THE INFLUENCE OF WEATHER AND ICE ON FERRY OPERATIONS: MODELLING PRESENT-DAY EFFECTS TO PREDICT FUTURE TRENDS

by

Andrew Sargeant

Submitted in partial fulfillment of the requirements
for the degree of Master of Applied Science

at

Dalhousie University
Halifax, Nova Scotia
July 2016

This thesis is dedicated to Tom

Table of Contents

List of Tables

List of Figures

Abstract

Ferry performance is influenced to a great degree by environmental factors. Adverse weather and ice conditions can severely restrict the ability to conduct operations in a safe, efficient, and financially viable manner. Furthermore, as regional conditions vary due to climate change influences, the effects may become more severe. In order to better understand how specific weather and ice factors influence ferry operations, a statistical analysis is conducted using a case study of historical Marine Atlantic Incorporated traffic data and historical weather and ice condition data from the National Centers for Environmental Prediction North American Regional Reanalysis and Optimum Interpolation Sea Surface Temperature analysis. Random Forest models are constructed to predict ferry sailing cancellations and delays, using selected environmental factors as inputs, to examine the influence and relationships of specific factors and combinations of factors, and to project rates of cancellation in the coming decades using Coupled Model Intercomparison Project Phase 5 data sets. Results show that (1) environmental factors are good predictors of cancellations and poor predictors of delays, (2) wind speed is the most important environmental factor for cancellation prediction, and air temperature the most important for delay prediction, and, (3) that the ratio of cancelled sailings to total sailings is projected to increase over the next three decades.

List of Abbreviations and Symbols Used

ANOVA	analysis of variance.
AUC	Area Under the Curve.
BP	Blue Puttees.
CART	Classification and Regression Tree.
CMIP5	Climate Model Intercomparison Project Phase 5.
CTree	Classification Tree.
ECDF	empirical cumulative distribution function.
GBTree	Gradient-Boosted Trees.
HL	Highlanders.
KNN	k-Nearest Neighbours.
kPa	kiloPascals.
kts	nautical miles per hour.
LDA	Linear Discriminant Analysis.
LogReg	Logistic Regression.
MAI	Marine Atlantic Incorporated.
MSE	mean squared error.
MV	Motor Vessel.
NARR	North American Regional Reanalysis.
NCEP	National Centers for Environmental Prediction.
NOAA	National Ocean and Atmospheric Administration.
OISST	Optimum Interpolation Sea Surface Temperature.
OOB	out-of-bag.
RCP	representative concentration pathways.
RF	Random Forest.

RMSE	root mean squared error.
ROC	Receiver Operating Characteristic.
RoPax	Roll-On Roll-Off Passenger.
RoRo	Roll-On Roll-Off.
SVM	Support Vector Machines.

Acknowledgements

This research would not have been possible without the support of key individuals from Marine Atlantic Incorporated, who, to preserve confidentiality, unfortunately cannot be mentioned here. I would like to gratefully acknowledge their time and energy in fostering a sound partnership, answering countless questions, providing data essential to the project, and enabling resources to enhance the study.

I am also grateful for the invaluable guidance, support, enthusiasm, and feedback from my thesis co-advisors, Dr. Ronald Pelot and Dr. Alireza Ghasemi, and from my thesis committee members, Dr. Claver Diallo and Dr. Ahsan Habib.

The historical climate data sets used in this research are publicly available from the National Centers for Environmental Prediction North American Regional Reanalysis project and the Optimum Interpolation Sea Surface Temperature analysis, both of which are accessible from the website of the National Ocean and Atmospheric Administration Earth Systems Research Library, Physical Sciences Division.

The climate change data sets used in this research are publicily available from the Coupled Model Intercomparison Project Phase 5, which is a result of the efforts of the World Climate Research Programme's Working Group on Coupled Modelling. Further thanks are extended to the climate modeling groups (listed in Chapter 6 of this paper) for producing and making available their model output.

I would also like to express my thanks to the Department of National Defence for the opportunity, time, and sponsorship afforded me to complete this thesis and the masters degree of which it is a significant part.

Chapter 1

Introduction

1.1 Problem Introduction

Ferry operations play a crucial role in transportation and logistics networks in regions that feature extended coastlines, islands, lakes, and rivers. In Canada, ferries provide an essential transportation link in the Pacific Northwest, the Great Lakes, Canada's North, and Atlantic Canada (Transport Canada, 2015). In 2014, ferries in Canada transported over 53 million passengers, almost 20 million vehicles, and billions of dollars of goods (Canadian Ferry Operators Association, 2015).

Ferry operations are subject to their environment in many ways. Ferry service can be interrupted by mechanical and electrical breakdowns on the vessels or port infrastructure, traffic congestion, labour disputes, computer system faults in reservation and check-in systems, and environmental factors. Performance measurement of ferry operations is a relatively new and growing field, and only large ferry companies have implemented performance measurement programs to varying degrees. Metrics typically include reliability of service, on-time departure and arrival, safety incident occurrence, cost-efficiency, and customer satisfaction (Bennion, 2010).

Many of the factors that affect ferry operations are mitigated by management and company policies, and indeed performance standards are often set for those factors over which the company exercises a degree of control, such as mechanical problems, staffing, and scheduling. Similar to the commercial airline industry, however, ferry companies tend to exclude environmental factors from their metrics due to the inherent unpredictability and limited ability to control them.

Environmental factors in the context of ferry performance inlcude various aspects of weather, sea state, and ice, which affect, and potentially interrupt, ferry service principally through their physical interactions with the ferry vessels. Wind, waves, and ice can reduce vessel speed and prolong the journey, or increase the navigation risk, particularly when entering or leaving harbours or when docking or undocking.

Extreme temperatures can have a detrimental effect on mechanical and electric components, increase the risk of freezing spray in winter, and increase work hazard and decrease productivity of employees. Atmospheric pressure is often an indicator of the presence of severe weather and passing storms.

The focus of this thesis is to examine the influence of environmental factors on ferry operations, and determine whether significant changes are to be expected in the future. In particular, the occurrence of ferry trip cancellations and delays are investigated, as well as the severity of delays. Ferry operations and environmental conditions are highly location dependent. A case study of the ferry operations of Marine Atlantic Incorporated (MAI), which operates in Atlanic Canada between Nova Scotia and Newfoundland, is undertaken. The methods and resulting models are proposed for application in other regions.

1.1.1 Overview of Marine Atlantic Incorporated

MAI is a Canadian Crown corporation that reports to the Government of Canada through the Minister of Transport. MAI guarantees a year-round ferry link between North Sydney, Nova Scotia and Port aux Basques, Newfoundland, in order to fulfill its constitutional mandate of providing ferry service between the island of Newfoundland and the province of Nova Scotia (Treasury Board Secretariat, 2016). During the summer months MAI also provides ferry service between North Sydney and Argentia, Newfoundland. Figure 1.1 provides a map of the MAI operating area, routes, and ports.

MAI operates four ice-class Roll-On Roll-Off Passenger (RoPax) vessels, the names and characteristics of which are described in Table 1.1. The Motor Vessel (MV) Blue Puttees, MV Highlanders, and MV Leif Ericson operate principally on the North Sydney - Port aux Basques route, and the MV Atlantic Vision operates on the North Sydney - Argentia route during the summer and on the North Sydney - Port aux Basques route on an as-required basis during the remainder of the year. The MV Leif Ericson is used mainly as a commercial carrier and transports most of the hazardous goods. Table 1.2 provides a summary description of the applicable ice-classes. Additional information on ice-class designations can be found from Veritas (2016).

Figure 1.1: MAI area of operations, ports, and routes.

MAI maintains a schedule that typically offers two sailings per day in each direction between North Sydney and Port aux Basques. This route is 178 km (96 nm) in length and approximately seven hours in duration. One day per week only one sailing is offered in each direction, which allows time for vessel management and maintenance activities. The North Sydney - Argentia route is only offered during the summer between June and September, normally three sailings per week in each direction. The route is 520 km (281 nm) in length and approximately sixteen hours in duration. Most of the MAI business occurs on the North Sydney - Port aux Basques route, and in order to focus on year-round environmental issues, the Argentia route is not included in the study.

MAI is the only year-round ferry service between Nova Scotia and Newfoundland and therefore provides an essential logistics link between the two provinces. The commercial trucking industry uses the service to transport a wide range of goods, including important commodities such as fruits and vegetables, dairy, meat, and

Table 1.1: MAI vessel characteristics.

Vessel	Principal Use	Length (metres)	Capacity (lane-metres†)	Passengers	Ice Class
MV Blue Puttees	Passenger, Commercial	199.5	2840	750	1A
MV Highlanders	Passenger, Commercial	199.5	2840	750	1A
MV Atlantic Vision	Passenger, Commercial	203.3	2425	700	1A*
MV Leif Ericson	Commercial	158	1550	380	1B

† Unit of deck area for RoRo vessels. One lane-metre is an area one metre long by two metres wide.

Table 1.2: Ice-class designations (Det Norske Veritas 2016)

Ice-class	Description
1A*	Normally capable of navigating in difficult ice conditions (thickness 0.5-1.0 m) without the assistance of icebreakers
1A	Capable of navigating in difficult ice conditions (thickness 0.5-1.0 m), with the assistance of icebreakers when necessary
1B	Capable of navigating in moderate ice conditions (thickness 0.3-0.5 m), with the assistance of icebreakers when necessary

medical supplies to businesses and institutions in Newfoundland. The service is also an essential component in the supply chain for Newfoundland-based industries by providing a means to export goods (Marine Atlantic Inc., 2015).

The main commercial shipping competition for MAI is OceanEx, which provides year-round container shipping service between St. John's, Newfoundland, and ports in North America (although mainly Halifax). OceanEx does not provide Roll-On Roll-Off (RoRo) service, however, so is often less convenient for commercial trucking customers.

1.1.2 Atlantic Canada Environmental Characteristics

Due to its temperate climate, Altantic Canada weather varies considerably and is highly seasonal. The many miles of coastline in Atlantic Canada means that weather is also heavily influenced by the presence of the ocean. Robichaud and Mullock (2001) provide a comprehensive synopsis of climate and weather conditions in the region, summarized here.

Summer is typically characterized by large, stable, high pressure air masses that move slowly through the region or remain stationary for periods of time. Storms are

less frequent than other seasons and the effect of the Bermuda High becomes more pronounced, causing the circulation to be southwesterly. Storms, when they do occur, are typically the result of a tropical depression originating in the southern latitudes. Advection sea fog is common due to moist air being pushed up by the southwesterly flow and cooling over the cooler Atlantic Canada waters. On-shore sea breezes are typical of warm, sunny days.

Winter sees increased storm activity, in both power and frequency, due to the greater difference in temperatures between northern and southern latitudes, as well as increased circulation, generally from the west or northwest. Freezing precipitation in all its forms is common and can linger even after a low pressure system has passed. Masses of cold arctic air typically are pushed down by northern high pressure systems in between passing lows, causing cold but clear conditions. Ice is common in the Gulf of St. Lawrence and Cabot Strait, but normally isn't present in large quantities until mid-winter.

1.1.3 Research Method

The problem of understanding how environmental factors affect MAI operations is approached by studying the occurrence of cancelled sailings and delayed sailings, as well as the extent to which sailings are delayed. Relationships between these three sub-problems and the presence of various environmental factors are investigated using statistical analysis and modelling techniques. A standard statistical analysis and modelling methodology is followed, which begins with data acquisition and formatting, followed by an exploratory analysis of the data, formulation of a model, model validation, generation of results, and discussion. The prediction models are then used to predict the extent to which changes can be expected in the coming decades by using data sets from recognized climate models. All of the analysis and modelling was completed using R, a popular open-source programming language and environment for statistical computing and graphics (R Core Team, 2016).

1.1.4 Thesis Outline

The remainder of this thesis is structured as follows: A literature review makes up the balance of Chapter 1 and includes past research relevant to this study in the areas

of maritime risk modelling, weather forecasting and ship routing, maritime incidents, and statistical modelling. Chapter 2 describes the sources and characteristics of data used for this thesis, the data formatting and pre-processing, as well as an exploratory data analysis of the environmental factors. Chapters 3 and 4 conduct a statistical analysis of the influence of environmental factors on ferry sailing cancellation as well as delay and delay length, respectively, including the development of a prediction model for each. Chapter 5 examines the relationships between cancelled and delayed sailings, projects potential variations due to climate change, offers ideas on future research, and provides concluding remarks.

1.2 Literature Review

The literature is sparse on the topic of environmental affects and ferry operations. Some work has been done in the areas of maritime transportation risk modelling, fisheries, naval architecture, and navigation, however the context tends to be safety-related, with the focus on incidents, accidents and collisions (O'Connor & O'Connor, 2006; Kelman, 2008; Grabowski, Ayyalasomayajula, Merrick, & Mccafferty, 2007; Rezaee, Pelot, & Finnis, 2016). Other studies explore the operational context of maritime transportation but typically from the perspctive of marine traffic and transportation network states and efficiency, with relatively little emphasis on environmental factors. This chapter reviews the past literature relative to this study, i.e., within the context of environmental factors and ferry operations. A summary of related works in the areas of maritime risk modelling, weather routing, maritime incidents, and transportation modelling, follows. As there are similarities in modelling approaches between air and marine transportation, related works in that area are considered as well. A review of statistical modelling techniques relevant to this research is also provided.

1.2.1 Maritime Risk Modelling

Risk modelling is becoming more common within the maritime transportation domain. Washington State Ferries, the largest passenger vessel ferry system in the United States, undertakes significant work and supports academic reasearch in this area. Athough not directly focused on environmental factors, the general approach

to risk analysis provides a solid framework to approach similar problems. Given that ferry accidents are low probability, high consequence events, Merrick, Dorp, Mazzuchi, and Harrald (2001) approach the problem by combining system simulation, expert judgment, and available data. The simulation component of the study captures the dynamic environment of traffic and weather in order to assess risk reduction policies and system-level decisions in terms of collisions and accidents. This research is used to underpin a comprehensive risk management framework for Washington State Ferries (Dorp, Merrick, Harrald, Mazzuchi, & Grabowski, 2001). They build further upon the risk modelling research by investigating the uncertainty involved in simulation-based maritime risk assessments, and propose a Bayesian simulation technique to model this uncertainty (Merrick, 2005). These efforts include environmental factors as part of their assessments but stop short of actually investigating the effects of these factors in detail.

1.2.2 Weather Routing

Weather routing of ships involves selecting the optimal route for a vessel with respect to the potential weather it will experience during the voyage. Given a set of possible routes, the predicted weather and the vessel's seakeeping characteristics in various weather and wave conditions, the best route can be determined depending on the goal (usually fuel efficiency, time, or safety of crew or cargo). These types of problems are typically approached as network optimization problems and several researchers have applied them across different geographic scales, such as trans-ocean voyages (Sen & Padhy, 2015; Shao, Zhou, & Thong, 2012) and coastal shipping (Takashima, Mezaoui, & Shoji, 2009; de Osés & la Castells, 2008). A variety of algorithms are employed in this regard, including Dijkstra's algorithm (Sen & Padhy, 2015; Takashima et al., 2009) and dynamic programming techniques (Shao et al., 2012; Fang & Lin, 2015). These studies look at the voyages on an individual basis and proceed with the assumption that they will not be cancelled.

de Osés and la Castells (2008) get closer to the issue at hand by considering the weather impacts on several short sea shipping routes in Europe. They approach this from the perspective of network expansion in the face of competition from speedy ground and air transportation. The probability of encountering heavy weather on

possible expansion routes is determined, as is the risk of cancellation and seasickness (as a measure of comfort and safety) for various types of ships. The study is limited, however, in that it uses only significant wave height (the average of the highest one-third of waves in an observation) for the descriptive heavy weather variable.

1.2.3 Maritime Incidents

Significant research has been conducted in Atlantic Canada on the occurrence and severity of maritime incidents in relation to weather factors. Rezaee et al. (2016) analysed weather and incident data in Alantic Canadian waters to determine relationships. A logistic regression model was created for each of incident occurrence and incident severity. The incident occurrence model found that Laplacian of pressure (cyclone intensity), wind speed, sea surface temperature, and darkness were significant factors. The incident severity model found that ice concentration, wind speed, sea surface temperature, and darkness were significant factors. This work was focused on providing insight to the fishing industry, marine traffic decision makers, and the Canadian Coast Guard.

Wu, Pelot, and Hilliard (2009) found similar results when looking at the effect of weather on the relative incident rate of fishing accidents. General results demonstrated an increase in the relative incident rate as weather factors deteriorated. Decision trees were used to determine the factors of greatest significance, indicating that ice concentration was the dominant factor. In the absence of ice, wave height was dominant. The decision tree methodology allowed the authors to determine variable importance and provide a visual tool for understanding the effects of combinations of factors. This ability to visualize variable relationships is a strength of decision trees, however they are limited in the predictive performance. Wu et al. (2009) also provide a thorough review of the fishing vessel accident literature.

1.2.4 Flight Delay Prediction

The complexity and rapid pace of airline operations drives substantial efforts to predict and mitigate disruptions to the air transportation system. Like marine transportation, airline operations are affected to a great degree by weather and other

causes of interruptions such as mechanical breakdowns. Advances in airline system operations modelling and optimization have increased system efficiency and resilience through problem incident prediction techniques and decision support tools that help mitigate impacts. Klein (2010) approaches airport delay prediction by using a modified form of the Weather-Impacted Traffic Index model, a well-known decision-support tool for predicting the effects of weather on real-time flight operations. Zhang (2008) proposes airline schedule recovery from interruptions caused by weather by employing real-time intermodal substitution and optimization using non-linear programming. Jarrah (1993) analysed flight cancellations and delays and developed a decision-support tool to assist air traffic controllers based on network optimization. Ground delay programming is a significant component of airline delay and cancellation management. Provan, Cook, and Cunningham (2011) offer a probabilistic model to predict aircraft arrival rates and airport capacity based on the weather forecast to improve ground delay program planning. The methods used in the airline industry to predict service disruptions tend to rely to a great deal on existing, industry-standard models and data sets, such as the Weather-Impacted Traffic Index and ground delay program systems, as well as network-optimization techniques due to the underlying system frameworks. Although there is some potential for application of these methods in the the ferry operations domain, the problem structure and complexity are fundementally different, and no established industry-wide data sets or established planning models exist.

1.2.5 Hazard-Based Duration Models

Hazard-based duration modelling involves various approaches to the analysis of incidents from the perspective of the end-of-duration occurrence (predicting the time that a particular incident will end), and is typically related to choices or behaviours that result in a specific occurrence. These approaches are often employed in transporation modelling to predict incident occurrence and duration, such as prediction of traffic incident duration, or prediction of travel time based on transportation mode and route selection. Specific examples in the literature include traffic incident clearance time and emergency vehicle arrival time prediction (Ji, 2014), highway incident duration analysis (Nam & Mannering, 2000; Boyles, Fajardo, & Waller, 2007; Garib, Radwan,

& Al-Deek, 1997), and activity choice behaviour while commuting (i.e., selection of transportation modes and routes, activities engaged in while commuting, such as shopping, etc.) (Bhat, 1996). Hensher and Mannering (1994) provide a summary of hazard-based duration models and their application to transportation problems. The current analysis, however, involving ferry operations with respect to environmental conditions, is approached from a systems point of view, involving specific input variables and response variables, with no requirement for choice or behaviour modelling, so hazard-based duration models are of limited use.

