NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

CHARACTERIZING SHIP NAVIGATION PATTERNS USING AUTOMATIC IDENTIFICATION SYSTEM (AIS) DATA IN THE BALTIC SEA

by

Janet S. von Eiff

March 2018

Thesis Advisor: Robert Koyak
Second Reader: Sam Huddleston

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704–0188*
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** March 2018	**3. REPORT TYPE AND DATES COVERED** Master's thesis	
4. TITLE AND SUBTITLE CHARACTERIZING SHIP NAVIGATION PATTERNS USING AUTOMATIC IDENTIFICATION SYSTEM (AIS) DATA IN THE BALTIC SEA			**5. FUNDING NUMBERS**
6. AUTHOR(S) Janet S. von Eiff			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A			**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB number ____N/A____.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.			**12b. DISTRIBUTION CODE**

13. ABSTRACT (maximum 200 words)

The Intelligence, Surveillance, and Reconnaissance (ISR) community is interested in developing a model that can assist in characterizing patterns of ship navigation. We examine techniques used to highlight those patterns using historical Automatic Identification System (AIS) data in the Baltic Sea from January to April 2014. A regression model is used to determine which factors influence the amount of time a cargo ship spends in a port in the Saint Petersburg, Russia, area. We find that the best model is able to explain about 29 percent of the variance of the length of time that a vessel is in the Saint Petersburg area. We use three random forest models, that differ in their use of past information, to predict a vessel's next port of visit. The random forest models we use in this analysis demonstrate that predicting a vessel's next port of call is not a Markov model but a higher-order network where past information is used to more accurately predict the future state. The transitional probabilities change when predictor variables are added that reach deeper into the past. Our findings suggest that successful prediction of the movement of a vessel depends on having accurate information on its recent history.

14. SUBJECT TERMS AIS, Baltic Sea, data analysis, regression, higher-order network, Bayes information criterion, random forest			**15. NUMBER OF PAGES** 85
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540–01-280-5500 Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

CHARACTERIZING SHIP NAVIGATION PATTERNS USING AUTOMATIC IDENTIFICATION SYSTEM (AIS) DATA IN THE BALTIC SEA

Janet S. von Eiff
Lieutenant, United States Navy
B.S., United States Naval Academy, 2011

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 2018

Approved by: Robert A. Koyak
Thesis Advisor

Sam Huddleston
Second Reader

Patricia Jacobs
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Intelligence, Surveillance, and Reconnaissance (ISR) community is interested in developing a model that can assist in characterizing patterns of ship navigation. We examine techniques used to highlight those patterns using historical Automatic Identification System (AIS) data in the Baltic Sea from January to April 2014. A regression model is used to determine which factors influence the amount of time a cargo ship spends in a port in the Saint Petersburg, Russia, area. We find that the best model is able to explain about 29 percent of the variance of the length of time that a vessel is in the Saint Petersburg area. We use three random forest models, that differ in their use of past information, to predict a vessel's next port of visit. The random forest models we use in this analysis demonstrate that predicting a vessel's next port of call is not a Markov model but a higher-order network where past information is used to more accurately predict the future state. The transitional probabilities change when predictor variables are added that reach deeper into the past. Our findings suggest that successful prediction of the movement of a vessel depends on having accurate information on its recent history.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

AIS	Automatic Identification System
MDA	Maritime Domain Awareness
AOI	Area of Interest
VOI	Vessel of Interest
COLREGS	International Regulations for Preventing Collisions at Sea
IMO	International Maritime Organization
ISR	Intelligence, Surveillance, and Reconnaissance
SOLAS	Safety of Life at Sea
GT	Gross Tonnage
CRBM	Conditional Restricted Boltzmann Machine
OPTICS	Ordering Points to Identify the Cluster Structures
TREAD	Traffic Route Extraction and Anomaly Detection
DBSCAN	Density Based Spatial Clustering of Applications with Noise
MMC	Mobility Markov Chain
HON	Higher-Order Network
MMSI	Maritime Mobile Service Identity
UTC	Coordinated Universal Time
MID	Maritime Identification Digits
BIC	Bayes Information Criterion

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Countries that have coastal borders in areas with high levels of maritime activity require an accurate maritime picture for their national security. Maritime Domain Awareness (MDA) for those particular countries can be enhanced by using historical Automated Identification System (AIS) data to identify ship navigation patterns in the maritime area. The primary purpose of AIS is collision avoidance through its autonomous ability to identify other vessels that are fitted with an AIS transponder. Although the International Maritime Organization (IMO) mandates that vessels of a certain size are equipped with an AIS transponder, there is no law enforcement in any part of the world that enforces that regulation. Thus, the data received from AIS transmissions is not always complete and is subject to error.

AIS data is big data that is publicly available for research to be conducted on topics such as analyzing maritime traffic patterns, predicting vessel movements, and other topics related to enhancing MDA. The purpose of this thesis is to characterize ship navigation patterns using AIS data in the Baltic Sea to better allocate surveillance assets for specific ships. The data we use in this thesis is limited to a bounding box with latitude between 53 and 60.5 degrees north and longitude between 13.4, and 29.4 degrees east. Our dataset contains over 25 million observations that were obtained during a four-month time period beginning in January 2014. We limit the scope of our analysis to vessels that are self-identified as cargo ships. Maritime transportation occurs with high density in the Baltic Sea due to its many ports, which often are in close proximity to other ports. To reduce complexity, we create port clusters by combining ports within fifty miles of each other and which belong to the same country to reduce the number of port stops. By doing this, we reduce 78 unique ports within the bounding box to 34 port clusters. For the purpose of analysis, we focus on a cluster of three ports in the Saint Petersburg, Russia area and the cargo ships that stop there. We choose this cluster because it has the highest inflow and outflow of cargo ship traffic compared to the other Baltic port clusters.

Before we begin our analysis, we clean the data by scanning for outliers that may affect our analysis. Obtaining stop points is important to understanding the movement

patterns of a vessel. With the cleaned data, we determine stop points using defined thresholds and process the information to describe segments for each unique vessel. The segments are separated by a vessel's Maritime Mobile Service Identity (MMSI) number, which is a nine-digit number that correspond to a vessel's transponder. One vessel typically has multiple segments which when linked define its route through the Baltic Sea. We consider only the MMSIs that stop in the Saint Petersburg area at least once and examine the behavior of the vessels within this group.

Our first objective is to analyze the effects of different factors on the length of port stays in the Baltic region, focusing on the Saint Petersburg area. We use regression to identify factors that influence the amount of time that a cargo ship spends in port. The explanatory variables that we consider are:

- Prior port that the vessel visited before Saint Petersburg area
- Next port that the vessel transits to after Saint Petersburg area
- Departure day of the week
- Arrival day of the week
- Arrival day in the year (1 = January 1, 2 = January 2, …, 130 = April 30)
- MID (a three-digit code for the country a vessel is registered to)

We use a logarithmic transformation on the response variable, which is the length of the port stay, to improve the fit of the model. The variable Maritime Identification Digits (MID) is able to explain about 25 percent of the variance of the response variable and is determined to be the best explanatory variable in the regression analysis. This emphasizes the importance of maritime law enforcement ensuring that a vessel's MMSI is updated when first installed and that the MID matches the country flag that is being flown by the vessel. The second strongest explanatory variable is prior port, which highlights the importance of having past knowledge of the ports that a vessel visits during its voyage. Overall, the regression analysis explains about 29 percent of the variance of the response variable. The final model that we use considers Bayes Information Criterion (BIC) to penalize the addition of extra variables in the regression and avoid overfitting. The original

model has 121 variables that we reduce to 23 variables using BIC. This analysis is beneficial because it allows organizations that conduct surveillance to better understand the factors that influence how long a vessel of interest is in port. Knowing the duration of a port stay for a vessel of interest gives the organization time to allocate resources to the vessel of interest's current port or next port of visit.

Our second objective is to predict with greater accuracy whether a vessel will depart the Baltic Sea or visit another port within the Baltic Sea. We construct three random forest models, that differ in their use of past information, to predict the next port of visit. We use a binomial response variable that accounts for whether a vessel leaves or remains in the Baltic Sea after departing a port in the Saint Petersburg, Russia area. Our predictor variables include MID, arrival day of the week, arrival day in the year, the length of time a vessel stays in port, and previous ports visited. The overall misclassification rate when considering zero previous ports is 31.5 percent. It decreases by about three percent with the addition of one previous port as a predictor variable, resulting in an overall misclassification rate of 28.2 percent. The third model has the lowest misclassification rate, in comparison to the first two models, with 27.2 percent. This analysis shows that there is potential to more accurately predict the next port of visit for a vessel given its previous destinations.

Our third objective is to predict, with greater accuracy, a vessel's next port of visit by considering its previous ports visited. We separate the data into a training set and a test set. However, to ensure that all 23 next ports of visit are included in both sets, we group the next ports into four categories. Those categories are: "SW ATLANTIC," "KOTKA FI AREA," "OTHER FI SE," and "ALL OTHERS." We use 20 percent of the dataset (160 observations), randomly selected without replacement, as our test set and the remaining 80 percent (640 observations) as our training set. We fit three random forest models, set up similarly as the models used in the previous objective, to predict the transitional probabilities of a vessel's next port of visit. The overall misclassification rates for each of the models decreases, similarly to the second objective. Model 0 has a misclassification rate of 38.5 percent using the training data and about 44 percent with the test set. Model 1 has a lower overall misclassification error with 36.3 percent using the training data and

about 37 percent with the test set. Model 2 has the lowest overall misclassification error in comparison to the first two models. With the training data, the overall misclassification error is 31.7 percent and with the test set it is about 30 percent. This analysis emphasizes the point that past information matters and is needed to more accurately predict a vessel's next port of visit. Although all possible next ports of visit were not included in this analysis, we consider them all in the next objective by estimating the transitional probabilities for all 23 ports.

The fourth objective is to estimate the transitional probabilities for a vessel's next port of visit to examine whether a vessel's next port of call is a Markov model or a higher-order network, where past states matter for accurate prediction of the future state. We construct three random forest models, setup similarly to the models used in the previous two objectives, to estimate the next port of visit. Even though MID is the best explanatory variable in the random forest analysis, the previous port variable increases with importance the further into the past we explore. The transitional probabilities change when explanatory variables are added that go deeper into the past, indicating that prediction of a vessel's next port of call is not a Markov model but a higher-order network where past information is needed to more accurately predict the future state.

The methodology used in this thesis is useful because it allows the user to more accurately predict a vessel of interest's next port of visit by considering its previous ports visited. Misclassification rates decreased with the addition of past information in the prediction models. By estimating the probabilities for all 23 possible next ports, there is a difference in the models and the transitional probabilities. This indicates that past information matters and is useful to consider in the allocation of surveillance assets. Incorporating past port visits into the model increases the probability of being in the right place to fully take advantage of the vessel of interest's full port stay.

ACKNOWLEDGMENTS

I have learned so much about data analysis from Professor Robert Koyak during the short time that we worked together. His mentorship and guidance enabled me to complete this thesis, and I am appreciative of his help! I would also like to thank LTC Sam Huddleston for his inputs that made my thesis a better story. Finally, thank you to my family and friends for all of your support throughout my time at the Naval Postgraduate School.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Maritime Domain Awareness (MDA) is important for countries that have coastal borders in areas with high levels of maritime traffic. This concept involves the safety and security of a countries mainland, economic prosperity, and the environment of a nation (Tetreault 2005). It deals specifically with the detection, identification, and observation of all vessels in an area of interest (AOI). Although tracking a vessel's movement is not difficult, it is challenging to determine intent. Without knowing a vessel's maneuvering intentions, it is hard to know if the vessel of interest (VOI) is behaving in a questionable way or if it is following a predictable pattern. Therefore, it is important to be able to characterize patterns of navigation, so that anomalous behavior being exhibited by a vessel can be examined more closely. This is especially important in waters near populous coastal regions, which are exposed to possible maritime threats. Given the current port of a vessel, knowing where it came from and where it is going to next is not enough to determine if its behavior is common for a vessel of its type. A vessel's route, consisting of sequential ports visited, can provide more insight into the vessel's behavior and choice of ports.

A key tool that has been used to enhance MDA globally is the Automated Identification System (AIS), which provides information on the movement of vessels at sea. The primary purpose of AIS is collision avoidance through its autonomous ability to identify other vessels that are fitted with an AIS transponder. A vessel that is AIS capable can receive information from other vessels in the area, such as the vessel's name, location and current status. That information could be used in the event of a possible collision, to take early action in accordance with International Regulations for Preventing Collisions at Sea (COLREGS). AIS provided information on the speed, course, destination, and position of a vessel. Other uses of AIS data include tracking ships in a certain area, or worldwide, and monitoring traffic through straits or channels. These instances require that the current position of a vessel be known. International Maritime Organization (IMO) requires that an AIS transponder be placed on all ships with gross tonnage (GT) greater than 300 that travel internationally, and all ships greater than 500 GT that travel domestically near their country of origin (International Maritime Organization 2002). It is also required for all passenger

ships, regardless of size. AIS transmissions can be useful in real time to track and monitor vessels or historically to determine various behaviors of vessels.

The data received from AIS transmissions is not perfect: it is not always complete and it is subject to error. Under the Safety of Life at Sea (SOLAS) Convention, there is a regulation that allows the master of a ship to make decisions that are necessary for safe navigation, such as turning off AIS equipment when it is working improperly or if the master believes the safety of his vessel is being compromised by providing the vessel's location to potential pirates (Tetreault 2005). These are the only times that the AIS transceiver can be turned off for vessels to be in accordance with IMO regulations. However, there are gaps in AIS data that cannot be accounted for by those reasons alone. Spoofing is a concern because ships can intentionally mask their intentions by providing information that tells a different story from what they are actually doing. This is why AIS data should not be the only means for vessel tracking. However, AIS data can provide valuable insights given that the pre-processing of the data eliminates potential outliers, which are observations that are not physically sensible. These outliers include observations that jump in location and would require speeds that are not possible for certain vessels. McAbee (2013) explains in detail how the receiving stations complete the timestamp that is transmitted from a vessel in the area. The timestamp in the transmission is in seconds and the receiver must use a local time reference to complete the timestamp with the appropriate hours and minutes. Errors often arise in the timestamps, but AIS data would not be particularly useful without it as a reference. Researchers are attempting to fill those gaps to gain useful insights in a specific area, or worldwide, that would increase the awareness of vessels operating near the shore of coastal regions.

For this thesis, the AIS data are narrowed down to a specific AOI. We focus on the Baltic Sea due to its containment of vessels, with the Danish Straits and the Kiel Canal being the only points of entry or exit. These passageways into the Atlantic Ocean aided in the construction of routes because once in the Baltic Sea, vessels are either transiting to a port or on their way out toward the Atlantic after departing from a port of call. The Baltic Sea is considered a difficult area for shipping because of the narrow Danish straits, shallow waters throughout, and seventeen major islands dispersed within the sea. These hazards in

such a small body of water do not leave much room for safe navigation. Thus, it would seem that there would be a need for the vessels to follow a predictable path in order to ensure the safety of maritime traffic. The AIS dataset is limited to an area bounded by latitude between 53 and 60.5 degrees north and longitude between 13.4 and 29.4 degrees east. The coordinates reference a box that covers the entire Baltic Sea with cut off points to the Southwest going out of the Baltic Sea through the Danish Straits or the Kiel Canal, to the north going into the Gulf of Bothnia, and Northeast in the Gulf of Finland before Saint Petersburg. A visual representation of this box can be seen in Figure 1. We account for the ports cutoff by the bounding box by labeling the vessels that go outside the bounds as visiting "SW ATLANTIC," "GULF OF BOTHINA," or "SAINT PETERSBURG RU AREA."

