NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

CLASSIFYING VESSELS OPERATING IN THE SOUTH CHINA SEA BY ORIGIN WITH THE AUTOMATIC IDENTIFICATION SYSTEM

by

Kimberly M. Cull

March 2018

Thesis Advisor: Lyn R. Whitaker
Second Reader: Andrew Anglemyer

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 CLASSIFYING VESSELS OPERATING IN THE SOUTH CHINA SEA BY ORIGIN WITH THE AUTOMATIC IDENTIFICATION SYSTEM			**5. FUNDING NUMBERS**
6. AUTHOR(S) Kimberly M. Cull			
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)

This research focuses on building classification models with multinomial responses based upon seven months of Automatic Identification System (AIS) data gathered from the South China Sea. The models, built using Gradient Boosted Machines (GBM), assess the validity of utilizing AIS to confirm an operating vessel's origin, by country and geographical region. Two types of models are built. The first model captures the naturally dependent nature of AIS signals and serves as a proof of concept for how well a global model trained over many years could perform The second model attempts to reduce the dependency between AIS signals in order to characterize maritime patterns of behavior by country and region. With relative accuracy, both types of models are able to predict a vessel's origin and provide insight into maritime patterns of behavior.

14. SUBJECT TERMS data analysis, classification, Automatic Identification System, South China Sea, gradient boosted models			**15. NUMBER OF PAGES** 81
			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.

CLASSIFYING VESSELS OPERATING IN THE SOUTH CHINA SEA BY ORIGIN WITH THE AUTOMATIC IDENTIFICATION SYSTEM

Kimberly M. Cull
Lieutenant, United States Navy
B.S., United States Naval Academy, 2012

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: Lyn R. Whitaker
Thesis Advisor

Andrew Anglemyer
Second Reader

Patricia Jacobs
Chair, Department of Operations Analysis

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

This research focuses on building classification models with multinomial responses based upon seven months of Automatic Identification System (AIS) data gathered from the South China Sea. The models, built using Gradient Boosted Machines (GBM), assess the validity of utilizing AIS to confirm an operating vessel's origin, by country and geographical region. Two types of models are built. The first model captures the naturally dependent nature of AIS signals and serves as a proof of concept for how well a global model trained over many years could perform The second model attempts to reduce the dependency between AIS signals in order to characterize maritime patterns of behavior by country and region. With relative accuracy, both types of models are able to predict a vessel's origin and provide insight into maritime patterns of behavior.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AIS	Automatic Identification System
AIS-SART	Automatic Identification System search and rescue transmission
AtoN	aids-to-navigation
CPA	closest point of approach
COG	course over ground
GB	gigabyte
GBM	gradient boosted model
GLRM	generalized low rank model
GPS	Global Positioning System
ID	identification
IMO	International Maritime Organization
MMSI	Maritime Mobile Service Identity
SAR	search and rescue
SOG	speed over ground

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

This thesis classifies vessels operating in the South China Sea by country of origin and geographical region utilizing Automatic Identification System (AIS) data. Predicting vessel origin can aid in verifying a ship's claimed identity by revealing underlying patterns of operating behavior found within AIS transmissions. Exploiting a predictive capability can lead to several benefits, including assisting law enforcement and the U.S. Navy in securing maritime borders and ensuring freedom of the seas.

Classification models are built with gradient boosted models to determine predictive capability. Two types of models are built and tested on four responses adapted from each vessel's Maritime Mobile Service Identity number. These responses are country, geographical region, a response that bins every vessel into a "USA," "China" or "Other" category, and a final response that differentiates vessels originating from Mainland China from Taiwan and Hong Kong. The first set of models, also named the dependent models, attempt to exploit the lack of independence between AIS observations. Due to the nature of maritime traffic, ships operating worldwide are likely to continue to do so under the same country for long periods of time, presumably with similar distinguishing behavioral patterns. Theoretically, models trained with several years of data would reflect a similar issue with independence. Therefore, the first set of models, which trains and validates on data from the end of November 2013 through the end of March, 2014 and tests on data collected over 13 weeks from April to June 2014, serves as a proof of concept for how well a larger all-encompassing model can perform in predicting vessel origin if current computational limitations did not exist. The second set of models, also known as the independent models, explicitly attempt to reduce the dependency between observations used to train the models and the observations used to assess the accuracy of these models capturing operational behaviors specific to countries. Dependency is reduced in the independent models by ensuring that the training, validation, and test sets do not

include the same vessels operating in the South China Sea, but do include a sample of vessels from every geographic origin.

Overall, models built with responses composed of few regions perform at a prediction accuracy level of approximately 85 percent regardless of model type. For models built with responses categorizing observations by country and geographical region, dependent models out perform the corresponding independent models by nearly 20 percent. The first model type performs at 50 percent accuracy, while the second preforms around 35 percent. While those numbers may seem unimpressive, if the model is given the leeway to provide a series of top predictions, the prediction accuracy of the models can be improved. There are 58 different countries with greater than or equal to 10 unique vessels operating in the South China Sea and 10 geographical regions are represented in the data. Figure ES-1 illustrates the prediction accuracy of these responses for each model given a range from one to ten and one to three prediction opportunities for country and geographical regions, respectively.

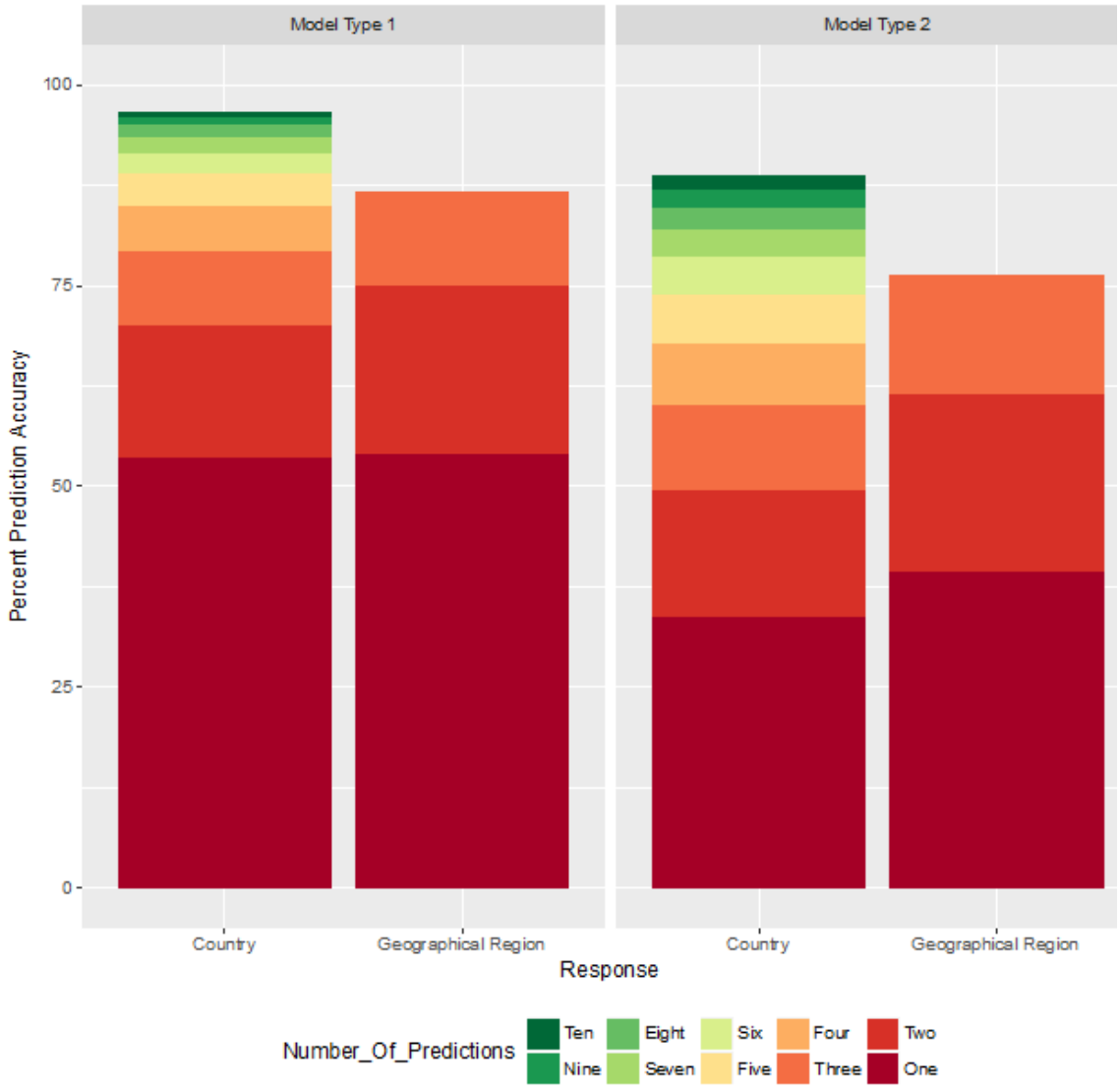


Figure ES-1. Hit Ratios for Country and Geographical Region Responses

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

Thank you Dr. Lyn Whitaker for all of the assistance and mentorship that you provided throughout this journey. Without you, none of this would be possible. A special thank you to my husband Danny, my family, and my friends who all supported me during my time as a graduate student.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Many companies and organizations are becoming increasingly interested in data science and its applications. From gaining insights to predicting most likely outcomes, the benefits of an organization's emphasis on data has seemingly unlimited possibilities. "Big data" is a colloquial term for massive amounts of stored information. It is often computationally difficult to prepare, process, and analyze. The advantage of its analysis, however, outweighs the challenges it poses to analysts. The United States' military has access to a vast amount of data. When properly analyzed, data has the potential to influence and improve operations domestically and abroad, providing the U.S. with an advantage over its adversaries.

One massive data set available to the military and significantly applicable to the U.S. Navy is the data generated by the Automatic Identification System (AIS). Originally created to mitigate the risk of collisions at sea, AIS has become an integral part of maritime operations. Since its implementation in 2002 by the International Maritime Organization (IMO), the installation and use of AIS aboard vessels has significantly increased and, in many cases, become a legal requirement for operation at sea (U.S. Department of Homeland Security, 2017). The purpose of AIS is to create an "autonomous, automatic, [and] continuous" (International Telecommunications Union [ITU], 2014, p. 5) means to track voyage and safety related information in real time. Over time, massive amounts of information related to maritime operations has been collected and stored thanks to the global employment of AIS.

This research builds classification models with a multinomial response based upon AIS data gathered from the South China Sea. The goal is to accurately predict a vessel's country of origin when country specific AIS data is missing, clearly in error, or as a method of confirming a ship's identification. Characterizing ship movement and behavior by country can help detect suspicious vessels, aid in maritime interdiction operations, and, potentially,

assist the U.S. in conducting clandestine operations of its own. Such research could provide insight into detection vulnerabilities and protect assets by learning to “hide in plain-sight.”

A. BACKGROUND

This section provides an overview of the Automatic Identification System. It also discusses the South China Sea, the primary focal region of this research.

1. Automatic Identification System (AIS)

AIS broadcasts two types of vessel data: static and dynamic. Each is separately transmitted at intervals technically specified by the International Telecommunications Union (ITU, 2014). Static transmissions occur every 6 minutes or when any portion of the data has been changed while AIS is activated (ITU, 2014). Dynamic data, on the other hand, is transmitted more frequently at intervals between 2 seconds and 3 minutes depending on the type of transponder installed and the relative speed of the vessel (ITU, 2014). The dynamic data requires more frequent transmissions because it captures a ship’s movement and includes information like location in latitudinal and longitudinal Global Positioning System (GPS) fixes, speed over ground, course over ground, heading, and the date and time associated. A vessel’s static profile includes less frequently changing information like the ship’s name, call sign, dimensions, ship type, intended destination, and the date and time of transmission. Both static and dynamic signals include a vessel’s Maritime Mobile Service Identity (MMSI) number, a unique identifier assigned to a vessel indicating its country of origin by the first three digits of the nine digit identifier. The MMSI, therefore, can be used to link a vessel’s static and dynamic data despite the differences in transmission intervals. Though the majority of AIS signals originate from vessels, the system also incorporates safety related messages from land based stations, navigation aids, and search and rescue (SAR) crews. All AIS messages are intended to encourage open lines of communication and increase situational awareness at sea.

2. South China Sea

One region of great economic and political interest to the United States is the South China Sea. Eight countries, China, Taiwan, the Philippines, Brunei, Indonesia, Malaysia, Singapore, and Vietnam, surround the almost one million square mile area of ocean (Google Maps, 2017). Kaplan (2011, P. 80) describes the South China Sea as the "throat of global sea routes." A majority of Asia's energy and crude oil supplies transit the South China Sea and its surrounding straits. In fact, according to Kaplan (2011, p. 80), more than one third of the world's traffic transits through the South China Sea and is responsible for "more than half the world's annual merchant fleet tonnage." The South China Sea also boasts a significant amount of natural resources, including natural gas and oil (Kaplan, 2011). Its significant economic and strategic value unsurprisingly makes the South China Sea an area plagued by territorial disputes; most famous, is China's claim to the Spratly Islands and its growing assertion of territorial waters. The dependency on the United States' diplomatic and military resources by the region's smaller nations and the U.S.'s desire to preserve freedom of the seas have led to increased U.S. military operations in 7th fleet.

Heavily trafficked sea-lanes in conjunction with political unrest create concerns of maritime safety and security. AIS can be used in this region not only in real time to reduce these concerns, but also by collecting and studying data in the region over time. Creating a means to classify behavioral patterns by country, ship type, or other attribute, can assist the U.S. in detecting anomalous AIS transmissions or vessel behavior.

B. ORGANIZATION OF THESIS

This thesis demonstrates the validity of using AIS data as a means to confirm a vessel's identity. It builds gradient boosted models (GBMs) to classify a ship's country of origin. The remainder of thesis is organized into four additional chapters. Chapter II provides a brief literature review of related AIS

research and analysis. Chapter III describes the data used to conduct this study, while Chapter IV details the analysis. Finally, Chapter V provides a conclusion and recommendations for future work.

II. LITERATURE REVIEW

With 90 percent of the world's trade by volume transported by sea (Pallotta, Vespe, & Bryan, 2013), it is not surprising that there is a growing interest in AIS data. In the last ten years, several studies have been conducted with many focusing on trajectory predictions and developing early warning collision systems in an effort to further prevent accidents at sea. Other researchers study elements of AIS beyond collision avoidance and vessel tracks. Specifically, some studies aim at identifying anomalous behavior at sea. Trajectory prediction and anomaly detection are both means of exploiting AIS to further secure maritime borders, national waterways, and international straits. Reducing accidents directly impacts maritime security by ensuring the fluidity of heavily trafficked waterways, safeguarding global commerce trade, and protecting ecosystems from bi-products of collisions, like oil spills. Other factors that threaten security are illicit activities, especially those that interrupt the flow of commerce or threaten a country's sovereignty. Anomaly detection can aid in identifying behavioral patterns associated with crime. With terrorism, piracy, and other criminal activities, like drug smuggling and human trafficking, on the rise, greater emphasis has been placed on interrupting these operations at sea.

A. VESSEL TRAJECTORY PREDICTION

AIS in conjunction with radar can significantly increase maritime safety by providing ships with real time data that can decrease a vessel's likelihood of a collision at sea. Research, like the Study on Collision Avoidance in Busy Waterways by using AIS Data (Mou, Tak, & Ligteringen, 2010), aims at developing algorithms to predict vessel trajectories and calculate closest point of approaches (CPAs). Many of these algorithmic approaches include clustering methods to identify patterns of movement in busy waterways followed by regression models (Bay, 2017), neural networks or random forests (Young, 2017) to predict vessel trajectory based on AIS dynamic data. Other research focuses

on the data science required prior to trajectory predictions. Such research emphasizes creating complete AIS databases by interpolation of missing track data, subsequently increasing the performance of learning algorithms (Mao, Tu, Zhang, Rachmawati, Rajabally, & Huang, 2016).

B. ANOMALY DETECTION

While trajectory prediction is an essential building block of AIS research, anomaly detection to identify abnormal behavior is the first step in targeting vessels engaged in unlawful operations. In Vessel Pattern Knowledge Discovery from AIS data: A Framework for Anomaly Detection and Route Prediction, Pallotta, Vespe, and Bryan (2013) use a rule-based system to help identify such anomalies. Rules for detection include attributes like “maximum speed allowed in port, presence in areas restricted to navigation, or inconsistencies between ship claimed and actual activity” (Pallotta, Vespe, & Bryan, 2013, p. 2220). Though a rule-based system can assist in detecting abnormal ship behavior, it is important to understand that not all anomalous behavior is attributed to illegal activity. One challenging aspect of working with AIS data is its inherent susceptibility to error. Error is most commonly caused by system malfunctions and human error. System malfunctions can be seen in AIS data when a vessel’s transmission includes an unidentifiable, but common MMSI number. MMSI numbers that are not unique to one vessel, are not 9 digits in length, or do not begin with a number between 2 and 7 are considered illegitimate identifiers. These common MMSI errors can sometimes be attributed to an incorrectly installed AIS transponder causing the equipment’s default MMSI to be displayed (Harati-Mokhtari, Wall, Brooks, & Wang, 2008). It is also possible for system malfunctions to cause errors in location data. Other inaccuracies can be attributed to human error, like typos or similar mistakes made when entering voyage information to include outdated reporting.

