention.

2) Data transformation

Anchorage duration varies from a few hours to a few months and the histogram in Fig. 1 suggests that it closely follows a log-normal distribution. In order to make the prediction task more manageable, we discretized the duration attribute into roughly equal-sized bins as shown in Table II. In addition, we discretized the numeric attributes of length and gross tonnage with equal entropy, which resulted in 10 and 7 classes respectively. This way, we were able to convert the prediction problem to a classification problem with all nominal attributes.

TABLE II: DISCRETIZED DURATION WITH FIVE INTERVALS

Representation	Description	Range (h)
VS	Very Short Duration	0 - 3.33
S	Short Duration	3.33 - 6.83
M	Medium Duration	6.83 - 11.55
L	Long Duration	11.55 - 18.49
VL	Very Long Duration	18.49 - ∞

3) Dimension reduction

Having a large number of attributes in a dataset will not only slow down the classification task and require excessive storage, but it will also result in poor generalization, reduce classification accuracy, and make it difficult to interpret the final model. Thus, a critical task is to identify the truly associated attributes and remove the irrelevant and redundant ones.

Various attribute ranking and selection methods have been proposed in the literature including information gain, gain ratio, symmetrical uncertainty, one-R and Chi-square tests. In this study, we consider information gain and Chi-square tests for the attribute selection task.

B. Classification Models

As discussed earlier, numerous data mining algorithms are available for extraction of information from a given dataset. In this study, due to the categorical nature of the (transformed) attributes and the large amount of observations, we considered the following three classifiers: Decision Tree, Nearest Neighbor, and Naïve Bayes, which are briefly described below.

1) Decision tree

Decision tree learning is a popular data mining method that employs the predictive model of decision trees by mapping groups of attribute levels to decisions about the predicted class label. A decision tree is a tree-like structure where each interior node is labelled with an input attribute and the emanating arcs from the node are marked with possible levels of other attributes, leading to multiple leaves representing a class label for the response variable.

Decision tree classifiers have several advantages compared to other data mining approaches: they are easy to comprehend and interpret, capable of handling both numerical and categorical data, able to perform proper classification for large datasets in reasonable time, and they typically require little data preprocessing. However, having too many leaves might generate overly complex trees with lower accuracy, resulting in overfitting. In order to avoid this problem, mechanisms such as pruning are employed to remove these problematic leaves, which decrease the complexity of the final classifier and improve classification accuracy.

2) Nearest neighbor

The algorithm of *k*-Nearest Neighbors (*k*-NN) is a non-parametric technique where the input consists of the *k* nearest instances of the observation to be classified with respect to a distance metric. This simple classification rule is based on a majority vote of these *k* nearest neighbors, with typically small *k* between 1 and 10. Sometimes it can be helpful to weight the neighbors' contributions in a way that the closer neighbors have more effect on the final decision that the farther ones.

3) Naïve Bayes

Naïve Bayes classifier is a probabilistic algorithm that makes use of the Bayes' theorem under the assumption of strong independency between the attributes. It assumes the significance of a specific attribute is isolated from the occurrence or absence of any other attribute regarding the class label. Despite this oversimplified assumption, Naïve Bayes has been shown to work reasonably well in many complex classification problems.

C. Performance Criteria for Model Evaluation

Model evaluation is an essential step in deciding on the final model. Using a pre-specified performance metric, we can compare various classification models and identify the one with the highest performance. There exist various performance metrics for model evaluation including *K*-fold cross-validation accuracy, area under the ROC curve, root mean squared error, and resubstitution and hold-out accuracies. In this study, we consider resubstitution and hold-out accuracy performance metrics due to their simplicity and ease-of-use. In the former, the assessment is performed on the training data used during the model generation process. In the latter, the available data is first partitioned into train and test data (which was 70% and 30% in our case). The model is built using the train data and then it is tested on the test data.

IV. RESULTS AND DISCUSSION

This section presents data preprocessing and classification results with the three data mining methods described above.

A. Data Preprocessing Results

The result of the attribute ranking process using the information gain criterion is presented in Table III. From this table, it can be inferred that the reason attribute has the strongest association with anchorage duration when compared against other attributes. Furthermore, attributes of flag and vessel type are among the lowest in terms of information gain, revealing their weak association with the response variable.