1.2.6 Statistical Modelling Techniques

Statistical modelling is often used in risk analysis. The wide variety of tools allows the modeller to tailor the approach to the application, as evidenced by previously cited examples for logistic regression (Rezaee et al., 2016), random parameters negative binomial regression (Rezaee et al., 2016), classification trees (Wu et al., 2009), and Bayesian techniques (Merrick, 2005). Random Forest (RF) models were chosen for this analysis for their blend of strong predictive power and useful methods for understanding variable relationships (further justification of this selection is found in Chapter 4).

RF is a machine learning model based on the ensemble methodology that takes a series of weak learners to combine results and increase predictive performance. Specifically, RF grow many decision trees (a forest) and allow each tree in the forest to have a say in the predicted response (a theoretical explanation of RF models and their formulation is provided in Chapter 4). They are particularly powerful in the presence of many predictors and/or predictors with complex interactions. RF have been used in many domains, including the classification of molecular compounds (Svetnik et al., 2003), predicting aquatic toxicity (Polishchuk et al., 2009), classification of plant species (Cutler et al., 2007), and biomedicine (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). Within the risk analysis domain, RF have been used to predict flood hazard risk (Wang et al., 2015; Albers, Dery, & Petticrew, 2015), model forest fire occurrence (Oliveira, Oehler, San-Miguel-Ayanz, Camia, & Pereira, 2012) and forest fire risk factors (Pierce, Farris, & Taylor, 2012), assist in aircraft system fault detection (Lee, Park, & Jung, 2014), and traffic accident prediction (Lin, Wang, &

Sadek, 2015) and analysis (Siddiqui, Abdel-Aty, & Huang, 2012).

The predictive power of RF, particularly their strength as classifiers and regression tools, shows great potential for application in the risk analysis and transportation modelling fields. Within the context of ferry operations, shipping operations, vessel scheduling, and transportation and marine risk assessment, there are no examples of RF models in the literature, and indeed there is a notable lack of machine learning techniques in general within these domains. Through the use of RF, this research attempts to highlight the predictive power and insight into variable relationships that can be obtained by employing machine learning techniques within the marine operations and risk domain.

Other statistical analysis tools used in this study are the analysis of variance (ANOVA) test, Tukey's honest significant difference test, and the Pearson product-moment correlation. The ANOVA tests the hypothesis that the means between two or more sets of observations are the same. By comparing the response variable means at different factor levels the test hypothesis is determined to be true or false. If false, a comparison method such as Tukey's honest significant difference test can be used to determine which factors cause the test hypothesis to fail. This method calculates confidence intervals for the pairwise differences between factor level means to determine response variable means that are significantly different from each other. The Pearson product-moment correlation measures the strength of linear association between two variables by attempting to draw a best-fit line between the data points and measuring how far all the points are from the line to provide a correlation coefficient. Detailed explanations and formulas for each of these statistical tools can be found in Hayter (2012).

Chapter 2

Data Preparation and Exploration

Three types of data were required to conduct this research: MAI ferry operations data, data on the observed weather factors in the area of the ferry route during the period of study, and data on ice concentration in the area of the ferry route during the period of study. This chapter is comprised of two sections: the first summarizes the source and characteristics of each data type, the key features that were added to enhance the statistical analysis, and the measures taken to handle errors and missing data; the second section conducts an exploratory data analysis on the environmental factors to provide an overview of the conditions in the area of study.

2.1 Data Sources and Preparation

2.1.1 MAI Operations Data

Data on MAI operations were provided by the company in the form of a "traffic data" set (Marine Atlantic Incorporated, 2015). The traffic data set contains the details of each sailing, such as scheduled arrival and departure times, actual arrival and departure times, cancellation status, vessel name, departure port, arrival port, number of passengers, numbers and types of vehicles, etc.

The traffic data set provided by MAI covers the period from 2000 to 2015, however only the period from January 2012 to August 2015 was used. This time window was chosen for two reasons. First, prior to March 2011 MAI was involved in an intense period of recapitalization. In March 2011 all of the new ferries that are in service today had been brought into service and the old ferries had been retired. The new vessels are more capable under challenging environmental conditions than their predecessors and the intention of this study is to assess current operating conditions, so the period in which now-retired vessels operated was omitted. Second, as will be seen in the description of weather data, a large amount of weather data for 2011 is missing, while

the data for 2012-2015 is fully intact.

Furthermore, all sailings on the North Sydney - Argentia route were omitted, as previously explained. Thus, the MAI operations under analysis for this research were limited to the North Sydney - Port aux Basques route for the period January 2012 to August 2015.

The traffic data set provided by MAI contained over 100 fields that described the characteristics of each sailing during the period. Many of these fields were not relevant to this study and were removed. Table 2.1 summarizes the data fields used for the analysis.

Table 2.1: Traffic data set fields relevant to analysis.

Feature	Description
sch_arriva	the scheduled date and time of arrival
sch_depart	the scheduled date and time of departure
act_arriva	the actual date and time of arrival
act_depart	the actual date and time of departure
from_to_po	code descriptor of departure port and arrival port
vessel_cod	code descriptor of vessel
missed	boolean indicator of cancelled sailing
on_time	boolean indicator of on-time sailing
depart_sta	primary reason for delay
delay_reas	sub-reason for delay

Several descriptive features of the traffic data were derived from the fields in Table 2.1, which were used to enhance the analysis. Table 2.2 summarizes the features that were added to the traffic data set.

Table 2.2: Features added to traffic data set

Feature	Description
delta_arr	difference between actual arrival time and scheduled arrival time
delta_dep	difference between actual departure time and scheduled departure times
year	calendar year
month	calendar month
month.pos	month and year
day	day of the month
wday	day of the week
canc	boolean indicator for cancelled sailing
late	boolean indicator for late arrival
ontime	boolean indicator for on-time arrival
status	one of “cancelled”, “late”, or “ontime”

All of the data in the traffic data set was input manually by MAI staff and is therefore prone to data input errors. Some of the errors were observable due to their not complying with the logic of ferry operations (i.e., an arrival time that was earlier than the departure time). Upon inspection it was found that some of these errors could be identified and corrected manually based on the logic of ferry operations, however the size of the data set made this impractical. These errors were therefore addressed by filtering the data through a set of rules based on the logic of ferry operations and deleting the records that were found to be in error (see Appendix A for a complete list of the filtering rules). 11.4% of the traffic data set records were deleted as a result of these rules.

A specific type of data input error was identified but not removed due to an inability to determine the set of affected records. This is an error in either or both of the `act_depart` or `act_arriva` fields, in which these values were input without consideration of the difference in time zone between Nova Scotia and Newfoundland (30 minutes). These errors remain in the data set because there is no practical method of determining which records contain erroneous data. The occurrence of this error is assumed to be random.

Once the fields of interest were extracted and the errors removed, there were no missing values in the data set. A modified, cleansed traffic data set with 5679 records and containing only the fields of interest for the period in question was created.

2.1.2 Weather Data

Weather data were obtained from the National Centers for Environmental Prediction (NCEP) North American Regional Reanalysis (NARR), which is an extension of the NCEP Global Reanalysis, that features very high resolution for the area covering North America (32 km, 45 levels, 8 x daily) (Mesinger, Dimego, Kalnay, & Mitchell, 2006). Data were obtained in NetCDF format (a common packaging algorithm for environmental data), which stores data in grids for a given geographical area. The grids in question for this research were extracted (the Cabot Strait between North Sydney and Port aux Basques). See Figure 2.1 for the region in question and the specific grid points used for this study.

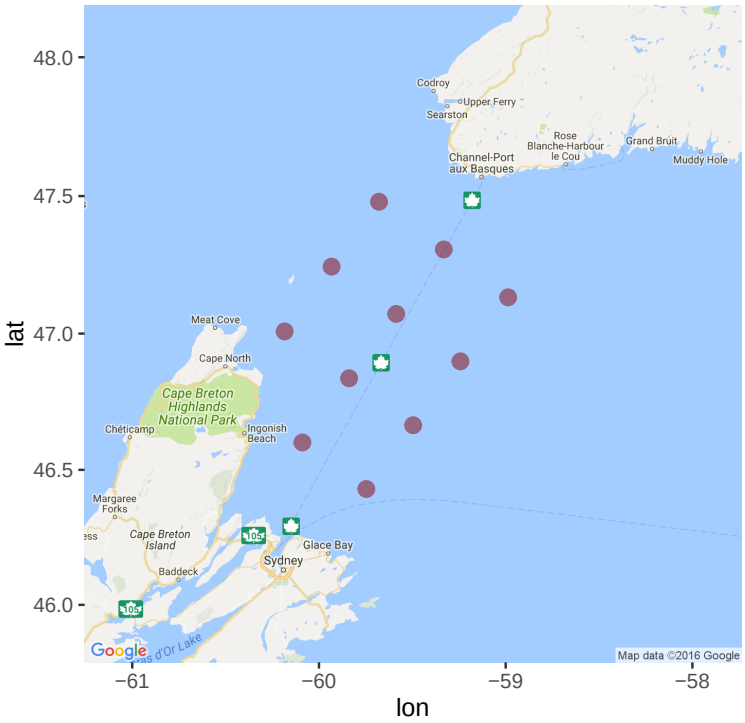

Figure 2.1: Location of grid points.

Weather factors of interest to this research are wind speed, wind direction, atmospheric pressure, air temperature, and precipitation, which, in consultation with the MAI ferry captains, are the factors that are most commonly observed and tracked by the company and forecasted by marine weather forecasts. Table 2.3 summarizes these factors and their units of measurement.

Table 2.3: Environmental factors used in the study.

Factor	Units
Wind speed	kts
Wind direction	degrees
Atmospheric pressure	kPa
Air temperature	°C
Precipitation	mm
Ice concentration	%

For each weather factor, the value at each grid point in the area of interest was extracted for the entire period of interest at intervals of three hours / eight times daily. The mean value across all grid points was then calculated or each time interval. This resulted in a data set with one record for each 3-hour time interval over the period of

study, consisting of one value for each factor.

With the exception of wind, the NARR weather factor values could be used at face value. Wind data required pre-processing due to the manner in which wind data are stored in NetCDF files. Wind data are provided in the form of two vectors, commonly known as **u** and **v**, which are the east-west and north-south components of the actual wind speed and direction, respectively, measured in m/s. To determine the actual wind direction in degrees with respect to due north, the vectors were resolved using Equation 2.1.

$$wind\ direction = \frac{180}{\pi} atan2(\mathbf{u}, \mathbf{v}) + 180 \tag{2.1}$$

Similarly, Equation 2.2 was used to determine the actual wind speed in nautical miles per hour (kts).

$$wind\ speed = \frac{3600}{1852}\sqrt{\mathbf{u}^2 + \mathbf{v}^2} \tag{2.2}$$

A derivation of Equations 2.1 and 2.2 is provided in Appendix B.

2.1.3 Ice Data

Ice data were obtained from the National Ocean and Atmospheric Administration (NOAA) Optimum Interpolation Sea Surface Temperature (OISST) analysis database (Reynolds et al., 2007). Similar to the weather factors listed above, ice data are stored in NetCDF files in gridded format.

The characteristic of ice used for this study was ice concentration, which is the percentage of a given area that is covered by ice. In the case of the Cabot Strait, this always refers to first year ice and not ice that has accumulated over consecutive years because the ice clears out each spring. The values provided by the OISST are daily means. These were extracted for the area of interest over the period of the study and the mean values across all grid points were computed. The OISST analysis provided a complete data set with no missing values. There were no obvious errors in the data set.

2.1.4 Data Matching

A master data set was compiled using the modified traffic data set, the extracted weather factors, and the extracted ice concentration. The weather and ice data records were linked to specific records in the traffic data set based on the scheduled time of departure of a given sailing. In order to account for changing conditions during the period of the crossing, the most adverse value of each environmental factor for the period of the crossing was used. This is to reflect that conditions may often be benign at the time of sailing, but degrade quickly after departure, and ensure the influence of adverse conditions was captured.

The result is a data set that contains 5679 records, one for each scheduled ferry sailing between January 2012 and August 2015, along with all of the relevant operational, weather, and ice information for that sailing.

2.2 Exploratory Data Analysis

This section provides an overview of the environmental conditions in the MAI operating area.

2.2.1 Wind Speed and Direction

Data exploration is more straightforward for wind speed than for wind direction. Wind speed measurements are continuous and linear, which allows for simple calculation of summary statistics and production of histograms, boxplots, etc. Wind direction data is based on compass bearings and is therefore circular in nature. 0° is not the minimum value and 360° is not the maximum value, they are the same. Furthermore, calculation of summary statistics such as mean wind direction may not provide value in many applications. The mean of a directly east wind and a directly west wind is useless, however if the wind is varying between southwest and west over a period of interest, the mean calculation may provide value, depending on the application.

For this application it was decided that choosing a specific point to measure wind direction would provide data more beneficial to the study. Thus, instead of taking the mean of the wind directions for each geographically displaced grid point for every

3-hour period, the grid point closest to Port aux Basques was chosen to provide the wind data. This was decided for two reasons. First, due to the geographical area the the grid points are spread over (almost 16000 km^2), the mean of the wind direction across all grid points for each 3-hour record would have little value. A grid point on the west side of the area could have a westerly wind while another point on the east side could have an easterly wind, the average of which would not yield anything useful. This is in contrast to using the mean of the other factors' grid values, which do provide a meaningful description of the those factors. The second reason for using only the Port aux Basques grid point is that the vessels are much more susceptible to hazards exacerbated by wind in Port aux Basques harbour due to the local geography, as compared to North Sydney or other points along the route. In general, the vessels can navigate in and out of North Sydney and across the Cabot Strait in much more adverse conditions than they can navigate within Port aux Basques harbour.

2.2.2 Other Environmental Factors

For the remaining environmental factors (atmospheric pressure, air temperature, precipitation, and ice concentration), the value for each time period of this analysis was determined by calculating the mean of the 11 grid points within the area of interest. This provides a representative value of each variable across the area of operations.

2.2.3 Summary Statistics

In order to provide an overview of the environmental conditions in the area of operations during the period of the study, summary statistics, histogram plots, and monthly summary plots are provided. Table 2.4 shows the summary statistics all of the enivronmental factors except for wind direction. Figure 2.2 shows the histograms for the same factors, and Figure 2.3 shows the monthly values in boxplot format. Wind direction was combined with wind speed in the polar histogram in Figure 2.4 to show how often the wind originates from each direction at various wind speeds. Additionally, ice concentration in the Cabot Strait varies significantly from year to year. Figure 2.5 shows the trend in ice concentration over the period of the study.

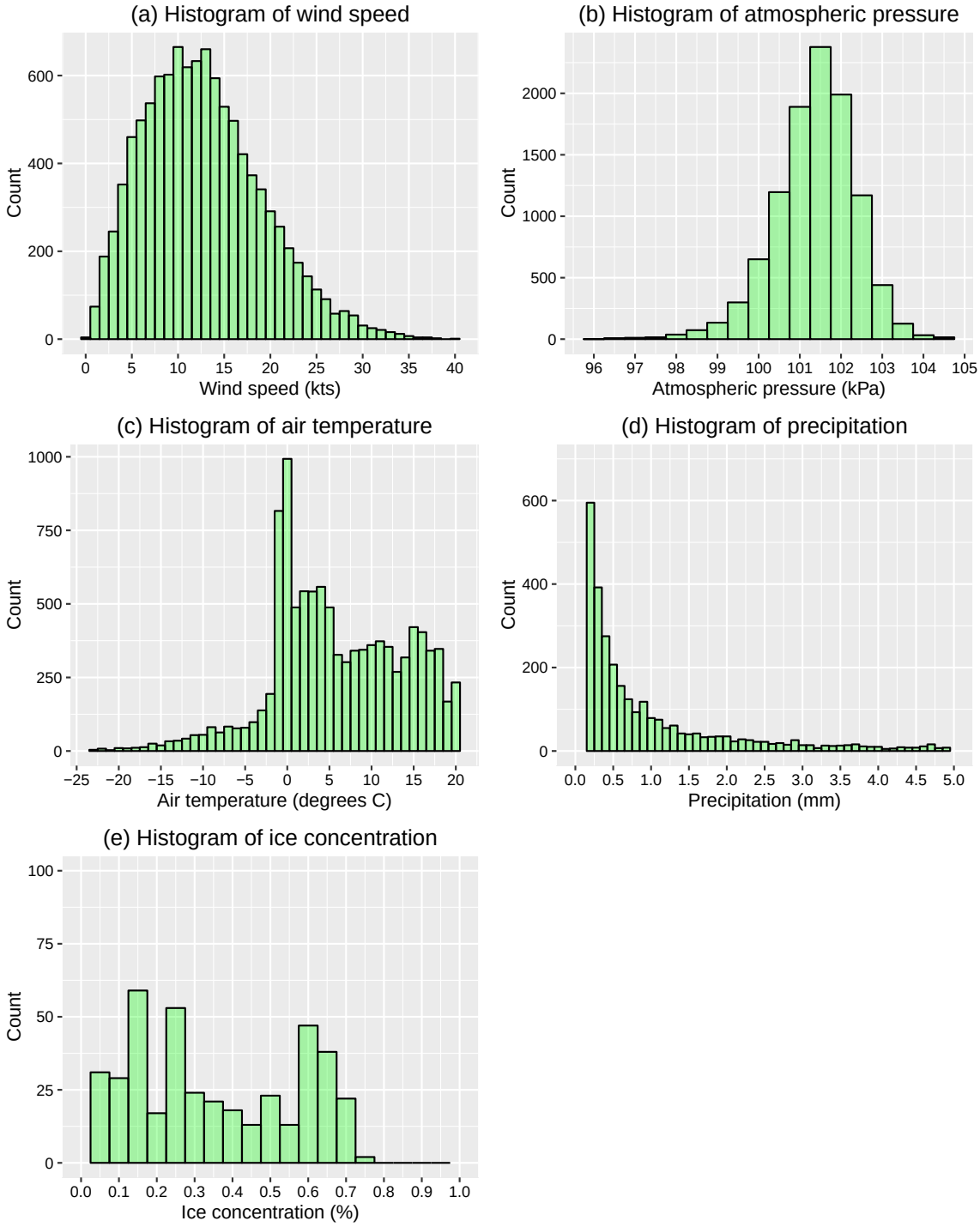

Figure 2.2: Histograms of environmental factors.