Figure 1. Bounding Box for the Area of Interest in the Baltic Sea

It has been estimated that up to 15 percent of the world's cargo traffic passes through the Baltic Sea, making it one of the busiest maritime areas in the world (Baltic LINes 2016). We determine that 40 percent of the ships in the 2014 AIS data used in this thesis were self-labeled as general cargo ships. The volume of maritime traffic in the Baltic Sea is likely to increase due to anticipated climate warming and the potential decrease in ice conditions that are expected during the winter seasons when the Baltic is usually

restrictive for large vessels. With the expected increase in shipping in the Baltic Sea, safety of navigation becomes more of a concern. The dataset that we examine in this thesis contains more than 25 million observations from a four-month period, beginning in January 2014. In that year, the Baltic Sea was not heavily covered in ice due to a mild winter and the northern Bay of Bothnia received most of the ice coverage (Vainio and Eriksson 2018). This means that the winter season should not have had much of an effect on the observations.

A. RESEARCH OBJECTIVES

The research presented in this thesis is focused on four main objectives that attempt to characterize patterns of navigation within the Baltic Sea. Our first objective is to examine the length of stay at a particular port for cargo vessels. The port visit can be as important as the past, present, and future location of the vessel for predicting its movement. Using regression, we explore different factors to determine their effects on a vessel's length of stay at port. Some of the specific factors that we will be exploring are arrival and departure day of the week, arrival day in the year, what country the vessel is registered to, and the past and future port visited. The insights provided by the port stay analysis would be beneficial to the ISR community to better understand the factors that influence how long a vessel of interest remains in port.

The second objective is to predict with greater accuracy whether a vessel will depart the Baltic Sea or visit another port within the Baltic Sea. We use random forest models to predict whether a vessel departs or remains in the Baltic Sea after visiting a port in the Saint Petersburg, Russia area. These models derive probabilities based on the addition of information pertaining to vessels that stop in Saint Petersburg area and the previous ports that were visited by those vessels. The insights provided by this binomial prediction analysis would highlight the worth of including knowledge of past information in the prediction of a vessel's next state after transitioning from a port.

The third objective is to predict, with greater accuracy, a vessel's next port of visit by considering its past navigation pattern. This objective builds off the second objective by now looking at how accurately we can predict ports rather than simply in or out of the

Baltic Sea. We separate the data into a training and test set to determine how well our random forest models predict the next port of visit for vessels that stop in the Saint Petersburg area. The predictor variables prior ports visited and the next ports of visit are grouped into four levels: SW ATLANTIC, KOTKA FI AREA, OTHER FI SE, and ALL OTHERS. We group these categorical variables to ensure that all ports are included in the training set and the test set. The three random forest models differ in the number of previous ports included in the models, ranging from no prior port to two prior ports. The insights provided by this multinomial prediction analysis would enhance MDA for countries near coastal waters and give maritime law enforcement additional statistical intelligence in regards to suspicious vessel's behavior.

The fourth objective is to estimate the probabilities for a vessel's next port of visit to determine if the probabilities follow a Markov model or a higher-order network. We estimate a probability distribution for all 23 possible next ports by using random forest models on four chosen scenarios. We use the random forest models to determine how knowledge of previous port visits influences the model predictions for the next port of visit for a vessel. There has been research conducted on this topic using AIS data, but not in a confined area such as the Baltic Sea. The Baltic is unique because there is a single point of entry and departure from the sea out toward the Atlantic Ocean. The insights provided by this analysis would be beneficial to the decisions that are being made in regards to the allocation of surveillance assets for specific ships.

B. THESIS STRUCTURE

This thesis is organized as follows. In Chapter II, we review the literature that is relevant to analyzing AIS data for maritime traffic routes to gain maritime domain awareness. The approaches vary in literature as analysts try to find the methodology that is the quickest and most accurate so that it may be applied to real-time AIS data. In Chapter III, we discuss the AIS data collection and preparation process that occurred in order to ensure that the outliers in the data were accounted for and that the data made sense when pictured visually. In Chapter IV, we present the methodology for the analysis of the

research objectives and discuss why the results are rational. Finally, in Chapter V, we summarize and discuss the conclusions and topics for future work.

II. LITERATURE REVIEW

The following literature review shows the growing importance of historical AIS data and how it can be used to enhance MDA.

A. VESSEL BEHAVIOR

Vessel behavior is used to recognize vessel activity that is considered suspicious, such as paralleling or following other vessels in the area. Research conducted in this area can be used as a basis for automated threat detection to improve MDA.

In an NPS thesis, Tester (2013) focuses on vessel behavior in the maritime domain in order to enhance MDA. His objective is to develop a method that would autonomously determine whether a VOI had any co-occurances with other vessels in the area during a period of time. To achieve this, spatiotemporal clustering is used to cluster groups of vessels based on proximity, course, and speed at each unique time-step. The behaviors that are focused on were paralleling or following the VOIs course and speed. As a result, the algorithm is able to track the clusters of vessels until continuity fell below the threshold.

Gutierrez Torre (2017) treats AIS data as a time series when using Conditional Restricted Boltzmann Machines (CRBMs) and k-means clustering to extract patterns of navigation in maritime traffic. The dataset from Spanish Port Authority records is processed in CRBM while maintaining the time factor and reducing dimensionality. The reduced output is then clustered to identify patterns which could be used to correct missing or erroneous data, trace ship behaviors, and recognize their activity. The patterns that were discovered were used to model air quality in coastal urban zones.

In an NPS thesis, Hintze (2017) demonstrates an approach to predict the membership of clusters for AIS observations, as a basis for an automated threat detection system in the Gulf of Mexico. The AOI is separated into zones of equal size for better results. He uses a technique called Ordering Points to Identify the Cluster Structures (OPTICS) to cluster by zone by combining geospatial coordinates with vessel attributes. Classification trees are used to predict cluster membership for new observations.

B. ANOMALY DETECTION AND MOTION PREDICTION

Vessel behavior is also used to identify anomalous behavior by defining motion patterns to compare to a live-feed of AIS data. These motion patterns are also used to predict the motion of vessels or probable destinations.

In an NPS thesis, McAbee (2013) uses the Hough transformation to extract linear patterns of density from the high density traffic areas that were identified. Once the routes or "highways" are established, they can be used to collectively create an atlas from AIS data to define what normal maritime behavior looks like for both open-ocean and coastal areas. With normality for vessels operating in a particular area defined, McAbee (2013) states that "anomaly detection is performed by comparing incoming ship position reports to the generated atlas." Her use of an image processing technique, the Hough transformation, to extract patterns of navigation is the first of its kind to provide insight into abnormal behavior. Those insights can then be analyzed to determine intentions of those abnormal behaviors from specific vessels.

Ristic, La Scala, Morelande, and Gordon (2008) present a basic solution to motion prediction using historical AIS data. They extract motion patterns from the data and use adaptive kernel density estimation to construct anomaly detectors, which are then applied on a live-feed of AIS data. Motion patterns that are extracted are assumed to operate in a normal manner. The authors use the extracted motion patterns and a Gaussian sum tracking filter to predict the motion of vessel.

Pallotta, Vespe, and Bryan (2013) propose an unsupervised methodology called Traffic Route Extraction and Anomaly Detection (TREAD) to gain an understanding of maritime traffic. They eliminate the problem of using turn points for a "vectorial" model in unregulated traffic areas by utilizing a Density Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to distinguish waypoints and the routes between them when a pattern is consistently observed. The main routes that show a minimum number of transits is then decomposed and organized into an atlas for future reference. Associated with these routes are spatial, temporal, and attribute information that allows for the

detection of anomolous behavior and the prediction of future positions or probable destinations for a vessel given real-time positional information.

In an NPS thesis, Bay (2017) examines the effect of clustering vessels in the Port Fourchon, LA area to identify patterns of movement in the Northern Gulf of Mexico maritime area. Due to the oil and natural gas platforms near Port Fourchon, clustering is not an effective technique to classify vessel movements in and out of the port. Using the Bureau of Ocean Energy Management dataset and buoy data from the National Data Buoy Center of U.S. Department of Commerce in conjunction with the historical AIS data, regression analysis is conducted to determine how closely vessels adhere to a great-circle route when taking weather and sea-state into consideration. Bay concludes that weather and sea-state should be included in prediction models of vessels at sea.

In an NPS thesis, Young (2017) presents two models that are used to predict the future location of a vessel along a clustered route using historical AIS data. Young finds that random forest produced prediction intervals that are close to the coverage to the nominal target probability for the true future position. Young finds that a neural networks approach produces less accurate prediction intervals than random forest. Additionally, random forests are easier to implement and computationally less demanding than neural networks.

C. THESIS FOCUS

The literature reviewed here shows the importance of extracting patterns of motion to gain an understanding of ship movement. With that knowledge, the patterns are compared to real-time AIS data to further the research being conducted on improving MDA. This thesis uses the patterns that were extracted to determine trends found in the Baltic Sea among cargo ships. We are particularly interested in destination prediction using past knowledge of where a vessel has been to more accurately predict where it will go next.

Fernandez Arguedas, Pallotta, and Vespe (2017) develop a geographical network based on vessel behavioral changes such as a change in speed or direction. The purpose of the network is to enhance maritime situational awareness applications such as track reconstruction, destination prediction, and anomaly detection. Their approach is based on

the underlying idea that vessels follow consistent routes due to regulated traffic or fuel efficiency considerations. To automatically represent maritime traffic in a graph-based topology, they propose the Maritime Traffic Representation System which is an unsupervised maritime network generator that uses historical AIS data to detect changes in a vessel's behavior. The detected changes, such as a change in course, are used to create segments in the routes and associated to create maritime lanes. The proposed method is used on the Baltic Sea's high traffic density area, which resulted in reducing over a million observations to about two thousand arcs.

Gambs, Killijian, and Del Prado (2012) develop a Mobility Markov Chain (MMC) model that they extend to predict the next location of an individual based on his movement over a period of time and the locations that were visited. Their data consists of mobility traces from GPS-enabled devices in the Shanghai area. K-means clustering is used to identify points of interest within the dataset. The memoryless property of the Markov models is proven to negatively impact the accuracy of predicting the next place that a vessel visits. The data is split into a training and a test set to evaluate the accuracy of the predictor. The authors show that accuracy improves as the past is included but it stabilizes as soon as the number of places remembered equals two. Accuracy and predictability of the model are improved with inclusion of the past but the accuracy ranged from 70 percent to 95 percent.

Xu, Wickramarathne, and Chawla (2016) focus on developing a network based on a set of interactions that are produced from the movement of the components or vessels. Historical AIS data is used to demonstrate that higher order networks (HON) are a more accurate representation of vessel movement from port to port, unlike simple Markov models. HON is also shown to be more scalable because first-order networks are maintained unless higher-order nodes are needed to increase accuracy. A higher-order node can represent multiple ports in arbitrary order rather than a single port. The authors claim that "HON can help random walkers simulate movements more accurately," leading to more accurate predictions of where the vessel will transit to next (Xu et al. 2016). The research revealed that a ships movement had up to a fifth order dependency, which means that the next port a ship visits can depend upon its five previous ports that it visited.

The literature mentioned in this section propose methods to use in order to determine a vessel's future location. We create segments in the routes to understand all of the ports visited previously to a vessel stopping in the Saint Petersburg area. Gambs et al. (2012) discuss how a Markov model negatively impacts the accuracy of predicting the next place of visit for a person. This statement holds true for vessels because there are maritime patterns of navigation that are followed, especially in an enclosed such as the Baltic Sea. Xu et al. introduce a HON and discuss how a single node can represent multiple ports in arbitrary order to better predict a vessel's next port of visit. We use the idea behind a HON to emphasize that past knowledge of several of a vessel's port of calls increases the accuracy of predicting a cargo ship's next port of visit.

THIS PAGE INTENTIONALLY LEFT BLANK

III. METHODOLOGY

The AIS data used in this thesis is a subset that contains only transmissions received in the Baltic Sea during the time period of January through April 2014. This chapter describes the process we use to set up the data for analysis and the methodology we use to gain insights for the research objectives mentioned in chapter I.

A. DATA DESCRIPTION

The AIS observations in the dataset have both dynamic and static information. Dynamic data includes information such as Maritime Mobile Service Identity (MMSI) number, speed, position, course over ground, heading, and timestamp. The updates to this information are automatic since the transponder is connected to ship equipment such as GPS. It is transmitted in near real time, according to Raymond (2016), "every two to ten seconds depending on the vessel's speed while it is underway, and every three minutes while the vessel is at anchor and stationary." Our data is subsampled to give one dynamic AIS measurement approximately every ten minutes. On average, there are roughly three million transmissions, worldwide, recorded for dynamic data in one day. Static data includes information such as MMSI number, IMO number, call sign, ship name, ship type, and destination. It is manually entered into the AIS transponder upon installation and only needs to be changed if there is a major conversion of the vessel, or if the name or call sign changes, or when entering the updated next port of destination (Harati-Mokhtari et al. 2007). Due to the fact that the static information does not change as often as the dynamic information, it is transmitted every six minutes (Raymond 2016).

The MMSI is a unique nine-digit number that is used to identify an AIS transponder onboard a ship. It can be compared to a cell phone number in that other vessels that receive a transmission are able to use that number to identify and contact a vessel in emergency situations. The position entry consists of the latitude and longitude of a particular vessel at the time of the transmission. The timestamp contains the date, hours, minutes, and seconds in Coordinated Universal Time (UTC) and is dependent on the station that is collecting and storing the transmissions. The IMO number is a unique seven-digit number that is specific

to a ship and assigned to merchant ships under SOLAS. The call sign and ship name are used in communications to contact other vessels in the area. The next port of destination for a vessel is included to give more situational awareness in regard to the direction in which it is heading.

Harati-Mokhtari, Wall, Brooks, and Wang (2007) estimate that one in every fourteen AIS transmissions has at least one erroneous data field. For the static data in particular, the trustworthiness of the static data is heavily dependent on human interaction with the system. For example, the MMSI number that identifies a transponder requires that the master of a vessel ensures that the number is accurately entered at the time of installation of the AIS unit on the bridge (Harati-Mokhtari et al. 2007). Otherwise, the transponder retains the default MMSI number that the AIS unit had from the manufacturer and causes confusion when looking at historical AIS data because a MMSI can then be in multiple locations at the same time since there are multiple ships using the MMSI number. The fields that we use include MMSI, ship type, position, and timestamp. We assume that the data accurately describes vessels present in the Baltic Sea.

B. DATA PROCESSING

Bay (2017) describes a method used to begin processing the data from its original AIVDM/AIVDO format to a file that is then decoded and converted into a spatial-points data frame using the statistical programming language R (R Core Team 2017). Daily dynamic and static reports, from January to April 2014, are sorted in chronological order and duplicate observations are removed. The dynamic data is combined with the static data by using the MMSI and timestamp entry as a reference to match the information correctly. Using the ship type entries to narrow down the data, we explore the vessels that are self-labeled as cargo ships. This group of vessels contains 4,170 unique MMSIs and is the largest vessel group in the Baltic Sea. According to Harati-Mokhtari et al. (2007), approximately 74 percent of the entries in the ship type field are vague or misleading. We assume that the self-identification of a ship as a cargo vessel is accurate for the purpose of this analysis.

The next step is to clean the data by scanning for outliers that may affect our analysis. We define a maximum velocity threshold of 1200 meters per minute, which is equivalent to about 45 miles per hour (MPH). This threshold is reasonable because most cargo ships operate normally at speeds ranging from 23 to 29 MPH (Rodrigue 2017). The data is then organized by MMSI in sequential order, according to its timestamp, to then calculate the incremental distances between observations using the latitude, longitude, and time entries. Velocities are then estimated by dividing distance by time. If there are only a few incremental velocities that are found to be at or above the threshold, they are removed from the data for that vessel. However, if over 10 percent of the data is found to exceed the threshold, then that MMSI is not processed. Koyak (2017) describes the algorithm used to detect infeasible speeds and to remove outliers in the data.