The most troubling issues caused by human factors, however, are those that are intentional. These anomalies can signal illicit activity and it is important to

distinguish these abnormalities from those rooted in human error or system malfunctions. Like any technology, AIS is susceptible to spoofing, hijacking, availability disruption (Balduzzi, Pasta, & Wilhoit, 2014), and deception tactics. Spoofing can include creating non-existent ships, fake or inaccurate aids-to-navigation (AtoN), and phony search and rescue transmissions (AIS-SART). All of these practices can encourage targeted vessels to maneuver a certain way or be lured to a particular location (Balduzzi, Pasta, & Wilhoit, 2014). Similarly, false weather messages can be transmitted to achieve similar means (Balduzzi, Pasta, & Wilhoit, 2014). Hijacking is another threat posed to vessel using AIS. This can occur when a false message is broadcasted to over-ride or change any information related to a vessel, AtoN, or port authority (Balduzzi, Pasta, & Wilhoit, 2014). Availability disruption can deceive vessels by delaying or disabling AIS transmissions, or tricking a vessel into changing transmission frequencies to render its AIS inoperable (Balduzzi, Pasta, & Wilhoit, 2014). Not everyone wishing to conduct prohibited operations at sea needs to be computer savvy, however. Deception can be as simple as purposefully manipulating ones own transmission to reflect another ship's information, or turning off AIS transmissions entirely.

THIS PAGE INTENTIONALLY LEFT BLANK

III. DATA

This chapter begins by describing the merging and cleaning of static and dynamic AIS records into a useable format for predictive analysis. This chapter also contains some initial exploratory analysis. The data used for this research consists of raw static and dynamic AIS signals collected from the South China Sea from the period of November 30, 2013, to June 30, 2014. The area chosen is between 105 and 121 degrees longitude and 2 to 23 degree latitude as shown in Figure 1.

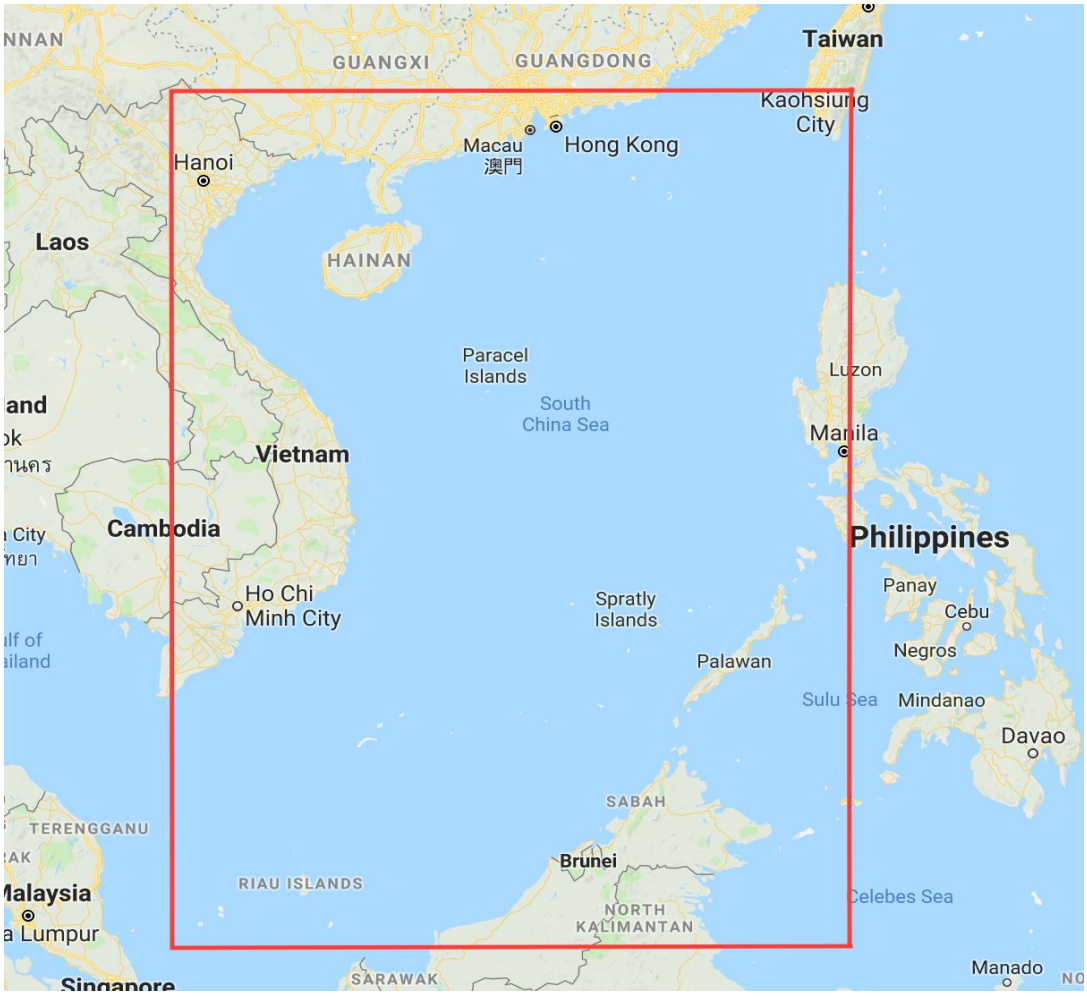
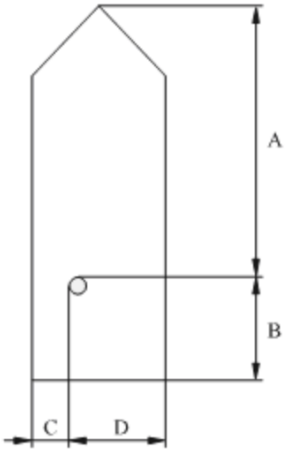


Figure 1. Geographic Map of Regional Area of Interest. Adapted from Google Maps (2017).

The combined size of the static and dynamic data prior to cleaning is approximately 24 gigabytes (GBs) with 264,530,120 AIS records. To overcome some of the computational challenges of storing, cleaning, and analyzing large amounts of data, this research is conducted through the programming language R (R Core Team, 2017) on a super computer with 4270 cores and more than 18 terabytes of memory (Naval Postgraduate School, 2018). The specific R packages used in this thesis are chosen based upon their capacity to work with large amounts of data.

A. PREPARATION AND CLEANING

The static and dynamic data needs to be combined into one useable data set and then cleaned for predictive analysis. The dynamic data is considered the primary data set, as it can be filtered by both location and time, and is the basis for constructing a combined static and dynamic South China Sea data set. Static records cannot be pulled for a specific location. Instead, these records are selected to have the same date and time range as the South China Sea dynamic data and, then, the static data is further reduced by removing duplicate static records and records that contain MMSIs not found within the South China Sea dynamic data's unique set of MMSIs. Dynamic records include columns for MMSI, latitudinal and longitudinal coordinates, recorded speed and course over ground, ship identification (ID), and time. The static data also includes MMSI, time, ID, as well as, ship type, destination, name, call sign, and four columns describing ship dimensions. The dimension columns, A, B, C, and D, are distances measured from a reference point on the vessel and are defined in Figure 2.



	Number of bits	Bit fields	Distance (m)
A	9	Bit 21-Bit 29	0-511 511 = 511 m or greater
B	9	Bit 12-Bit 20	0-511 511 = 511 m or greater
C	6	Bit 6-Bit 11	0-63; 63 = 63 m or greater
D	6	Bit 0-Bit 5	0-63; 63 = 63 m or greater

The dimension A should be in the direction of the transmitted heading information (bow)
Reference point of reported position not available, but dimensions of ship are available: A = C = 0 and B ≠ 0 and D ≠ 0.
Neither reference point of reported position nor dimensions of ship available; A = B = C = D = 0 (= default).
For use in the message table, A = most significant field, D = least significant field.

Figure 2. Reference Point for Reported Position and Overall Dimensions of Ship. Source: ITU (2014).

Dynamic and static records that do not include time are removed. The dynamic data set includes several records that do not contain valid MMSI identifiers. An invalid MMSI is one that does not meet the required 9-digit length and/or does not contain a recognizable country identifier. Using the first three digits of the each MMSI, a new column is constructed identifying the country associated with each dynamic entry. Records that are assigned “NA” (i.e., those with an invalid MMSI) in this column are removed from further analysis along with any MMSI that does not meet the 9-digit length requirement. Some countries represented by the data, have fewer than ten unique MMSIs operating within the South China Sea region over the seven-month period the data captures. Those records, approximately 10,000,000, are removed from the data and not included in modeling. This leaves only those records for MMSIs that can be identified as belonging to one of 58 countries. Additionally, some MMSIs are not vessels and are identified by ship type as navigation aids or SAR aircraft. These records are also removed from further analysis.

To reduce the number of records further, a sample of the dynamic data is taken to create a smaller, more manageable data set for cleaning and analysis.

Two percent of the dynamic signals from every MMSI are randomly sampled to capture the scope of vessels operating in the South China Sea. This sample of dynamic signals, approximately 4,102,000 records, is merged by MMSI and matched by nearest time with the static records, so that each dynamic signal is augmented with columns of static information. Only the dynamic data's time stamp is retained for further analysis. The resulting "sample" data set, when compressed using the R package feather (Wickham, 2016), is 425 MBs and its approximately 4,102,000 records serve as the observations for the analysis that follows.

The sample data must be cleaned and formatted prior to conducting analysis. Categorical variables in the sample data are coerced into R objects called factors (R Core Team, 2017). The destination column is also coerced into a factor after removing extra white space, digits, and special characters from strings. Entries with missing values or with string lengths less than or equal to one are assigned "UNKNOWN." The destination column contains text entries that are manually entered by the crew of each vessel; therefore, it has many unique entries, corresponding to the destination factor. This particular sample of data contains 53,983 unique inputs of destinations. Two vessels could be traveling to the same destination and have entered the information differently. For example, if two vessels are en-route to Gladstone, Australia one may enter "GLADSTONE" and the other "GLADSTONEAUS." In an effort to reduce the number of levels and still preserve information, text analysis is conducted on the destinations. Using the Jaro similarity distance metric (van der Loo, 2014), destination levels are grouped by their text similarity. Every entry associated with a group is assigned the group's longest character string. The new number of destination levels is significantly reduced to 9,420. Ship type is given by a numeric code used to identify ships as belonging to 25 broad ship types. This column is transformed into a factor and renamed Category. Time is reformatted into two columns, date and time. The date column is further transformed to reflect the day of the week each signal is transmitted, while the time column groups every signal

by the appropriate hour of a 24-hour day. Both columns are treated as factors. MMSI, name, and call sign information is excluded from analysis because of its connection to a vessel's country of origin. Table 1 summarizes the signal data variables used in analysis and their associated class.

Table 1. Predictor Variables

Predictor	**Class (Type)**
Week Days	Factor
Time (24-Hour)	Factor
Latitude	Numerical
Longitude	Numerical
Speed Over Ground (SOG)	Numerical
Course Over Ground (COG)	Numerical
Ship Type (Category)	Factor
Destination	Factor
Dimension A	Numerical
Dimension B	Numerical
Dimension C	Numerical
Dimension D	Numerical

Four response variables are constructed for exploratory analysis and modeling. The "Country" column serves as the primary response variable and contains 58 levels. An additional column, named "GeoRegion," identifies each country by geographical region based upon United Nations' world regions categorizations (Internet World Stats, 2017). Another column, "Big3," is adapted from the Country column to identify each record as one of three levels: "USA," "China," or "Other." A similar response column, called "SplitOneChina," further differentiates Mainland China from Taiwan and Hong Kong. The data's response columns are found in Table 2.

Table 2. Response Variables

Response	Class (Type)
Country	Factor
GeoRegion	Factor
Big3	Factor
SplitOneChina	Factor

B. INITIAL EXPLORATION

Initial exploration, modeling, and analysis is conducted on the sample data set. Initial exploration of the data includes computation of simple descriptive statistics. The number of vessels appearing monthly in the South China Sea is graphed in Figure 3 and belongs to one of 58 countries included in the sample data. Since 2 percent of signals were sampled for each vessel, Figure 3 approximates the number of vessels from the 58 countries represented within the sample that are operating in every month from December 2013 to June 2014 in the South China Sea.

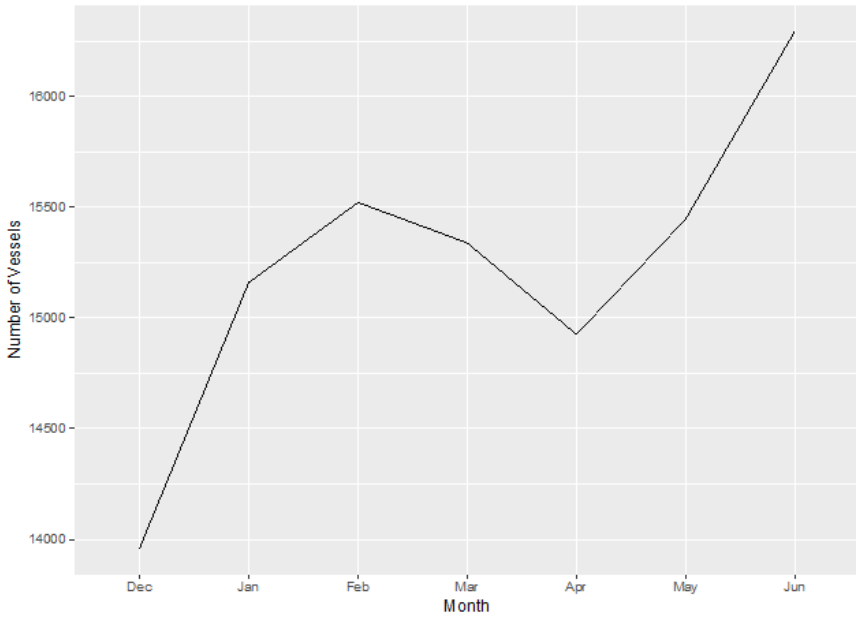


Figure 3. Number of Vessels by Month from the Sample Data in the South China Sea from December 2013 to June 2014

The majority of the vessels appearing within the sample data set have AIS records in every month of the seven-month time span. The turnover of vessels within the region is relatively low. No more than 17 percent of the vessels are absent for any given month. June, having the greatest number of operating vessels, is only missing 3 percent of vessels operating at some point from December to May.

Pareto Charts are constructed for each response variable. Figure 4 gives the distribution of vessels by countries operating in the South China Sea, Figure 5 provides the vessel distribution by geographical region, Figure 6 and Figure 7 gives the vessel distribution by the Big3 response and SplitOneChina response, respectively. Most ships identify with an Asian or Central American country. The United States has relatively few ships operating in the South China Sea, approximately 60, making up less than one percent of the data.