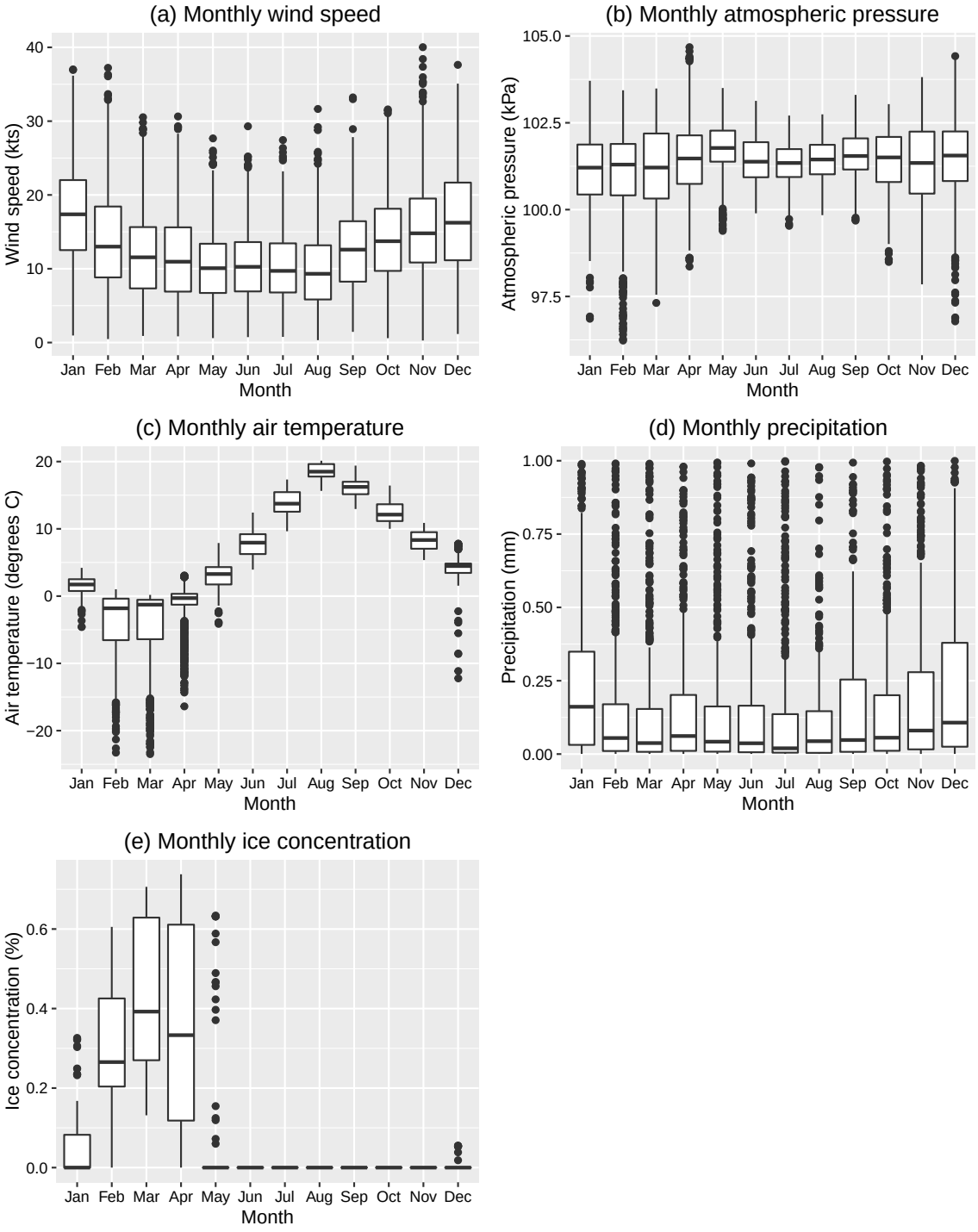

Figure 2.3: Boxplots of environmental factors by month.

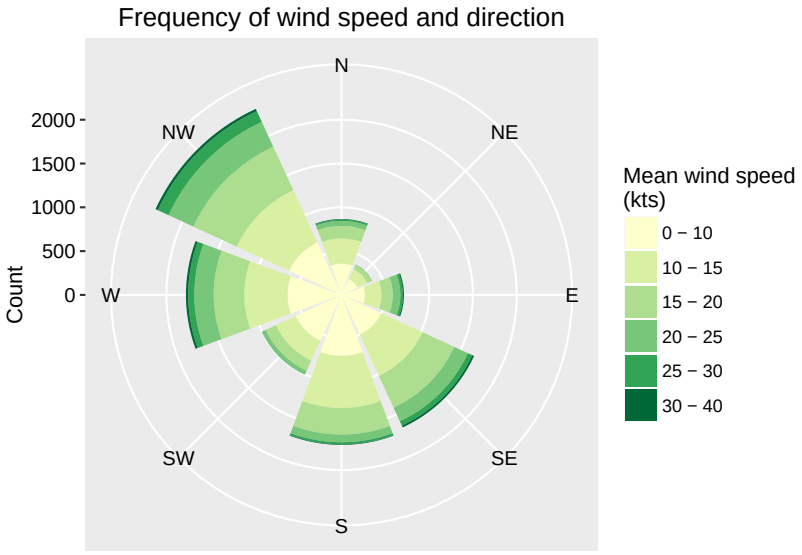

Figure 2.4: Polar histogram of wind speed and direction.

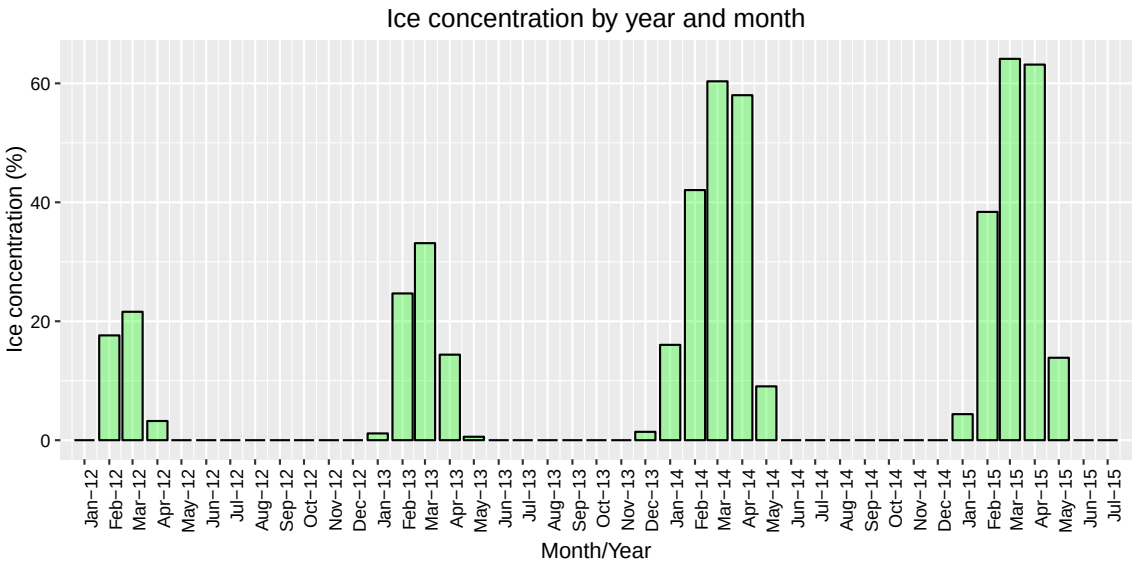

Figure 2.5: Ice concentration, 2012-2015.

Table 2.4: Summary statistics of environmental factors.

	Units	Min	Q1	Median	Mean	Q3	Max
Wind speed	kts	0.274	7.909	12.097	12.738	16.728	40.021
Wind direction	degrees	0.034	147.924	224.154	216.250	297.721	359.928
Atmospheric pressure	kPa	96.231	100.807	101.449	101.371	102.020	104.677
Air temperature	degrees C	-23.483	0.023	4.690	5.846	12.212	20.126
Precipitation	mm	-0.000	-0.000	0.017	0.384	0.214	16.770
Ice concentration	%	0.000	0.000	0.000	10.002	11.769	73.769

2.2.4 Relationships Between Independent Variables

Figure 2.6 shows a heatmap of the correlation between descriptive variables, as well as the correlation coefficients. There is a moderate-strong negative correlation between air temperature and ice concentration, which is most likely explained by the coincidence of low temperatures and the presence of sea ice. Wind speed and pressure both show a weak correlation with precipitation, and the remaining correlations are negligible.

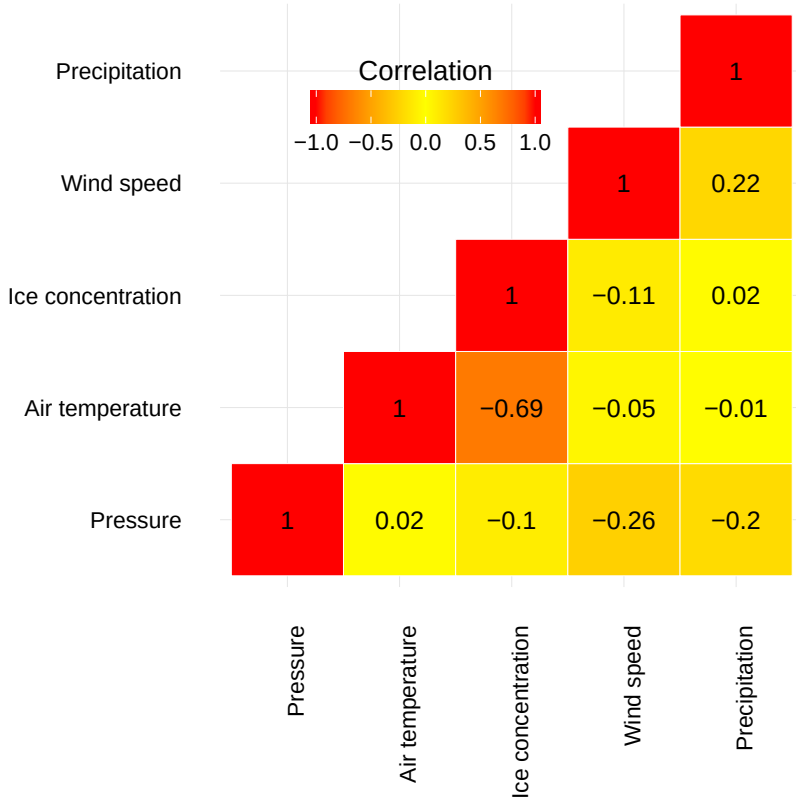

Figure 2.6: Correlation between descriptive variables.

Chapter 3

The Influence of Environmental Factors on Cancellation Occurrence

3.1 Introduction

MAI ferry sailings are cancelled for reasons that make the vessels either incapable of completing a crossing, unsafe to do so, or impractical from a business perspective. According to MAI decision-makers, technical breakdowns cause cancellations but not in great numbers. Staffing and labour issues have also caused cancellations, but these occurrences are rare and often can be predicted within a reasonable time horizon. Environmental factors are the dominant cause of cancellation. Data provided by MAI shows that 90% of the cancellations in the 2014-2015 period were due to environmental reasons (the remaining 10% were for mechanical reasons). High winds can increase the risk of collision or grounding while navigating within tight harbours or during docking. High ice concentrations can cause vessels to become stuck in the ice. Large waves can cause increased discomfort and safety issues onboard the vessel. The presence of precipitation, cold temperatures, fog, and other factors can exacerbate hazardous conditions and increase risk. The presence of these factors and combinations of these factors are all potential reasons for decision-makers to cancel sailings.

3.1.1 Company Expertise and Experience

Discussions were held with MAI vessel captains and operations staff to gain an understanding of company practices with respect to environmental factors and cancellations. In most cases the conditions under which cancelling is warranted are well understood and a sailing will be cancelled if any of those conditions are likely. MAI staff provided expert opinions on various environmental factors and their influence on cancellation, which are summarized in this section.

Wind Speed

Wind speed is the the most common reason for cancelling a sailing. The vessels are designed as ocean-going ferries and can safely navigate in open ocean in almost any wind speed, however entering and leaving harbour, and docking and undocking can be very hazardous in high wind. The risk of collision and grounding increases as the wind increases due to the high sides of the vessels that act like sails, pushing the vessel off course. This phenomenon is especially problematic at the slower speeds used in harbour and while approaching docks. Due to the local geography specific to each harbour and the nature and layout of the ferry docks, different considerations are evaluated for each port. For example, the wind speed is more of a factor in Port aux Basques, due to the small size and narrowness of the harbour, and the presence of a small island in close proximity to the dock. The vessels are designed to be as maneouverable as possible for their size by being fitted with bow thrusters and specialized rudders that assist with stern movement, however tugboat assitance is not available and the vessels must be able to dock and undock independently.

Wind Direction

Certain wind directions are more problematic than others based on the local geography and port setup. For example, MAI staff stated that sailings will typically be cancelled if the the wind is 30 kts or more from a southerly direction, due to the manner in which high winds from that direction are prone to push the vessel off course within Port aux Basques harbour. However, if the wind is from a northerly direction, the threshold is higher, around 40 kts. Typically sailings are cancelled if the wind is 40 kts or higher from any direction.

Wave Height

Wave height was indicated as a factor in decision-making due to its ability to cause discomfort on the vessels. MAI vessels are very seaworthy and capable of safely navigating in large sea states, however the comfort and safety of the passengers, crew, and cargo is reduced as wave height increases. In general wave heights of three metres or more cause a sufficient reduction in onboard safety and comfort to warrant

a cancellation. Wave height is highly correlated with wind speed and direction, so the reasons for cancellation in the presence of large waves is often cited as high winds.

Atmospheric Pressure

Atmospheric pressure is not used as a specific data point when making cancellation decisions. It is a general indicator of weather that is currently occurring or may soon occur, and can be used as a characteristic in assessing the severity of passing storms or the stability of favourable weather. Typically, lower pressure is associated with storms or other adverse conditions, while higher pressure indicates pleasant weather and decent conditions. Rezaee et al. (2016) found that the Laplacian of pressure, which is an indicator of the presence of a passing extratropical cyclone, is a factor in the severity of maritime incidents. It is reasonable to hypothesize that the same factors that increase the severity of maritime accidents may also negatively influence ferry operations, albeit in a different manner. From a decision-making perspective for MAI, if a storm with high winds causes a cancellation, the cancellation will not be due to the low pressure associated with the storm, but due to the result of the storm causing high winds. Therefore, although atmospheric pressure is not specifically used in decision-making, it may be an indicator of cancellation.

Air Temperature

Air temperature is not used as a specific data point when making cancellation decisions. Like atmospheric pressure it often correlates to other weather conditions such as the presence of storms or favourable conditions, but it has no direct effect on MAI operations. Rezaee et al. (2016) found that temperature was a factor in the severity of fishing incidents because of the effect it can have on people doing manual labour. It is reasonable to assume that MAI deckhands are susceptible to the same influences, however they are much better protected onboard large ferries than on small fishing vessels. Low temperatures combined with wind and waves causing sea spray can cause icing on the vessel, which can reduce stability if allowed to build up over time, but MAI staff indicate that this is not a problem onboard MAI ferries. Although air temperature is not specifically used in decision-making, it may yet be an indicator of cancellation due to its correlation with other weather patterns.

Precipitation

Precipitation is typically not used as a specific data point when making cancellation decision. The adverse effects of rain, snow, ice, etc., tend to be limited to a decrease in visibility that can affect vessel navigation, however the ferries are fitted with navigation systems that allow them to navigate in any condition of visibility.

Ice Concentration

Ice concentration was stated as a factor in cancellation decision-making, although it is uncommon for a sailing to be cancelled due to ice. If ice concentrations build up to a point that significantly increases the risk of a vessel becoming stuck in an ice flow, the sailing may be cancelled. The vessels are capable of navigating in first year ice, which is the only ice they encounter in the Cabot Strait, however it is possible for the prevailing conditions to cause areas of very dense ice that can mire the vessels, which happened most recently in March 2015 (Ayers, 2015). In such a case Canadian Coast Guard icebreakers are called to assist. This is not a frequent occurrence because the vessel captains use ice charts for their navigation and the Coast Guard strives to keep paths clear, but in years with large amounts of ice it is sometimes unavoidable. Therefore, if it is likely that long delays will be caused by ice, the sailing will be cancelled until the ice clears enough to allow passage.

Ice concentration was selected to represent the presence of ice for this analysis for two reasons. First, ice concentration data are relatively easy to collect, have good accuracy compared to other ice characteristics (such as ice thickness, which is much more difficult to determine) and are easily available in NetCDF format from various databases. Second, the Cabot Strait experiences the build up of only first-year ice, so the variability of ice thickness is low. Areas with increased ice thickness are typically a result of the wind pushing ice against a coast and causing it to pile up, which can impede navigation but is very difficult to forecast and detect.

Other Factors

MAI staff reported that other environmental factors that are typically forecasted and tracked within the maritime environment are not used as factors in cancellation

decision-making, because their effects on navigation and safety are negligible. These factors include relative humidity, dew point, boundary layer, and tide.

3.1.2 Decision-Making

The decision to cancel is a balancing of two criteria: the ability of the vessel to safely complete the sailing and the comfort onboard the vessel during the sailing. Safety involves avoiding collisions and groundings as well as maintaining the well-being of the people onboard. Onboard comfort is considered because in adverse conditions the passengers and crew may be safe, but extremely uncomfortable, which is important to consider from the context of customer experience (a stated priority of MAI).

Within the context of environmental factors, data used to make the cancellation decision are obtained from various weather prediction and observation services. The principal source for weather data is the Environment Canada Marine Forecast for the Cabot Strait, which covers the area between Cape Breton and the southwest of Newfoundland. The marine forecast provides predicted wind speed and direction, wave height, precipitation, visibility, pressure, and temperature for the region. A secondary weather data source is a local weather station in Port aux Basques harbour owned by MAI. This provides real-time detailed conditions within the harbour that are dependent on the local geography and assists vessel captains with planning their harbour entrance or exit. Ice conditions are provided by the Canadian Ice Sercvice in the form of ice charts, which detail the observed ice characteristics for a given area.

Ice conditions and the weather forecast are continually tracked and evaluated every morning during the company operations meeting, attended by vessel captains and terminal managers. The decision-making is collaborative, but ultimately the decision to sail or not is the responsible of the vessel captain, in accordance with Canadian law. If conditions warrant cancellation, the decision will typically be made 24-48 hours prior to sailing. This allows time to alert commercial customers and passengers that have reservations to make alternate plans. Normally if one sailing is cancelled all sailings in that timeframe are cancelled, which prevents the buildup of multiple vessels in one port.

3.1.3 Cancellation Impacts

According to MAI, cancellation of sailings has financial, traffic congestion, and customer experience impacts. Financial impacts include lost revenue due to not sailing, or the risk of high operating expense if the decision to sail is made but the vessel is long delayed due to adverse conditions. The impact on traffic is manifested by the buildup of commercial traffic waiting to board the vessels, which can sometimes take days to recover from. The impact to customer experience involves both the comfort onboard the vessel and the inconvenience of a sailing being cancelled. The latter has been significantly reduced in recent years, however, since MAI instituted a "Red Alert" system that automatically alerts commercial customers and passengers with reservations by email and text when a cancellation is likely or has occurred. This allows alternate plans to be made and relieves traffic congestion in the terminal parking lot.

3.2 Exploratory Data Analysis

This section explores the nature of cancelled sailings with respect to environmental factors independently.