For the purpose of reducing the number of attributes, association analysis is performed through correlation and Chi-squared tests. The former resulted in a high correlation of 0.931 for the attributes of length and gross tonnage, and the latter revealed a high level of association between the reason and zone attributes. Since the reason attribute has the highest

information gain, we kept this attribute and removed the zone attribute. In addition, based on the higher information gain measure of gross tonnage compared to length, we decided to keep gross tonnage and remove length.

TABLE III: ATTRIBUTES RANKED BY INFORMATION GAIN

Rank	Attribute	Information Gain
1	Reason	0.153
2	Zone	0.039
3	Arrival Port	0.031
4	Departure Port	0.031
5	Month	0.015
6	Arrival Country	0.012
7	Departure Country	0.011
8	Gross Tonnage	0.008
9	Year	0.008
10	Length	0.008
11	Flag	0.008
12	Vessel Type	0.003

Attributes with very large number of levels such as departure and arrival ports (with 1397 and 1778 levels respectively) are likely to might make the final model more complicated and less accurate. One approach to mitigate this problem is to combine their levels and generate a higher level aggregation. In this particular case, there exist natural higher level attributes in the form of departure and arrival countries that are good representations of departure and arrival ports.

Subsequent to the data preprocessing described above, we are left with the 6 attributes in Table IV to be used for classification.

TABLE IV: ATTRIBUTES USED FOR CLASSIFICATION

Rank	Variable	Explanation
1	Reason	Cause of anchor
2	Month	Month of anchor
3	Arrival Country	Country the vessel will arrive at
4	Departure Country	Country the vessel departed from
5	Gross Tonnage	Weight of vessel
6	Year	Year of anchor

TABLE V: PREDICTION ACCURACIES

Method	Resubstitution accuracy	Hold-out accuracy
1-Nearest Neighbor	0.525	0.274
5-Nearest Neighbor	0.475	0.350
Naïve Bayes	0.345	0.353
Decision Tree (unprunned)	**0.750**	0.310
Decision Tree (prunned)	0.480	**0.380**

B. Classification Results

The effect of pruning in Decision Trees and number of neighbors in Nearest Neighbor were evaluated using the performance metrics of resubstitution and hold-out accuracies. Performance comparison of these classifiers as well as Naïve Bayes is displayed in Table V. It can be seen from this table that by increasing the number of neighbors in *k*-NN, resubstitution accuracy decreases yet hold-out accuracy improves. An intuition behind this observation is that 1-NN overfits the data (as expected) and gives a better resubstitution accuracy. On the other hand, 5-NN exhibits less overfitting and yields better generalization, resulting in better hold-out accuracy. Similarly, Decision Tree pruning increases the hold-out accuracy due to less overfitting and better generalization. We observe that among these five classifiers, Decision Tree has the highest performance and it is the recommended classifier for anchorage duration prediction.

V. SUMMARY AND CONCLUSIONS

In this study, we present a data mining framework for vessel anchorage duration prediction. We chose Istanbul Strait anchorages as a case study and we used the data for these anchorages between 2006 and 2014 for building and testing our prediction algorithms. For this analysis, first we explored the statistical relationships between the available attibutes for vessels and their duration of stay. This analysis indicated that length and gross tonnage as well as reason and zone have high correlations. Subsequent to a data preprocessing step, we observed that the key attributes for duration prediction are reason of anchor, month, arrival and departure countries, gross tonnage, and year. With the goal of identifying the best prediction method for anchorage duration, we compared several methods and we found that decision tree is the superior technique. With this classifier, we achieved a resubstitution accuracy of 75% and a hold-out accuracy of 38%. In this work, it was also observed that reason of anchorage is the dominant attribute and has the highest association with anchorage duration.

ACKNOWLEDGMENT

This work was supported by TUBITAK (The Scientific And Technological Research Council of Turkey), Grant Number 113M489.