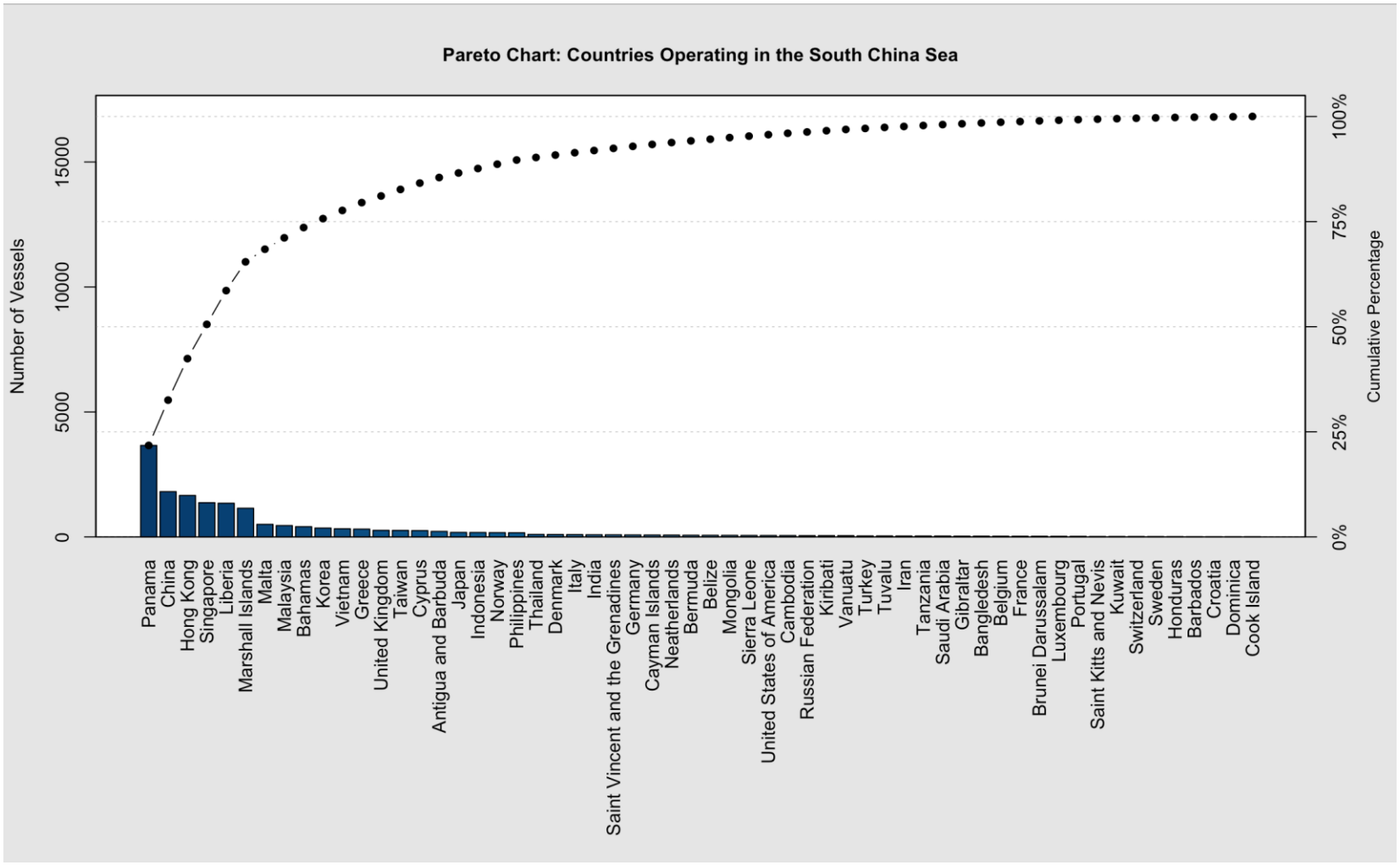


Figure 4. Number of Vessels by Country Operating in the South China Sea from December 2013 to June 2014

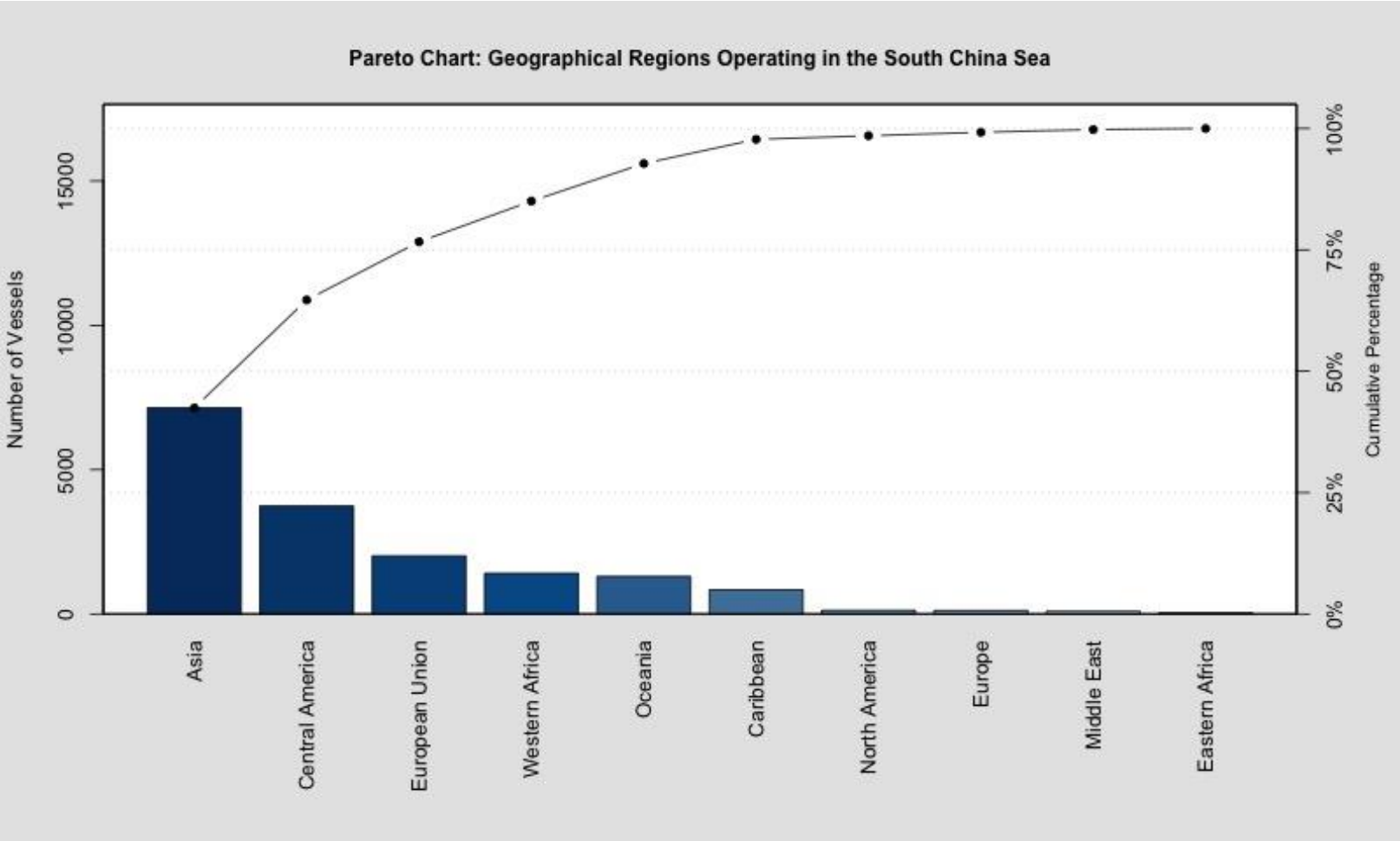


Figure 5. Number of Vessels by Geographical Region Operating in the South China Sea from December 2013 to June 2014

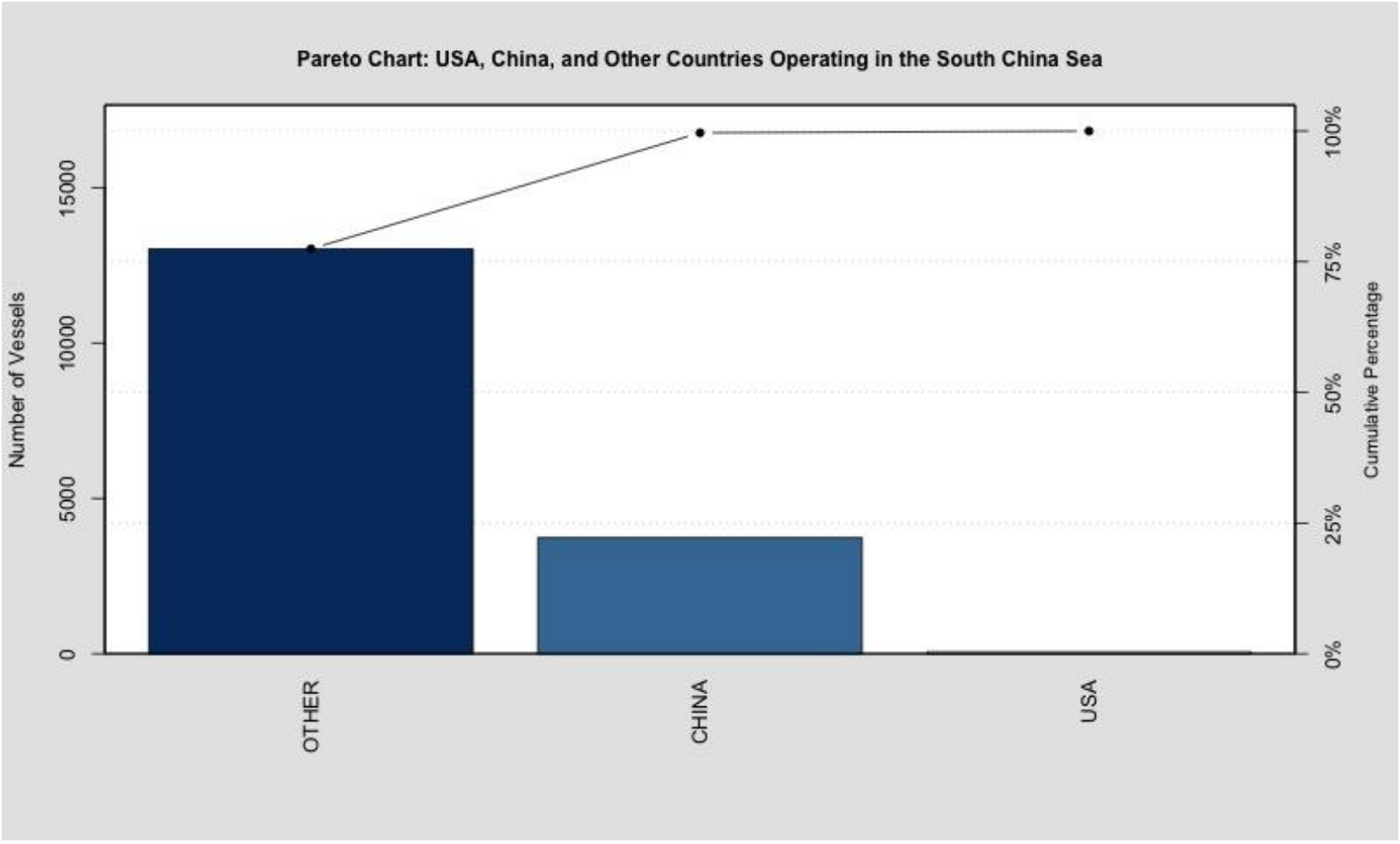


Figure 6. Number of Vessels from USA, China, and Other Countries Operating in the South China Sea from December 2013 to June 2014

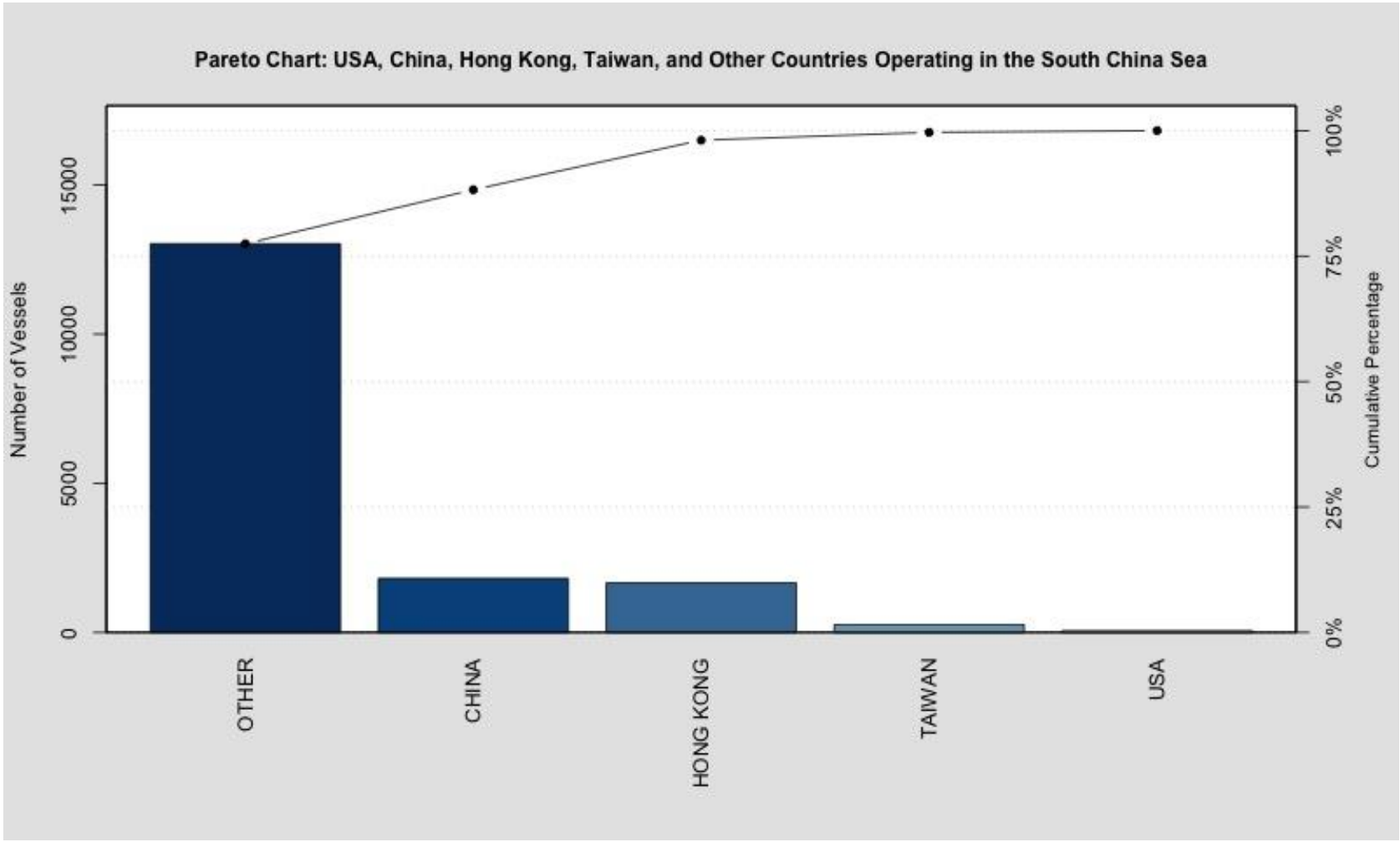


Figure 7. Number of Vessels from USA, China, Hong Kong, Taiwan, and Other Countries Operating in the South China Sea from December 2013 to June 2014

THIS PAGE INTENTIONALLY LEFT BLANK

IV. MODELING AND ANALYSIS

Two separate types of models are built using GBMs from the xgboost R package (Chen, He, Benesty, Khotilovich, & Tang, 2018). We briefly describe GBMs and the two types trained for analysis. We choose GBMs because they perform remarkably well. While training GBMs may be computationally intensive, they are relatively straightforward and scale to very large data sets. GBMs can easily accommodate data containing both numeric and categorical variables and automatically account for interactions. They are also invariant to choice of scale for numerical variables. Furthermore, although the sample data contains no missing values, once built, a GBM is capable of predicting on records with missing values. Hastie, Tibshirani, and Friedman (2016) provide a discussion of GBMs and their training.

The xgboost package uses a stochastic gradient descent algorithm (Friedman, Hastie, & Tibshirani, 2000) where gradients are approximated by (classification) trees and where each tree is fit (or trained) using a random sample of the observations and a random sample of the predictor variables. The resulting model is a linear function of trees. These GBM models have a number of hyper-parameters that must be chosen. The most important are: the learning rate for gradient descent, the number of iterations for the gradient descent algorithm, and the depth of the trees. The other hyper-parameters are the numbers of observations and predictor variables used to grow trees.

As with all such algorithmic models, the training accuracy of a trained GBM is usually greater than its prediction accuracy, i.e., the ability of a GBM to predict or correctly classify the data used to train the GBM is greater than its ability to predict with a new “future” set of data. See Hastie, Tibshirani, and Friedman (2016) for a discussion of training and prediction accuracy. The goal when training is to avoid choices of hyper-parameters, such as tree depth, that over-fit where the training accuracy is much greater than the prediction accuracy. Even when a trained GBM is not over-fit, the training accuracy usually gives an

overly optimistic estimate of prediction accuracy. An approach to guard against over-fitting and to assess the accuracy of the final model is to partition a data set into a training set, used to train models, a validation set, used to choose model hyper-parameters, and a test set, used to simulate a future data set in order to assess the final model.

AIS records are highly dependent because each vessel is associated with several rows of signal transmissions. The first model variety, called the "dependent model," attempts to leverage this dependency. Vessels operating in a specific region, or even globally, have a relatively small turnover over time, therefore, a model trained on all data encompassed over a period of time cannot be truly independent from a test set comprised of future AIS data. The dependent models are trained on all signals occurring between November 30th and February 28th, validated on those appearing in the month of March, and tested on the remaining data appearing through June 30th one week at a time. The first type of model provides a proof of concept for the development of a larger scale model trained globally over many years worth of data. It also serves to simulate the effect of a quickly deployable vessel origin confirmation system developed for a specific region.