3.2.1 Data Sources and Preparation

The data sources and preparation used for this analysis are as explained in Chapter 2. The resulting data set consisted of 5679 records, one for each scheduled sailing for the period of the study on the North Sydney - Port aux Basques route. Table 3.1 lists the independent variables used in this analysis. The year and month fields were used to establish trends but were removed for the modelling portion of the study in order to focus solely on the presence of environmental factors. The vessel identifier and departure and arrival port identidier were not used in the analysis based on the way cancellations decisions are made, i.e., they are made for all vessels and routes within a particular period.

Table 3.1: Independent variables used in analysis of cancellations.

IV	Description
`year`	calendar year
`month`	calendar month
`month.pos`	calendar year and month
`ws`	mean wind speed (kts)
`wd`	mean wind direction (degrees)
`pres`	mean atmospheric pressure (kPa)
`air`	mean air temperature (°C)
`precip`	mean precipitation (mm)
`ice`	mean ice concentration (%)

3.2.2 Observations by Month and Year

Of the 5679 sailings, 603 were cancelled, which is 10.62% of all sailings in the data set. Figure 3.1 shows the percentage of cancelled sailings by month over the entire study period. Months with no cancellations are rare while some months have cancellations rates higher than 35%. Figure 3.2 shows the percentage of cancelled sailings aggregated by month over the study period. As expected, the percentage of cancelled sailings is higher during months when adverse environmental conditions are expected.

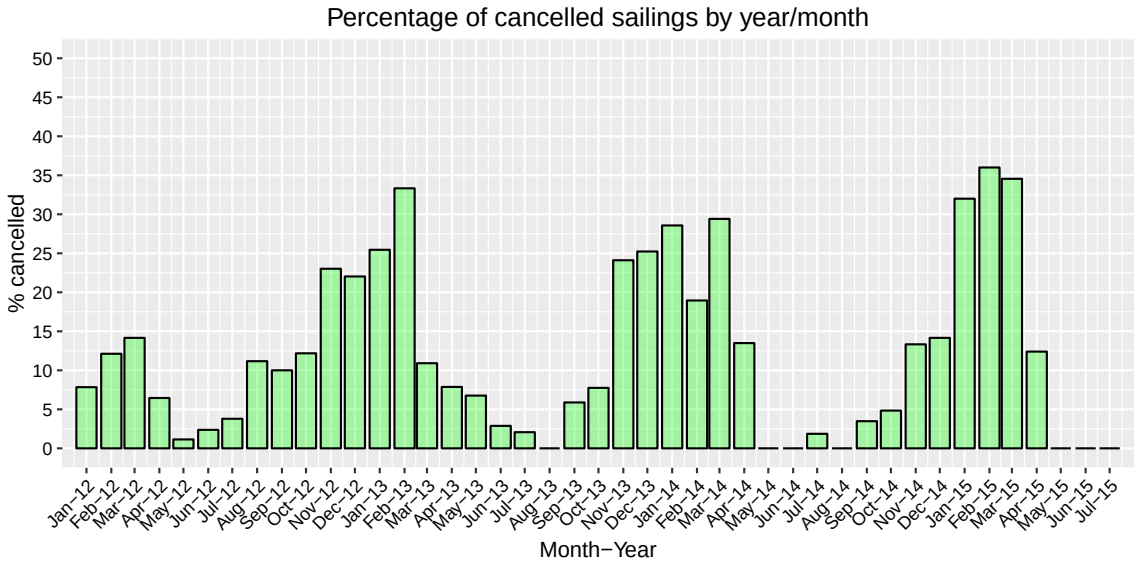

Figure 3.1: Percentage of cancelled sailings by month, 2012-2015.

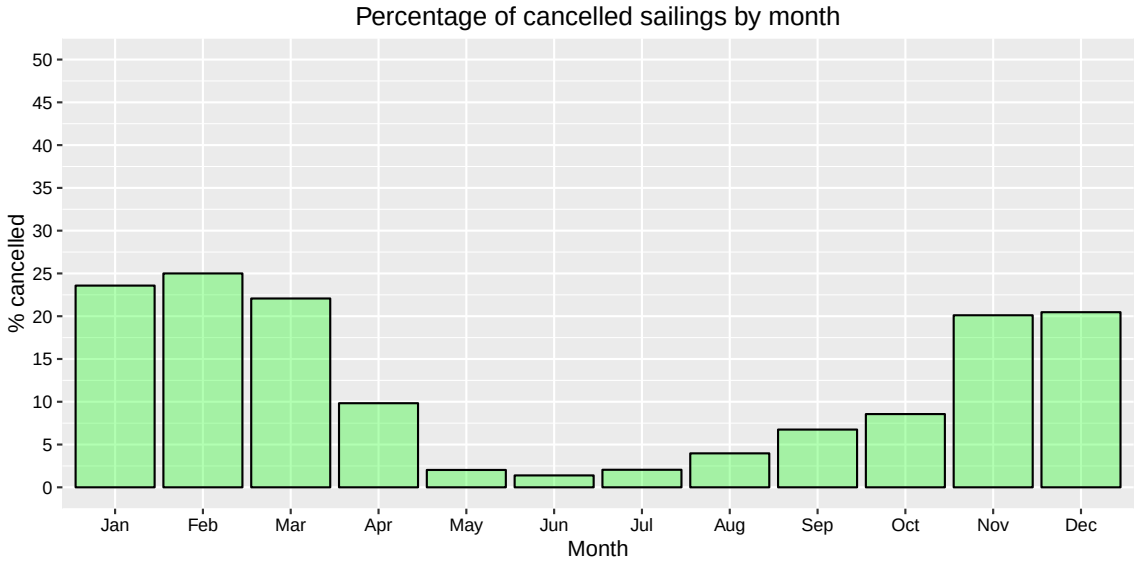

Figure 3.2: Percentage of cancelled sailings aggregated by month.

3.2.3 Observations by Environmental Factors

The plots in Figure 3.3 show the cumulative percentages of cancelled sailings aggregated over intervals spanning the range of each environmental factor. As expected, the percentage increases with wind speed, to the point of all sailings being cancelled once the wind rises above 35 kts, which ties in with statements made by MAI staff. The percentage of cancellations aggregated by wind direction ranges from approximately 6-19%. The higher percentage occurring when the wind is eastlery may be explained by the fact that easterly is not a prevailing wind direction in Atlantic Canada, and typically is only observed in the presence of approaching storms, thus a higher proportion of sailings may be cancelled when the wind is from that direction. Atmospheric pressure appears to have dramatic effect. Almost all sailings are cancelled when the pressure is below 99 kPa, followed by a decreasing trend in cancellatinos as pressure rises. This is explained by the lower pressures that are typically observed in the presence of storms that bring higher winds. The plot of air temperature shows a higher proportion of cancelled sailings when the temperature is between -15°C and 0°C. This is most likely due to the the increase in cancellations in the presence of winter storms, which typically are accompanied by temperatures slightly below 0 in this region, as opposed to the stable winter high pressure systems that are typically accompanied by much colder temperatures but otherwise moderate weather. With

respect to precipitation, typically, low pressure systems that bring strong winds also bring significant precipitation, which may explain the upward trend in percentage of cancellations as precipitation increases. The general upward trend in cancellations as ice concentration increases is to be expected and ties in with statements made by MAI staff.

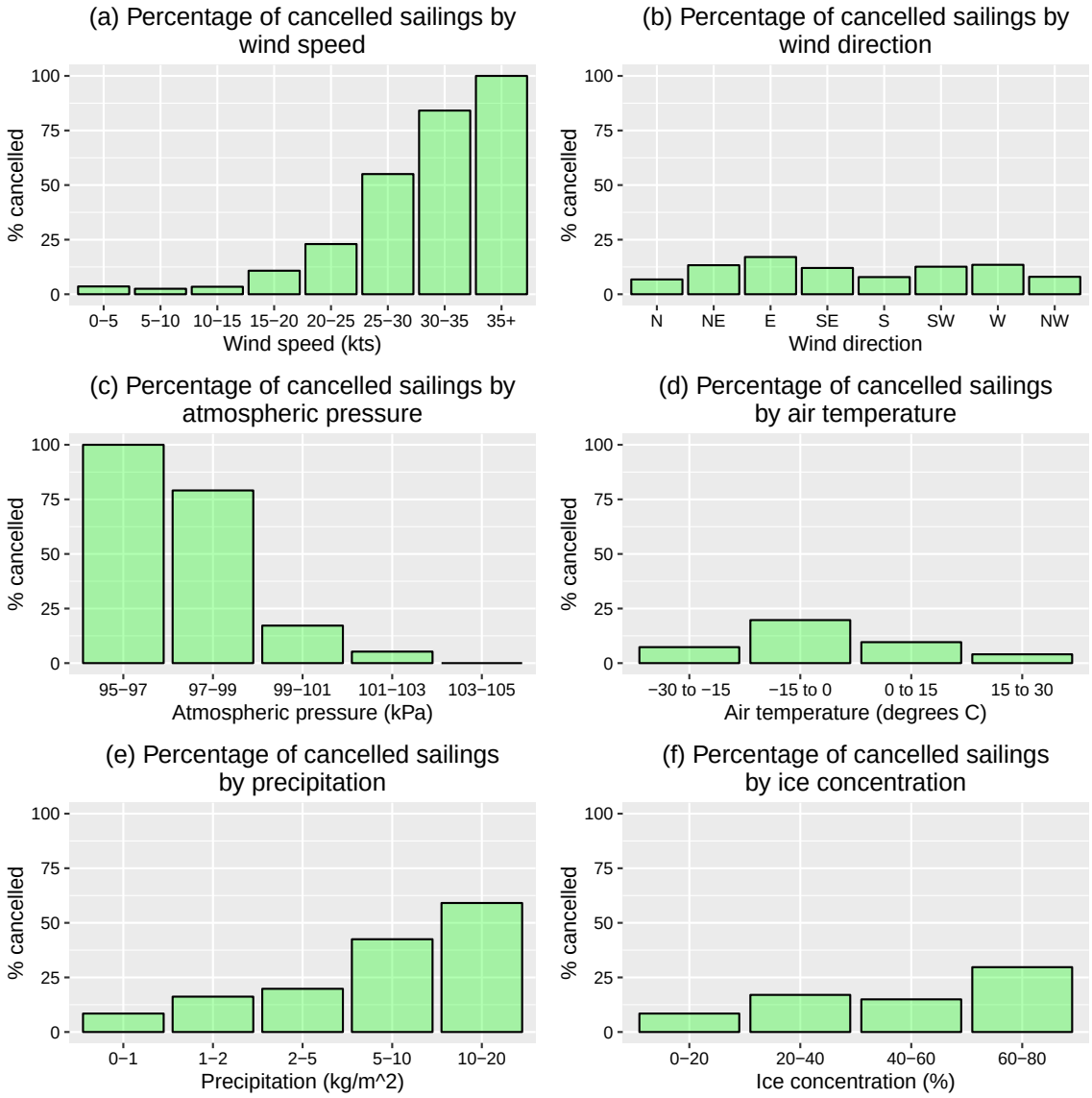

Figure 3.3: Percentage of cancelled sailings aggregated by environmental factors.

Figure 3.4 shows the percentage of cancelled sailings by wind speed and direction in heatmap format. The heatmap supports statements by MAI staff about cancelling sailings as the wind speed approaches 30 kts from southerly directions, and as the

wind surpasses 30 kts from any direction. Note that blank elements in the heatmap reflect a lack of data points for those wind speed and direction intervals.

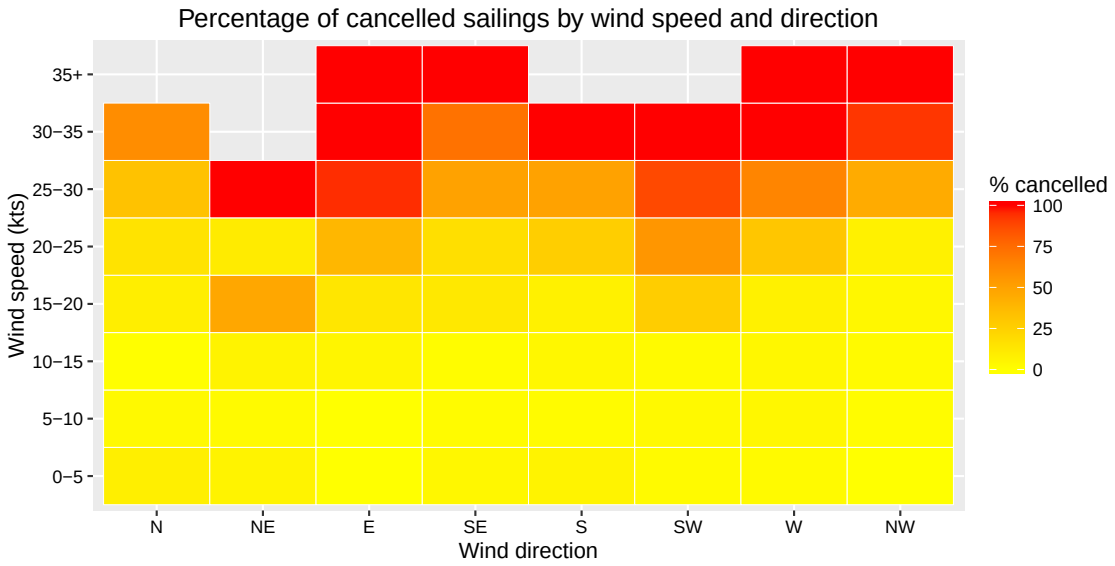

Figure 3.4: Percentage of cancelled sailings by wind direction and speed.

3.3 Modelling

This analysis examines the relationship between the dependent variable `canc` (binary descriptor of whether a sailing is cancelled or not) and the environmental factors described above as independent variables. The exploratory data analysis established trends between the dependent and independent variables as well as limited interactions between variables. In order to analyse the response of the entire set of independent variables, classification modelling techniques were employed.

Several classification models were investigated in order to determine the most suitable approach to adopt, including Logistic Regression (LogReg), Classification Tree (CTree), Gradient-Boosted Trees (GBTree), Linear Discriminant Analysis (LDA), k-Nearest Neighbours (KNN), Support Vector Machines (SVM), and Random Forest (RF). To make the selection, various performance metrics were measured. Several authors have analysed classification model performance metrics, including strengths, weaknesses, and suitable applications of each (Sokolova, Japkowicz, and Szpakowicz (2006), Sokolova and Lapalme (2009)). A summary is provided here in terms of modelling cancellation occurrence.

Accuracy (Equation 3.1) is the ratio of correct predictions to all predictions. The data set is quite unbalanced, however, so the utility of this metric is limited because it is easy for the model to choose "not cancelled" and have a high likelihood of being correct.

$$accuracy = \frac{\#\ correct\ predictions}{total\ \#\ predictions} \tag{3.1}$$

Sensitivity (Equation 3.2) is the ratio of true positive predictions (i.e., "cancelled") to the sum of the true positive and false negative predictions, which provides a measure of model performance in correctly predicting the positive ("cancelled") class.

$$sensitivity = \frac{true\ positives}{true\ positives + false\ negatives} \tag{3.2}$$

Similarly, *specificity* (Equation 3.3) provides a measure of performance for predicting the negative class ("not cancelled"). A comparison of sensitivity and specificity provides insight into the model performance from a class perspective.

$$specificity = \frac{true\ negatives}{true\ negatives + false\ positives} \tag{3.3}$$

Balanced accuracy (Equation 3.4) is the average of the class error rates and provides better insight into model performance for unbalanced data sets.

$$balanced\ accuracy = \frac{sensitivity + specificity}{2} \tag{3.4}$$

The *Kappa* statistic (Equation 3.5) provides more insight into model accuracy. It is a comparison between the predicted accuracy and the expected accuracy (random chance accuracy) and provides a measure of the model's predictive ability compared to predictions made randomly.

$$kappa = \frac{accuracy - expected\ accuracy}{1 - expected\ accuracy} \tag{3.5}$$

AUC is a measure of the model's ability to avoid false classification and is one of the standard metrics for evaluating classification model performance. AUC is the area under the ROC curve, which is a plot of sensitivity (y-axis) against 1-specificity (x-axis). A value of 0.5 represents a random prediction and of 1 is a perfect prediction. Similar to kappa, it provides a metric for evaluating model performance across classes and against random chance.

The *caret* package in R (Kuhn et al., 2016) was used to design and implement a standardized model evaluation to select the best-performing model for this study. Each model was trained on the same training data set with the same random seed. 10-fold cross-validation was used to improve each model's performance and reduce overfitting. Several metrics were measured for each model and are summarized in Table 3.2. Based on the metrics of AUC and kappa, the Random Forest (RF) model exhibited the best performance. The RF model also exhibited the best sensitivity and is therefore the best model at predicting the positive class (in this case, cancelled sailings), which is important for imbalanced data sets.

Table 3.2: Performance metrics used for cancellation occurrence model selection

Model	Accuracy	Sensitivity	Specificity	BalAcc	Kappa	AUC
CTree	0.9240	0.5193	0.9724	0.7459	0.5520	0.8870
LogReg	0.9270	0.4309	0.9862	0.7086	0.5210	0.8950
LDA	0.9290	0.5193	0.9777	0.7485	0.5710	0.8820
GBTree	0.9480	0.6630	0.9793	0.8226	0.7030	0.9010
KNN	0.9300	0.4530	0.9862	0.7196	0.5420	0.8990
SVM	0.9370	0.5138	0.9869	0.7503	0.6000	0.8790
RF	0.9540	0.7238	0.9810	0.8524	0.7430	0.9060

3.3.1 Random Forests

Random Forest (RF) is an ensemble machine learning method developed by Breiman (2001) that takes multiple weak tree-based learners and combines them into a strong learner in the form of an ensemble, or forest, of trees. The result has a synergistic effect, i.e., the final model is stronger than the sum of its parts. In the case of RF, many classification trees are produced, each one with a vote as to the predicted class of the input vector. The votes are aggregated to determine the predicted class of the forest. Whereas individual classification trees are fast and provide easily interpretable output, their predictive performance is low compared to more advanced machine learning methods. RF is typically considered a "black box" method, meaning that gaining understanding of how variables affect predictions and interact is less straightforward, however the predictive performance is much higher than for individual classification trees. Furthermore, methods have been developed to estimate variable relationships to gain knowledge from RF models, more so than for other black box techniques

like support vector machines and neural networks. The RF method developed by Breiman (2001) is summarized here, along with relevant features of the model used in this research.

An ensemble of k tree-based classifiers is created $T_1(X, \theta_1), T_2(X, \theta_2), ..., T_k(X, \theta_k)$ where X is the input vector of independent variables and $\theta_1, ..., \theta_k$ is a set of idependent identically distributed random vectors (taken from the training data set). Each tree determines the output class for the input vector X based on the random vector θ_i, thereby voting for that class. The votes from all the trees are aggregated to determine the output class for the ensemble.

Each tree T_i is grown (or trained) first by taking a random sample (with replacement) of the training data as the bootstrap sample. For each bootstrap, each tree T_i is grown using the CART algorithm developed by Breiman (1984) (i.e., finding the best split at each node from among the predictor variables, X), but using only a random subset of predictor variables at each node (known as random feature selection). Each tree is grown to its maximum extent with no pruning. This is repeated for each tree until a forest of sufficient size has been grown.