REFERENCES

[1] S. Toyoda and F. Yahei, "Marine traffic engineering," *Journal of Navigation*, vol. 24, pp. 24-34, 1971.
[2] G. E. Bijwaard and S. Knapp, "Analysis of ship life cycles — the impact of economic cycles and ship inspections," *Marine Policy*, vol. 33, pp. 350–369, 2009.
[3] S. Y. Huang, W. J. Hsu, and Y. He, "Assessing capacity and improving utilization of anchorages," *Transportation Research Part E: Logistics and Transportation Review*, vol. 47, pp. 216–227, 2011.
[4] I. Ari, V. Aksakalli, V. Aydogdu, and S. Kum, "Optimal ship navigation with safety distance and realistic turn constraints," *European Journal of Operational Research*, vol. 229, pp. 707–717, 2013.
[5] P. A. M. Silveira, A. P. Teixeira, and C. G. Soares, "Use of AIS data to characterize marine traffic patterns and ship collision risk off the Coast of Portugal," *The Journal of Navigation*, vol. 66, pp. 879–898, 2013.
[6] O. S. Uluscu, B. Ozbas, T. Altok, I. Or, and T. Yilmaz, "Transit vessel scheduling in the Strait of Istanbul," *The Journal of Navigation*, vol. 62, pp. 59–77, 2009.
[7] M. A. Yazici and E. N. Otay, "A navigation safety support model for the Strait of Istanbul," *The Journal of Navigation*, vol. 62, pp. 609–630, 2009.
[8] Y. V. Aydogdu, C. Yurtoren, J. S. Park, and Y. S. Park, "A study on local traffic management to improve marine traffic safety in the Istanbul Strait," *The Journal of Navigation*, vol. 65, pp. 99–112, 2012.
[9] R. M. Alkan and T. Ocalan, "Usability of the GPS precise point positioning technique in marine applications," *The Journal of Navigation*, vol. 66, pp. 579–588, 2013.
[10] Y. V. Aydogdu, "A comparison of maritime risk perception and accident statistics in the Istanbul Straight," *The Journal of Navigation*, vol. 67, pp. 129–144, 2014.
[11] R. Lagerweij, "Learning a model of ship movements," Master's thesis, University of Amsterdam, 2009.

[12] C. Tang and Z. Shao, "Data mining platform based on AIS data," in *Proc. International Conference on Transportation Engineering*, China, pp. 4465–4470, 2009.
[13] M. C. Tsou, "Discovering knowledge from AIS database for application in VTS," *The Journal of Navigation*, vol. 63, pp. 449–469, 2010.
[14] K. M. S. Oo, C. Shi, H. Qinyou, and A. Weintrit, "Clustering analysis and identification of marine traffic congested zones at Wusongkou, Shanghai," *Zeszyty Naukowe Akademii Morskiej W Gdyni*, vol. 67, pp. 101–113, 2010.
[15] D. Özdemir, "Strategic choice for Istanbul: A domestic or international orientation for logistics?" *Cities*, vol. 27, pp. 154-163, 2010.
[16] J. R. Morgan, "The Turkish straits: Christos L. Rozakis and Petros N. Stagos Martinus Nijhoff Publishers, Dordrecht, 1987," *Marine Policy*, pp. 173-174, 1989.
[17] M. Malekipirbazari, V. Aksakalli, and V. Aydogdu, "A Temporal analysis of vessel type traffic in Istanbul Strait anchorages," in *Proc. International Conference on Industrial Engineering and Operations Management*, Dubai, UAE, 2015.
[18] D. Oz, V. Aksakalli, A. F. Alkaya, and V. Aydogdu, "Anchorage planning with utilization and safety considerations," *Computers and Operations Research*, vol. 62, pp. 12-22, 2015.

M. Malekipirbazari received his B.Sc. and M.Sc. degrees in chemical engineering from Iran. He also received a M.Sc. degree in industrial and systems engineering from Istanbul Sehir University, Turkey, where he is currently employed as a research assistant. His research interests include machine learning, data mining, statistical analysis, and optimization.

V. Aksakalli received his B.Sc. degree in mathematics from Middle East Technical University in Ankara, Turkey, his M.Sc. degree in industrial engineering and operations research from North Carolina State University in Raleigh, NC, and M.Sc. and Ph.D. degrees in applied mathematics & statistics from Johns Hopkins University in Baltimore, Maryland. He is currently an Associate Professor of Industrial Engineering at Istanbul Sehir University in Turkey. His research interests are in stochastic optimization, machine learning, and applied probability and statistics.

Y. V. Aydogdu is an Associate Professor in Maritime Faculty of Istanbul Technical University, Istanbul, Turkey. He earned his Ph.D. degree in Maritime Traffic Information Engineering from Graduate School of Korea Maritime University. He received his B.Sc. degree from Istanbul Technical University, Deck Department of Maritime Faculty and also earned his M.Sc. degree in the Department of Maritime Transportation Engineering from Graduate School of Istanbul Technical University. He has experience as navigation officer, flag state inspector and training manager in the maritime field. His research areas of interest are local marine traffic management and port management.