C. CREATING ROUTE SEGMENTS

During the period of time that a vessel is in the Baltic Sea, it makes multiple stops at various locations that we define as stop points. To decide if a vessel is stopped, we define a time and distance threshold. A vessel is designated as stopped if it has moved less than 5,000 meters over a time period of at least 120 minutes. This means that if the vessel has not traveled a little over three miles in two hours, then it is considered to be stopped at its current location. These time and distance thresholds are realistic because in order for a vessel to be at or below the threshold it would have to moving at a speed of about one knot. For a large ship to maintain bare steerageway, it would have to sustain a minimum speed of at least three knots. Once we have calculated the stop points using our defined thresholds, we can process the information to describe segments for each of the unique MMSIs. Obtaining stop points is important to understanding the movement patterns of a vessel.

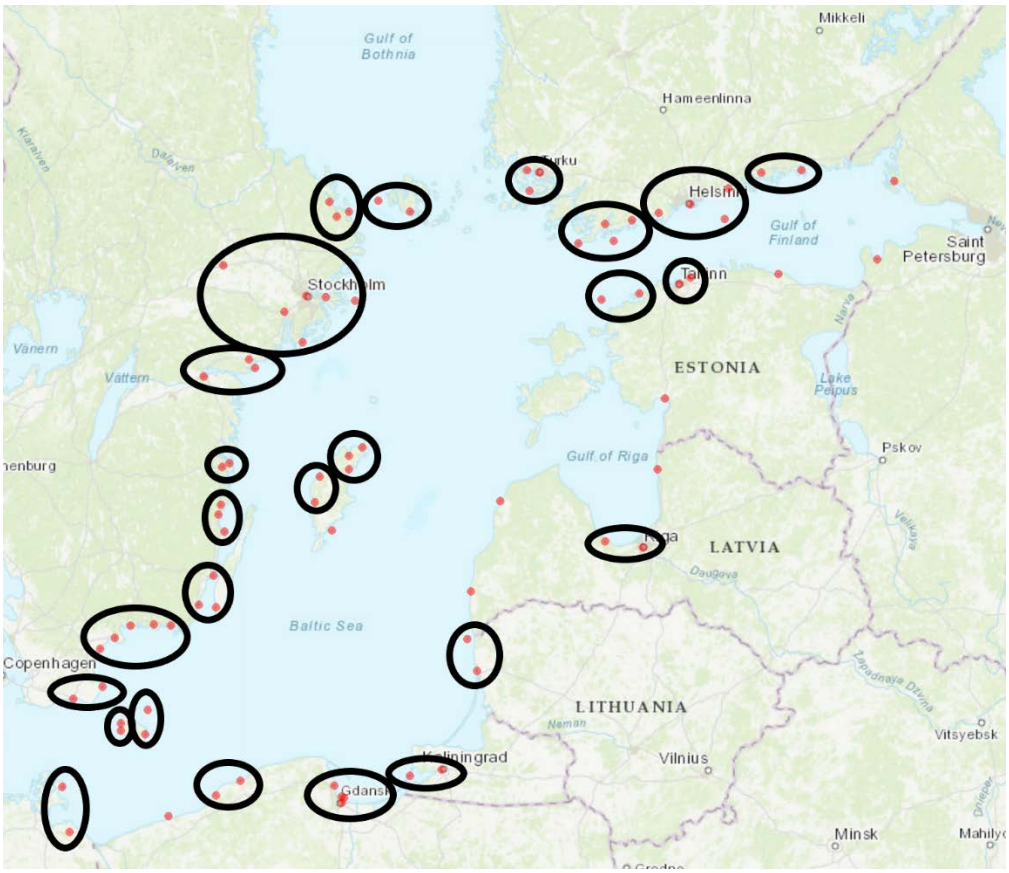
The dataset of segments contains a vessel's MMSI, the start type, start latitude and longitude, start time, end type, end latitude and longitude, end time, and total distance traveled in the segment. The start type is constructed to determine where the segment is starting from. If a vessel is starting from a defined stop point, then it is assigned a value of minus one; a non-stop point within the Baltic Sea is assigned a value of zero; entry from

the west boundary (Denmark-Sweden Strait) is assigned a value of one; entry from the east boundary is assigned a value of two; entry from the south boundary is assigned a value of three; and entry from the north boundary (Gulf of Bothnia) is assigned a value of four. The end type is defined in a similar manner. A defined stop point is a recognized port in the Baltic Sea region. A vessel is considered to be stopped and in port if it is inside a 5000-meter radius of the port, using coordinates for the port listed in the World Port Index (National Geo-Spatial Intelligence Agency 2017). The total distance traveled is the sum of incremental distances of AIS-reported positions throughout the segment. One MMSI can have multiple segments that should be sequential and when put together constitute its route throughout the Baltic Sea.

D. ASSIGNING ROUTE SEGMENT LABELS

We label the segments according to where the vessel is transiting to and from. Harati-Mokhtari et al. (2007) approximates that "49 percent of the destination entries are erroneous or misleading" in referring to the static AIS reports. Therefore, we use the start and end points of the route segments that we determine from the AIS positional messages to identify which ports, if any, the vessel visits. We convert the World Port Index into an R dataset that contains the ports name, country, latitude, longitude, harbor size, and harbor type. We then merge the segment dataset with the port locations. The port with the minimum distance is then assigned to the route label for the segment.

There are 78 identified ports within our bounding box in the Baltic Sea and many occur in close proximity to other ports, which makes it difficult to distinguish the exact port that the vessel is visiting. To minimize error, we create port clusters by combining ports that are within fifty miles of each other and belong to the same country into one entity to reduce the number of potential port stops. By doing this, we reduce 78 unique port names to 34 port clusters. In Figure 2, there are 25 port clusters that have more than one port, leaving nine ports with their original port name. Port clusters are used to simplify the analysis. We add three other ports to account for the ports cutoff by the bounding box. Those ports are named: "SW ATLANTIC," "GULF OF BOTHINA," and "SAINT PETERSBURG RU AREA."



Circled ports are combined into a port cluster.

Figure 2. Baltic Ports within Bounding Box

For the purpose of analysis, we focus on a cluster of ports in the Saint Petersburg, Russia area. This cluster consists of three different ports, as listed in Table 1, and has the highest inflow and outflow of cargo vessel traffic compared to the other Baltic ports. The port stays that were less than one hour were removed from the data to exclude potential erroneous data. Cargo ships would need more than an hour to moor and offload its cargo.

Table 1. Ports by Name within Port Clusters

CLUSTER NAMES	PORT NAME	CLUSTER NAME	PORT NAME
WOLGAST DE AREA	WOLGAST	KALINGRAD RU AREA	KALINGRAD
	SASSNITZ		BALTIYSK
E. BORNHOLM DK AREA	CHRISTIANSO HARBOR	ST. PETERSBURG RU AREA	ST. PETERSBURG
	NEKSO		LOMONOSOV
W. BORNHOLM DK AREA	RONNE		KRONSHTADT
	HASLE	HALLSTAVIK SE AREA	HALLSTAVIK
TALLINN EE AREA	TALLINN		GRISSLEHAMN
	MUUGA-PORT OF TALLINN		HARGSHAMN
PALDISKI EE AREA	PALDISKI	STOCKHOLM SE AREA	STOCKHOLM
	OSMUSSAAR		SODERTALJE
TURKU FI AREA	TURKU		NYNASHAMN
	PARGAS		GUSTAVSBERG
	NAANTALI		SANDHAMN
MARIEHAMN FI AREA	MARIEHAMN		VASTERAS
	SIGNILSKAR	NYKOPING SE AREA	NYKOPING
HANKO FI AREA	HANKO		NORRKOPING
	EKENAS		OXELOSUND
	JUSSARO	VASTERVIK SE AREA	VASTERVIK
	INKOO		VERKEBACK
HELSINKI FI AREA	HELSINKI	OSKARSHAMN SE AREA	OSKARSHAMN
	PORKKALA		STORA JATTERSON
	TOLKKINEN		FIGHOLM
	PORVOO	KALMAR SE AREA	KALMAR
KOTKA FI AREA	KOTKA		BERGKVARA
	LOVIISA		DEGERHAMN
KLAIPEDA LT AREA	KLAIPEDA	KARLSKRONA SE AREA	KARLSKRONA
	BUTINGE OIL TERMINAL		RONNEBY
RIGA LV AREA	RIGA		KARLSHAMN
	LIELUPE		SOLVESBORG
GDANSK PL AREA	GDANSK		AHUS
	PORT POLNOCHNY	YSTAD SE AREA	YSTAD
	NOWY PORT		SIMRISHAMN
	GDYNIA	E. GOTLAND SE AREA	SLITE
USTKA PL AREA	USTKA		FAROSUND
	DARLOWO		STORUGNS
		W. GOTLAND SE AREA	VISBY
			KLINTEHAMN

E. REGRESSION ANALYSIS

Our first research objective is to analyze the effects of different factors on the duration of port stays in the Baltic region, focusing on the Saint Petersburg, Russia area. This port cluster is of strategic interest and has the most data with 997 observations. With the new subset of data, we use regression to explain, per Faraway (2015), how a response variable Y (Time in Port) can be analyzed by the explanatory variables $X_1, \ldots, X_{p-1}$, where $p-1$ is the number of factors used in the regression. This analysis is used to examine the relationship between the response variable and the predictor variables to determine which factors are the best to explain the variation in the length of port stays.

The explanatory variables that we consider are the country that the vessel is registered to, the day of the week that the vessel arrives and departs from port, the day in the year that the vessel arrives in port (1 = January 1, 2 = January 2, …, 130 = April 30), and the previous and next port of destination. We calculate length of stay in port by taking the start time of the next route segment and subtracting the time between then and when the vessel arrived in port. The country that the vessel is registered to can be obtained from the first three digits in the MMSI number, which are also known as the maritime identification digits (MID). For example, the MIDs for the United States are 338, 366, 367, 368, and 369. This means that vessels registered in the United States will have MMSIs that begin with one of those five MIDs. The days of the week that a vessel arrives and departs from port are obtained from the start and end time in the route segments. The "weekdays" function in R takes the start or end times as input and outputs the day of the week that it occurred. The previous and next port of destination are based on the route segments for each particular MMSI and the port that it came from and where it went to next after its current port. If there is no data concerning where the vessel came from or went to next, it is left blank.

In Chapter IV, we analyze and evaluate the explanatory variables from the regression model to determine if they are useful in explaining the response variable, which is the number of hours that a vessel stays in port in the Saint Petersburg area.

F. INTRODUCTION TO HIGHER-ORDER NETWORKS

A higher-order network (HON) discovers variable orders of dependencies that can be missed in conventional network representations due to the limiting Markov model properties (Xu et al. 2016). According to Lin (2017, p. 34), "if someone wants to find the probability of some future event related to a Markov chain, then he only needs to know the present state of the Markov chain." This is equivalent to stating that Markov models are first-order dependent, which means that they depend only on the current state when determining a future state. This assumption may lead to inaccuracies if the past does matter and important information is lost (Xu et al. 2016). Therefore, for shipping data, a HON may provide better estimation for movement patterns with the flow of maritime traffic being an aggregation of port stops.

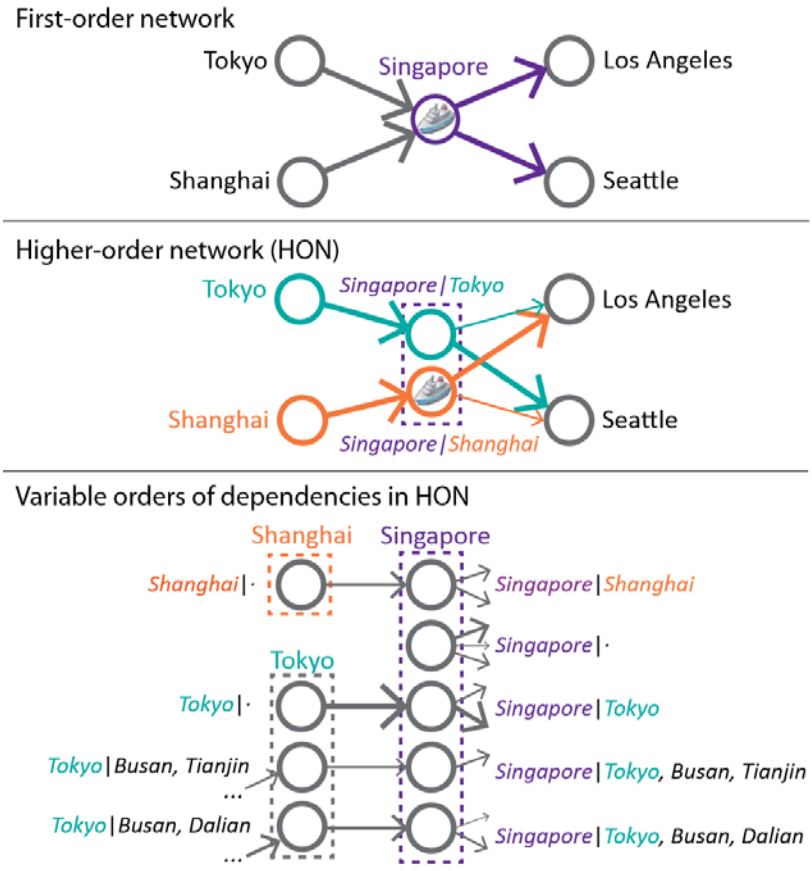


Figure 3. Example of a HON for a vessel departing Singapore and transiting to either Los Angeles or Seattle. Source: Xu et al. (2016).

Citing the example of vessels currently in port in Singapore, Xu et al. (2016) examine the dependency of a vessel's past port visited prior to Singapore on whether the next port of visit will be Los Angeles or Seattle. Under a first-order model the probabilities that a vessel in Singapore next visits Los Angeles or Seattle is proportional to the number of trips between Singapore and each of these ports. Xu et al. (2016) demonstrate, however, that if a vessel had previously visited Tokyo, then its next port of visit was more likely to be Seattle than Los Angeles than what a first-order model would predict. Similarly, if the vessel's previous port of visit was Shanghai, it had a higher probability of its next port of visit being Los Angeles. Variable order of dependencies in HON illustrates that knowledge of a vessels past, regardless of sequential order, can influence its future location.

To explore the interconnectedness of the cargo ships on their routes within the Baltic Sea, we use port clusters as our set of states that the vessels transition to. We estimate transitional probabilities for vessels moving from state to state by determining the number of transitions from one state to all of the other states and dividing each frequency by the total number of transitions from that particular state. We use random forests to estimate the probability of a vessel's next port of visit using factors such as the arrival day of the week in the Saint Petersburg area, the day in year, length of port stay in hours, MID, and previous ports visited.

G. RANDOM FOREST ANALYSIS

Our second research objective is to predict with greater accuracy whether a vessel would depart or remain in the Baltic Sea to visit another port after stopping in the Saint Petersburg area. Young (2017) describes the concept of partition trees and how they can be used in regression or classification. Random forests partition the feature space by creating uncorrelated trees that bootstrap different versions of the data (Young 2017). It draws a sample with replacement and fits a regression tree to the bootstrapped data. The set of cases that are not selected in a bootstrap sample is used to compute the classification errors of prediction by comparing the predicted values to the observed values. This comparison provides a measure of prediction performance that avoids overfitting by not using data that is used in the construction of the tree (Faraway 2016).