Using an independent bootstrap sample for each tree in this context is known as *bagging*. Bagging allows for a method of cross-validation in parallel with the training of each tree. The instances in the bootstrap sample are considered *in the bag* (about two-thirds of the training data), and the remainder are out-of-bag (OOB) (about one-third of the instances). Each tree is grown on its bootstrap (*in-the-bag* sample) and validated on the OOB sample, from which the misclassification error rate can be calculated (known as OOB error). As the forest grows, the OOB error from individual trees is aggregated and a running unbiased estimate of the classification error for the ensemble is maintained. Breiman (2001) demonstrated that as the number of trees increases, OOB error converges, which is why RF models don't overfit as more trees are added. He also demonstrated that this method of validation may occasionally overestimate the error (but only by small amounts), but performs as well as, and in some cases out-performs, other established cross-validation methods.

RF model tuning consists of determining the number of features *mtry* (input variables) to randomly select at each node, and the number of trees *ntrees* to grow in

the forest. If M is the total number of features, $mtry$ is typically $\sqrt{M}$ for classification and $M/3$ for regression, however this value can be modified. Breiman (2001) demonstrated that increasing $mtry$ increases the correlation between trees (and thus increases the OOB error) but also increases the strength of the classifier, which in turn is related to a decrease in OOB error. Thus, an optimum value of $mtry$ can be determined. Cutler et al. (2007) found that $mtry$ had little effect on classifier performance in their work on classification in ecology, while Strobl et al. (2008) found that it had a large effect on their work in bioinformatics and genetic markers. Given the lack of consensus in the literature, $mtry$ will be optimized for this study.

The number of trees $ntrees$ needs to be sufficiently large to allow the OOB error to converge. The default is 500 trees, however there is no penalty for larger numbers aside from processing time. Svetnik et al. (2003) and Polishchuk et al. (2009) found that most often 500 trees is more than sufficient.

Individual classification tree algorithms are known for their ability to identify important features among the independent variables and for producing an easily interpretable model that exlains the interactions between independent and dependent variables. RF retain the ability to identify important features but are limited in their ability to explain variable relationships compared to single tree methods, due the lack of an explicit, interpretable model. The tradeoff, however, is a vast increase in predictive performance.

RF have two methods of determining feature importance. The first is based on the mean decrease in accuracy that is observed as variables are permuted. For each tree in the forest, each variable is randomly permuted in the OOB cases one at a time, and the differences in prediction accuracy from the non-permuted baseline are aggregated to determine the decrease in accuracy caused by the permutation. Important features are identified by a larger reduction of prediction accuracy, while less important features have smaller or negligible reductions in prediction accuracy.

The second method of determining feature importance is based on the mean decrease in gini. In the context of classification trees, gini (also known as gini impurity) is a measure of the purity of the nodes in a tree. The set of predicted outcomes are distributed across the terminal nodes of the tree and the gini for each node is the probability that a randomly selected element of a node is mistakenly classified, which

reaches its minimum (zero) when all the elements in the node are correctly classified. The reduction in gini observed at every split in the tree is added for each variable to determine the variable most effective at reducing gini. For RF models, the gini reduction values are aggregated across all trees to determine variable importance for the ensemble.

Feature importance identifies independent variables that have a greater effect on the model, however they provide no information about how the model reacts across each variable's range. In order to gain further knowledge of the impacts of independent variables three methods are used: comparing the predicted versus actual responses across the range of each variable, determining the partial dependence of each variable, and determining the partial dependence of two variables at once (bivariate partial dependence).

Comparing the predicted versus actual responses across each variable range is easily done with a set of simple plots. For each independent variable of interest two plots are made with the range of the variable on the x-axis: one with the actual values of the dependent variable from the testing data set on the y-axis, and the other with the predicted values of the dependent variable based on the testing data set on the y-axix. A smoothing function is applied to each for ease of viewing, especially when the number of predictions is large, and the two plots can be compared for each variable to demonstrate how closely the model tracks the actual data on a per variable basis.

Partial dependence provides an estimate of the marginal effect of a variable on the class probability for classification models, and on the response for regression models. Plotting the partial dependence for independent variables of interest is a method of visualizing the effects of variables on the prediction. Cutler et al. (2007) provide an excellent explanation of partial dependence plots, which is based on the work of Hastie, Tibshirani, and Friedman (2001), and apply it to RF classification models used in the field of ecology. In summary, a classification or regression function f depends on m predictor variables $X = (X_1, X_2, ..., X_m)$, such that $f(X) = f(X_1, X_2, ..., X_m)$. The partial dependence of f on variable X_j is defined as the expectation of f with respect to all variables except X_j, or, $f_j(X_j) = E_{X_{-j}}[f(X)]$, where X_{-j} is all variables except X_j. This is estimated by iteratively fixing values of X_j over the range of X_j and averaging the prediction function over all the combinations of observed values of

the remaining predictors.

For partial dependence of classification models there is a prediction function for each class. If $p_k(X)$ is the probability of prediction of the k^{th} class, then the response function for class k is

$$f_k(X) = \log p_k(X) - \frac{1}{K}\sum_{i=1}^{K} \log p_i(X) \tag{3.6}$$

which, for the case when there are two classes (such as the cancellation occurrence model) and p is the probability of a sailing being cancelled, reduces to

$$f(X) = 0.5 \log(\frac{p(X)}{(1-p(X)}) = 0.5 \operatorname{logit} p(X) \tag{3.7}$$

The scale of the y-axis on the partial dependence plot for classification is then half of the logit probability of the class, or for the cancellation occurrence model, half of the logit probability of a sailing being cancelled.

Finally, partial dependence can be extended to two variables to provide the bivariate partial dependence, i.e. the conditional expectation of function $f(X)$ with respect to all variables except X_j and X_l. Bivariate partial dependence plots are three-dimensional plots to estimate the interactions between two variables on the response. In theory higher order partial dependence can be determined, however visualization becomes extremely challenging and very computationally intensive.

3.3.2 Model Development

To examine the relationship between environmental factors and the cancellation of ferry sailings, a RF model was formulated in R using the *randomForest* package (Liaw & Wiener, 2002), the environmental factors previously described, and the binary response variable *canc*, which denotes that a sailing was "not cancelled" (0) or "cancelled" (1). The model variables are summarized in Table 3.3.

The model was trained on a training set consisting of 90% of the original data set and tested on a testing set consisting of the remaining 10%. Breiman (2001) maintains that the procedure for OOB error calculation means that RF models are not at risk of over-training. Furthermore, Millard (2015) found that for RF models the OOB error decreases and classification accuracy increases as the size of the training data

Table 3.3: Variables used in cancellation model formulation.

`canc`	dependent	categorical (binary)
`wind speed`	independent	continuous
`wind direction`	independent	categorical
`atmospheric pressure`	independent	continuous
`air temperature`	independent	continuous
`precipitation`	independent	continuous
`ice concentration`	independent	continuous

set increases. Therefore, a high ratio of training data set size to testing data set size was used for the model.

The number of trees *ntrees* was set as 100 and *mtry* (number of features to randomly select at each node) was determined by finding the minimal OOB error for the ensemble over the range of possible values. The plots in Figure 3.5 are representative plots based on a specific random seed that show (1) for this model the minimum OOB error occurs when *mtry*=5, and (2) the effect of the number of trees on the OOB error. Values of *ntree* greater than 50 produce a stable minimum error.

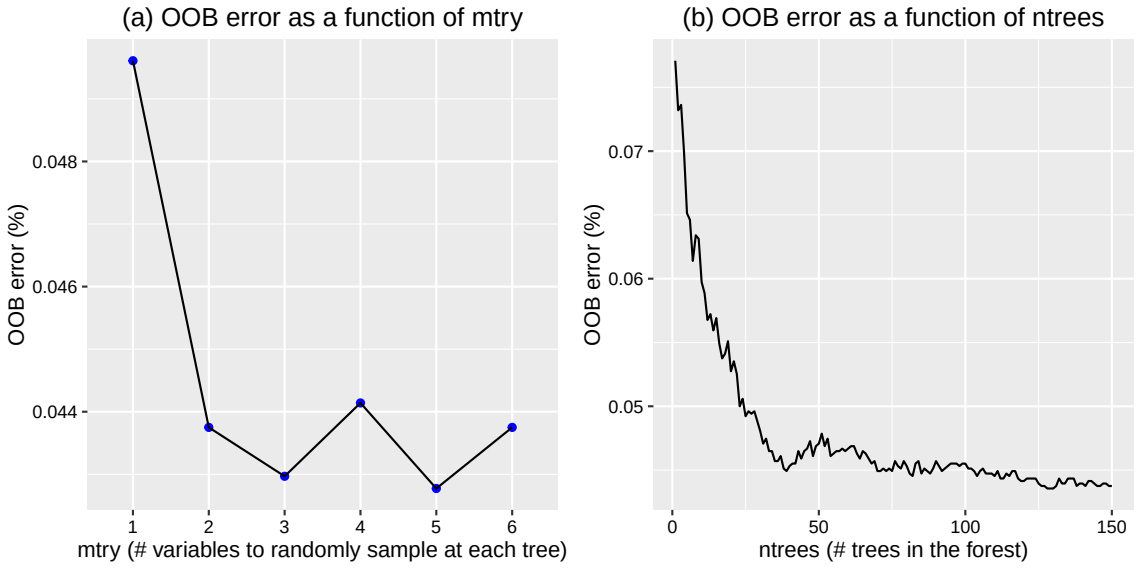

Figure 3.5: Cancellation occurrence model OOB error as a function of number of variables and of number of trees in the forest.

3.4 Model Performance

Table 3.4 shows the key performance metrics of the constructed model. The high measure of specificity is no surprise given the imbalance of the data set. The sensitivity score of 0.7333 is satisfactory given that only environmental factors are being considered in this model, and cancellations also occur for reasons that are not related to environmental conditions.

Table 3.4: Cancellation occurrence model performance.

	Accuracy	Sensitivity	Specificity	BalAcc	Kappa	AUC
Model 1	0.9613	0.7333	0.9882	0.8608	0.7788	0.9079

With this in mind, a new feature was added to the data set. The `canc` field was examined for cases where it was 1 and recoded to 0 if the environmental conditions during the planned voyage duration were benign. This decision was made based on the extreme improbability of benign environmental conditions being the cause of a cancellation. In other words, cancellations that occurred during benign conditions were assumed to be caused by another, non-environmentally related reason. These recoded records were kept in the data set as non-cancelled sailings (as opposed to deleting them) because the model is based purely on environmental conditions and it is important for the model to "learn" that moderate or benign environmental conditions are not a cause of cancellation. For example, if a cancellation occurred because of a mechanical breakdown, but would the sailing would have otherwise not been cancelled given the environmental conditions at the time, it is important for the model to understand that those conditions are satisfactory and would not have caused a cancellation.

Based on MAI staff, environmental reasons for cancelling a sailing are wind and/or ice conditions. Thus, minimum thresholds of 20 kts and 10% ice concentration (Model 2), and 25 kts and 10% ice concentration (Model 3) and were used to recode *canc*. Specifically, *canc* was recoded to 0 if $canc = 1$ and $wind\ speed < 20$ kts and $ice\ concentration < 10\%$ for Model 2, and recoded to 0 if $canc = 1$ and $wind\ speed < 25$ kts and $ice\ concentration < 10\%$ for Model 3. Table 3.5 shows the improved model performance of Model 2 over Model 1, and Model 3 over both

Models 1 and 2, based on the recoded "reduction" in sailings that were cancelled for non-evironmental reasons.

Table 3.5: Cancellation occurrence model performance using increased wind speed and ice concentration thresholds.

	Accuracy	Sensitivity	Specificity	BalAcc	Kappa	AUC
Model 1	0.9613	0.7333	0.9882	0.8608	0.7788	0.9079
Model 2	0.9771	0.8163	0.9923	0.9043	0.8478	0.9866
Model 3	0.9824	0.8750	0.9888	0.9319	0.8391	0.9956

3.5 Results and Discussion

In this section the various aspects of the effects of the independent variables on model outcomes are explored.

3.5.1 Variable Importance

As discussed previously, RF have two methods of evaluating variable importance, one that results in a score for the mean decrease in accuracy, and the other a score for the mean decrease of gini. Figure 3.6 shows the variable importance evaluated by both methods.

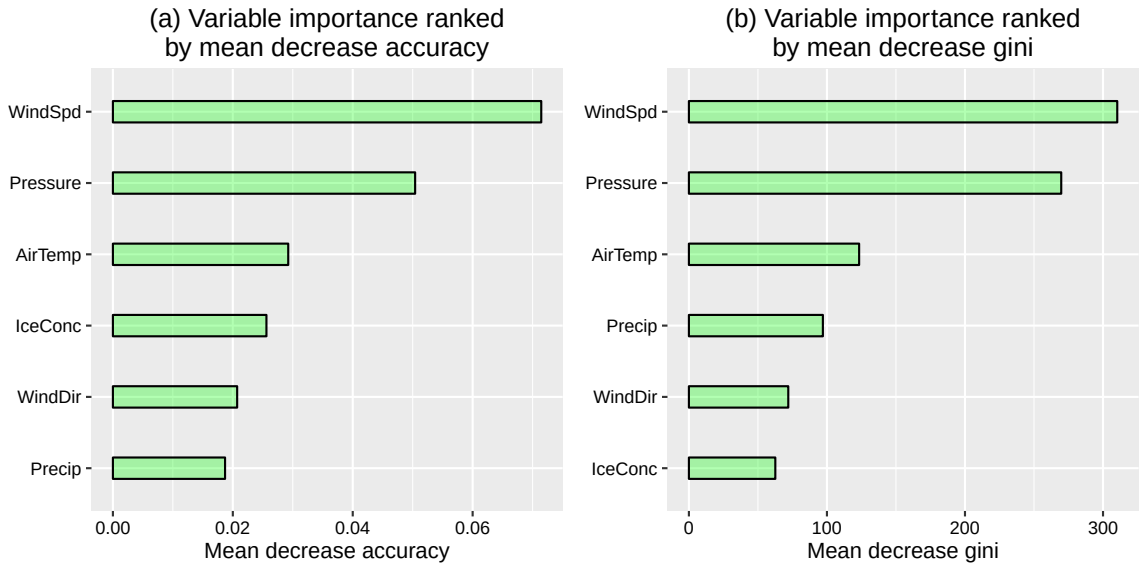

Figure 3.6: Cancellation occurrence model variable importance by mean decrease accuracy and mean decrease gini.

The methods are in agreement on the order of importance and the relative differences in importance between the three most important variables. There is some disagreement in the order of the three least important variables, however both methods agree on which variables are of lesser importance. This discrepancy is most likely due to the method in which each importance calculation is made (described previously). The mean decrease of gini method is known as a quick estimation that generally has good agreement with the mean decrease in accuracy method for well-performing models, however the mean decrease in accuracy method is the standard to be used in the case of discrepencies (Breiman, 2001).

The model found that wind speed is the most important environmental variable in predicting cancellations, which supports the statements made by MAI staff on how wind speed affects operations. Pressure is the second most important variable, most likely due to its indication of the presence of storms and associated higher winds, however its prediction ability is limited because high winds can occur at any pressure level. Air temperature, precipitation, and ice concentration were found to be less important, i.e., sailings are generally not cancelled for extreme values of these factors. Ice concentration and wind direction were expected to have higher importance. However, high ice concentrations can cause a cancellation but according to MAI staff, this is rare and could be difficult for the model to detect. Similarly for wind direction, high winds can originate from any direction and, as a storm passes, the wind typically clocks through various directions (sometimes up to 270°), which could make it difficult for the model to detect.

3.5.2 Variable Responses

Figure 3.7 provides a comparison of the predicted and actual responses over the range of each variable. The curves were generated by applying a smoothing function to the actual and predicted responses (either 0 or 1) over the range of each variable for ease of viewing. This provides an estimation of how closely the predicted responses are to the actual responses on a per variable basis. The variables with higher importance (wind speed and pressure) track more closely to the actual predictions, however even the less important variables track closely, indicating a well-performing model.

3.5.3 Variable Partial Dependence

Figure 3.8 shows the partial dependence of each of the independent variables, in order of importance, referenced to the "cancelled" class. The plots show the range of each environmental factor for which the probability of predicting the cancelled class are highest, independently (i.e., not accounting for interactions).

The probability is higher as the wind increases above 20 kts, the pressure drops below 99 kPa, the air temperature is less than 5°C, or if the precipitation is greater than 5 mm. The wind direction shows slightly higher chances of predicting the cancelled class when the wind is from the northeast, however no wind direction stands out from the others. As would be expected, the ice concentration shows a sharp increase as it increases from 0, however the subsequent slowly increasing trend illustrates only low to moderate partial dependence of ice concentration until the values are quite high.

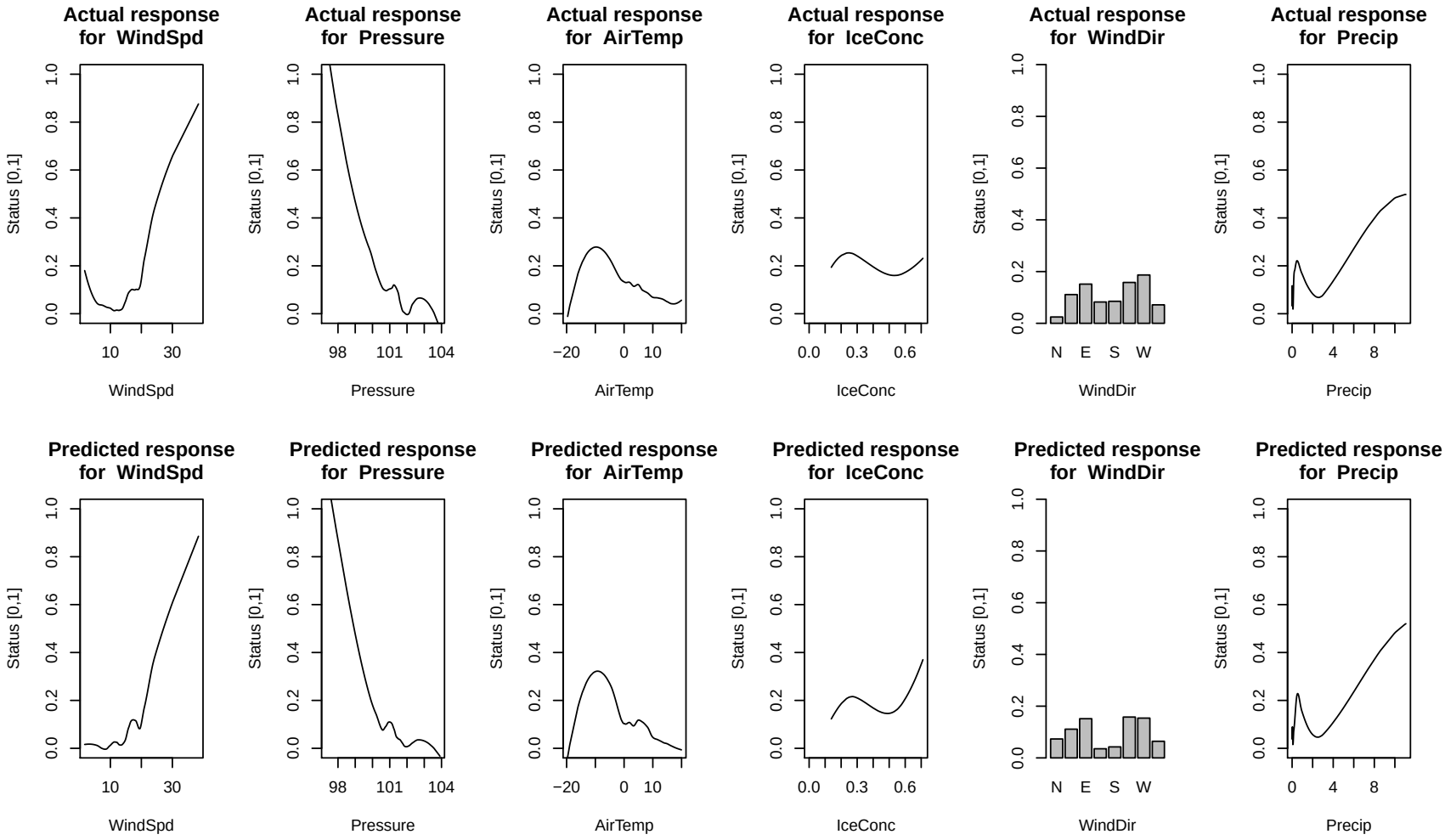

Figure 3.7: Cancellation occurrence model predicted vs actual responses for each variable.