A second model variation, referred to as the "independent model" for simplicity's sake, is built to reduce the dependency between the trained model and the test and validation sets. This model type trains on a random sample of 60 percent of every country's unique vessels over the span of the entire data set. Another 15 percent is taken to validate the model with the remaining vessels appearing in the test set. This model type is intentional in that it allows no vessel contained in the training or validation sets to also appear in the test set. Independent models serve to evaluate whether or not countries exhibit patterns of operating behavior in the South China Sea.

Several R packages are available to build the GBMs in this analysis. The GBM function from the package h2o (The H2O.ai team, 2017) is not sufficient to process the large amounts of data, required in this thesis. Instead, xgboost is

chosen for its ability to work with big data and parallelize computations. It uses a fast, efficient algorithm to build gradient boosted models. Part of its efficiency is derived from the explicit formatting requirements for data passed into its functions. All data used in xgboost is required to be numerical (Chen, He, Benesty, Khotilovich, & Tang, 2018). Since AIS data contains many factors, factor variables are one-hot encoded and stored as a sparse data set. Response variables are also converted into numerical representation. Each level of the response factor is assigned a unique number beginning with 0. This is performed on all training, validation, and test sets.

A. DEPENDENT MODELING

The first model type is trained with a learning rate of 0.1 and a maximum tree depth of 3. It uses 50 percent random row sampling and 80 percent column sampling when building each tree to prevent over-fitting. Models developed with categorical responses that have more levels require more computational time. This is particularly true of the model built with the Country response. Due to time constraints and computational resources, the Country response model is trained for 38,000 iterations. This means that this particular model could improve and reach optimality if allowed a longer time to train. Three other response models, GeoRegion, Big3, and SplitOneChina, are run for 80,000 iterations or until the stopping condition of 20 iterations with no improvement to the validation classification accuracy is met al.l three models stop before reaching 80,000 iterations. Table 3 gives the number of iterations per response type for the first model type.

Table 3. Dependent Modeling: Number of Iterations by Response

Response	Number of Iterations
Country	38000
GeoRegion	76993
Big3	15201
SplitOneChina	15038

The xgboost model output includes a measure of the relative importance of each predictor in classifying vessels. A percentage of influence is also calculated for each predictor variable with all predictor influences summing to 100 percent. Predictors with the highest relative influence scores are considered the most important factors in determining a vessel's classification. Each of the four response models determines a similar relative influence for predictor variables varying slightly in actual percentages of relative influence. Figure 8 provides the average percentage influence of predictor variables between the four models.

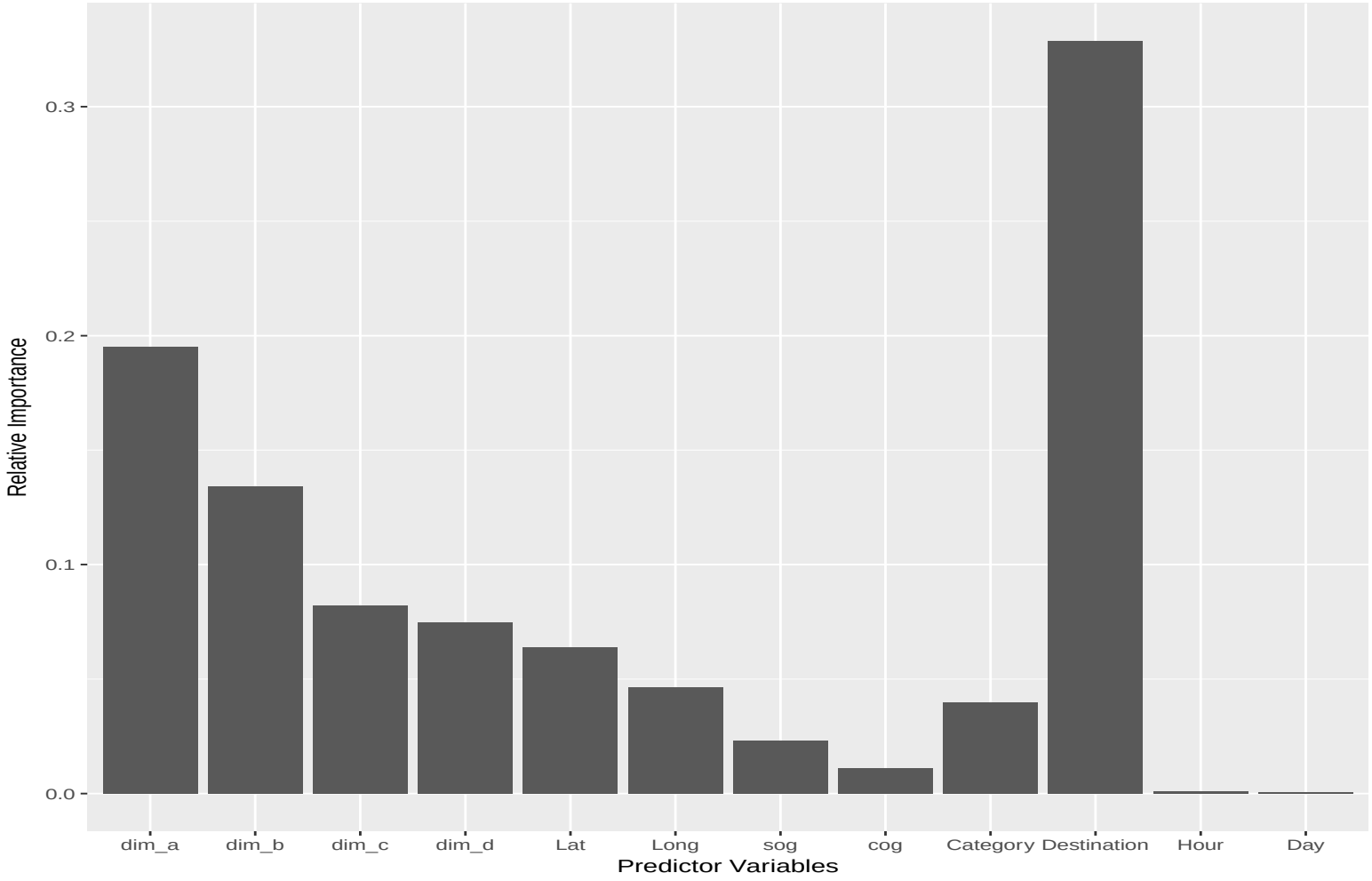


Figure 8. Average Predictor Variable Relative Influence for Dependent Models

Destination is the most important variable in determining vessel origin, followed by the dimensions of the vessel. The least important variables are the dynamic variables associated with location, speed, course, and time. This means that the most relevant information in determining a vessel's origin based on one dynamic transmission combined with a static transmission relies heavily on its final destination and its structure in relation to the location of its AIS transmitter.

The dependent model's responses are predicted over a 13-week period of time following the four full months of data used to train and validate the model. Figure 9 shows the performance of each response variable over the 13 weeks.

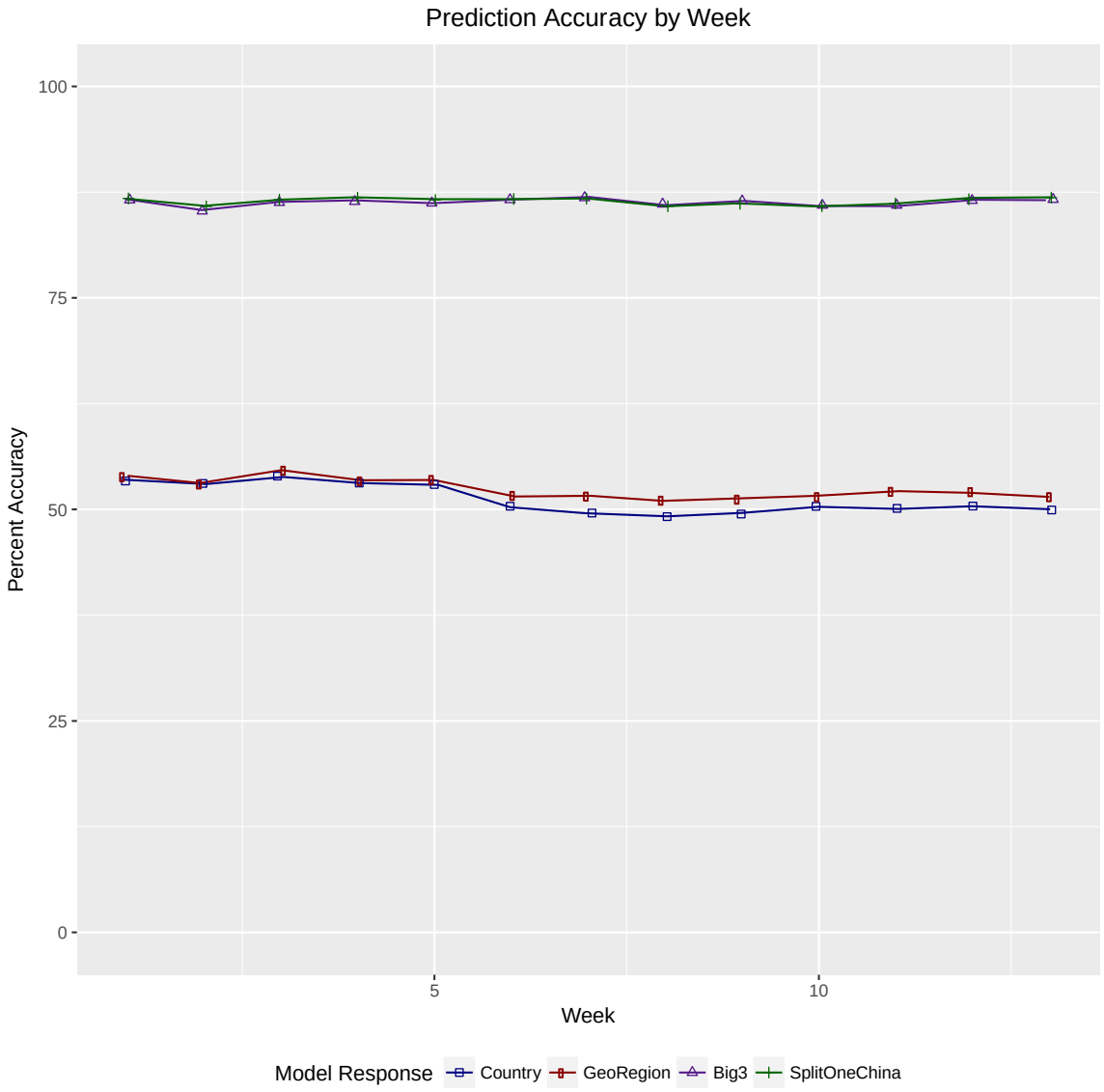


Figure 9. Dependent Model Prediction Accuracy over 13 Weeks

The models built with responses that have fewer classes perform better over the 13 weeks than the models built with more complex responses. The Big3 and SplitOneChina models are able to consistently predict vessel origin with around 86 percent accuracy while the Country and GeoRegion models accurately predict origin around 53 percent of the time. Counterintuitive to what might be expected, the prediction accuracy for the dependent models remains fairly constant over time with the Big3 and SplitOneChina models demonstrating

the most consistency. The models built with the Country and GeoRegion responses show only a slight degradation in prediction accuracy over the 13 weeks. This is due in part to the low turnover of vessels operating in the South China Sea. It suggests that, once deployed, model prediction accuracy will not degrade much over time. This is particularly important for these models, which cannot be trained quickly. Each GBM model requires on the order of days to weeks of computational time to train using the Naval Postgraduate School's super computer.

The output of a GBM model with a categorical response gives a score for each possible level of the categorical response. The classification with the highest score is the model's prediction for the observation. Hit ratios are calculations of prediction accuracy when more than the highest scoring level is considered. Since the Country and GeoRegion models predict correctly only approximately 50 percent of the time, a hit ratio for the models is analyzed to determine how well the models can perform when allowed to predict more than once. Figure 10 displays the Country model's performance over 13 weeks when the correct class is among m highest scores where m=1,2,…,10.

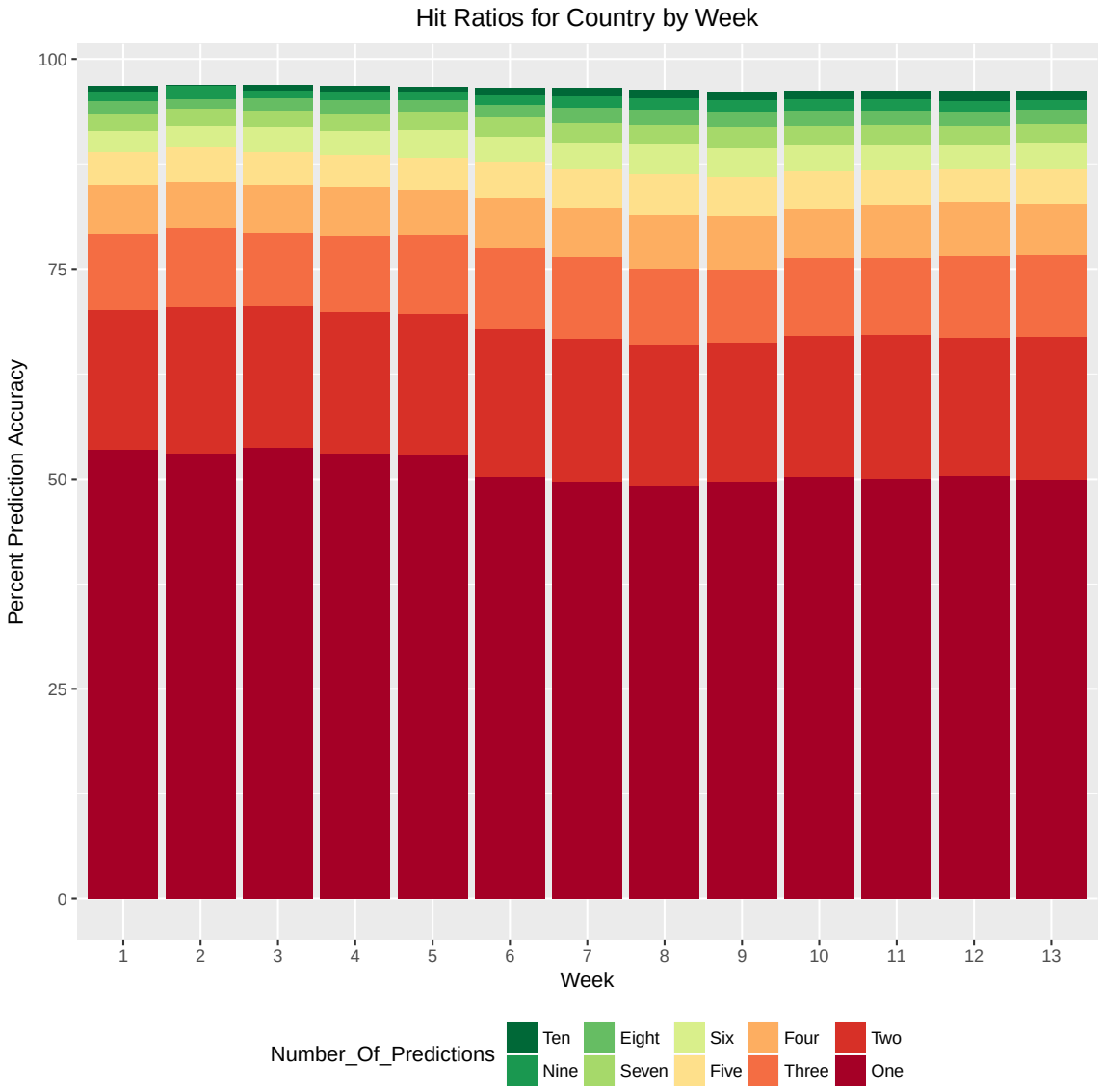


Figure 10. Prediction Accuracy of Country Dependent Model When given Ten Opportunities to Predict Vessel Origin

The Country model is able to predict at nearly 97 percent accuracy within ten predictions. In other words, one of the top ten highest scoring categorical levels the GBM returns contains the correct country 97 percent of the time. Given only one additional prediction, or within the top two highest scoring levels, the model's accuracy improves by almost 20 percent. Given three total predictions, or within the top three highest scoring levels, the model will have correctly determined the vessel's country of origin 80 percent of the time.

Similarly, Figure 11 demonstrates how well the GeoRegion model's accuracy can improve if the three geographical regions with the highest scores are considered.