We construct three random forest models that predict the next state for a vessel, differing in their use of past information. The first model uses the predictor variables **ArrivalDayOfWeek** to represent the weekday of arrival for a vessel in Saint Petersburg area, **ArrivalDayInYear** (1 = January 1, 2 = January 2, …, 130 = April 30), **PortStay** is the number of hours that a vessel is in port in the Saint Petersburg area, and **MID** for the maritime identification digits representing the country of registration. The second model adds **PrevPort1**, the most recent port visited by a vessel prior to Saint Petersburg. The third model adds **PrevPort2**, a port previous to the prior port visited, on to the second model.

Table 2. Predictor Variables in Chosen Scenarios for Classification of the Next Port Visited Using Random Forests

	Arrival DayOfWeek	**Arrival DayInYear**	**PortStay**	**MID**	**PrevPort1**	**PrevPort2**
Scenario 1	Tuesday	84	21	209	SW Atlantic	St. Petersburg RU
Scenario 2	Sunday	12	19	209	SW Atlantic	Reka Luga RU
Scenario 3	Thursday	100	29	636	Reka Luga RU	SW Atlantic
Scenario 4	Sunday	103	109	377	SW Atlantic	St. Petersburg RU

Our third objective is to predict with greater accuracy a vessel's next port of visit. This objective is similar to the second except that we take a closer look at specific ports, rather than a binomial response variable that indicates whether a vessel stays or leaves the Baltic Sea. We examine random forest models that are set up similarly to the three models mentioned, but which predict transitional probabilities for which port will be visited next. For this analysis, the next ports of visit are defined by four categories: "SW ATLANTIC," "KOTKA FI AREA," "OTHER FI SE," and "ALL OTHERS."

Our fourth objective is to estimate the transitional probabilities for all 23 possible next ports of visit to examine how past information affects the probabilities. We expand on the third objective and consider all port clusters instead of grouping the port clusters into four categories. We utilize three random forest models set up similarly to the previous objectives and examine the different models to highlight the effects of including past information about a vessel when predicting its next port of visit.

In Chapter IV, we present the analysis for the HON using cargo ship data in the Baltic Sea to determine where a given vessel may be transiting to next, given prior knowledge of the vessel's past ports.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. ANALYSIS AND EVALUATION

Our analysis consists of four parts in order to answer our four research objectives. First, we present the findings from the regression analysis to examine the effects of certain factors on the response variable, which is time in port. The explanatory variables we use in the regression model include the MID for the country that the vessel is registered to, a vessel's prior and next port of visit, the day in the year (1 = January 1, 2 = January 2, …, 130 = April 30), and the arrival and departure day of the week. As explained in Chapter III we consider only the cargo vessels that visited the Saint Petersburg, Russia area between January and April of 2014. Second, we present the results from the predictive model that we obtain using random forests. The response variable for this regression is a binomial variable that indicates if a vessel is departing or remaining in the Baltic Sea. This prediction is based on different factors and past knowledge of what ports the vessel has been to prior to stopping in the Saint Petersburg area. Third, we expand our analysis and predict a vessel's next port of visit by using groupings of the possible next ports. This analysis also uses random forest with the same factors considered. Fourth, we consider all 23 possible next ports of visit and estimate transitional probabilities. This analysis compares the different random forest models to examine the usefulness of including past information about a vessel.

A. REGRESSION RESULTS

For the analysis of port-stay durations, we identify the cargo ships that stopped at a port in the St. Petersburg area. As mentioned in Chapter III, when we refer to Saint Petersburg area, we are referring to the Saint Petersburg cluster which consists of three different ports. We obtain the following variables:

- Prior port that vessel visited before Saint Petersburg area
- Next port that the vessel transits to after Saint Petersburg area
- Departure day of the week
- Arrival day of the week

- Arrival day in the year
- MID that represents the code of the country that a vessel is registered to

Our objective is to examine the effects that the variables listed above have on the length of time that a vessel spends in port. Figure 4 shows the distribution for the hours a cargo ships spends in a port in the Saint Petersburg area. The length of time ranges from a little over an hour to about 413 hours. We remove one observation from the dataset with a time in port of over one thousand hours due to its potential as a possible outlier.

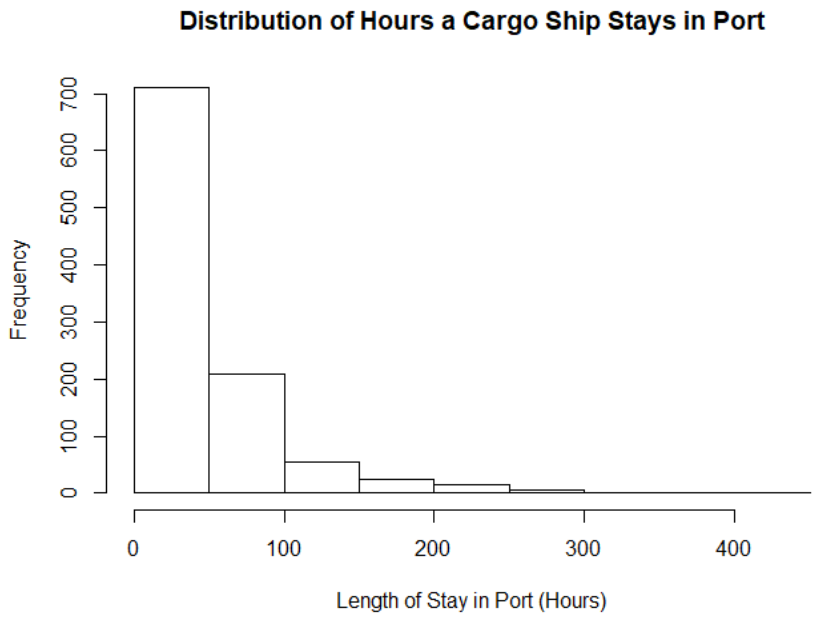


Figure 4. Distribution of Hours a Cargo Ships Stays in a Port in the Saint Petersburg, Russia Area

Length of port stay and the day in the year are both numeric, while all of the other explanatory variables are categorical. Because several of the explanatory variables are categorical with many levels, there are 121 parameters in the regression, including the constant term, with sample size of 1019 observations. We examine the use of transformations on the response variable, which is the amount of hours that a vessel spends in port, to determine how to best address the heteroscedasticity and non-normality shown in Figure 5. Heteroscedasticity is non-constant variance, which can be analyzed in a plot

of residuals versus fitted values (Faraway 2015). For there to be homoscedasticity, there should be equal variance in the residuals across the explanatory variables. According to Faraway (2015, p. 64), "residuals can be assessed for normality using a Q-Q plot." The residuals appear to have a long right-tailed error distribution.

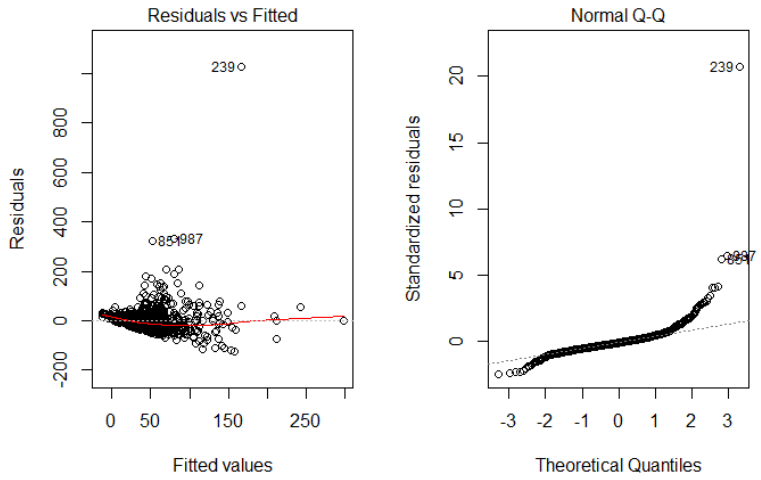


Figure 5. Plot of Regression Analysis Residuals

We explore the use of Box-Cox transformations on the time in port regression to improve the fit of the model. Figure 6 shows a log-likelihood plot, where the confidence interval for λ is about 0.13 to about 0.21. The Box-Cox analysis of the regression model recommends a λ of 0.18 for the value of the exponent that maximizes the likelihood. Although the confidence interval does not contain the value zero for λ, where zero implies a logarithmic transformation, it is close enough to be considered as the preferred choice. Per Faraway (2015, p. 118), "If explaining the model is important, you should round λ to the nearest interpretable value." Thus, for this thesis, we use a logarithmic transformation for the response variable.

Figure 7 shows plots of residuals for the regression analysis after the logarithmic transformation on the response variable. The plot of residuals versus fitted values shows more of an equal variance in the residuals across the explanatory variables, addressing the

problem of heteroscedasticity. The Q-Q plot appears to have a shorter right-tailed error distribution than in Figure 5.

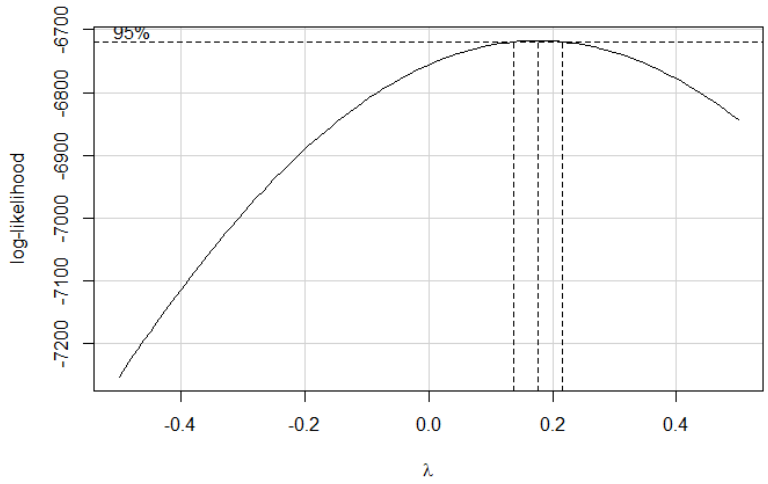


Figure 6. Log-likelihood Plot for Box-Cox Transformation of the Time in Port Regression

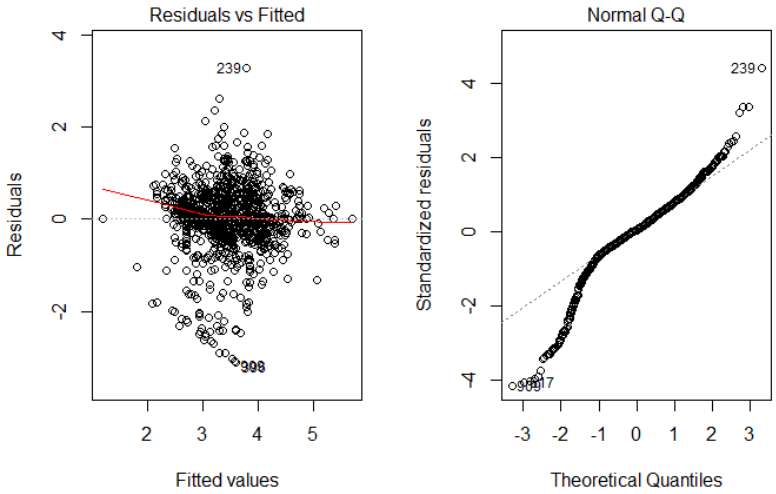


Figure 7. Plot of Regression Analysis Residuals with Log Transformation

Figure 8 is a plot of the MIDs with the highest frequency of stops at ports in the Saint Petersburg cluster. The horizontal axis is converted from the MIDs to the names of

the country that the vessels are registered to and the y-axis is the amount of time that the vessels spent in one of the ports in the Saint Petersburg area before departing to their next port of visit. There are some countries that have more than one box plot which is due to the fact that countries can have multiple MIDs assigned to them. This is usually for countries with widespread maritime activity. We note in this plot that Antigua and Antigua2 have nearly the same variance with a slightly different median. The same can be said for Cyprus and Cyprus2, which means that the MID provides useful information in regards to the amount of time that the vessels spend in port. However, Cyprus3 differs in its variance and median. This observation is interesting and can depend on whether these vessels are different from those in the other two groups of country codes. If the ships are larger, they carry more cargo and can take longer to unload in port.

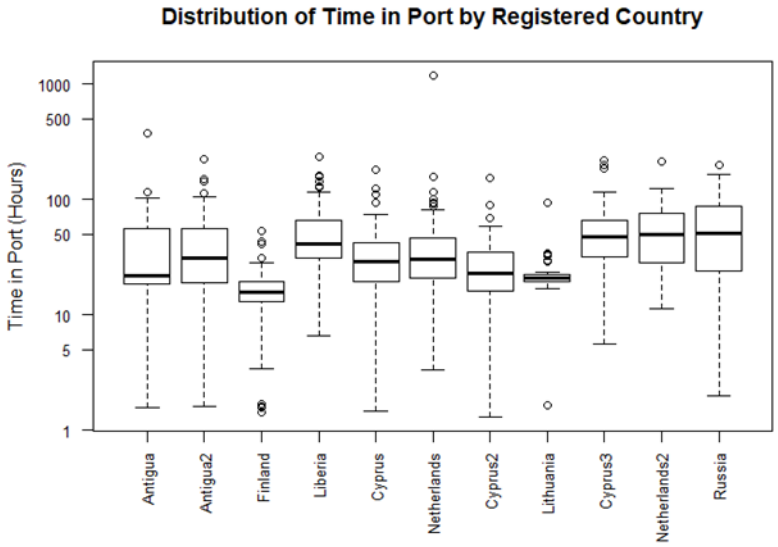


Figure 8. MIDs by Country with Highest Frequency of Port Stops in the Saint Petersburg Area

The R-squared value for our regression model after we apply a logarithmic transformation to the response variable, which is the number of hours that a cargo ship spends in a port in the Saint Petersburg area, is 0.3929 and the adjusted R-squared value is 0.3117. R-squared is a measure of how well the regression model fits the response variable using the same data that was used to estimate the regression parameters; as a

result, it is optimistically biased. R-squared never decreases when another explanatory variable is added to the regression model, regardless of how useful it is in predicting the response variable. Adjusted R-squared modifies R-squared so that the addition of explanatory variables to the regression model exacts a penalty. The addition of a weak explanatory variable that increases R-squared only marginally may result in adjusted R-squared being decreased. The large discrepancy between these two values in the present analysis is an indication of overfitting, and points to the need to simplify the regression model. Additionally, we are unable to compare goodness-of-fit measures such as R-squared to the original model because of the logarithmic scale that we apply to the response variable. The strongest explanatory variable is a vessel's MID, which has an R-squared of 0.2565 if used to predict logarithm of time in port without any other explanatory variables. Prior ports visited are also strong explanatory variable in this model. However, with 121 coefficients, this model is overfit and does not explain how well it could predict new instances.

We use the root mean squared error (RMSE) as a measure of performance to examine how well the explanatory variables in the model explain the variance in the response varaible. RMSE is the metric that we use and it is defined by the equation:

$$\sqrt{(\sum_{i=1}^{n}(\hat{y}-y_i)^2/n)}$$

where *n* is the number of observations in the dataset (1019) and $\hat{y}_i$ is the prediction for the logarithm of time in port for observation i that is exponentiated to reverse the logarithmic transformation. RMSE for our model is 38.1 hours.