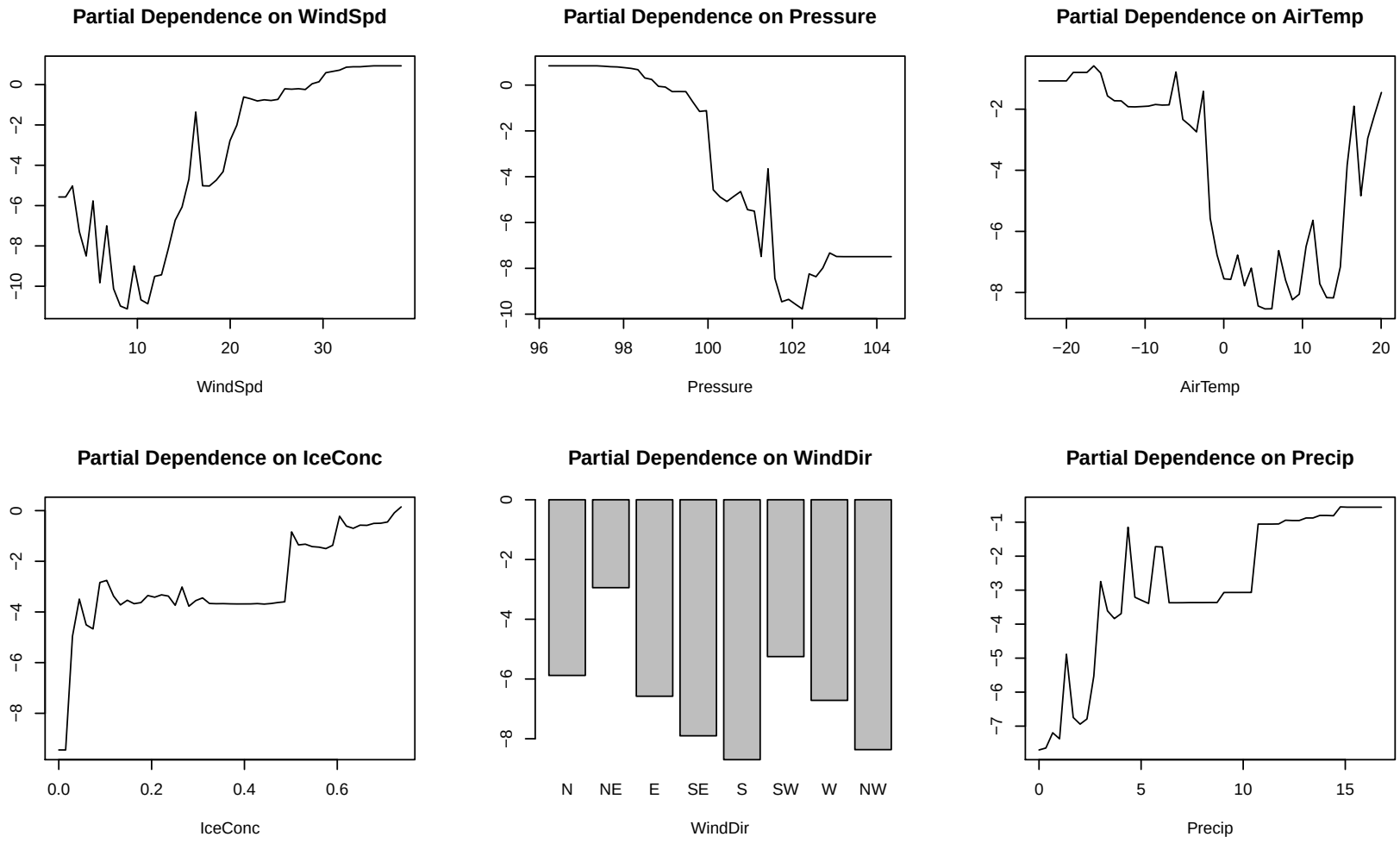

Figure 3.8: Cancellation occurrence model partial dependence of independent variables.

3.5.4 Bivariate Partial Dependence

Figure 3.9 shows the bivariate partial dependence plots, which estimate interactions between two variables, for combinations of the three variables of highest importance (wind speed, pressure, and air temperature).

The plot of wind speed and pressure shows a high probability of predicting a cancellation when pressure is low even when the wind speed is low-moderate (i.e., wind speed that would normally not cause a cancellation alone). This may be explained by a drop in pressure being an indicator of oncoming adverse weather. The plot also shows that when winds are low-moderate and pressure is moderate-high, the probability of predicting a cancellation is much lower, but increases quickly as wind speed increases, regardless of pressure.

The plot of wind speed and air temperature has a similar form but lower maximum probabilities. The importance of wind speed remains evident, and low air temperature increases the probability even over lower wind speeds, perhaps due the correlation with the presence of ice. Again, over low-moderate wind speeds and moderate-high temperatures, the probability of predicting a cancellations is lower. The probability is highest when temperature is 5-10°C and the wind is 35-40kts, however it drops off slightly as the temperature either increases or decreases.

The moderately strong influences of air temperature and pressure are again reflected in the third bivariate plot, as is the lower probability at more moderate levels. The spike at the higher range of air temperature may indicate cancellations during the busy summer months that are caused for reasons linked to an increased demand on equipment, infrastructure, and personnel.

3.5.5 Model Run-Time

The RF cancellation model was run on an Apple Macbook Air with a 1.7 GHz Intel Core i7 processor and 8 GB of RAM, which was more than capable of running the model. Model training times of less than 20 seconds were typical, and the next longest processing times were due to the bivariate partial dependence plot computations, which each took approximately 8 seconds.

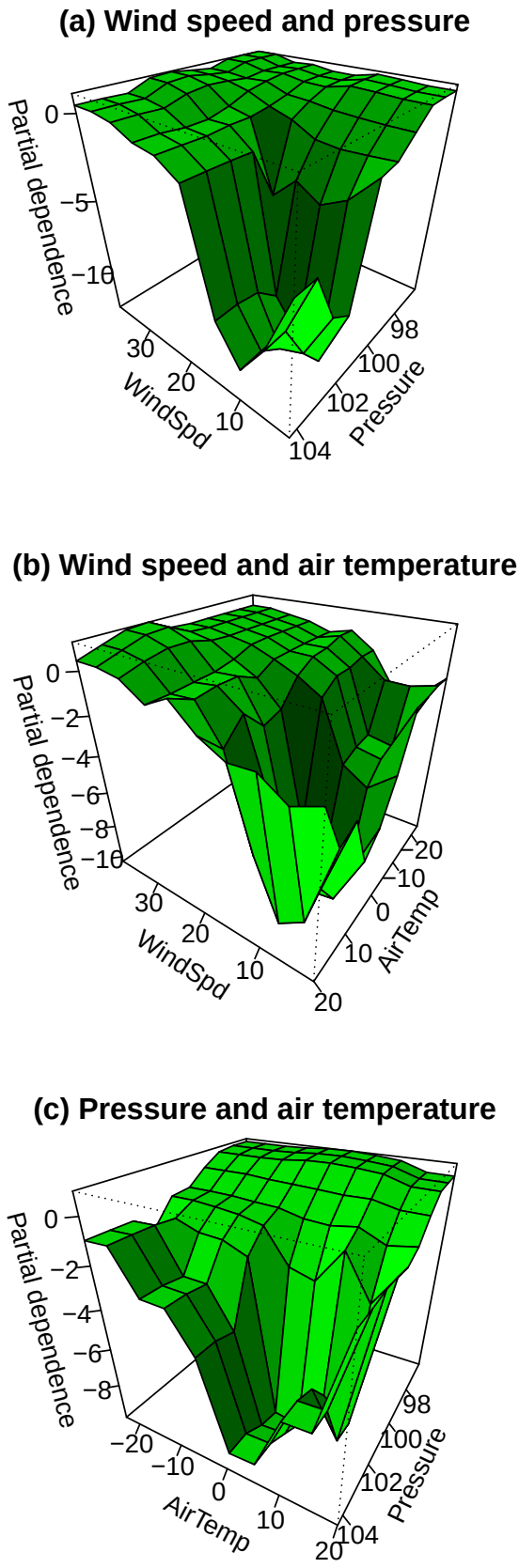

Figure 3.9: Cancellation occurrence model bivariate partial dependence of wind speed, pressure, and air temperature.

Chapter 4

The Influence of Environmental Factors on Delay Occurrence and Length

4.1 Introduction

MAI ferry sailings that are not cancelled are either early, on-time, or late departing and/or arriving. The arrival status is of greater interest to this study because that is the final status result of a particular sailing. The company defines late or delayed sailings as departing and/or arriving more than 15 minutes (0.25 hours) later than the scheduled departure or arrival time. Sailings can be delayed for many reasons and the company tracks these in the traffic data set when a delay occurs. Table 4.1 summarizes the reasons, divided into reasons and sub-reasons.

Table 4.1: Reasons for late departure and arrival.

Reason	Sub-reason
IT Systems	Other, IT Systems
Port Operations	Other, Loading Delays, Extra Lashings, Discharging Delays, Bunkering Operations, Waiting for Passenger Count, Security Related Delays
Assets	Vessel Ramps, Vessel Machinery, Other, Terminal Ramps
Environmental Factors	High Winds, Heavy Sea Conditions, Heavy Sea Ice, Heavy Snow, Other, Poor Visibility
Human Factors	Other, Vessel Staff not Available, Medical Emergencies, Unruly Passengers
Safety	Other, Vessel Detained (Safety Inspection Certificate)

The previous chapter focused on the effect of environmental factors on sailing cancellation. This chapter focuses on the status of non-cancelled sailings, so the cancelled sailings were removed from the data set. Of the remaining 5083 records in the data set, 1197 were delayed (23.55%). Of these delayed sailings, a significant number are believed to be caused by environmental factors. The traffic data provided by MAI contains some degree of tracking of the reasons for delays, but this data field was found to be somewhat inconsistent and was omitted from the modelling portion

of this analysis (for example, some delayed sailings had no delay reason while some on-time sailings had delay reasons). As a rough guide, however, Figure 4.1 shows the relative frequency of delay reasons, for the tracking that does exist. As compared to the analysis on cancellations, in which environmental factors accounted for 90% of the cancellations, Figure 4.1 shows that environmental reasons may account for only about one third of delay reasons. Almost two-thirds of the delays are caused by other reasons, so it is expected that the analysis of delays with respect to environmental factors on sailings being on time or delayed will be more challenging than the analysis of cancellations.

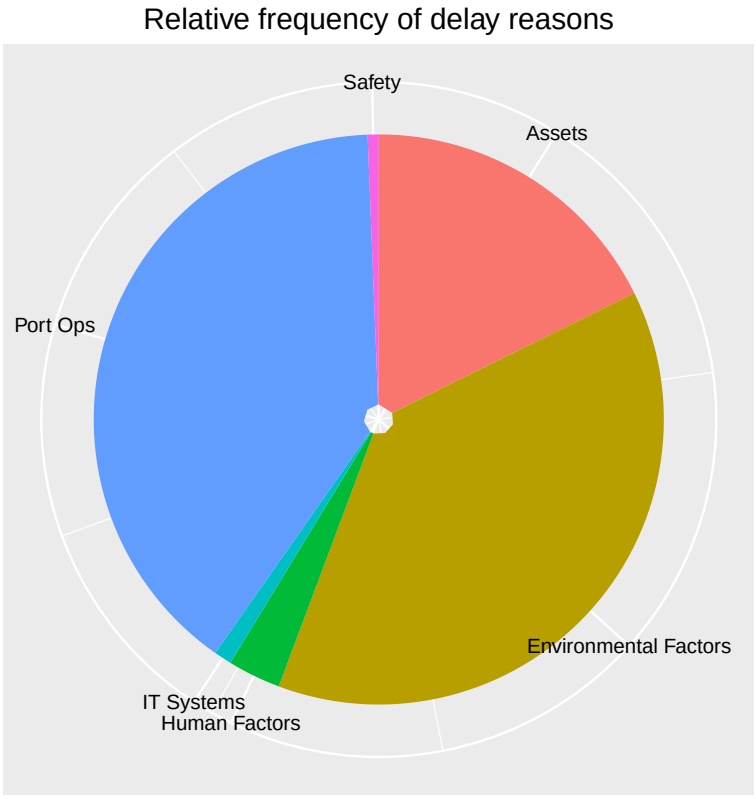

Figure 4.1: Relative frequency of delay reasons.

4.1.1 Company Expertise and Experience

Similar to the approach taken for the analysis of sailing cancellations, MAI staff were consulted on the effects of environmental factors on sailings being delayed or on time. Unlike sailing cancellations, however, in which there is a concsious decision-making

process by MAI staff, delays occur despite the company's best efforts to have ferries run on time. In general the company does not assess environmental conditions and choose to delay a ferry, (however this does happen occasionally if the benefits of delaying outweigh those of cancelling), but will endeavour to have the sailing be on time, and it is then subject to the external forces.

The environmental factors that can cause delays are the same as those that cause cancellations, albeit with different effects. These are explained here in terms of the experience of MAI staff.

Wind Speed

Wind speed can cause delays by increasing the difficulty of navigation in departing or entering harbours and maneouvering when undocking or docking in the ports. Once the vessels are at sea, however, wind speed itself becomes less of a factor because the the vessels are of sufficient displacement and have sufficient power to maintain an intended course and speed under most conditions. At sea, however, wind speed can indirectly cause delays if the conditions support the formation of large waves.

Wind Direction

Like wind speed, wind direction may be a factor in causing delays in departing or entering harbour or undocking and docking. Due to local geography, wind direction affects each port differently, which makes these activities more challenging, especially as the wind speed increases. At sea, wind direction is only an indirect factor in that certain directions support the formation of large waves.

Wave height

Wave height can cause delays through the hydrodynamic effects of larger waves on a ship at sea. Larger waves require more power to maintain a desired course and speed, and increasing speed may not be a suitable option due to the increase in onboard discomfort that can result as the ship's speed increases. Wave height is effected by the wind speed (higher wind, higher waves), the wind direction (in relation to the local geography: fewer land obstructions, higher waves), the fetch (the area over which waves have to build up: longer distance, higher waves), and the time that

favourable conditions exist for the formation of waves (longer time, higher waves). For this analysis wave height was not used as a specific data point due to the lack of historical data. Wind speed was used as a proxy given its relationship to wave height.

Atmospheric Pressure

Atmospheric pressure itself does not directly cause delays, however the weather that occurs during the presence of low pressure systems, such as high wind and waves, can cause delays.

Air Temperature

Air temperature may cause delays due to the effects of extreme temperatures on labouring personnel and equipment. For example, in order to maintain safety personnel working on the loading dock in very hot or very cold temperatures may work at a reduced rate, which could delay departure. Similarly, mechanical equipment may have trouble functioning in very cold temperatures, and electrical equipment can malfunction in hot, humid conditions. In particular, the hydraulic ramps used for loading and unloading vehicles are prone to malfunctions in cold conditions.

Precipitation

Similar to air temperature, precipitation may cause delays if large amounts of precipitation have a detrimental effect on the personnel or equipment.

Ice Concentration

Ice concentration causes delays either through the ship navigating through ice itself, or trying to find alternative routes around areas of high ice concentration. The vessels are all designed to navigate in first year ice, however forward progress is slowed as ice concentration increases. In rare cases, the vessels can even become trapped in the ice for periods of time. These cases often require assitance from a Canadian Coast Guard icebreaker to become free. First year ice can be a challenge to predict because it can form and move quickly and can be pushed around by the wind and can pile up in certain areas (usually near land obstructions).

Other Factors

Poor visibility due to fog, mist, precipitation, etc., was stated as not being a signficant cause of delay. Vessels may occasionally proceed at slower speeds in conditions of severely reduced visibility, however they are designed to operate in all conditions of visibility without restriction. Humidity, dew point, boundary layer, and tide also have no effect in practice and are thus omitted from the study.

4.1.2 Impacts of Delayed Sailings

The impacts of delayed sailings tend to be less severe than for cancelled sailings. This is because the revenue stream is not interrupted and customers are not required to rebook. In some cases a delayed sailing can cause a chain reaction; if it arrives late it may not have time to offload and reload completely before its next scheduled departure time, but this was not expressed as a major concern. The most important impact of delayed sailings is a reduction in customer satisfaction that can result from late arrival. Customers tend to be more accepting of this when the delay is caused by environmental reasons, because it is somewhat out of the company's ability to control.

The remainder of this chapter is divided into three sections. The first is an exploratory data analysis of the occurrence of delays and delay length. The second section explores delay occurrence in detail in a similar fashion to the analysis of the occurrence of cancellations conducted in Chapter 3, whereby RF modelling was used to explore deeper relationships and interactions between variables. The third section explores the length of delays using the same method, however a slightly different approach is taken because delay length is continuous variable.

4.2 Exploratory Data Analysis

The data sources and preparation used for this analysis are the same as presented in Chapter 2 and are similar to those used in the analysis of cancelled sailings in Chapter 3. After removing the cancelled sailings from the data set, the resulting data set consisted of 5083 records, one for each scheduled (and completed) sailing for the period of the study on the North Sydney - Port aux Basques route. Table 4.2 lists

the independent variable used in this analysis. For this analysis the identifier of each vessel was included because delays are specific to the scheduled sailing and vessel. This identifier was not used for the cancellation analysis because the cancellation of one sailing would cause the cancellation of all sailings on a certain route within that time period due to limitations on port capacity. Delays are different in that a delay on one vessel/sailing does not imply a delay on another vessel/sailing.