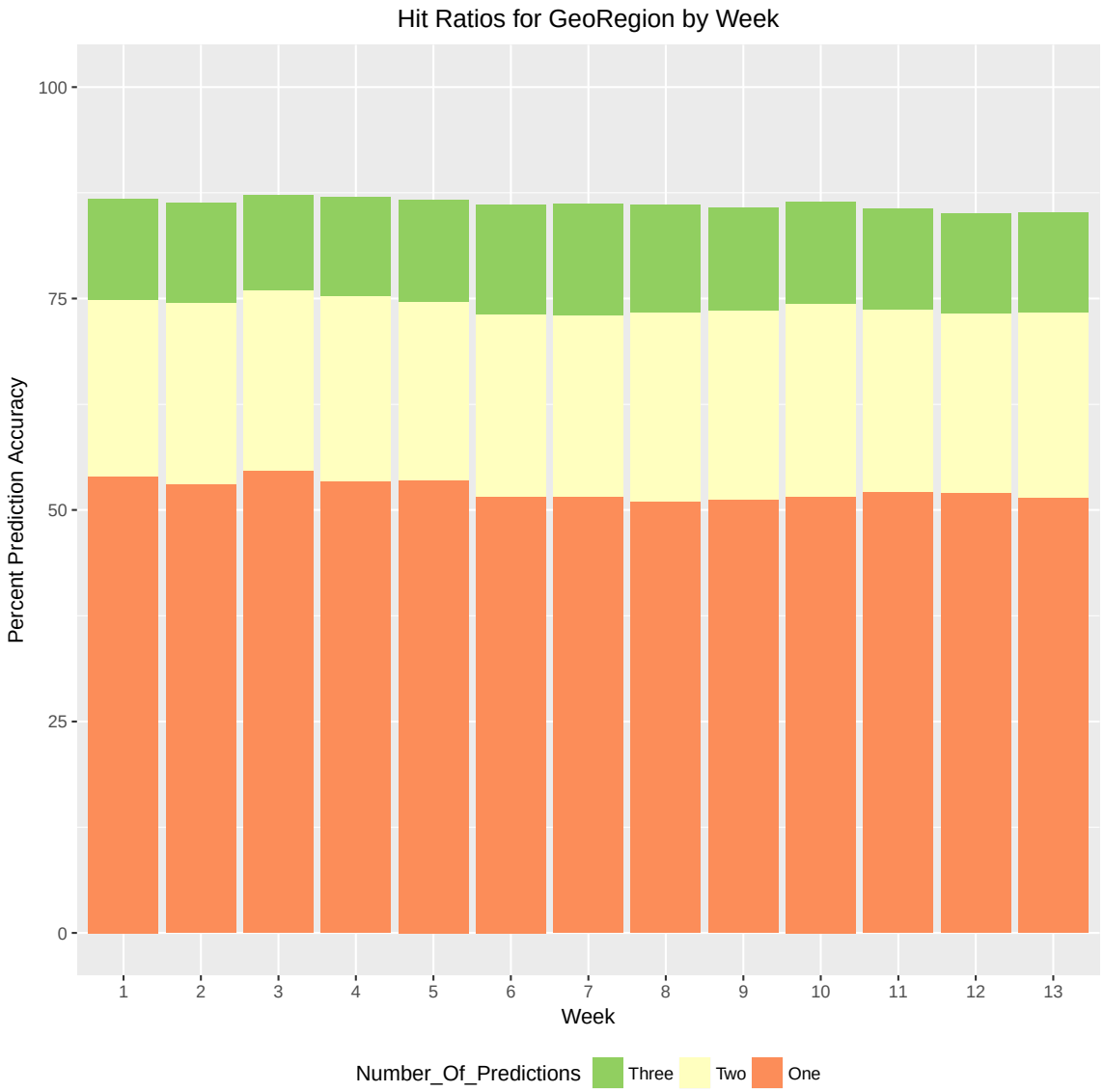


Figure 11. Prediction Accuracy of GeoRegion Dependent Model when given Three Opportunities to Predict Vessel Origin

Given the top three most likely geographical regions, the GeoRegion model is able to accurately predict a vessel's geographical origin with 85 percent accuracy.

Some countries are easier to identify than others. Figure 12 shows the 12 countries most easily predicted by the Country response model. All of these countries are predicted correctly the first time at least 90 percent of the time. Considering the overall prediction accuracy for this model is only around 50 percent, it may be concerning that the United States finds itself on this list.

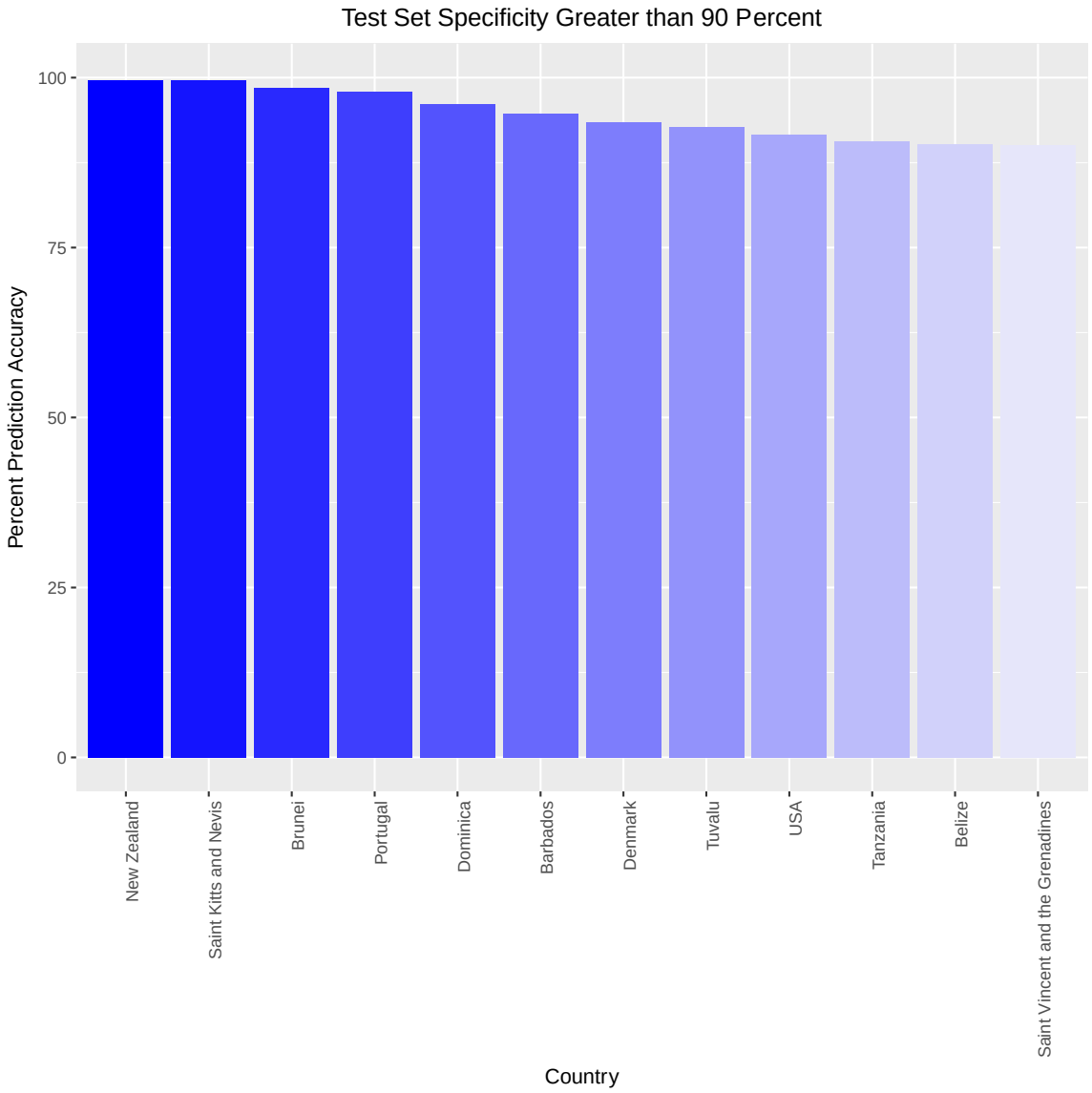


Figure 12. Countries Correctly Predicted in the Test Set over 90 Percent of the Time

There are some countries that are predicted accurately in the test set fewer than half of the time given the model's first prediction. Figure 13 graphs these seven countries.

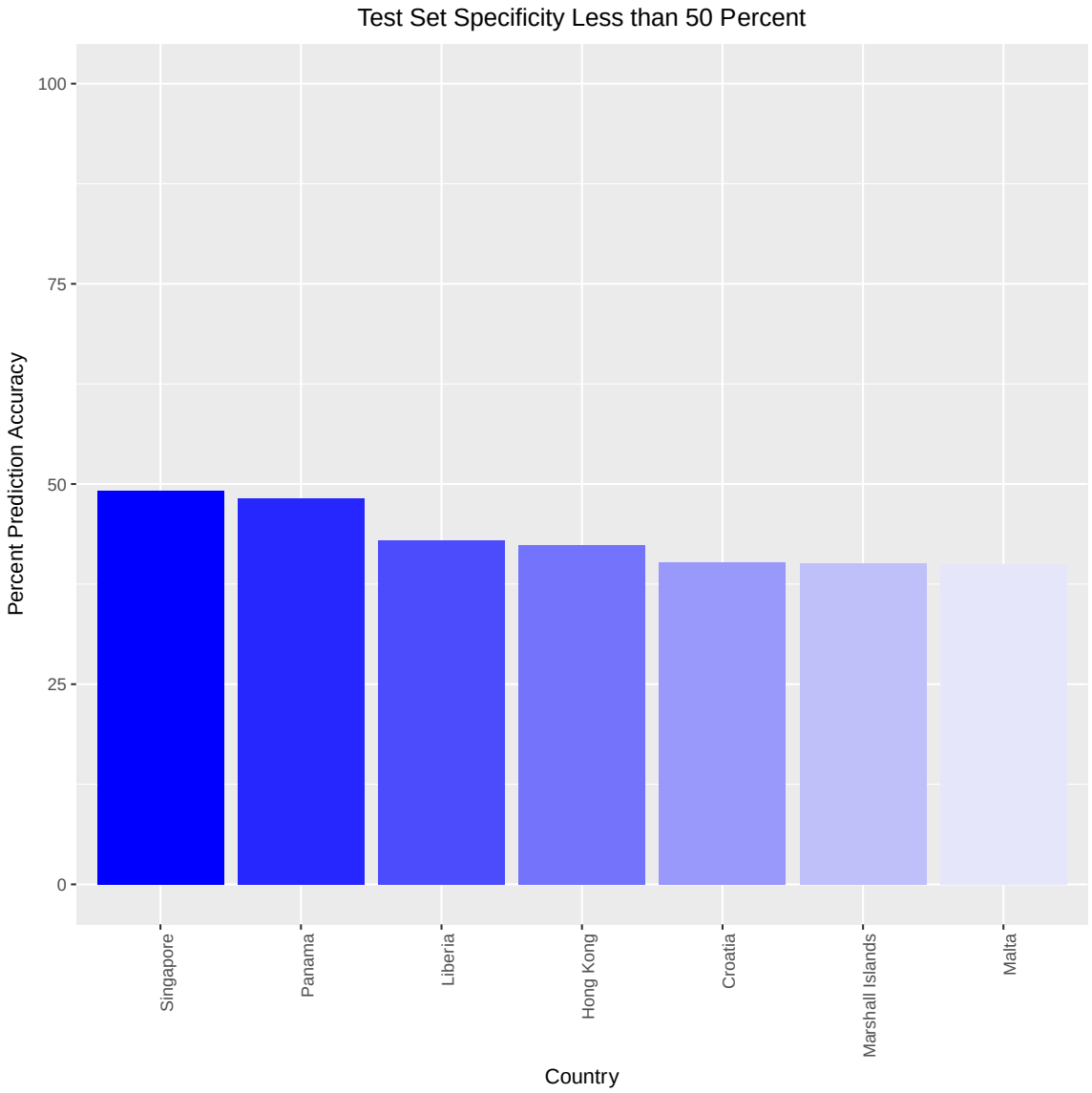


Figure 13. Countries Correctly Predicted in the Test Set Less than 50 Percent of the Time

Referring back to the Pareto Chart in Figure 4, six of the seven most difficult countries to identify are in the top seven most prevalent countries with vessels operating in the South China Sea. Figure 14 serves to better visualize the relationship between vessel frequency and the prediction accuracy of the model given only the first prediction.

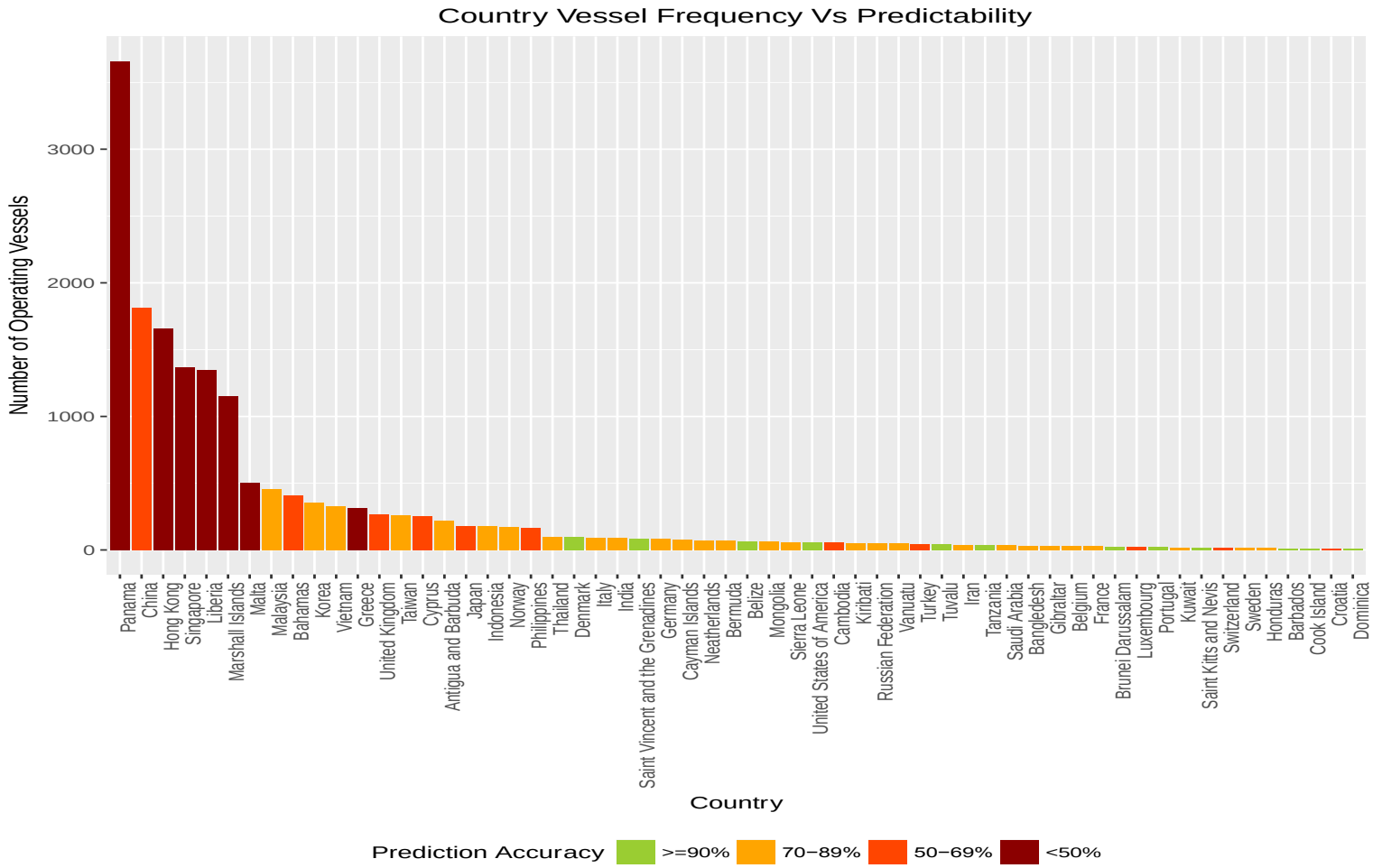


Figure 14. Dependent Model: Relationship between a Country's Number of Operating Vessels and Prediction Accuracy

Figure 14 illustrates a relationship between the number of operating vessels and the model's prediction accuracy. The more operating vessels a country has in the region the less accurately the model is able to predict vessels belonging to that nation. Mainland China, which ranks second in regional prevalence, is neither one of the easiest nor most difficult countries to predict. China is identifiable approximately 57 percent of the time, compared to the United States, whose few vessels operating in the region are recognized 91 percent of the time. Russia, who has slightly fewer vessels operating in the region than the United Sates, is accurately predicted just below 84 percent of the time.