B. GROUPING CATEGORICAL VARIABLES

We find the explanatory variables that were successful in creating a better model when grouped are the arrival and departure day of the week. In the groups for arrival day of the week, we consider end of the week to include Friday, Saturday, Sunday, and Monday. The middle of the week includes Tuesday, Wednesday, and Thursday. We consider the end of the week for the departure days to be Thursday, Friday, Saturday, and

Sunday. The beginning of the week for the departure days includes Monday, Tuesday, and Wednesday. We find that these groupings of arrival and departure day of the week strike a reasonable balance between explanatory power and model simplicity. The model with grouped arrival days of the week resulted with an R-squared value of 0.3913 and an adjusted R-squared value of 0.3138. The model with the grouped departure days of the week resulted with an R-squared value of 0.391 and adjusted R-squared value of 0.3134, which is practically unchanged. The relative mean squared error for these models are 38 hours for arrival and 38.1 hours for departure.

Next we examine the effects of collapsing both the arrival and departure days of the week. The final model with groupings of the arrival and departure days is the overall best model to explain the variance in the time that a cargo ship spends in a port in the Saint Petersburg area. MID remained the best explanatory variable with prior ports still being a strong explanatory variable. The R-squared value for the final model is 0.3893 and the adjusted R-squared value is 0.3154. Although the final model has a slightly higher adjusted R-squared value, the grouping of categorical variables did not significantly change the root mean squared error, which practically unchanged at 38 hours. The final model is not much better than the original model in terms of R-square and root mean squared error. This model has 111 variables, which is still too many to reduce the problem of overfitting the model.

C. STEPWISE MODEL SELECTION USING BAYES' INFORMATION CRITERION

We conduct model selection using Bayes Information Criterion (BIC), which penalizes additional variables in the model to prevent overfitting. BIC is defined as follows:

$$\text{BIC} = n\log(RSS / n) + p\log(n),$$

where *RSS* is the residual sum of squares, *p* is the number of coefficients in the regression model (number of explanatory variables plus one), and *n* is the number of observations (Faraway 2015). The closely-related Akaike Information Criterion (AIC) has the same form as BIC but with log(*n*) replaced by a smaller value which penalizes the addition of variables to the regression less severely. The use of BIC as a model-selection criterion in stepwise regression results in models that have fewer variables than with AIC, and many

fewer than with adjusted R-squared. We use BIC as an alternative to extracting a training and test set from the data due to the use of several categorical variables with large numbers of levels as possible explanatory variables.

Applying BIC to the original regression model results in reducing the variables from 121 to 23 variables. Figure 9 shows the variables that were selected, all of which are statistically significant. The explanatory variables that are shown to be strong predictors are MID, prior port, and next port. The R-squared value from this model is 0.2857 and the adjusted R-squared is 0.2692. These R-squared values are lower than our best model but the problem of overfitting has been addressed and the significant variables have been identified for predicting the number of hours that a cargo vessel spends in a port in the Saint Petersburg area. The root mean squared error for this model is 41.7 hours.

The regression analysis provides insights about cargo ships that stop in a port in the Saint Petersburg area. Figure 9 highlights the MIDs that are considered statistically significant, with a majority of them belonging to countries in Europe. These MIDs can be identified by the first digit which is a “2” for Europe. Russia, Germany, and Finland are the only countries within the Baltic that are included in the significant MIDs. The other European countries are either from just outside the Denmark straits or the Mediterranean Sea. The longest port stays of the significant European MIDs, are indicated by the larger coefficients of the Prefixes. Three of the MIDs (Prefix248, Prefix249, Prefix256) are registered to Malta, who has the longest port stay of all of the European countries. This makes sense since the cargo ships from this country have a long transit to the Saint Petersburg area, but Hong Kong (Prefix 477) has a longer transit and stays in port about half the time as cargo ships from Malta. The country with the longest port stay, of all of the MIDs represented in Figure 9, is Belize (Prefix312) with a coefficient of 1.697 indicating that on average there is an increase of 1.697 of log(Time) if the cargo ship has an MID of 312. The two MIDs with the shortest port stay belong to Germany (Prefix211) and Finland (Prefix230), which are two of the three MIDs that were registered to countries within the Baltic Sea. On the contrary, Russia (Prefix273) has a positive coefficient which means that log(Time) increases if the cargo ship is registered to Russia. This is not unusual since the ports in the Saint Petersburg area are Russian ports.