Table 4.2: Independent variables used in analysis of delays.

IV	Description
`year`	calendar year
`month`	calendar month
`month.pos`	calendar year and month
`vessel_cod`	vessel identifier
`ws`	wind speed (kts)
`wd`	wind direction (degrees)
`pres`	atmospheric pressure (kPa)
`air`	air temperature (°C)
`precip`	precipitation (mm)
`ice`	ice concentration (%)

The year and month fields were used to establish trends but were removed for the modelling portion of the study in order to focus solely on the effects of environmental factors.

The analysis of delays is approached in two ways. The first is the occurrence of a delay, which is represented as a binary factor in the data set. A sailing is considered to be delayed if it arrives more than 15 minutes (0.25 hours) later than the scheduled arrival time. The second is the length of a delay, which is a continuous variable measured in hours. A positive value means the vessel is late departing or arriving, and a negative value means the vessel was early departing or arriving (0 means exactly on-time).

General Observations

Of the 5083 non-cancelled sailings, 1197 were delayed, representing 23.55% of all sailings during the period. Figure 4.2 shows the percentage of delayed sailings and boxplots of the length of delays, aggregated by month and year over the entire study period. Figure 4.3 further aggregates the percentage of delayed sailings and delay

lengths by month only. The percentage of delayed sailings is highly variable on a month to month basis; some months have delay rates higher than 80%. The monthly trend has similarities to the monthly trends of environmental factors shown in the Chapter 2, which identified more adverse conditions during the colder months. The plots also demonstrate an increase in variance and median delay length during months when adverse environmental conditions are expected. Note that the boxplots are zoomed into the range [-1,5] hours for ease of viewing. A small number of outlier data points exist beyond this range.

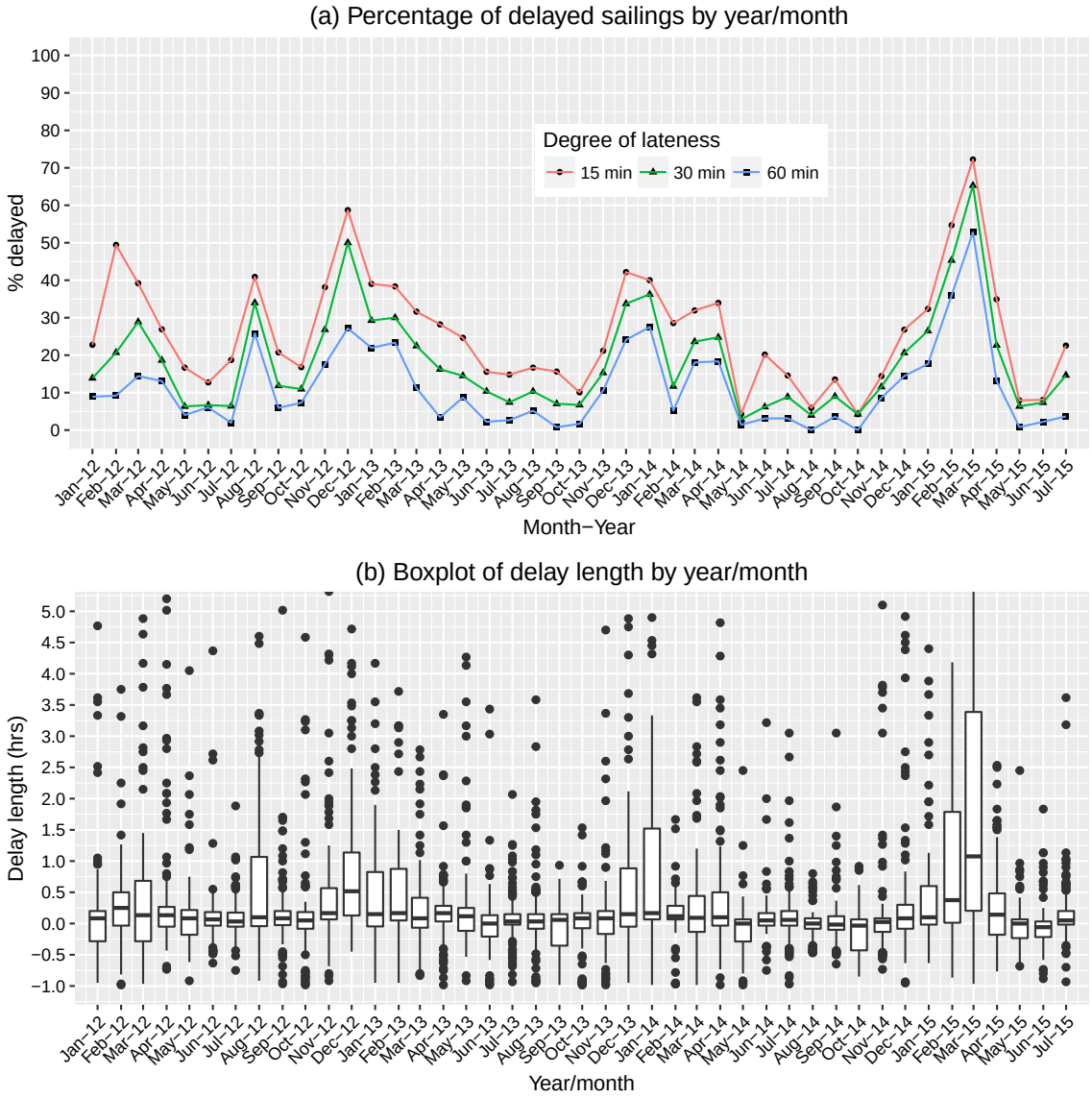

Figure 4.2: Occurrence and length of delay by month, 2012-2015.

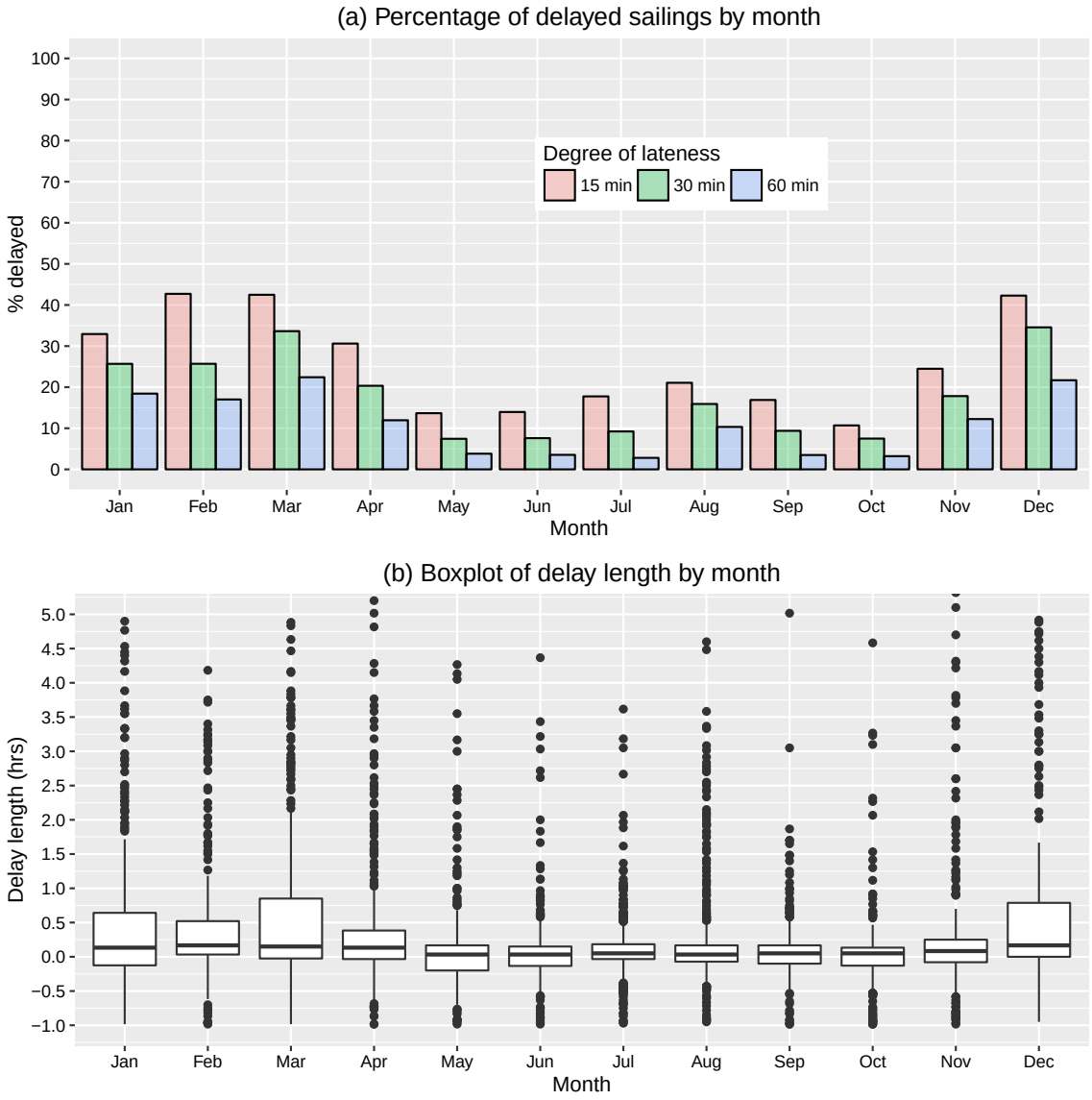

Figure 4.3: Occurrence and length of delay aggregated by month.

Figure 4.4 shows the histogram and ECDF of delay length for all of the records in the data set (2012-2015 period). The cutoff of 0.25 hours for being considered delayed or not is identified with a vertical red line. The long tail of the histogram shows the rare but non-zero frequency of longer delays. The maximum delay in the data set is approximately 25 hours, however this histogram was zoomed to span -1 hours to 6 hours for ease of viewing. From the ECDF it can be seen that there is an approximately 75% probability of not being delayed during the 2012-2015 period.

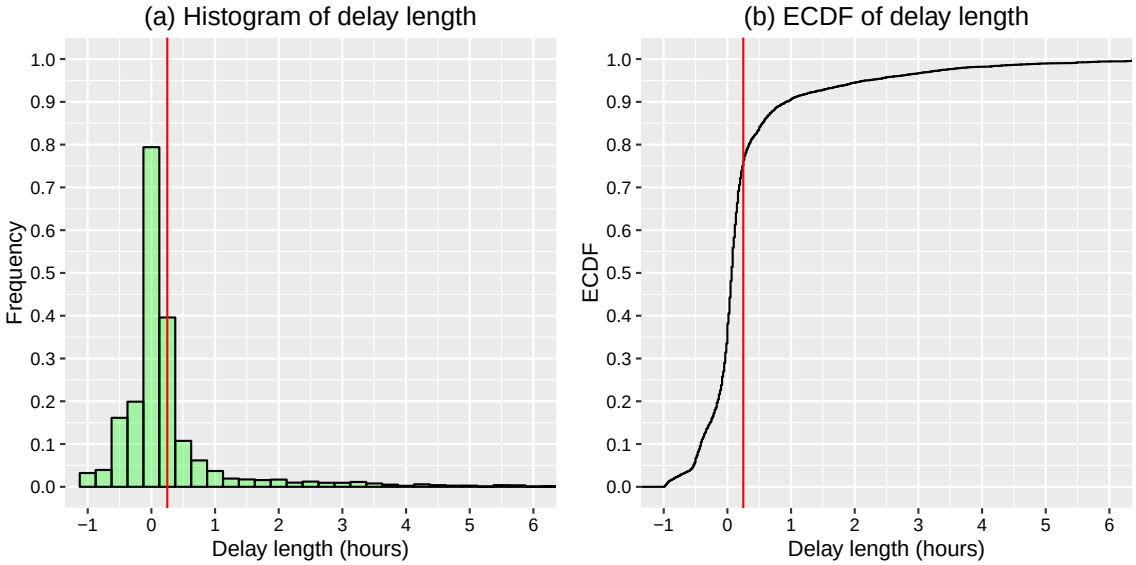

Figure 4.4: Histogram and ECDF of delay length.

Figure 4.5 shows the histograms of delay length by month, which highlights the more frequent and longer delays in the colder months, and a reduction in both delay occurrence and delay length in the warmer months. Figure 4.6 shows the ECDF plots of delay length by month. The probability of not being delayed is significantly higher in the warmer months than the colder months.

Obsevations by Environmental Factors

The plots in Figure 4.7 show the cumulative percentages of delayed sailings aggregated over intervals spanning the range of each environmental factor, and the plots in Figure 4.8 show the boxplots of the delay lengths aggregated over the same intervals. Increasing wind speed appears to be associated with a small increase in delay occurrence and in delay length. Wind direction appears to have little effect except for a

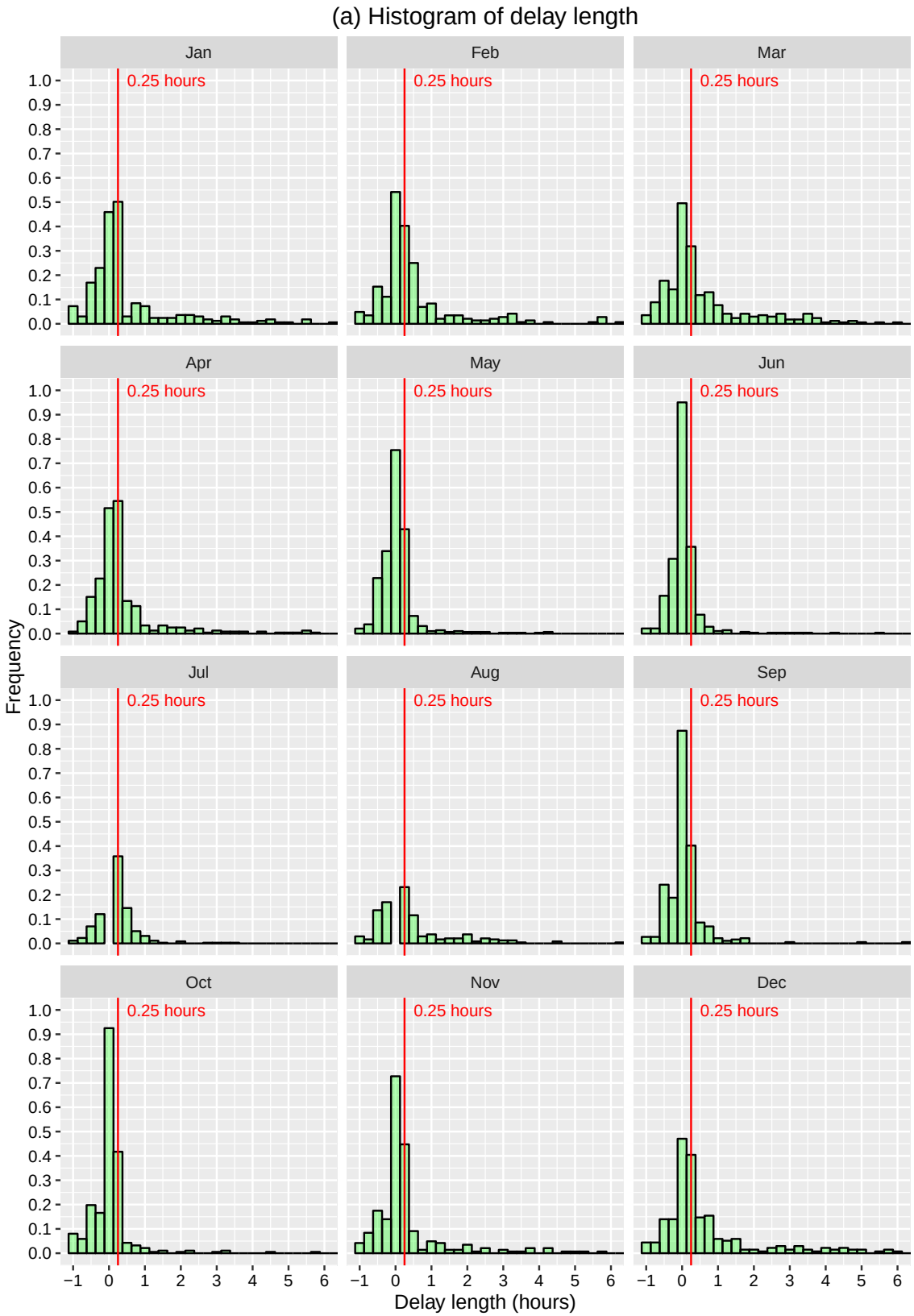

Figure 4.5: Histograms of delay length by month.

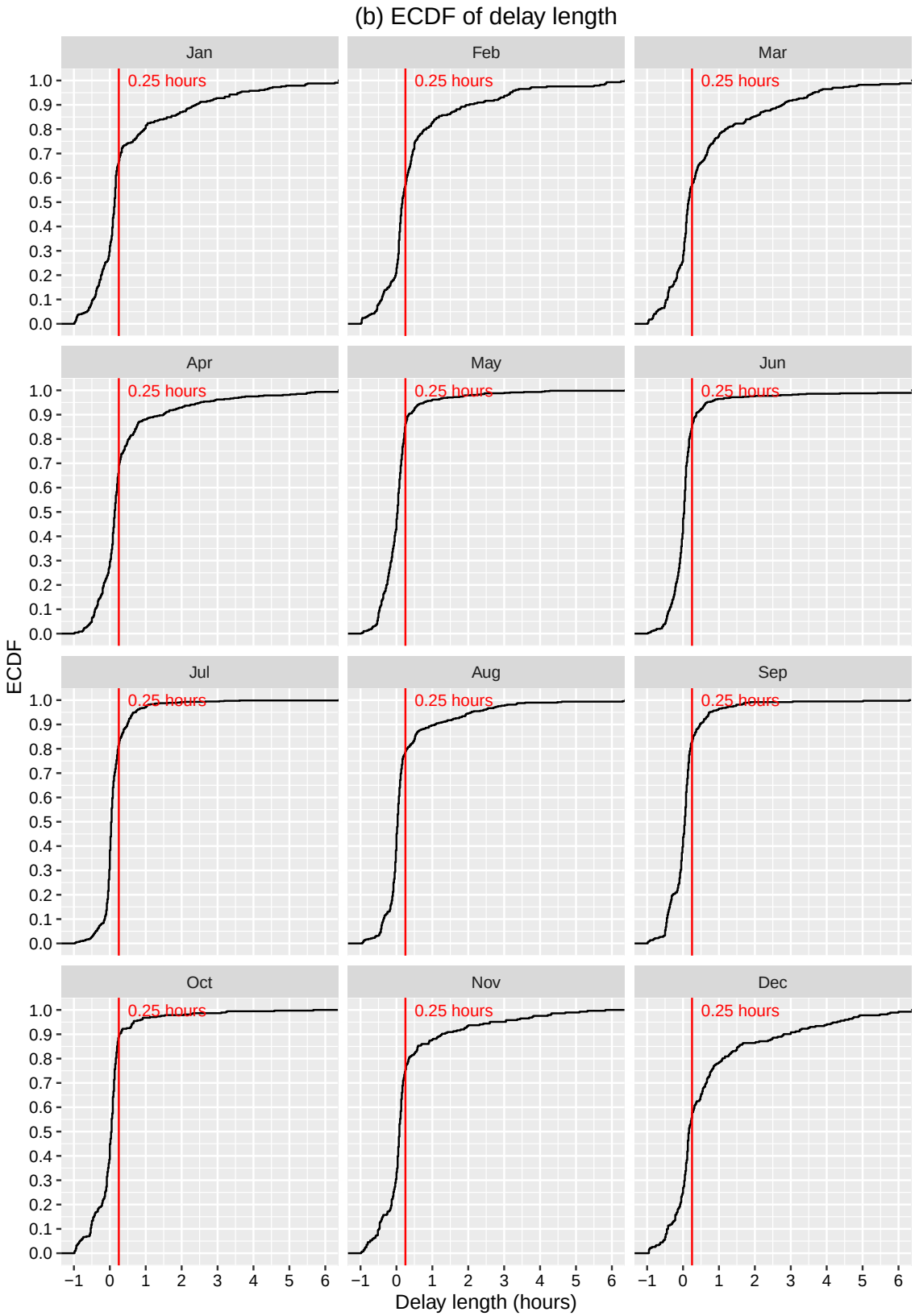

Figure 4.6: ECDFs of delay length by month.

slight peak in delay occurrence and delay length variance when the wind is from the west. There are significantly more and longer delays when the atmospheric pressure is low, probably due to low presence being an indicator of adverse weather. Similarly for air temperature there are more and longer delays at the lower range, perhaps due to the increased presence of sea ice. Precipitation does not appear to affect delay occurrence or length to a significant degree. Increasing ice concentration appears to increase the occurrence and length of delays as well.