The Big3 and SplitOneChina responses allow further exploration into U.S. and Chinese vessels operating in the region. Since both of these responses contain few categorization levels, only the highest scoring level is considered when measuring prediction accuracy. Figure 15 visualizes the ability of the Big3 model to differentiate Chinese and U.S. vessels apart from others operating in the South China Sea.

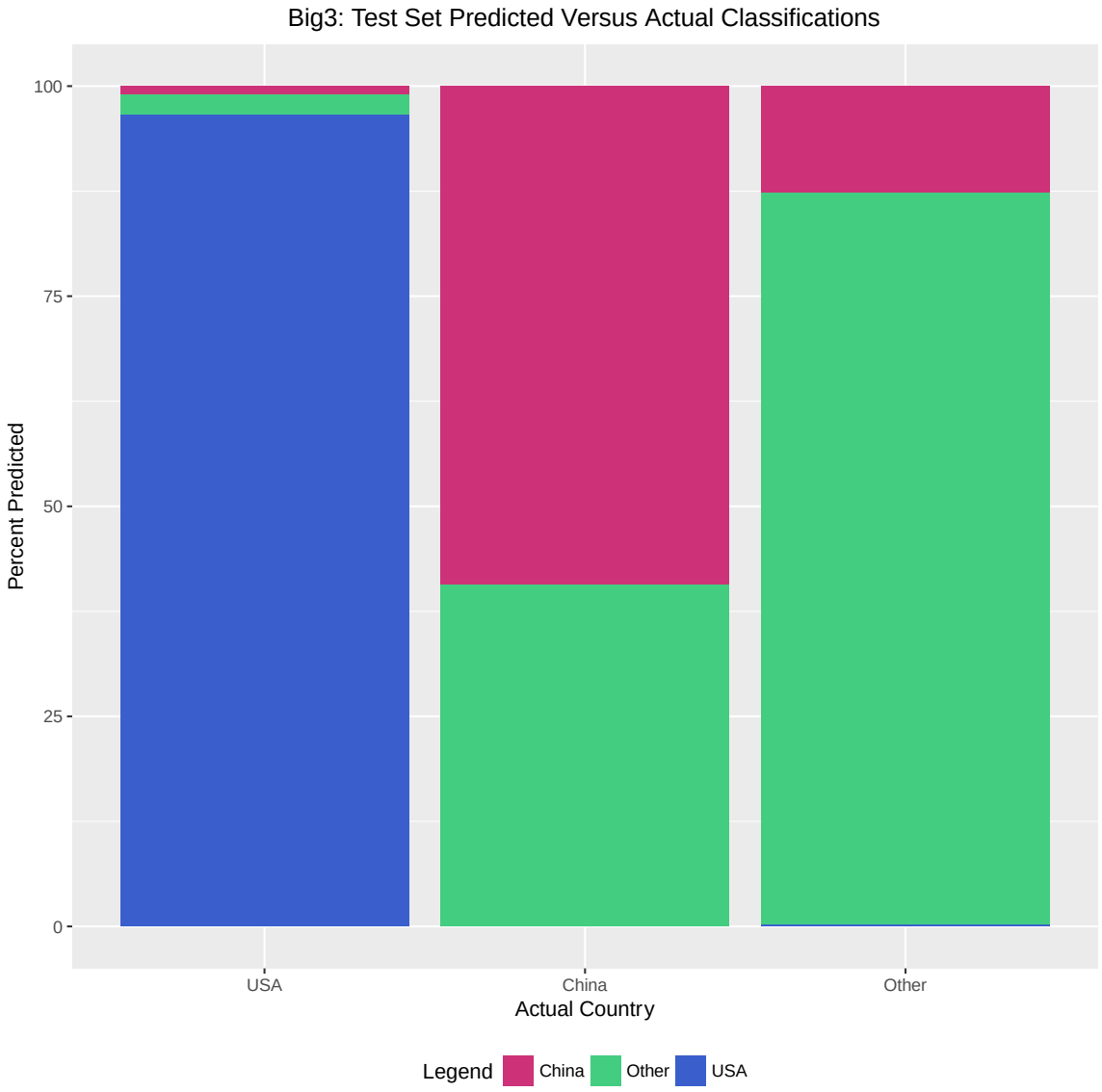


Figure 15. Dependent Model: Test Set Predicted Versus Actual Classifications for Big3

Similar to the Country model, U.S. vessels are accurately predicted greater than 90 percent of the time, approximately 97 percent of the time for the Big3 model. They are seldom confused for Chinese or Other vessels. Chinese vessels, however, are identified as Other nearly 50 percent of the time. Both Chinese and Other ships are rarely confused for those belonging to the United States, both with fewer than one percent predicted as U.S. vessels. Other ships

have a prediction accuracy of 89 percent, perhaps because majority of ships in the data set belong to this category.

Confusion for vessels grouped as belonging to China could be attributed to different operating behaviors between the three distinct regions comprised by the One China grouping. Figure 16 visualizes the response SplitOneChina in a similar manner to Figure 15. The category China is broken into Mainland China, Hong Kong, and Taiwan.

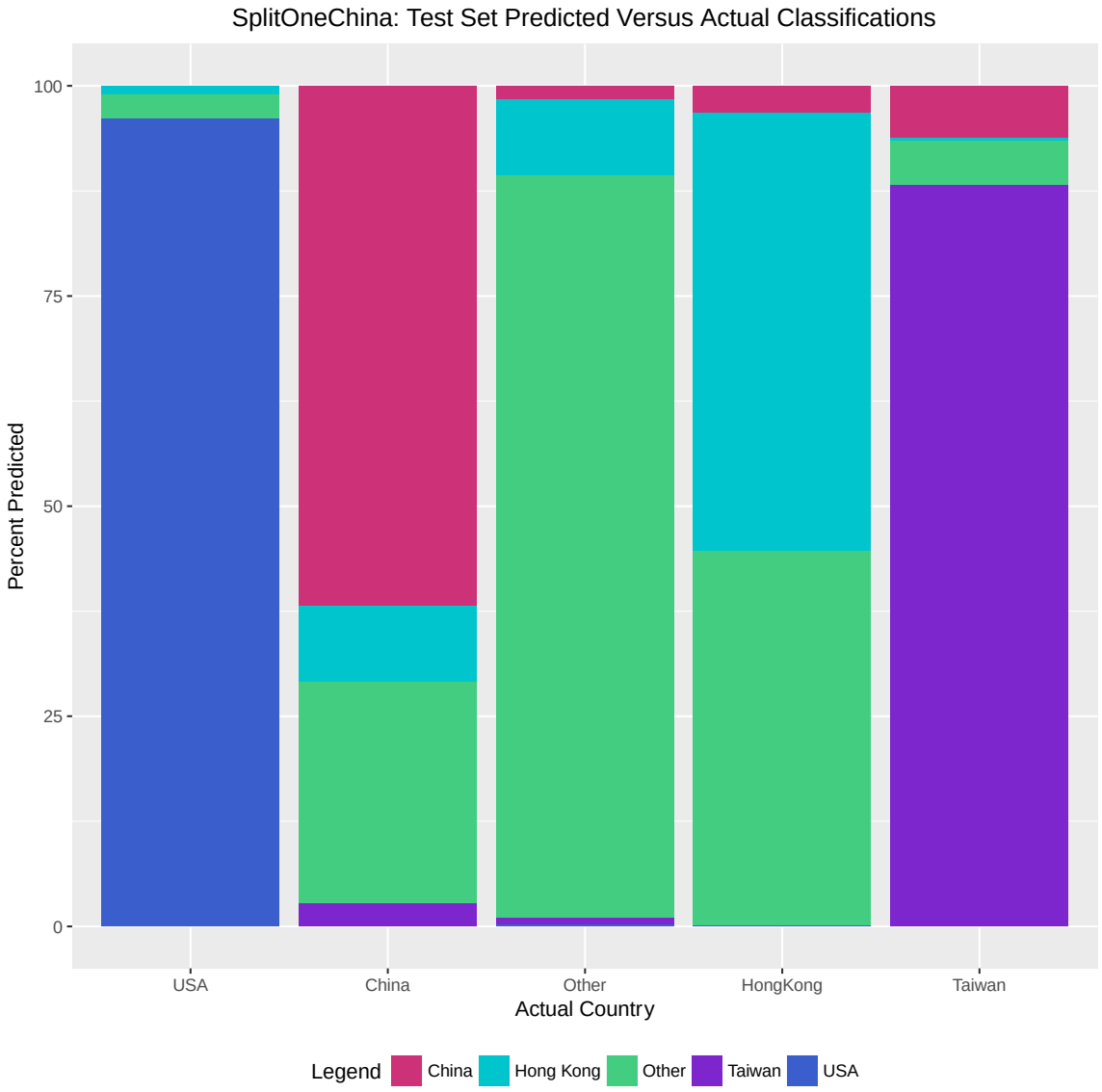


Figure 16. Dependent Model: Test Set Predicted Versus Actual Classifications for SplitOneChina

Separating Mainland China from Taiwan and Hong Kong shows marginal improvement with accuracy increasing from 59 percent to 62 percent. Mainland China is still confused with Other, Taiwan, and Hong Kong. Hong Kong vessels are difficult to distinguish from Other, but are rarely classified as originating from mainland China. Taiwan, however, is much easier for the model to correctly identify rivaling the accuracy the model is able to produce on the United States.

B. INDEPENDENT MODELING

Similar to the first model type, the independent models are also trained with a learning curve of 0.1, maximum tree depth of three, and use 50 percent random row sampling and 80 percent column sampling to prevent over-fitting. The models built with the three smaller responses, GeoRegion, Big3, and SplitOneChina, are trained to 80,000 iterations with the stopping condition of 20 rounds with no improvement to the validation classification accuracy. Due to the computational difficulty of training a model whose categorical response has a large number of levels, the model built with the Country response is only trained to 20,000 iterations. Table 4 provides the number of iterations for each independent model.

Table 4. Independent Modeling: Number of Iterations by Response

Response	Number of Iterations
Country	20000
GeoRegion	44425
Big3	4149
SplitOneChina	6032

The relative predictor importances for the second model type are very similar to those of the first. All models give a similar percent importance between the predictor variables. Figure 17 gives the percent importance by variable averaged over the four models. The independent models also find destination and the ship dimension variables to be the most important when predicting vessel origin.

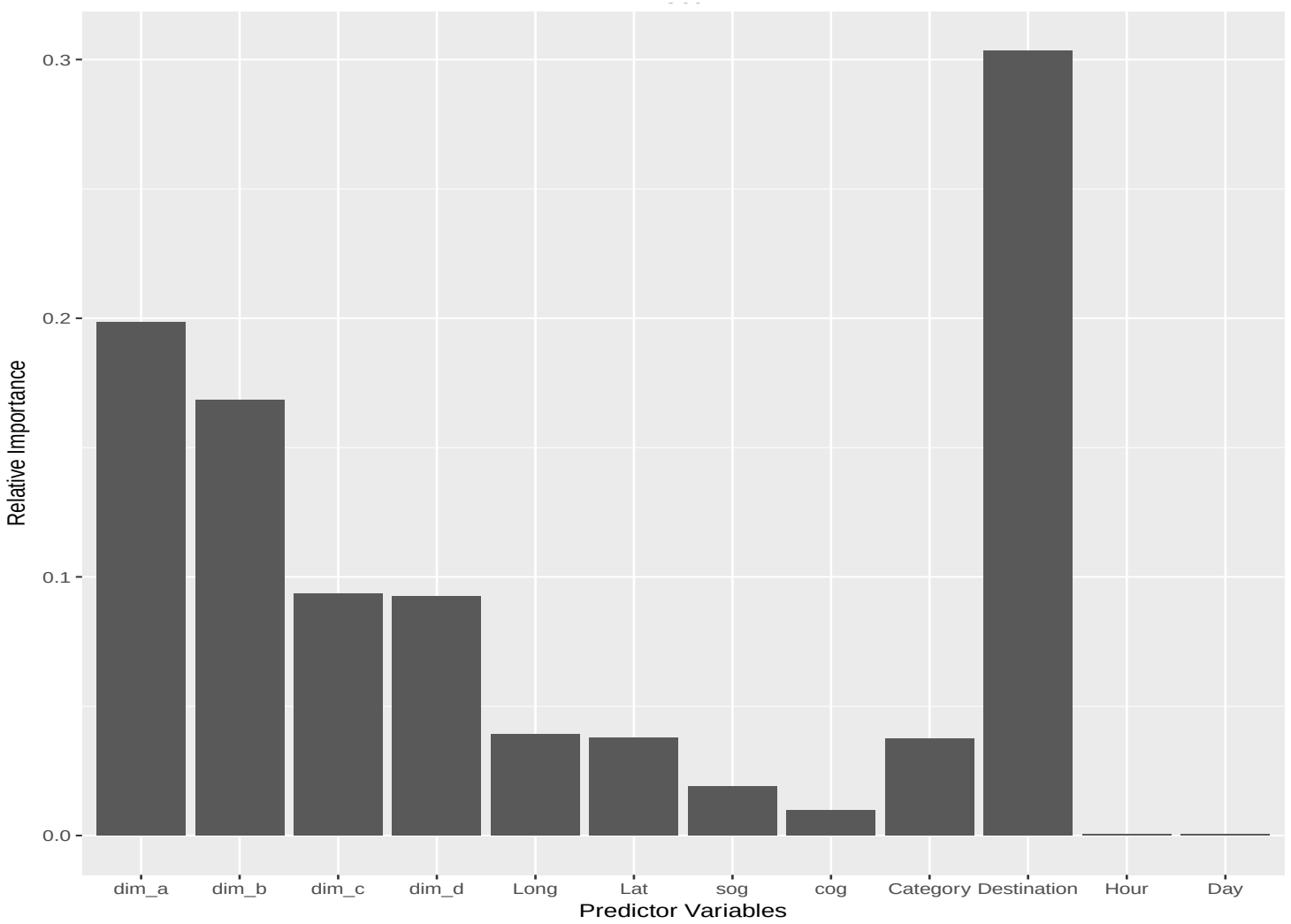


Figure 17. Average Predictor Variable Relative Influence for Independent Models

Not surprisingly, all independent models perform less accurately than their dependent counterparts because the independent models' training, validation, and test sets do not contain any overlap of unique vessels. Figure 18 illustrates the performances of each response for the second model type.

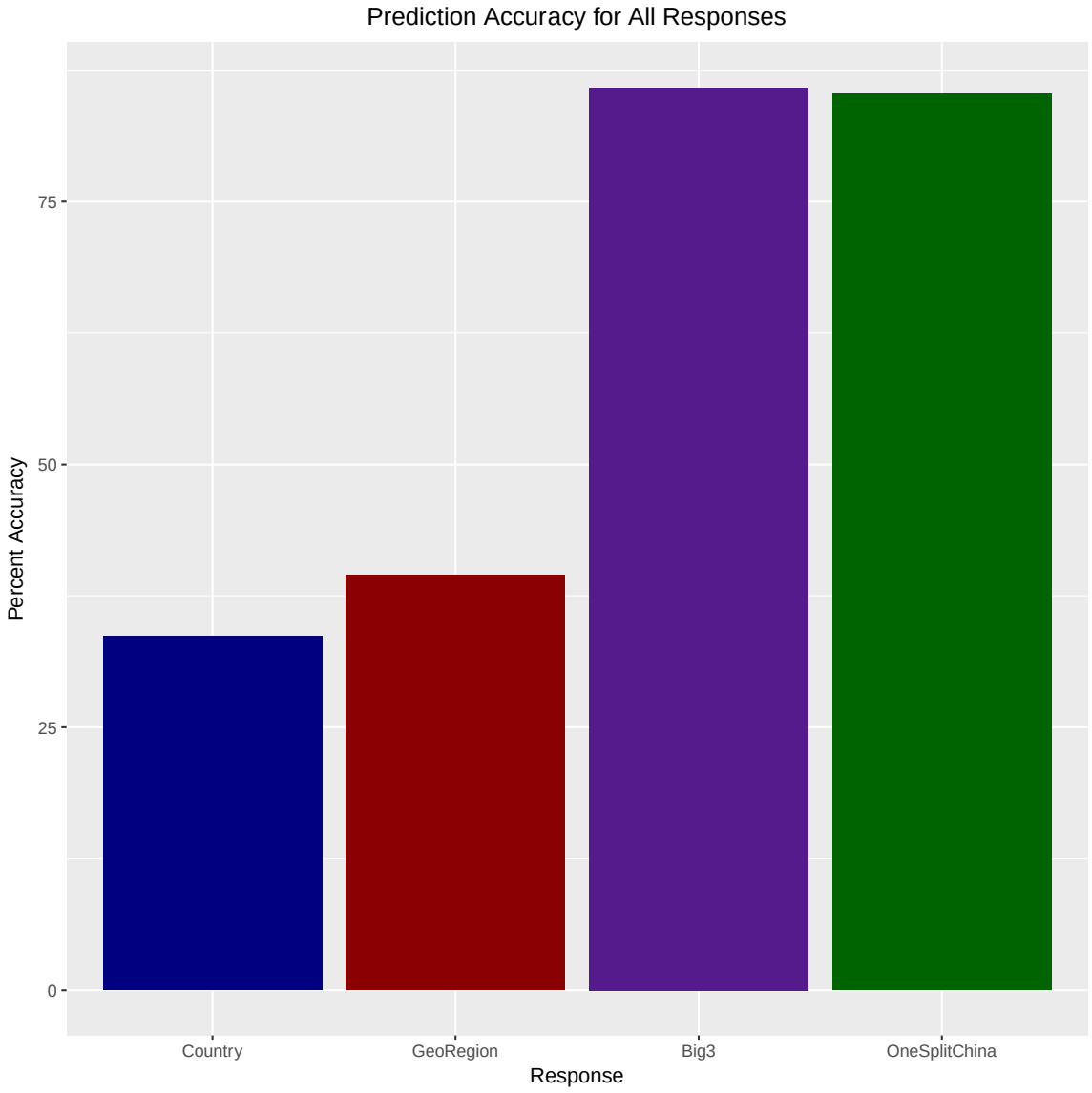


Figure 18. Independent Model Prediction Accuracy

The independent models built with the Big3 and SplitOneChina responses perform just under the prediction accuracy of their dependent model compliments. The Country and GeoRegion responses do not perform nearly as well, both demonstrating a reduction in prediction accuracy from the dependent models. The Country response prediction accuracy falls to approximately 34 percent, while the GeoRegion response performs slightly better at 39 percent. Allowing multiple predictions does improve the independent models. Figure 19 gives the hit ratios for the Country and GeoRegion responses for up to ten and three highest scoring levels, respectively.