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      3.65156    0.03818  95.637  < 2e-16 ***
`PriorPortGDANSK PL AREA`       -0.47581    0.17617  -2.701 0.007032 **
`PriorPortKOTKA FI AREA`        -0.38450    0.11391  -3.375 0.000765 ***
`PriorPortLIEPAJA LV`           -1.71706    0.57585  -2.982 0.002936 **
`PriorPortOSKARSHAMN SE AREA`   -2.89610    0.81400  -3.558 0.000392 ***
`PriorPortPALDISKI EE AREA`     -0.73424    0.15712  -4.673 3.37e-06 ***
`PriorPortREKA LUGA RU`         -0.45475    0.10106  -4.500 7.61e-06 ***
`NextPortGULF OF BOTHNIA`       -0.40205    0.09921  -4.052 5.46e-05 ***
`NextPortHELSINKI FI AREA`      -0.53167    0.11885  -4.473 8.59e-06 ***
`NextPortKLAIPEDA LT AREA`      -0.51663    0.18211  -2.837 0.004648 **
`NextPortKOTKA FI AREA`         -0.38946    0.07918  -4.918 1.02e-06 ***
`NextPortREKA LUGA RU`          -0.79081    0.14143  -5.591 2.91e-08 ***
`NextPortTALLINN EE AREA`       -0.74783    0.25906  -2.887 0.003978 **
Prefix211                       -0.59613    0.20104  -2.965 0.003097 **
Prefix230                       -0.55212    0.10028  -5.506 4.67e-08 ***
Prefix244                        0.44585    0.14380   3.100 0.001987 **
Prefix248                        0.81436    0.25814   3.155 0.001655 **
Prefix249                        0.80063    0.25814   3.102 0.001980 **
Prefix256                        0.82944    0.21857   3.795 0.000157 ***
Prefix273                        0.43001    0.10694   4.021 6.23e-05 ***
Prefix312                        1.69689    0.46768   3.628 0.000300 ***
Prefix375                        1.16802    0.30753   3.798 0.000155 ***
Prefix477                        0.49788    0.16876   2.950 0.003250 **
Prefix636                        0.28228    0.10082   2.800 0.005210 **
```

Figure 9. Regression Analysis Coefficients After BIC Model Selection

Although the MID is the best explanatory variable, there are other insights to be gained from the other explanatory variables in Figure 9. The prior port of the cargo ships stopping in Saint Petersburg is another strong explanatory variable. The shortest port stay occurred when the cargo ship had previously stopped in Oskarshamn, Sweden. The longest port stay occurred when the cargo ship had previously stopped in the Kotka, Finland area. There is no obvious correlation between the distance of travel within the Baltic Sea to Saint Petersburg and the duration of a cargo ship's stay in port. Next port visited by cargo ships in the Saint Petersburg area is another strong explanatory variable. Kotka, Finland and Reka Luga, Russia are both significant variables for the prior port and next port explanatory variables. This could indicate that there are substantial transitions between those two ports and Saint Petersburg.

D. RANDOM FOREST BINOMIAL RESPONSE VARIABLE RESULTS

Our second research objective is to more accurately predict whether a vessel that departs Saint Petersburg, Russia area will depart the Baltic Sea toward the Atlantic Ocean or visit another port within the Baltic. In total, there are 425 cases where the vessel exits out the Southwest Atlantic and 375 cases where the vessel visits another port within the Baltic Sea. We use three random forest models to predict our binomial response variable, which is basically "yes" if the vessel departed the Baltic Sea toward the Atlantic Ocean or a "no" if it visited another port within the Baltic. In the first model, the only predictor variables that we use are the day of the week that the vessel arrives in port, the day in the year that the vessel arrives in port, length of port stay, and the MID that represents the country of registration for the vessel. In the second model we add the previous port that the vessels visited as a predictor variable. In the third model we add the port previous to the prior port of visit as a predictor variable. This thesis utilizes the R package "randomForest" (Liaw and Wiener 2002) for both parts of the analysis. In Random Forest, classification errors are calculated using out of bag values that are not included in the tree used to predict.

Model 0 does not consider any previous ports visited by the vessel. As shown in Table 3, the misclassification rate for the prediction of whether a vessel is leaving the Baltic is about 30 percent and 35 percent for whether it remains in the Baltic and visits another port. The overall misclassification rate is 31.5 percent.

Table 3. Confusion Matrix and Misclassification Rate for Model with No Prior Ports

Model 0 Confusion Matrix

	Pred. No	Pred. Yes	Class. Error	Overall Error
Actual No	244	131	0.349	31.5%
Actual Yes	128	297	0.301	

The addition of one previous port to Model 1, changes the misclassification rate for the prediction of whether a vessel leaves the Baltic Sea. Table 4 shows about a seven percent decrease, resulting in a classification error of almost 23 percent. The classification error for the vessel remaining in the Baltic is practically unchanged at 34 percent. The overall misclassification rate decreases by about three percent to 28.2 percent by including one prior port into the prediction model.

Table 4. Confusion Matrix and Misclassification Rate for Model with One Prior Port

Model 1 Confusion Matrix

	Pred. No	Pred. Yes	Class. Error	Overall Error
Actual No	247	128	0.341	28.2 %
Actual Yes	97	328	0.228	

In the third model, two previous ports are added and there is a change in both of the misclassification rates. Table 5 shows the classification errors, with about a three percent decrease in the both the prediction of whether the vessel departed the Baltic or visited another port within the Baltic Sea. This model has the lowest misclassification rate with 27.2 percent, which emphasizes the importance of past information.

Table 5. Confusion Matrix and Misclassification Rate for Model with Two Prior Ports

Model 2 Confusion Matrix

	Pred. No	Pred. Yes	Class. Error	Overall Error
Actual No	259	116	0.309	27.2%
Actual Yes	86	339	0.202	

This approach highlights the value of including past information in the prediction models to predict the next port of visit. The overall misclassification rate decreased with the addition of prior ports, which indicates dependence on past information in order to more accurately predict where a vessel transitions to next. Figure 10 shows the number of vessels

that depart or remain in the Baltic Sea in general from the dataset and the predictions of those numbers from the three random forest models. Both values are increasing in accuracy as the model adds more predictor variables about where the vessel has been.

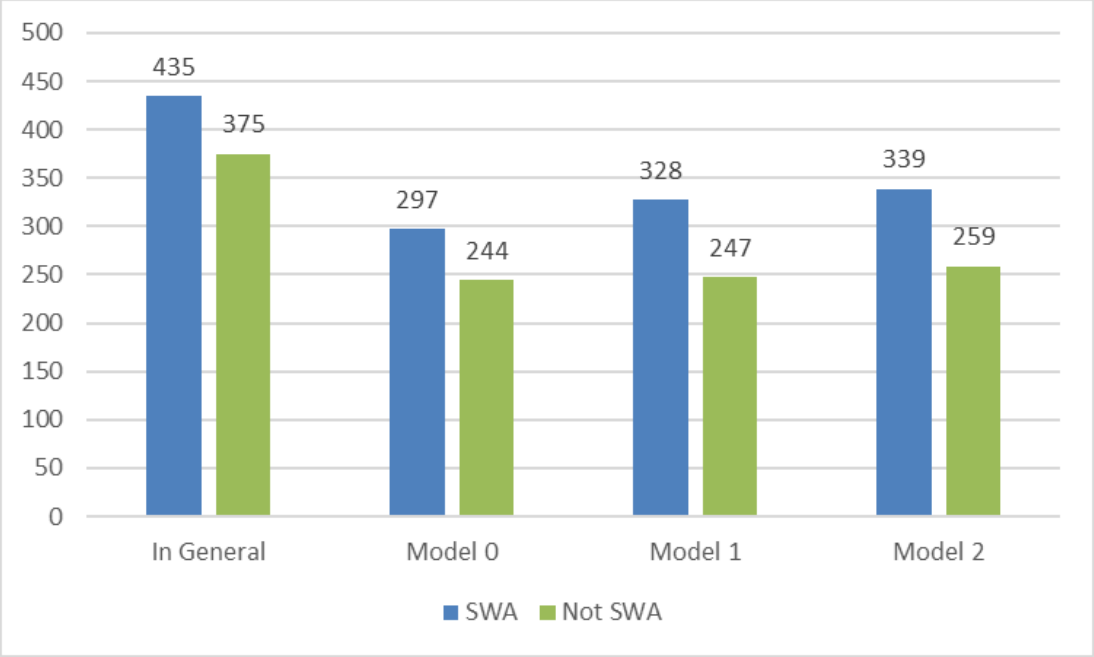


Figure 10. Comparison of Correct Predictions for SW Atlantic or Not SW Atlantic between Models and the Actual Dataset (In General)

E. RANDOM FOREST MULTINOMIAL RESPONSE VARIABLE RESULTS

Our third research objective is to predict with greater accuracy the next port of visit for a vessel that is currently in Saint Petersburg area. Our response variable changes from binomial to multinomial since we are now predicting transitional probabilities for the next port of visit. In order to separate the data into a training and test set, we group the 23 possible next ports of visit into four categories. These categories are named: "SW ATLANTIC," "KOTKA FI AREA," "OTHER FI SE," and "ALL OTHERS." The Southwest Atlantic and Kotak, Finland had the most observations and remained as they were. The rest of Finland and Sweden were grouped together, while the rest of the ports formed the last category of "ALL OTHERS."

For the test set, we randomly select 20 percent of the data without replacement. This results in a test set of 160 observations, while the training set has 240 observations. We fit three random forest models to the training set in order to predict the transitional probabilities of a vessel's next port of visit. These random forest models are set up similarly to the models used in the prediction of the binomial response variable. Our predictor variables include the arrival day of the week, arrival day in the year, duration of port stay, and the vessel's MID. One prior port is added to the second model as a predictor variable and two ports are added to the third model.

Table 6 shows the predicted next ports of visit versus the actual next ports of visit. The overall misclassification error is about 38.5 percent for a model that only considers information about the vessel and no prior port information. The lowest classification error is for the SW ATLANTIC with about 20 percent and the highest is ALL OTHERS with about 66 percent. This is reasonable considering that the SW ATLANTIC has the most observations and ALL OTHERS is comprised of many different next ports of visit. The test set has an overall misclassification error of about 44 percent. Although it is higher than the training set, the test set only consists of 160 observations.

Table 6. Confusion Matrix and Misclassification Rate for Training Data in Model 0 with No Prior Ports

		Actual					
	Model 0 (Training Set)	SW ATLANTIC	KOTKA FI AREA	OTHER FI SE	ALL OTHERS	Class. Error	Overall Error
Predicted	SW ATLANTIC	271	16	26	26	0.201	38.5%
	KOTKA FI AREA	34	48	8	5	0.495	
	OTHER FI SE	54	14	46	4	0.610	
	ALL OTHERS	46	7	5	30	0.659	

Table 7 shows the results for the training set in model 1, where one prior port is included as a predictor variable. The overall misclassification error for this model is about 36.3 percent. The lowest classification error is again the SW ATLANTIC, with an even lower percentage at about 16 percent. The highest classification error is again ALL OTHERS with an increase resulting in about 68 percent. The test set has an overall misclassification error of about 37 percent, which is close to the training set and better than in Model 1.

Table 7. Confusion Matrix and Misclassification Rate for Training Data in Model 1 with One Prior Port

	Model 1 (Training Set)	Actual: SW ATLANTIC	KOTKA FI AREA	OTHER FI SE	ALL OTHERS	Class. Error	Overall Error
Predicted	SW ATLANTIC	285	15	18	21	0.159	36.3%
	KOTKA FI AREA	34	49	8	4	0.484	
	OTHER FI SE	55	11	47	5	0.602	
	ALL OTHERS	48	8	4	28	0.682	

Table 8 shows the results for the training set in Model 2, where two ports are included as predictor variables. The overall misclassification error of about 31.7 percent. This the lowest of all three models, highlighting the importance of including past information in the prediction of the next port of visit for a vessel. The test set has an overall misclassification error of about 30 percent, which is the lowest of all models and lower than the training set. The lowest classification error is still the SW ATLANTIC with about 13 percent and the highest is still ALL OTHERS with about 64 percent.

Table 8. Confusion Matrix and Misclassification Rate for Training Data in Model 2 with Two Prior Ports

		Actual					
	Model 2 (Training Set)	SW ATLANTIC	KOTKA FI AREA	OTHER FI SE	ALL OTHERS	Class. Error	Overall Error
Predicted	SW ATLANTIC	296	7	17	19	0.127	31.7%
	KOTKA FI AREA	30	55	9	1	0.421	
	OTHER FI SE	49	10	55	4	0.534	
	ALL OTHERS	51	2	3	32	0.636	

This approach agrees with the prediction of the binomial response variable by emphasizing the value of including past information in the prediction models. Figure 11 shows performance improvement with the addition of previous ports in the models. The largest decrease in the classification error is between model 1 and model 2, indicating that there is more to be gained by going deeper into the past of a vessel when predicting its next port of visit.

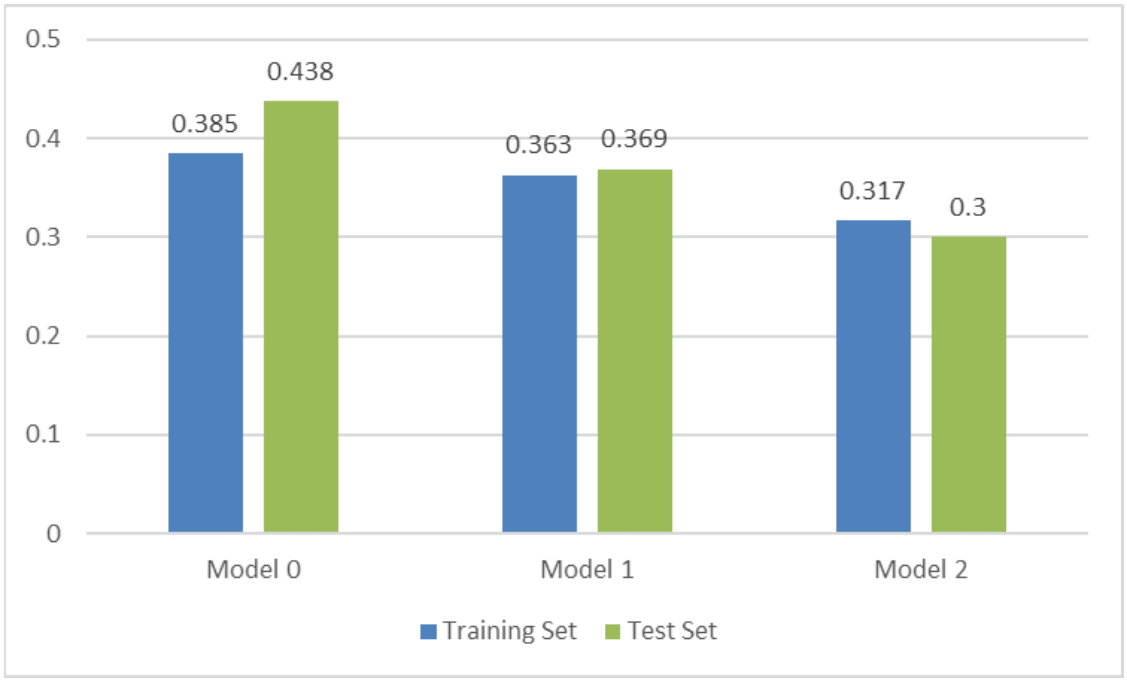


Figure 11. Comparison of Training and Test Set Misclassification Rates Between Models

F. RANDOM FOREST ESTIMATION RESULTS

Our fourth research objective is to estimate the transitional probabilities of a vessel's next port of visit based on information about the vessel and its voyage. We select four scenarios from the dataset for evaluation of the random forest models in estimating a vessel's next port of visit. We develop three random forest models, set up similarly as the last two objectives, to obtain estimated probabilities for the next port visited. We include all 23 possible next ports of visit that a cargo ship can transition to after visiting the Saint Petersburg area in our analysis. These ports are the only ports in the data visited after stopping in Saint Petersburg, Russia.

1. FIRST SCENARIO

For the first scenario, we set the vessel and port arrival attributes as follows:

- Arrival Day of Week: Tuesday
- Arrival Day in Year: 84
- Length of Port Stay: 21 hours
- MID (Country code of vessel): 209
- First Prior Port: SW Atlantic
- Second Prior Port: Saint Petersburg RU Area

Using this information, we apply our three random forest models to estimate the transitional probabilities for the vessel's next port of visit. We examine how the probabilities change as more information, in regards to previous ports, is added to the models. For the first scenario, Table 9 shows the probabilities of the ports listed being the next port of visit after the Saint Petersburg area. The model number represents how much information is known about the prior ports of a vessel's stopping in the Saint Petersburg area, ranging from no prior ports known to two prior ports known.

Table 9. Estimated Transitional Probabilities and Standard Errors for Each Random Forest Model with Different Number of Known Prior Ports for Scenario 1

NEXT PORT	RF Model 0	Std. Error	RF Model 1	Std. Error	RF Model 2	Std. Error
E.GOTLAND SE AREA	< 0.001	0.001	< 0.001	0.001	0.002	0.002
GDANSK PL AREA	0.098	0.008	0.122	0.009	0.128	0.009
GULF OF BOTHNIA	0.042	0.022	0.032	0.022	0.056	0.031
HAMINA FI	0.002	0.001	< 0.001	0.001	< 0.001	0.004
HANKO FI AREA	0.010	0.005	0.004	0.004	0.008	0.007
HELSINKI FI AREA	0.060	0.023	0.068	0.024	0.052	0.030
KALININGRAD RU AREA	< 0.001	< 0.001	< 0.001	0.002	0.002	0.001
KALMAR SE AREA	0.006	0.006	0.002	0.004	0.004	0.003
KLAIPEDA LT AREA	0.038	0.027	0.032	0.014	0.040	0.016
KOTKA FI AREA	0.244	0.035	0.272	0.044	0.184	0.028
KUNDA EE	< 0.001	0.002	< 0.001	0.002	0.002	0.004
NYKOPING SE AREA	0.002	0.007	0.002	0.009	0.012	0.012
PALDISKI EE AREA	0.010	0.002	0.030	0.008	0.030	0.011
PARNU EE	< 0.001	0.001	< 0.001	0.001	< 0.001	0.001
PRIMORSK RU	0.008	0.002	< 0.001	0.002	0.002	0.003
REKA LUGA RU	0.060	0.010	0.050	0.014	0.052	0.008
RIGA LV AREA	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.002
STOCKHOLM SE AREA	< 0.001	0.001	< 0.001	0.001	< 0.001	0.001
SW ATLANTIC	0.418	0.044	0.384	0.051	0.418	0.041
TALLINN EE AREA	< 0.001	0.004	0.002	0.007	0.006	0.009
VYBORG RU	< 0.001	0.002	< 0.001	0.001	< 0.001	0.002
W.GOTLAND SE AREA	0.002	0.011	< 0.001	0.004	< 0.001	0.004
WOLGAST DE AREA	< 0.001	< 0.001	< 0.001	< 0.001	0.002	< 0.001

In Table 10, the probabilities of the 23 possible next ports are influenced by the addition of predictor variables that go deeper into the past. The most likely next ports of visit, given two prior ports visited, are the SW ATLANTIC (0.42), KOTKA FI AREA (0.18), and GDANSK PL AREA (0.13). We calculate estimated standard errors by bootstrapping each random forest model 30 times. The same sample is used to fit all three random forest models each time. The resulting standard errors for the probability estimates are shown in Table 10. The resulting standard errors for the probability estimates shown in Table 10 are calculated in a similar way to the individual models, but take into account the difference between two models that are compared. A rough 95 percent confidence interval

for the true difference in probabilities is calculated by taking plus or minus two times the standard errors of the predicted probabilities.

Table 10. Standard Errors for the Differences in Probability Estimates between Random Forest Models in Scenario 1.

NEXT PORT	Model 0 - Model 1	Std. Error	Model 0 - Model 2	Std. Error	Model 1 - Model 2	Std. Error
E.GOTLAND SE AREA	0.000	0.001	-0.002	0.002	-0.002	0.002
GDANSK PL AREA	-0.024	0.009	-0.030	0.009	-0.006	0.011
GULF OF BOTHNIA	0.010	0.019	-0.014	0.024	-0.024	0.027
HAMINA FI	0.002	0.001	0.002	0.004	0.000	0.004
HANKO FI AREA	0.006	0.005	0.002	0.006	-0.004	0.007
HELSINKI FI AREA	-0.008	0.019	0.008	0.031	0.016	0.029