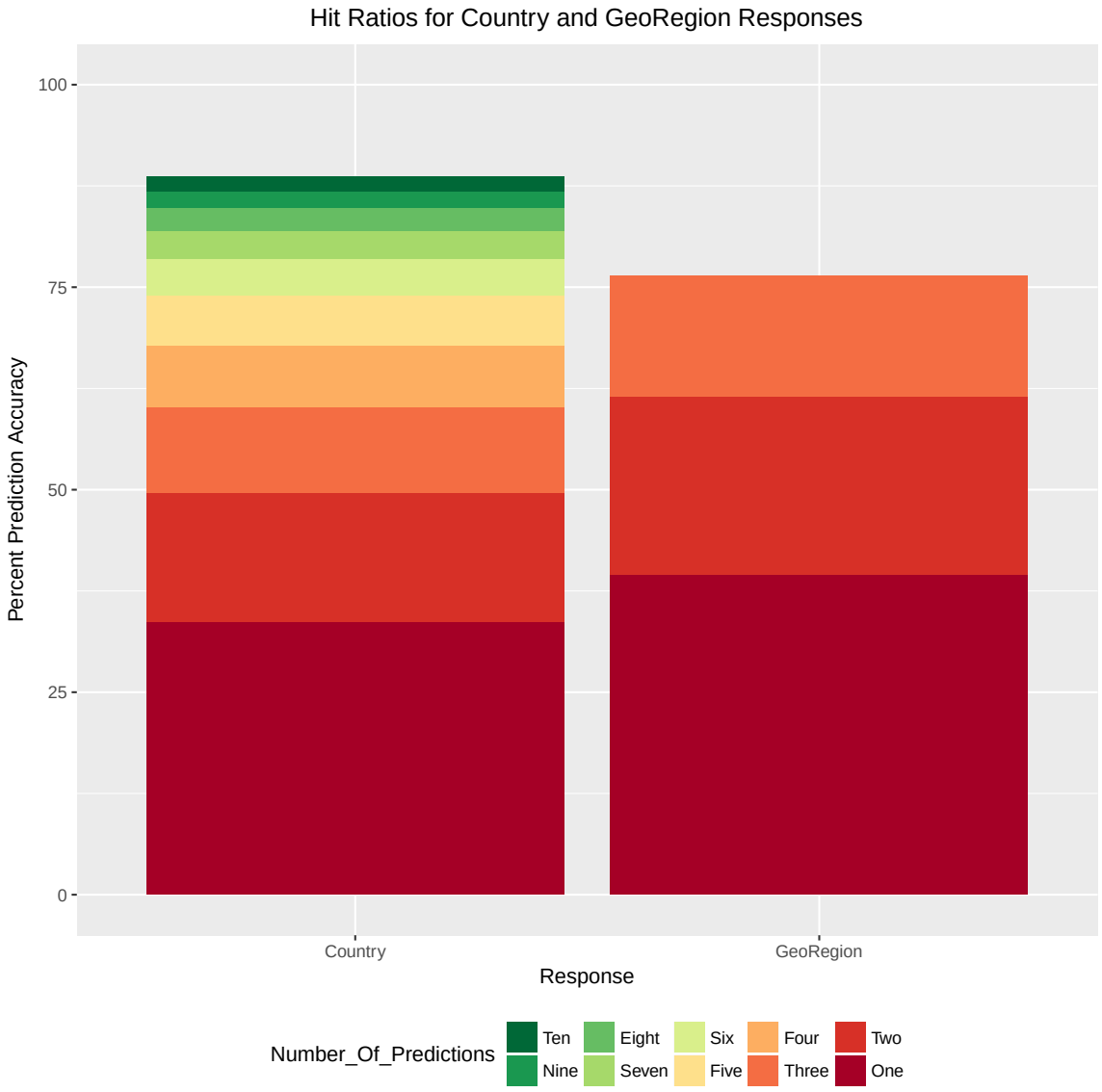


Figure 19. Prediction Accuracy of Country and GeoRegion Independent Models when given Multiple Prediction Opportunities

One additional prediction for the Country independent model improves its accuracy enough to be on par with the accuracy of the first prediction by the Country dependent model. Given the top ten highest scoring levels, the Country independent model is able to correctly predict a vessel's country of origin 88 percent of the time. Given three prediction opportunities, GeoRegion is able to predict with 76 percent accuracy.

Unlike the Country dependent model, only two countries in the second model type are predicted accurately over 90 percent of the time. Tuvalu's predictability remains consistent between the two model types only falling from 92 to 90 percent prediction accuracy. Brunei's prediction accuracy actually improves from 98 to 100 percent accuracy. The thirteen countries accurately predicted greater than 50 percent of the time given only the model's first prediction are graphed in Figure 20.

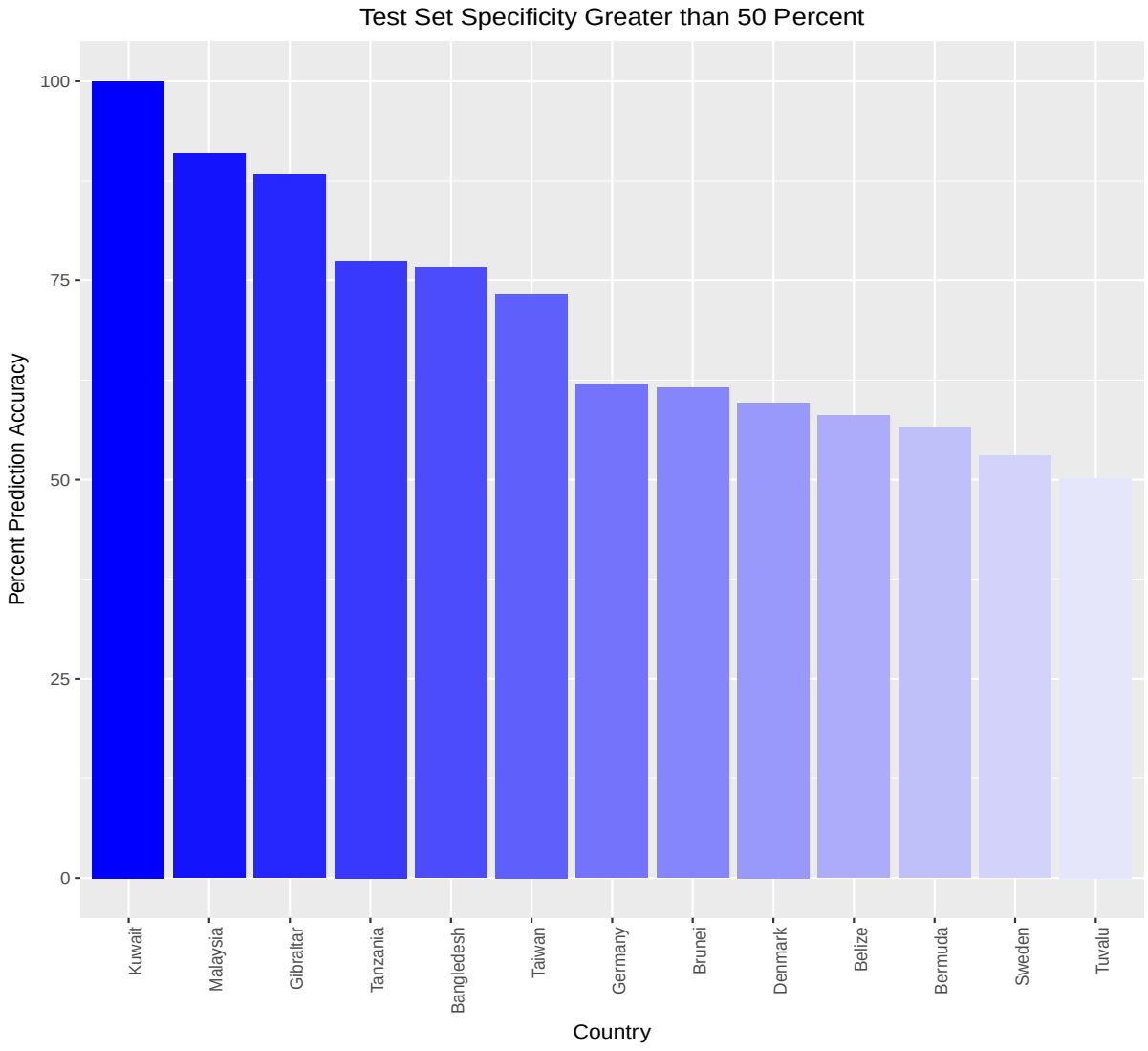


Figure 20. Countries Correctly Predicted in the Test Set over 50 Percent of the Time

Since most countries fall below 50 percent accuracy, Figure 21 graphs the relationship between the number of operating vessels and prediction accuracy with the percentage bins modified to reflect the lower performance rate of the independent models. Prediction accuracy in Figure 21 is calculated using the model's first prediction.

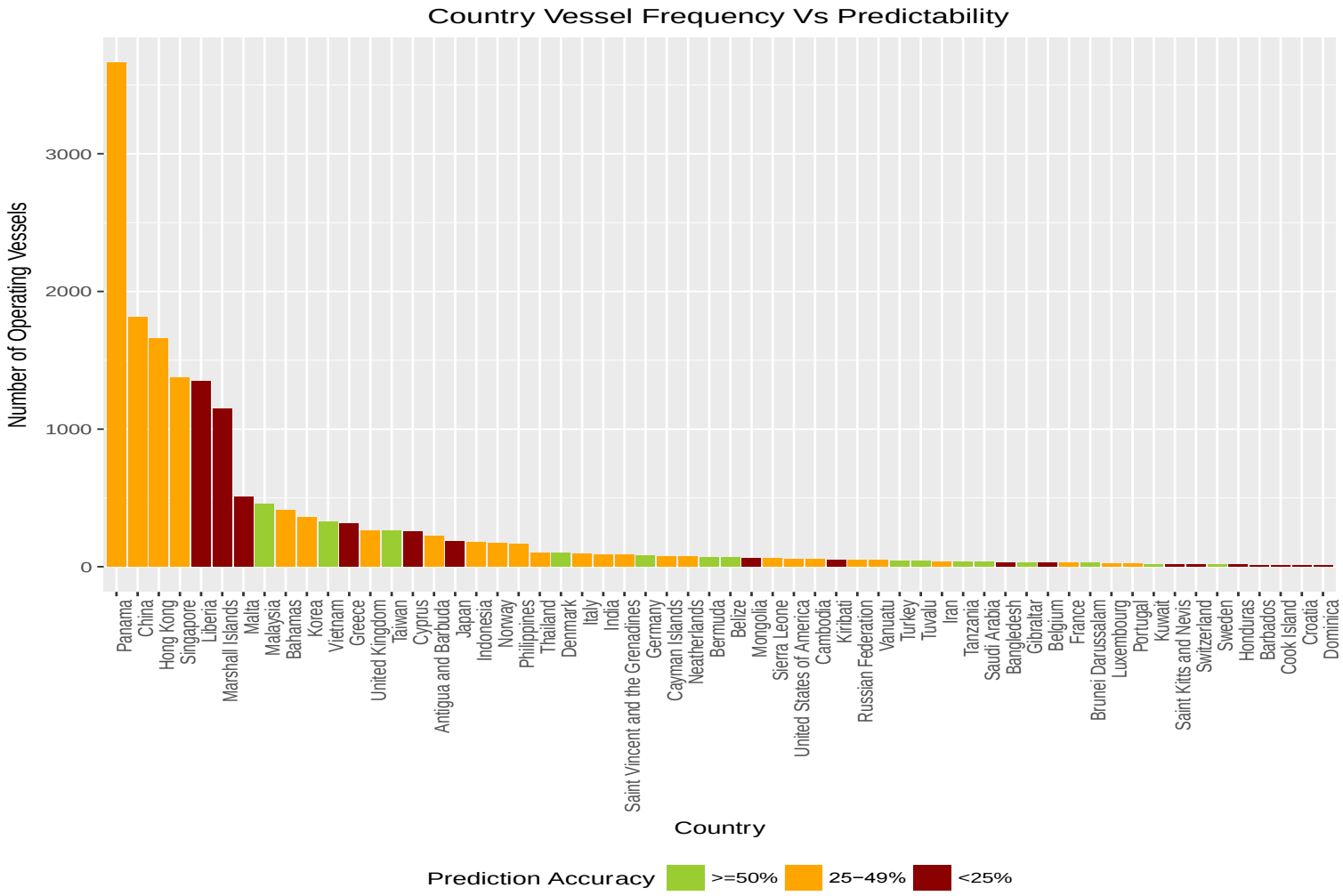


Figure 21. Independent Model: Relationship between a Country's Number of Operating Vessels and Prediction Accuracy

In Figure 21, the United States falls into the prediction accuracy category between 25 and 49 percent. U.S. vessels are accurately predicted by the model 48 percent of the time. Russia and Mainland China also fall into this category with prediction accuracies of 47 percent and 31 percent, respectively.

The Big3 and SplitOneChina independent models perform at a similar level of accuracy to the dependent models built with the same response given only the highest scoring level. Figure 22 gives a visualization of predicted versus actual classifications for the Big3 independent model. The misclassification ratios are significantly different from its dependent model counterpart, especially for the United States. U.S. vessels are predicted accurately only 37 percent of the time, often confused with Other vessels and China. China's predictability increases, but is still confused with Other vessels, however, at a much lower rate. Other categorized vessels are predicted with 87 percent accuracy. In general, the results of the independent models suggest that the vessel origin of new vessel operating in the region can be predicted to some extent.

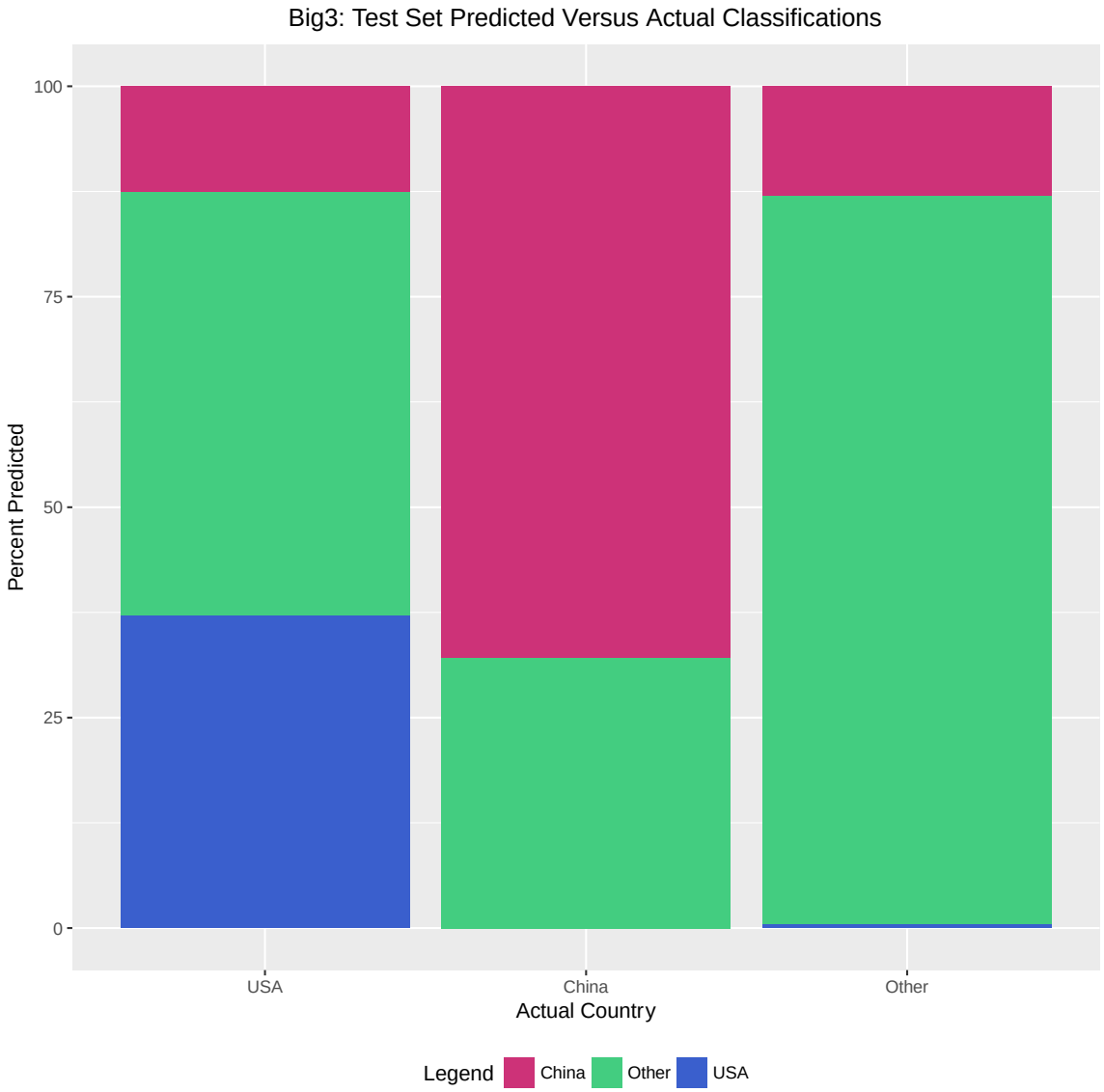


Figure 22. Independent Model: Test Set Predicted Versus Actual Classifications for Big3

Figure 23 partitions the China level into China, Hong Kong, and Taiwan. Other vessels are the easiest to categorize followed by Taiwanese vessels. U.S. vessels are most often mistaken for Other vessels and vessels from Hong Kong. China is the most difficult to predict, with an overall accuracy of 35 percent, however, their vessels are never misclassified as ones belonging to the United States.

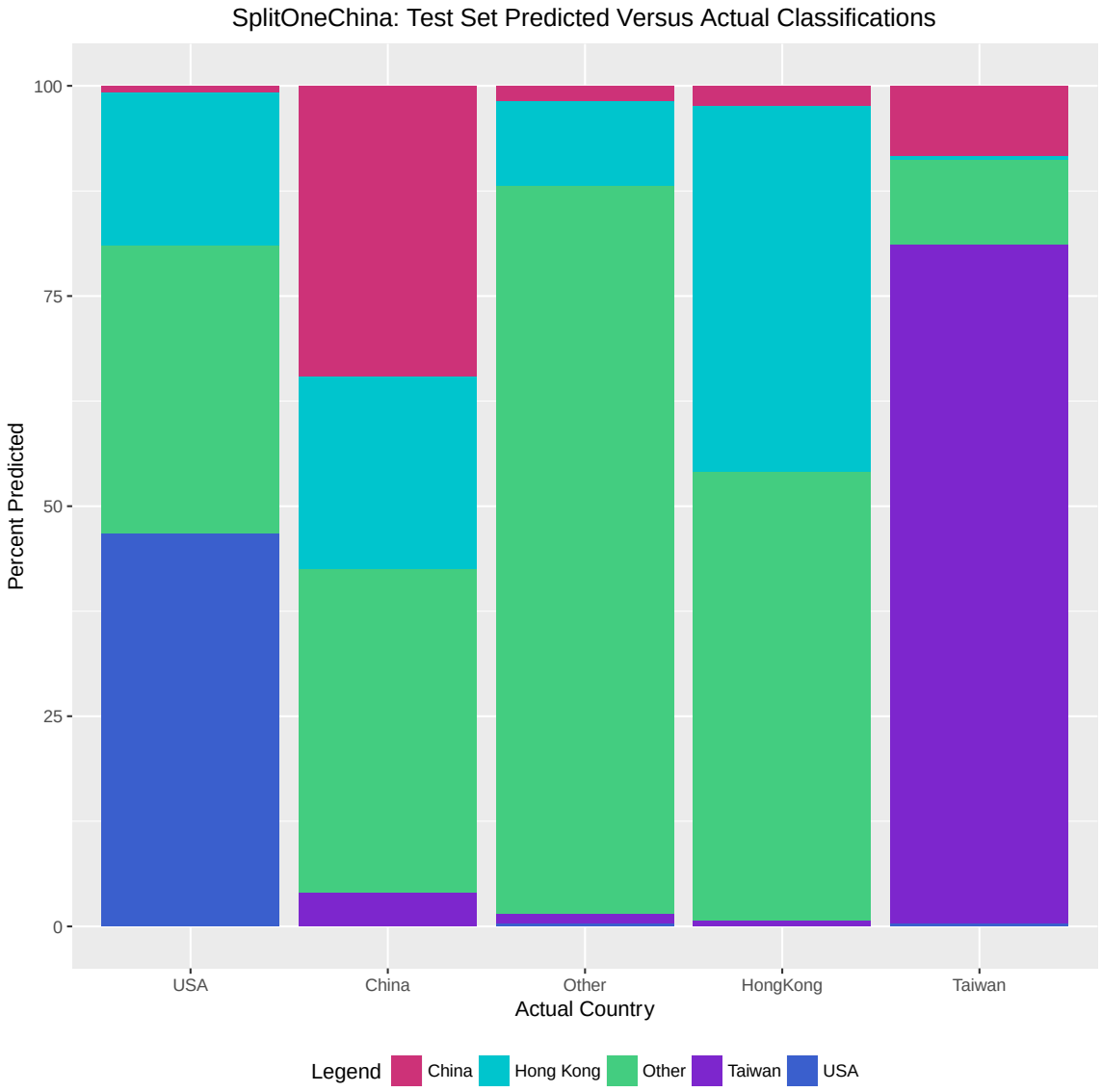


Figure 23. Independent Model: Test Set Predicted Versus Actual Classifications for SplitOneChina

C. OTHER APPROACHES

Two other approaches are considered and attempted during this analysis. Each categorical predictor variable yields one binary predictor per level. Because destination has over 9,000 levels, the number of input variables used to train the GBMs is quite large. The first additional approach reduces the number of input predictor variables using generalized low rank models (GLRMs), an extension of

principal components analysis that accommodates data sets with mixed categorical and numerical predictor variables (Udell, Horn, Zadeh, & Boyd, 2016). For exploratory purposes, we reduce the number of predictors using a GLRM to 2, 3, and 10 predictors. These are then clustered to uncover any obvious relationships between the lower rank versions of the data and country of origin. This approach is conducted using the R package h2o (The H2O.ai team, 2017). In all clustering attempts, no apparent connection is made to country, geographical region, location, or any factor variables, like ship type (category) or destination.

A second alternative approach attempts to extend the sample data that is based on a single AIS dynamic transmission into a data set summarizing tracks. This data set takes all transmissions associated with a single vessel over a 24-hour period or any time the vessel's static information changes and converts this information into a single observation by summarizing the dynamic portion of the data. New columns are constructed that attempt to detail a vessels dynamic movements over each day. These new columns are detailed in Table 5. The static columns of ship type, destination, and ship dimensions remain the same.

Table 5. Track Data Set Predictor Variables

Predictor	Description	Class (Type)
Week Days	Day of the week the signals occur	Factor
Max Time	Time of first signal	Factor
Min Time	Time of last signal	Factor
Average Latitude	Average of all recorded latitudes	Numerical
Max Latitude	Most north latitude recorded	Numerical
Min Latitude	Most south latitude recorded	Numerical
Average Longitude	Average of all recorded longitudes	Numerical
Max Longitude	Most east longitude recorded	Numerical
Min Longitude	Most west longitude recorded	Numerical
Average Speed Over Ground (SOG)	Average of all recorded SOGs	Numerical
Average Course Over Ground (COG)	Average of all recorded COGs	Numerical
Total Distance	Total distance traveled over day calculated	Numerical
Birds-eye Distance	Distance from first location and last location recorded calculated	Numerical
Total Number of Signals	Number of dynamic signals released over the course of the day	Numerical

The reconstructed track data set is analyzed with the same responses of Country, GeoRegion, Big3, and SplitOneChina. Models are built with the xgboost package's GBM, similar to the independent and dependent models built previously. In all cases, the track data set does not perform as well as the sample signal data. It is determined that summarizing track information in this manner is not sufficient enough for capturing the dynamic movement of vessels or improving prediction accuracy.

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSION

In this thesis we show that a vessel's country of origin can be predicted using single AIS dynamic transmissions combined with a single static transmission in the South China Sea. We train GBMs using two approaches. The first, the dependent models, partition the data into training, validation, and tests sets based upon time. This simulates the way such models will be used in practice. The second set of models, the independent models, partition each country's unique set of vessels randomly into training, validation, and test sets so that no two sets contain transmissions produced by the same vessel. This shows how such models can perform when classifying new vessels.

Dependent models perform better than independent models in all cases, though models built with the responses Big3 and SplitOneChina perform similarly at approximately 85 percent prediction accuracy regardless of model type. From the analysis, it can be concluded that destination, ship dimensions, and ship type are distinguishing features of vessels operating in the South China Sea.

In regard to the larger classification responses of Country and GeoRegion, the dependent models outperform the independent models significantly. It is possible with more time and computational resources that models built with the Country response could perform at a better rate given the opportunity to reach optimality. Since this is unachievable in this research because each model required weeks to build, the independent Country response model is unable to characterize maritime patterns of behaviors for many countries operating in the region. Most countries fall within a predictability range of 25–49 percent for this model. It is possible that if this model is trained for more iterations it could produce more prediction accuracies in the over 50 percent range.

Since the dependent models perform better, require less time to train, and, ultimately, serve as proof of concept for a larger, global model trained over

several years of AIS data, it is the most reasonable model to consider employing as a method to assist with informing decisions at sea. More research is needed to determine if the results of this research can be replicated throughout different regions of the world and over different time frames. Computing power is the limiting factor in assessing and training this type of model on vessels, but as computing power increases over time and algorithms become more sophisticated it may become easier to conduct related research on larger amounts of AIS data that produce similar or more accurate results. Once the models are trained, however, deploying and using these models require very little computational time. Furthermore, because there is little turnover of vessels in the South China Sea, the prediction accuracy of these models degrades little over time. We suspect that the same will be true for models trained in other regions.

A. FUTURE WORK

This research is really the tip of the iceberg for validating the usefulness of AIS data in predicting vessel origin and identifying patterns of behavior. Increasing the number of iterations or even changing certain xgboost parameters, like the learning rate or sampling rates (Chen, He, Benesty, Khotilovich, & Tang, 2018), could provide better, more accurate prediction results. Similar research could also be conducted on a different geographical area of interest like the Baltic or a particularly busy seaport to validate the usefulness of AIS in securing national ports and waterways by identifying vessels exhibiting operating patterns not in line with its claimed country of origin. Since many vessels operate in many different regions, one can envision a suite of such models. Identifying vessels whose origin is regularly misclassified in different regions or at different times, may indicate anomalous behavior. It may also be useful to build models with worldwide AIS data or data that can encompass several years to determine if seasonality exists.

Other research could potentially focus on a more efficient way to organize AIS signals prior to conducting model analysis. A different method of

reconstructing AIS signals to capture ship tracks could be developed and employed. One possible method would be to create a sparse data set that records each latitudinal and longitudinal coordinate released by one vessel over a certain period of time. Other methods of text analysis for destination could be explored when prepping the data, as well as determining the usefulness of using ship type as defined in the AIS static records instead of binning vessels into basic vessel types, like cargo and fishing as performed in this thesis.

Further research into the usefulness of AIS to detect anomalous behavior and secure waterways is not only warranted, but has the potential to positively impact the safety of operations at sea. The amount of AIS data available makes the possibilities of leveraging it seemingly endless. More research should focus on leveraging the data available and exploring patterns of behavior and prediction performances in different regions and worldwide.

THIS PAGE IS INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Bay, S. M. (2017). *Evaluation of factors on the patterns of ship movement and predictability of future ship location in the Gulf of Mexico* (Master's thesis). Retrieved from https://calhoun.nps.edu/handle/10945/53021

Balduzzi, M., Pasta, A., & Wilhoit, K. (2014). A security evaluation of AIS Automated Identification System. *Proceedings from the 30th Annual Computer Security Applications Conference*. https://doi.org/10.1145/2664243.2664257

Burgess, J. P. (2003). The politics of the South China Sea: Territoriality and international law. *Security Dialogue*, *34*(1), 7–10. https://doi.org/10.1177/09670106030341002

Chen, T., He, T., Benesty M., Khotilovich, V., & Tang, Y. (2018). xgboost: Extreme gradient boosting (R package version 0.6.4.1) [Computer software]. Retrieved from https://CRAN.R-project.org/package=xgboost

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2), 337–407.

Google Maps. (2017). South China Sea. Retrieved from https://www.google.com/maps/@15.9113154,112.9675641,4z

The H2O.ai team. (2017). h2o: R interface for h2o. R package version 3.16.0.2. Retrieved from https://CRAN.R-project.org/package=h2o

Harati-Mokhtari, A., Wall, A., Brooks, A., & Wang, J. (2008). *Automatic Identification System (AIS): A Human Factors Approach*. Retrieved from https://www.researchgate.net/publication/254062770_Automatic_Identification_System_AIS_A_Human_Factors_Approach

Hastie, T., Tibshirani, R., & Friedman, J. (2016) The elements of statistical learning: Data mining, inference, and prediction. *Springer Series in Statistics*, *2*.

International Telecommunications Union. (2014). M.1371: Technical characteristics for an automatic identification system using time division multiple access in VHF maritime mobile frequency band. Retrieved from https://www.itu.int/rec/R-REC-M.1371/en

Internet World Stats. (2017, December 4). *World Regions*. Retrieved from http://www.internetworldstats.com/list1.htm#geo

Kaplan, R. D. (2011). The South China Sea is the future of conflict. *Foreign Policy*, *188*, 76–85.

Lane, R., Nevell, D., Hayward, S., & Beaney, T. (2010). Maritime anomaly detection and threat assessment. *Proceedings from The 13th International Conference on Information Fusion*. https://doi.org/10.1109/ICIF.2010.5711998

Mao, S., Tu, E., Zhang, G., Rachmawati, L., Rajabally, E., & Huang, G. (2016). An automatic identification system (AIS) database for maritime trajectory prediction and data mining. ArXiv e-prints. Retrieved from https://arxiv.org/abs/1607.03306

Mou, J., Tak, C., & Ligteringen, H. (2010). Study on collision avoidance in busy waterways by using AIS data. *Ocean Engineering*, *37*(5). https://doi.org/10.1016/j.oceaneng.2010.01.012

Naval Postgraduate School. (2018, January 6). Retrieved on February 14, 2018 from NPS wiki: https://wiki.nps.edu/pages/viewpage.action? title= Home&spaceKey=HPC

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. https://doi.org/10.3390/e15062218

R Core Team. (2017). R: A Language and Environment for Statistical Computing [Computer software]. Retrieved from https://www.R-project.org/

Udell, M., Horn, C., Zedah, R., Boyd, S. (2016). Generalized low rank models. *Foundations of Trends in Machine Learning*. *9*(1), 1–118.

U.S. Department of Homeland Security. (2017). AIS frequently asked questions. Retrieved from https://ww.navcen.uscg.gov/?pageName=AISFAQ.

van der Loo M (2014). The stringdist package for approximate string matching. *The R Journal*, *6*, 111–122. Retrieved from https://CRAN.R-project.org/package=stringdist

Wickham, H. (2016). Feather: R bindings to the feather 'API' (R package version 0.3.1) [Computer software]. Retrieved from https://CRAN.R-project.org/package=feather

Young, B. L. (2017). *Predicting vessel trajectories from AIS data using R* (Master's thesis). Retrieved from https://calhoun.nps.edu/handle/10945/55564

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California