KALININGRAD RU AREA	0.000	0.002	-0.002	0.001	-0.002	0.002
KALMAR SE AREA	0.004	0.006	0.002	0.005	-0.002	0.004
KLAIPEDA LT AREA	0.006	0.024	-0.002	0.024	-0.008	0.016
KOTKA FI AREA	-0.028	0.032	0.060	0.029	0.088	0.036
KUNDA EE	0.000	0.002	-0.002	0.003	-0.002	0.004
NYKOPING SE AREA	0.000	0.007	-0.010	0.010	-0.010	0.009
PALDISKI EE AREA	-0.020	0.007	-0.020	0.011	0.000	0.009
PARNU EE	0.000	0.001	0.000	0.001	0.000	0.001
PRIMORSK RU	0.008	0.002	0.006	0.003	-0.002	0.003
REKA LUGA RU	0.010	0.014	0.008	0.012	-0.002	0.014
RIGA LV AREA	0.000	< 0.001	0.000	0.002	0.000	0.002
STOCKHOLM SE AREA	0.000	0.001	0.000	0.001	0.000	0.001
SW ATLANTIC	0.034	0.041	0.000	0.037	-0.034	0.046
TALLINN EE AREA	-0.002	0.006	-0.006	0.008	-0.004	0.009
VYBORG RU	0.000	0.002	0.000	0.003	0.000	0.002
W.GOTLAND SE AREA	0.002	0.009	0.002	0.009	0.000	0.003
WOLGAST DE AREA	0.000	< 0.001	-0.002	< 0.001	-0.002	< 0.001

Differences highlighted in Green are more than two standard errors in magnitude.

Table 10 presents the SEs for the differences between the three random forest models. We compare differences of predicted probabilities between models because a zero difference means that including prior ports does not change the probability of going to a particular port next. The numbers that are highlighted in green in Table 10 indicate differences that are greater in magnitude than two standard errors, which implies that the

probabilities in each model are statistically different. These differences indicate that past information does change the probability of a vessel's next port of visit. With four occurrences, the largest difference is between model 0 and model 2. Model 0 has no prior ports included in the random forest model and model 2 has two prior ports included.

2. SECOND SCENARIO

The second scenario that we select has the following information for the specific vessel:

- Arrival Day of Week: Sunday
- Arrival Day in Year: 12
- Length of Port Stay: 19 hours
- MID (Country code of vessel): 209
- First Prior Port: SW Atlantic
- Second Prior Port: Reka Luga RU

This scenario differs from the first because now we are exploring the 12th day of the year which is in January, vice late March. It is also a weekend day compared to a Tuesday in the first scenario.

In Table 11, the 23 possible next ports are given estimated transitional probabilities based on the information for this scenario and the three random forest models. The most likely ports to be visited next are SW ATLANTIC (0.31), KOTKA FI AREA (0.22), and HELSINKI FI AREA (0.16). The only port, of the three most likely, to have an increased probability after the addition of prior ports as predictor variables is Helsinki, Finland area. We calculate standard errors by bootstrapping each random forest model 30 times and use the same sample to fit all three random forest models each time. The standard errors for each model in scenario 2 are shown in Table 11.

Table 11. Estimated Transitional Probabilities and Standard Errors for Each Random Forest Model in Scenario 2

NEXT PORT	RF Model 0	Std. Error	RF Model 1	Std. Error	RF Model 2	Std. Error
E.GOTLAND SE AREA	< 0.001	0.003	< 0.001	0.003	< 0.001	0.005
GDANSK PL AREA	0.008	0.018	0.024	0.019	0.020	0.014
GULF OF BOTHNIA	0.052	0.052	0.022	0.015	0.086	0.015
HAMINA FI	< 0.001	0.013	< 0.001	0.016	0.002	0.016
HANKO FI AREA	0.004	0.004	0.002	0.004	0.010	0.005
HELSINKI FI AREA	0.136	0.015	0.182	0.017	0.160	0.013
KALININGRAD RU AREA	< 0.001	0.002	< 0.001	0.003	< 0.001	0.003
KALMAR SE AREA	0.014	0.002	0.006	0.002	0.018	0.003
KLAIPEDA LT AREA	0.040	0.033	0.028	0.038	0.024	0.038
KOTKA FI AREA	0.246	0.040	0.310	0.038	0.216	0.029
KUNDA EE	0.012	0.004	0.006	0.001	0.004	0.004
NYKOPING SE AREA	0.002	0.018	0.004	0.014	0.006	0.012
PALDISKI EE AREA	0.004	< 0.001	0.012	0.001	0.044	0.003
PARNU EE	< 0.001	0.002	< 0.001	0.001	< 0.001	0.003
PRIMORSK RU	0.008	0.007	0.004	0.009	0.004	0.012
REKA LUGA RU	0.080	0.016	0.074	0.016	0.084	0.013
RIGA LV AREA	< 0.001	< 0.001	< 0.001	< 0.001	0.002	0.001
STOCKHOLM SE AREA	< 0.001	0.001	< 0.001	0.001	< 0.001	0.004
SW ATLANTIC	0.390	0.070	0.320	0.077	0.308	0.060
TALLINN EE AREA	< 0.001	0.009	0.006	0.010	0.010	0.008
VYBORG RU	0.002	0.008	< 0.001	0.005	< 0.001	0.007
W.GOTLAND SE AREA	0.002	< 0.001	< 0.001	< 0.001	< 0.001	0.001
WOLGAST DE AREA	< 0.001	0.001	< 0.001	0.001	0.002	0.002

Table 12 contains the standard errors for the differences between the three random forest models. The numbers that are highlighted in green in Table 12 indicate differences that are greater in magnitude than two standard errors, which means that the probabilities in each model are statistically different. With five occurrences, the largest difference is between model 0 and model 1. This is different than the first scenario where model 0 and model 2 had the largest differences. The difference between model 1 and model 2 has four differences, which indicates significance in the addition of a second prior port to the random forest model.

Table 12. Standard Errors for the Differences in Probability Estimates Between Random Forest Models in Scenario 2.

NEXT PORT	Model 0 - Model 1	Std. Error	Model 0 - Model 2	Std. Error	Model 1 - Model 2	Std. Error
E.GOTLAND SE AREA	0.000	0.002	0.000	0.004	0.000	0.004
GDANSK PL AREA	-0.016	0.015	-0.012	0.015	0.004	0.012
GULF OF BOTHNIA	0.030	0.047	-0.034	0.044	-0.064	0.017
HAMINA FI	0.000	0.007	-0.002	0.012	-0.002	0.011
HANKO FI AREA	0.002	0.004	-0.006	0.005	-0.008	0.005
HELSINKI FI AREA	-0.046	0.010	-0.024	0.014	0.022	0.016
KALININGRAD RU AREA	0.000	0.003	0.000	0.004	0.000	0.004
KALMAR SE AREA	0.008	0.002	-0.004	0.003	-0.012	0.003
KLAIPEDA LT AREA	0.012	0.018	0.016	0.020	0.004	0.017
KOTKA FI AREA	-0.064	0.027	0.030	0.039	0.094	0.026
KUNDA EE	0.006	0.004	0.008	0.003	0.002	0.003
NYKOPING SE AREA	-0.002	0.011	-0.004	0.016	-0.002	0.009
PALDISKI EE AREA	-0.008	0.001	-0.040	0.003	-0.032	0.003
PARNU EE	0.000	0.002	0.000	0.003	0.000	0.003
PRIMORSK RU	0.004	0.006	0.004	0.009	0.000	0.008
REKA LUGA RU	0.006	0.016	-0.004	0.016	-0.010	0.016
RIGA LV AREA	0.000	< 0.001	-0.002	0.001	-0.002	0.001
STOCKHOLM SE AREA	0.000	0.002	0.000	0.004	0.000	0.004
SW ATLANTIC	0.070	0.044	0.082	0.051	0.012	0.045
TALLINN EE AREA	-0.006	0.008	-0.010	0.008	-0.004	0.009
VYBORG RU	0.002	0.006	0.002	0.007	0.000	0.006
W.GOTLAND SE AREA	0.002	< 0.001	0.002	0.001	0.000	0.001
WOLGAST DE AREA	0.000	0.001	-0.002	0.002	-0.002	0.001

Differences highlighted in Green are more than two standard errors in magnitude.

3. THIRD SCENARIO

For the third scenario, we examine a vessel with the following information:

- Arrival Day of Week: Thursday
- Arrival Day in Year: 100
- Length of Port Stay: 29 hours
- MID (Country code of vessel): 636
- First Prior Port: Reka Luga RU
- Second Prior Port: SW Atlantic

This scenario examines a day later in the week as well as seasonally by using a Thursday in April. The MID of the vessel is completely different because we are now using a vessel that is from Africa, Liberia to be exact, rather than a vessel from Europe as we did in the first two scenarios. The first prior port is also different than the first two scenarios since we are using an actual port instead of the southwest Atlantic. Table 13 shows the 23 potential next ports of visit, with the most likely ports being SW ATLANTIC (0.35), KOTKA FI AREA (0.25), and GULF OF BOTHNIA (0.11).

Table 13. Transitional Probabilities and Standard Errors for Each Random Forest Model in Scenario 3

NEXT PORT	RF Model 0	Std. Error	RF Model 1	Std. Error	RF Model 2	Std. Error
E.GOTLAND SE AREA	0.004	0.001	0.008	0.001	0.002	0.001
GDANSK PL AREA	0.040	0.011	0.082	0.011	0.064	0.010
GULF OF BOTHNIA	0.054	0.023	0.100	0.035	0.108	0.027
HAMINA FI	< 0.001	< 0.001	< 0.001	< 0.001	0.002	0.002
HANKO FI AREA	0.012	0.007	0.010	0.005	0.012	0.007
HELSINKI FI AREA	0.110	0.022	0.060	0.017	0.060	0.021
KALININGRAD RU AREA	< 0.001	0.001	< 0.001	0.001	< 0.001	0.001
KALMAR SE AREA	0.012	0.005	0.014	0.008	0.008	0.008
KLAIPEDA LT AREA	0.020	0.033	0.038	0.019	0.030	0.018
KOTKA FI AREA	0.182	0.042	0.232	0.031	0.252	0.031
KUNDA EE	0.008	< 0.001	0.004	0.009	0.004	0.009
NYKOPING SE AREA	0.024	0.005	0.016	0.003	0.008	0.007
PALDISKI EE AREA	0.018	0.006	0.026	0.007	0.020	0.007
PARNU EE	< 0.001	< 0.001	0.002	0.002	< 0.001	0.002
PRIMORSK RU	0.002	0.032	0.002	0.019	0.002	0.012
REKA LUGA RU	0.032	0.009	0.064	0.010	0.064	0.007
RIGA LV AREA	< 0.001	< 0.001	< 0.001	0.001	< 0.001	0.002
STOCKHOLM SE AREA	0.012	0.002	0.002	0.002	0.004	0.002
SW ATLANTIC	0.452	0.052	0.330	0.076	0.354	0.054
TALLINN EE AREA	0.012	0.009	0.006	0.038	0.006	0.022
VYBORG RU	< 0.001	0.001	< 0.001	0.001	< 0.001	0.002
W.GOTLAND SE AREA	0.002	0.001	< 0.001	0.002	< 0.001	0.002
WOLGAST DE AREA	0.004	< 0.001	0.004	< 0.001	< 0.001	< 0.001

We calculate standard errors by bootstrapping all three random forest models 30 times and using the same sample to fit each model. The standard errors for each model in

scenario 3 are shown in Table 14. Table 14 contains the standard errors for the differences between the three random forest models. The rows highlighted in green in Table 14 indicate differences that are greater in magnitude than two standard errors, which means that the probabilities in each model are statistically different. With eight occurrences, the largest difference is between model 0 and model 2. These results are similar to the first scenario and different than the second scenario. This scenario results in some significance by adding one prior port and more significance by adding two prior ports, which indicates dependence on past information in order to more accurately predict a vessel's next port of visit.

Table 14. Standard Errors for the Differences in Probability Estimates between Random Forest Models in Scenario 3

NEXT PORT	Model 0 - Model 1	Std. Error	Model 0 - Model 2	Std. Error	Model 1 - Model 2	Std. Error
E.GOTLAND SE AREA	-0.004	0.001	0.002	0.001	0.006	0.001
GDANSK PL AREA	-0.042	0.012	-0.024	0.009	0.018	0.010
GULF OF BOTHNIA	-0.046	0.031	-0.054	0.027	-0.008	0.028
HAMINA FI	0.000	0.000	-0.002	0.002	-0.002	0.002
HANKO FI AREA	0.002	0.006	0.000	0.007	-0.002	0.006
HELSINKI FI AREA	0.050	0.025	0.050	0.024	0.000	0.020
KALININGRAD RU AREA	0.000	0.001	0.000	0.001	0.000	0.001
KALMAR SE AREA	-0.002	0.009	0.004	0.009	0.006	0.006
KLAIPEDA LT AREA	-0.018	0.021	-0.010	0.028	0.008	0.018
KOTKA FI AREA	-0.050	0.035	-0.070	0.035	-0.020	0.019
KUNDA EE	0.004	0.009	0.004	0.009	0.000	0.003
NYKOPING SE AREA	0.008	0.004	0.016	0.007	0.008	0.006
PALDISKI EE AREA	-0.008	0.006	-0.002	0.007	0.006	0.007
PARNU EE	-0.002	0.002	0.000	0.002	0.002	0.002
PRIMORSK RU	0.000	0.017	0.000	0.023	0.000	0.011
REKA LUGA RU	-0.032	0.010	-0.032	0.010	0.000	0.010
RIGA LV AREA	0.000	0.001	0.000	0.002	0.000	0.002
STOCKHOLM SE AREA	0.010	0.002	0.008	0.003	-0.002	0.003
SW ATLANTIC	0.122	0.062	0.098	0.054	-0.024	0.040
TALLINN EE AREA	0.006	0.034	0.006	0.018	0.000	0.023
VYBORG RU	0.000	0.002	0.000	0.002	0.000	0.002
W.GOTLAND SE AREA	0.002	0.002	0.002	0.002	0.000	0.002
WOLGAST DE AREA	0.000	0.000	0.004	0.000	0.004	0.000

Differences highlighted in Green are more than two standard errors in magnitude.

4. FOURTH SCENARIO

The fourth scenario that we select is a vessel with the following information:

- Arrival Day of Week: Sunday
- Arrival Day in Year: 103
- Length of Port Stay: 109 hours
- MID (Country code of vessel): 377
- First Prior Port: SW Atlantic
- Second Prior Port: Saint Petersburg RU Area

This scenario considers a day in March where a vessel spent over one hundred hours in a port in the Saint Petersburg area. This amount of time is larger than the ones used in the previous three scenarios. The vessel also differs because it is registered to a country in the Caribbean, which means that it had to make a long journey to get to the Baltic Sea. The second prior port is the same location that the vessel is currently.

Table 15 shows the 23 potential next ports and their respective transitional probabilities given zero prior ports, one prior port, and two prior ports. The next ports of visit with the highest transitional probabilities are SW ATLANTIC (0.52), PRIMORSK RU (0.13), and KOTKA FI AREA (0.09). We calculate the standard errors in Table 15 by bootstrapping each random forest model 30 times. The same sample is used to fit all three random forest models each time.

Table 15. Transitional Probabilities and Standard Errors for Each Random Forest Model in Scenario 4

NEXT PORT	RF Model 0	Std. Error	RF Model 1	Std. Error	RF Model 2	Std. Error
E.GOTLAND SE AREA	0.006	0.002	0.010	0.001	0.016	0.002
GDANSK PL AREA	0.032	0.058	0.016	0.065	0.036	0.062
GULF OF BOTHNIA	0.048	0.025	0.018	0.028	0.050	0.021
HAMINA FI	0.004	< 0.001	0.002	0.001	0.016	0.004
HANKO FI AREA	0.002	0.014	< 0.001	0.022	0.006	0.013
HELSINKI FI AREA	0.006	0.025	0.004	0.023	0.024	0.020
KALININGRAD RU AREA	0.044	< 0.001	0.056	0.001	0.030	0.002
KALMAR SE AREA	< 0.001	0.007	< 0.001	0.003	0.000	0.004
KLAIPEDA LT AREA	0.030	0.023	0.034	0.024	0.024	0.025
KOTKA FI AREA	0.050	0.044	0.070	0.040	0.088	0.025
KUNDA EE	0.004	0.002	0.004	0.001	0.004	0.002
NYKOPING SE AREA	0.004	0.006	0.012	0.007	0.012	0.006
PALDISKI EE AREA	< 0.001	0.007	0.002	0.012	< 0.001	0.011
PARNU EE	0.004	< 0.001	< 0.001	< 0.001	0.002	0.001
PRIMORSK RU	0.138	0.016	0.166	0.007	0.134	0.008
REKA LUGA RU	0.008	0.021	0.008	0.017	0.012	0.013
RIGA LV AREA	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.001
STOCKHOLM SE AREA	0.010	0.003	< 0.001	0.003	0.002	0.003
SW ATLANTIC	0.582	0.043	0.550	0.044	0.518	0.059
TALLINN EE AREA	< 0.001	0.008	< 0.001	0.005	0.002	0.006
VYBORG RU	< 0.001	0.001	< 0.001	0.001	0.002	0.002
W.GOTLAND SE AREA	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
WOLGAST DE AREA	0.028	0.001	0.048	0.001	0.022	0.004

Table 16 presents the standard errors for the differences between the three random forest models. The numbers that are highlighted in green in Table 16 indicate differences that are greater in magnitude than two standard errors, which means that these differences indicate that past information does change the probability of a vessel's next port of visit. With seven occurrences, the largest difference is between model 0 and model 1. These results are similar to the second scenario and different than the first and third scenarios, which were similar to each other.

Table 16. Standard Errors for the Differences in Probability Estimates between Random Forest Models in Scenario 4.

NEXT PORT	Model 0 - Model 1	Std. Error	Model 0 - Model 2	Std. Error	Model 1 - Model 2	Std. Error
E.GOTLAND SE AREA	-0.004	0.002	-0.010	0.002	-0.006	0.003
GDANSK PL AREA	0.016	0.021	-0.004	0.022	-0.020	0.020
GULF OF BOTHNIA	0.030	0.015	-0.002	0.020	-0.032	0.024
HAMINA FI	0.002	0.001	-0.012	0.004	-0.014	0.004
HANKO FI AREA	0.002	0.011	-0.004	0.008	-0.006	0.012
HELSINKI FI AREA	0.002	0.021	-0.018	0.025	-0.020	0.021
KALININGRAD RU AREA	-0.012	0.001	0.014	0.001	0.026	0.002
KALMAR SE AREA	0.000	0.006	0.000	0.006	0.000	0.004
KLAIPEDA LT AREA	-0.004	0.021	0.006	0.021	0.010	0.013
KOTKA FI AREA	-0.020	0.030	-0.038	0.035	-0.018	0.031
KUNDA EE	0.000	0.002	0.000	0.002	0.000	0.002
NYKOPING SE AREA	-0.008	0.006	-0.008	0.005	0.000	0.006
PALDISKI EE AREA	-0.002	0.009	0.000	0.009	0.002	0.009
PARNU EE	0.004	0.000	0.002	0.001	-0.002	0.001
PRIMORSK RU	-0.028	0.011	0.004	0.012	0.032	0.007
REKA LUGA RU	0.000	0.017	-0.004	0.018	-0.004	0.015
RIGA LV AREA	0.000	0.000	0.000	0.001	0.000	0.001
STOCKHOLM SE AREA	0.010	0.003	0.008	0.004	-0.002	0.004
SW ATLANTIC	0.032	0.041	0.064	0.052	0.032	0.037
TALLINN EE AREA	0.000	0.006	-0.002	0.006	-0.002	0.006
VYBORG RU	0.000	0.001	-0.002	0.002	-0.002	0.002
W.GOTLAND SE AREA	0.000	0.000	0.000	0.000	0.000	0.000
WOLGAST DE AREA	-0.020	0.001	0.006	0.004	0.026	0.004

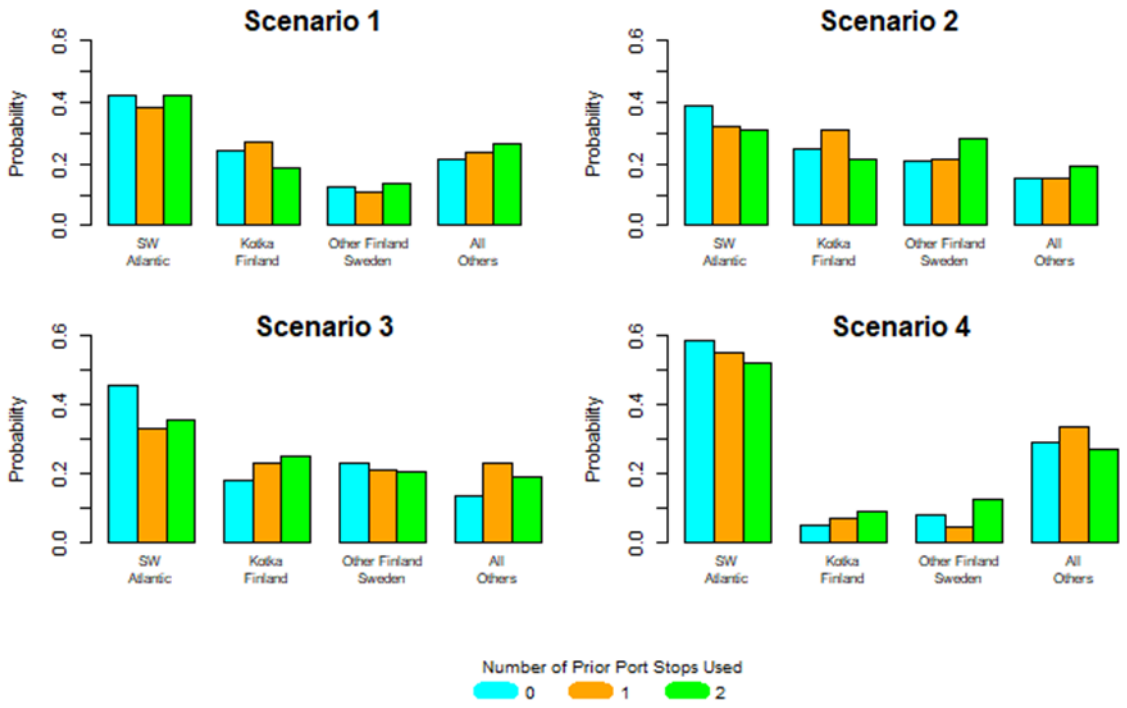
Differences highlighted in Green are more than two standard errors in magnitude.

G. DISCUSSION OF SCENARIOS

Across all four scenarios, there is not one port that is always statistically different among the three models. Scenario 2 and 4 deemed model 0 and model 1 to have the most differences between the potential next ports of visit. Of those differences, none of the ports between the two scenarios were the same. The two scenarios also agree that the next largest difference is between model 1 and model 2. This indicates that the knowledge of one prior port can significantly change the transitional probabilities for the next port of visit.

Scenario 1 and 3 considered model 0 and 2 to have the largest difference between the models. Gdansk, Kotka, and Wolgast are the three next ports of visit that the two scenarios have in common. The two scenarios concur with the next largest difference is between model 0 and model 1. While scenarios 2 and 4 both agreed that one prior port can make a difference in the transitional probabilities, these scenarios show that the knowledge of two prior ports will benefit the accuracy of probabilities the most. They also indicate that the although the addition of one prior port is beneficial, it is more beneficial in comparison to a model that does not have any prior ports included than to a model that already has one prior port.

Figure 12 is a summary of the changes in estimated probabilities for a vessel's next port of visit. The 23 possible next ports of visit are grouped to show the changes in probabilities. The groupings are consistent with the groupings used in the third objective to predict a vessel's next port of visit. The two categories, OTHER FI SE and ALL OTHERS, are cumulative probabilities of the ports that are in those two groups. In all scenarios there are changes in probabilities which indicates that the knowledge of prior ports visited is useful in predicting the next port of a vessel. In Scenario 1 there is an increase in transitional probabilities of the group ALL OTHERS. In Scenario 2 there is a decrease in the probability of SW ATLANTIC being the next port of visit after leaving the Saint Petersburg, Russia area. Scenario 3 shows an increase in probability for KOTKA FI AREA and a decrease in probability for OTHER FI SE. Scenario 4 also has an increase in transitional probability KOTKA FI but a decrease in SW ATLANTIC.



The addition of prior ports in the models changes the estimated probabilities.

Figure 12. Effects of Models in Estimating the Probability of a Vessel's Next Port of Visit

V. SUMMARY AND RECOMMENDATIONS

This thesis examines historical AIS data that focuses on the Baltic Sea during the time period of January to April 2014. Specifically, we use the three ports located in the Saint Petersburg cluster, as listed in Chapter III, as the port of interest and consider only the cargo ships that had a port stop there in our analysis. We process the data to gather attributes of vessels to determine if there is any effect on the vessel's pattern of navigation as it transits through the Baltic Sea. The goal of this research is to characterize these patterns of ship navigation to better allocate surveillance assets for specific ships. Although we focus this analysis on cargo ships, it can be implemented for all types of vessels. In the following two sections, we discuss the practicality and usefulness of the two methods that were explored in response to the four research objectives laid out in Chapter I. The last section discusses recommendations to improve upon the analysis presented in this thesis.

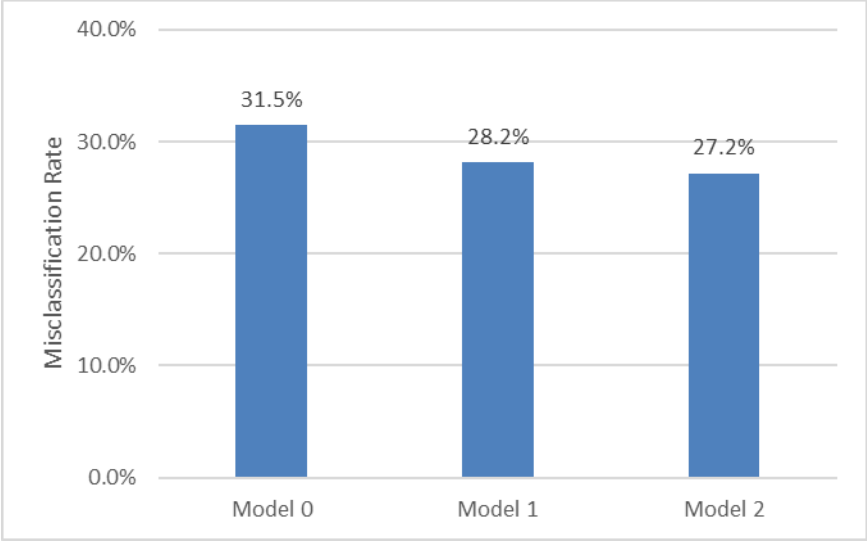
A. EFFECTIVENESS OF REGRESSION ANALYSIS

The first research objective is to explore the factors that influence the amount of time that a vessel spends in port. We predict the length of time based on the variables: MID (country code of vessel), prior port visited, next port of visit, arrival day of the week, departure day of the week, and the arrival day in the year. To correct for heteroscedasticity and non-normality, we use a logarithmic transformation to the response variable, which is the length of time in hours that a vessel spends in a port in the Saint Petersburg area. MID is determined to be the best explanatory variable in the regression analysis. It is able to explain about 25 percent of the variance in the response variable. This emphasizes the importance of maritime law enforcement ensuring that a vessel's MMSI is updated when first installed and that the MID matches the country flag that is being flown by the vessel. The second best explanatory variable is prior port, which highlights the importance of having past knowledge of the ports that a vessel visits during its voyage. Overall, the regression analysis is able to explain about 29 percent of the variance in the response variable after using BIC to penalize the addition of extra variables in the regression to avoid overfitting. The original model has 121 variables that we reduce to 23 variables using BIC.

Overall, the final regression model predictions are on average 41.7 hours off from the actual number of hours a cargo ship spent in port. We recommend using a vessel's MID, previous port visited, and next port of visit to predict the number of hours a vessel will spend in port. To improve on the final model, previous ports that go further into the past are supported. This analysis is beneficial to organizations that are following a specific vessel in its route because knowing how long a vessel is in port for allows for the setup of surveillance assets. Knowing the duration of a port stay for a VOI gives the organizations time to allocate resources to the VOI's current port or next port of visit.

B. EFFECTIVENESS OF RANDOM FOREST ANALYSIS

The second research objective is to predict with greater accuracy whether a vessel will depart or remain in the Baltic Sea after stopping in the Saint Petersburg, Russia area. We approach this objective by using a binomial response variable in three random forest models, which vary in the amount of past information that is included in the model. This approach illustrates the importance of incorporating prior ports in a vessel's voyage by showing a decrease in classification error as the models included prior ports. The first model with no prior ports included has misclassification errors of about 35 percent for vessels that remained in the Baltic Sea and about 30 percent for vessels that actually left the Baltic Sea. There is an eight percent improvement in the classification error for the second model in predicting that the vessel would depart the Baltic and no improvement for the prediction of the vessel visiting another port. The third model showed a three percent decrease in classification errors for the prediction of leaving and staying in the Baltic Sea. Figure 13 shows the misclassification rates for the three random forest models. The classification errors decrease as more past information is included in the prediction model. This analysis shows that there is potential to more accurately predict the next port of visit for a vessel given its previous destinations.



The three models vary in number of prior ports included in model.

Figure 13. Misclassification Rates for Departing or Remaining in the Baltic Sea.

The third research objective is to more accurately predict a vessel's next port of visit after departing from the Saint Petersburg area. The response variable changes from binomial to multinomial as we consider actual ports. To ensure that all 23 possible next ports of visit are included in the training and test set, we group the ports into four categories. Those categories are: "SW ATLANTIC," "KOTKA FI AREA," OTHER FI SE," and "ALL OTHERS." We use 20 percent of the dataset (160 observations) as our test set and the remaining 80 percent (640 observations) as our training set. We fit three random forest models, set up similarly to the second objective, to the training data to predict the transitional probabilities of a vessel's next port of visit. The overall misclassification error for Model 0 using the training data is 38.5 percent and with the test set it is about 44 percent. Model 1 has a lower overall misclassification error with 36.3 percent using the training data and about 37 percent with the test set. Model 2 has the lowest overall misclassification error in comparison to the first two models. With the training data, the overall misclassification error is 31.7 percent and with the test set it is about 30 percent. This analysis emphasizes the point that past information matters and is needed to more accurately predict a vessel's next port of visit. Although all possible next ports of visit were

not included in this analysis, we consider them all in the next objective by estimating the transitional probabilities for all 23 ports.

The fourth research objective is to explore the dependency that a vessel has on the previous ports visited in determining the next port of visit. We predict the next port of visit using the variables: MID (country code of vessel), arrival day of the week, arrival day in the year (1 = January 1, 2 = January 2, …, 130 = April 30), length of time vessel stays in port, first previous port, and second previous port. The three random forest models we use in this analysis are intended to show that predicting a vessel's next port of call is not a simple Markov model but more of a HON where the past matters to accurately predict the future state. Therefore, the first model has zero prior ports known, with the second model having one prior port and the third model having two prior ports known. Even though MID is the best estimator in the random forest analysis, the previous port variables increase with importance the further into the past we explore. The transitional probabilities changed when explanatory variables were added that go deeper into the past, indicating that prediction of a vessel's next port of call is not a simple Markov model but a higher-order network where more past information is needed to more accurately predict the future state.

This methodology is useful in surveillance allocation because it allows the user to more accurately predict a VOI's next port of visit by considering its previous ports visited. Misclassification rates decreased with the addition of past information in the prediction models. By estimating the probabilities for all 23 possible next ports, there is a difference in the models and the transitional probabilities.

C. RECOMMENDATIONS

The port of interest we use in this thesis is chosen to simplify the complexity of the network. This research can be scaled up to include all ports in the Baltic Sea, rather than just focusing on the three ports in the Saint Petersburg area. Scaling up the analysis would provide a means of comparing all ports to each other in terms of the time that vessels spend in port. It would not have too much of an effect on determining a vessel's next port of visit. Increasing the size of the bounding box or eliminating a bounding box and using worldwide data would help improve the transitional probabilities for a vessel going from one port to

the next. The bounding box limits where a vessel may have been prior to entering the Baltic Sea and gives more weight to the port SW ATLANTIC, which is the sole entrance and exit for the Baltic Sea. By using a global dataset, the previous ports encompass the actual ports that were visited by a vessel rather than simply labeling it the Southwest Atlantic. A global dataset that encompasses a full year is sufficient enough to develop consistent routes and capture the effects of the changing seasons.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Baltic LINes (2016) Shipping in the Baltic Sea–Past, present and future developments relevant for Maritime Spatial Planning. Project Report I. 35 p, VASAB, Retrieved from http://www.vasab.org/index.php/documents/doc_download/1275-baltic-lines-report-on-shipping-in-the-baltic-sea.

Bay S (2017) Evaluation of factors on the patterns of ship movement and predictability of future ship location in the Gulf of Mexico. Master's thesis, Dept. of Operations Research, Naval Postgraduate School, Monterey, CA. http://calhoun.nps.edu/handle/10945/53021.

Cheng J, Karambelkar B, Xie Y (2017) Leaflet: Create interactive web maps with the JavaScript 'Leaflet' Library. R package version 1.1.0. https://CRAN.Rproject.org/package=leaflet.

Faraway JJ (2015) *Linear models with R* 2nd ed. (CRC Press, Boca Raton, FL).

Faraway JJ (2016) *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models* 2nd ed. (CRC Press, Boca Raton, FL).

Fernandez AV, Pallotta G, Vespe M (2017) Maritime traffic networks: From historical positioning data to unsupervised maritime traffic monitoring. *IEEE Transactions on Intelligent Transportation Systems*, PP(99), 1–11, doi: 10.1109/TITS.2017.2699635.

Gambs S, Killijian M, Del Prado CM (2012) Next place prediction using mobility Markov chains. *Proceedings of the First Workshop on Measurement, Privacy, and Mobility* (Bern Switzerland). doi: 10.1145/2181196.2181199.

Gutierrez Torre, A. (2017). Discovering Ship Navigation Patterns towards Environmental Impact Modeling. (Master's thesis). Retrieved from http://hdl.handle.net/2117/111171.

Harati-Mokhtari A, Wall A, Brooks P, Wang J (2007) Automatic Identification System (AIS): Data reliability and human error implications. *The Journal of Navigation* 60(3): 373–389.

Hijmans RJ (2015) Geosphere: Spherical trigonometry. R package version 1.5-1. https://CRAN.R-project.org/package=geosphere.

Hintze J (2017) An analysis of vessel waypoint behavior through data clustering. Master's thesis, Dept. of Operations Research, Naval Postgraduate School, Monterey, CA. https://calhoun.nps.edu/handle/10945/56135.

International Maritime Organization (2002) SOLAS Chapter V safety of navigation. Accessed February 21, 2018, http://www.imo.org/en/OurWork/facilitation/documents/solas%20v%20on%20safety%20of%20navigation.pdf.

Koyak R (2017) Statistical tool for the analysis of Automated Information System (AIS) data final report. Unpublished manuscript. Naval Postgraduate School, Monterey, CA.

Liaw A, Wiener M (2002) Classification and Regression by randomForest. R News 2(3), 18–22.

Lin K (2017) Probability Models for Practitioners. Unpublished manuscript, Naval Postgraduate School, Monterey, CA.

McAbee A (2013) Traffic pattern detection using the Hough transformation for anomaly detection to improve maritime domain awareness. Master's thesis, Dept. of Electrical Engineering, Naval Postgraduate School, Monterey, CA. http://calhoun.nps.edu/handle/10945/38977.

National Geo-Spatial Intelligence Agency (2017) Pub 150 World Port Index twenty-sixth edition. Retrieved from https://msi.nga.mil/MSISiteContent/StaticFiles/NAV_PUBS/WPI/Pub150bk.pdf.

Pallotta G, Vespe M, Bryan K (2013) Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction. Entropy, 15(6), 2218–2245. Retrieved from http://www.mdpi.com/1099-4300/15/6/2218.

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.Rproject.org/.

Raymond ES (2016) AIVDM/AIVDO Protocol decoding. Retrieved from http://catb.org/gpsd/AIVDM.html.

Ristic B, La Scala B, Morelande M, Gordon N (2008) Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction. Information Fusion, 40–46. Retrieved from http://ieeexplore.ieee.org/ document/4632190/.

Rodrigue J (2017) The geography of transport systems. Fuel consumption by containership size and speed. Accessed February 21, 2018, https://people.hofstra.edu/geotrans/eng/ch8en/conc8en/fuel_consumption_containerships.html.

Tester KA (2013) A spatiotemporal clustering approach to maritime domain awareness. Master's thesis, Dept. of Electrical Engineering, Naval Postgraduate School, Monterey, CA. http://calhoun.nps.edu/handle/ 10945/37731.

Tetreault BJ (2005) Use of the Automatic Identification System (AIS) for maritime domain awareness (MDA). *Proceedings of MTS/IEEE*, doi: 10.1109/OCEANS.2005.1639983.

Vainio J Eriksson P (2018) The ice season 2013/2014 was mild. Finnish Meteorological Institute. Accessed February 21, 2018, http://en.ilmatieteenlaitos.fi/ice-winter-2013-2014.

Xu J, Wickramarathne T, Chawla N (2016) Representing higher-order dependencies in networks. *Science Advances* 2(5), http://doi.org/10.1126/sciadv.1600028.

Young BL (2017) Predicting vessel trajectories from AIS data using R. Master's thesis, Dept. of Operations Research, Naval Postgraduate School, Monterey, CA. https://calhoun.nps.edu/handle/10945/55564.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California