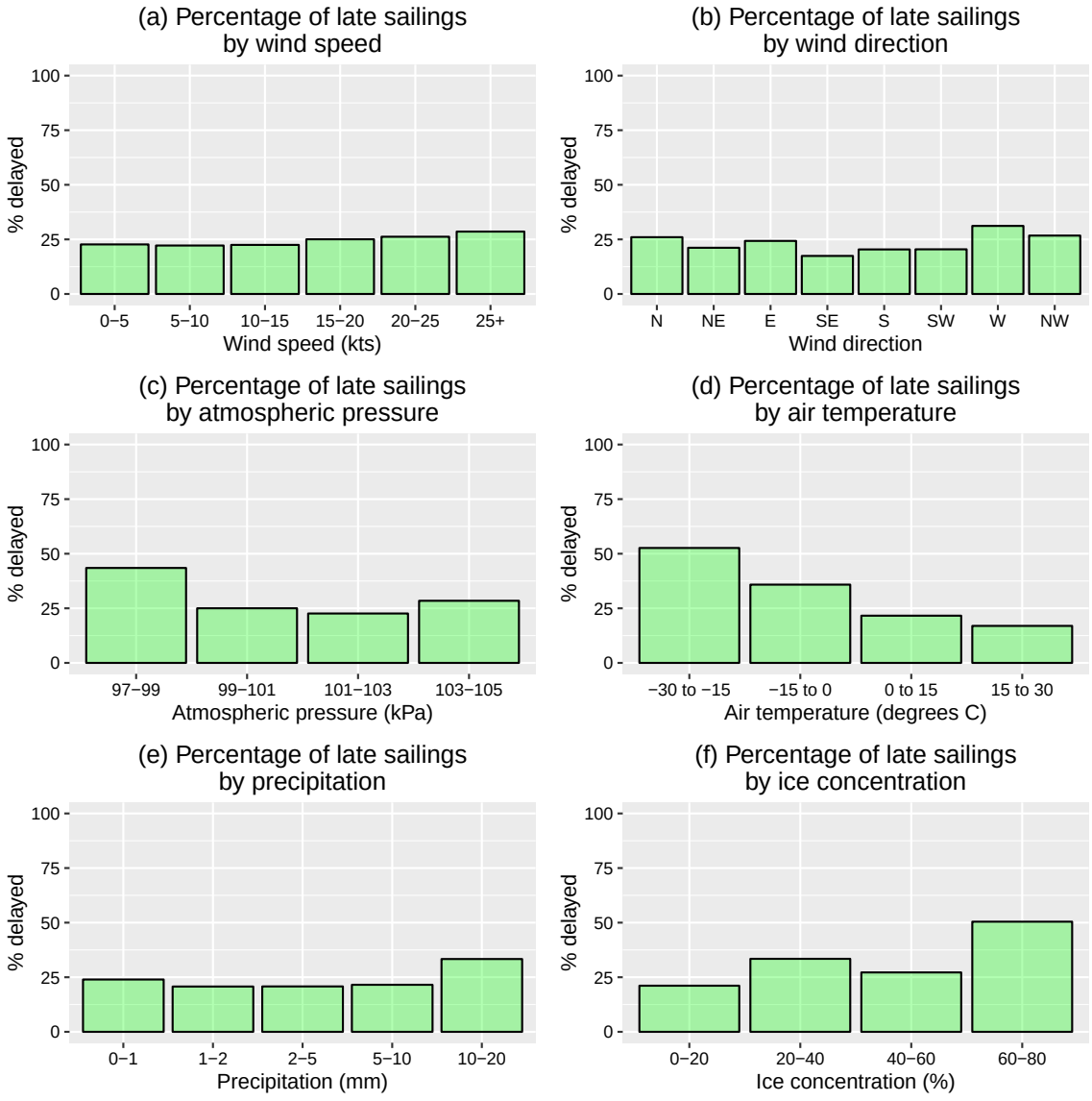

Figure 4.7: Percentage of delayed sailings aggregated by environmental factors.

Figure 4.9 shows the percentage of delayed sailings by wind speed and direction

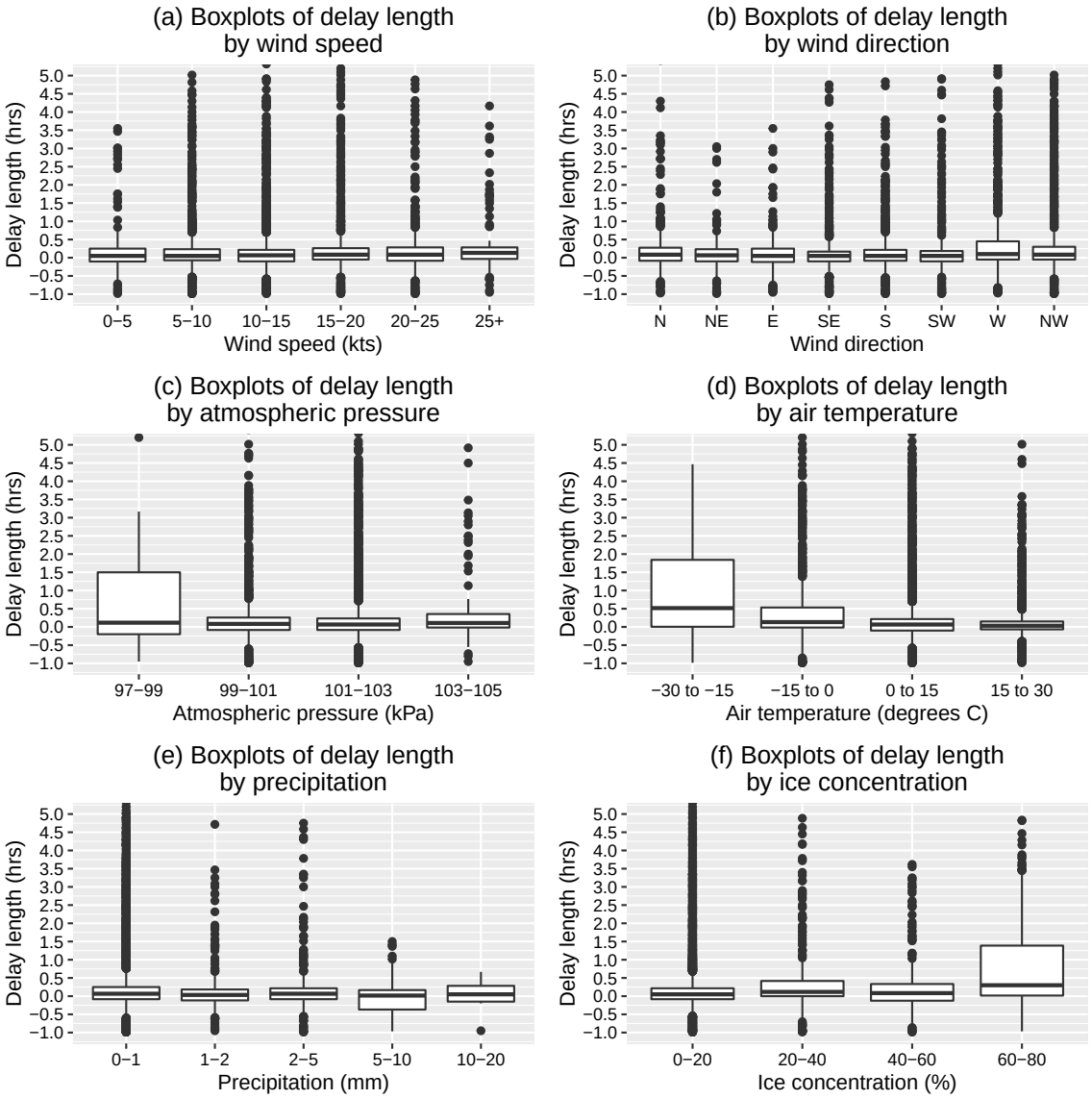

Figure 4.8: Delay length aggregated by environmental factors.

in heatmap format. Unlike the same heatmap for cancellations that demonstrated a clear trend, little can be discerned from this heatmap in terms of delays. Note that blank element in the heatmap reflect a lack of data points for those wind speed and direction intervals.

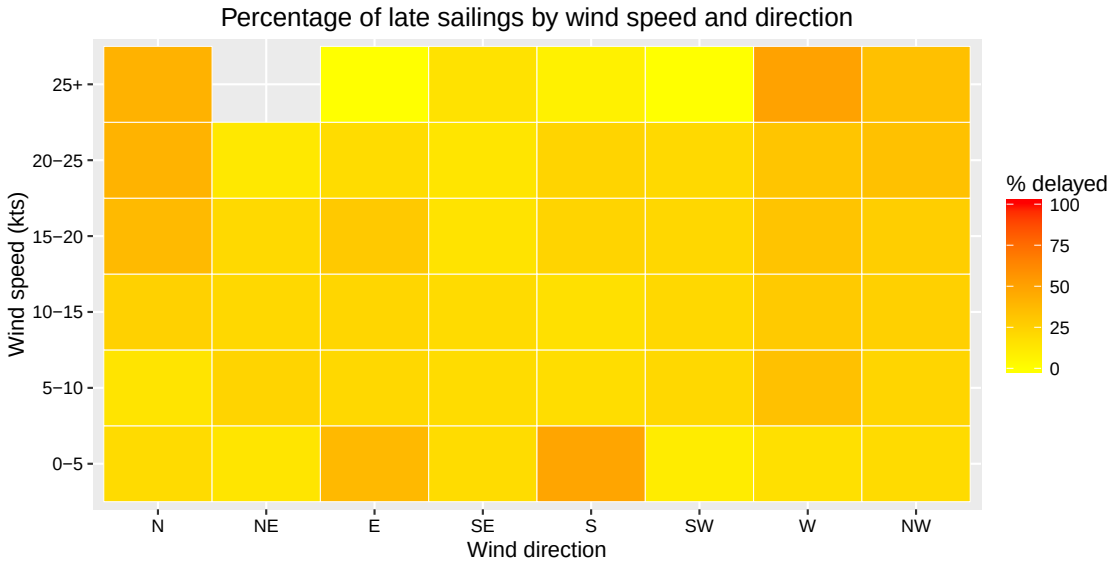

Figure 4.9: Percentage of delayed sailings by wind direction and speed.

Figure 4.10 shows the correlation between delay length and the continuous independent variables, as well as between the continuous independent variables. Delay length shows a weak correlation with air temperature and ice concentration and very weak or neglible correlations with the remaining variables. Air temperature and ice concentration shows a strong correlation, while pressure and wind speed are weak to moderately correlated with precipitation. The remaining combinations show weak or neglible correlations.

An ANOVA test was conducted to determine the statistical significance of the cardinal wind direction on delay length. The results, summarized in Appendix C, show a strong statistical significance. The Tukey Honest Significant Differences was also calculated to determine the confidence intervals on the differences between the means of the wind directions. The results show that only westerly winds are significant (see Appendix C for complete results).

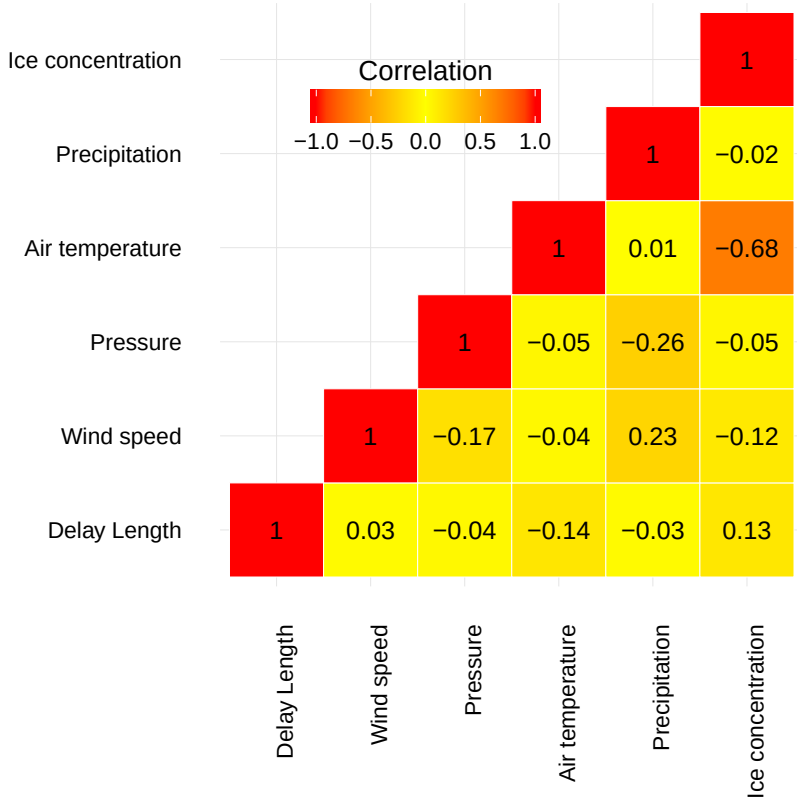

Figure 4.10: Correlation between delay model variables.

Observations by Vessel

Figure 4.11 shows the percentage of delayed sailings by vessel, as well as boxplots of the delay length for each vessel. The MV HL has the best performance in terms of both delay occurrence and delay length, and also appears to be early the most often. The MV BP is the sistership to MV HL but appears to have noticabley worse performance in terms of delays.

An ANOVA test was conducted to determine the statistical significance of the vessel on delay length. The results, summarized in Appendix C, show a strong significance. The relationship between the vessel and environmental conditions is not clear, however, and the significance may be partially attributed to non-environmental factors. The Tukey Honest Significant Differences was also calculated to determine the confidence intervals on the differences between the vessel means. The results verified that the HL is statistically significant against all other vessels. The remaining vessels only have partial or no pairwise signifiance (see Appendix C for complete results).

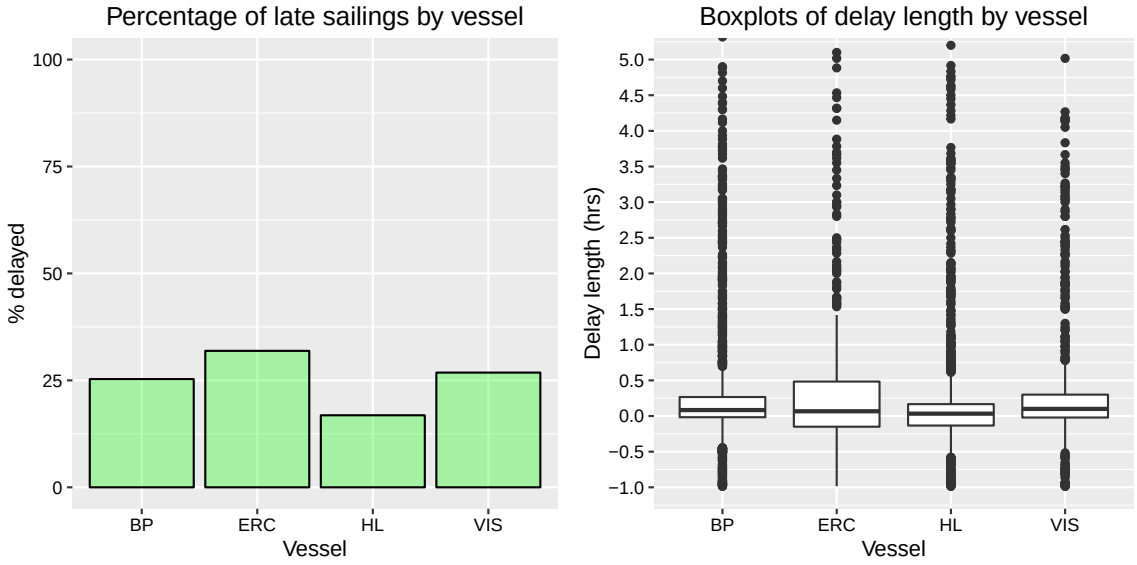

Figure 4.11: Occurrence and length of delay by vessel.

4.3 Delay Occurrence Modelling

This analysis examines the relationship between the dependent variable `late` (binary descriptor of whether a sailing is delayed or not) and the environmental factors described above as independent variables. The exploratory data analysis established trends between the dependent and independent variables as well as limited interactions between variables (which were inconclusive). In order to analyse the response of the entire set of independent variables, classification modelling techniques were employed.

4.3.1 Model Development

Several classification models were investigated in order to determine the most suitable approach to proceed, including logistic regression, classification tree, gradient boosted trees, linear discriminant analysis, k-nearest neighbours, support vector machines, and random forests.

The *caret* package in R (Kuhn et al., 2016) was used to design and implement a standardized test for model selection. Each model was trained on the same training data set with the same random seed. Repeated 10-fold cross-validation was used to

improve each model's performance and reduce overfitting. Several metrics were measured for each model and are summarized in Table 4.3. Based on the metrics of AUC and kappa, the Random Forest (RF) model exhibited the best performance. All of the models had low sensitivity scores, which reinforces the hypothesis that predicting delays will be more challenging than predicting cancellations. The sensitivity of the RF model is not indicative of a well-performing model, however it is much higher than the other models and is therefore also the best model at predicting the positive class (in this case, delayed sailings), which is important for imbalanced data sets.

Table 4.3: Performance metrics used for delay occurrence model selection

Model	Accuracy	Sensitivity	Specificity	BalAcc	Kappa	AUC
Ctree	0.7511	0.2134	0.8942	0.5538	0.1255	0.5708
LogReg	0.7872	0.0042	0.9955	0.4999	0.0417	0.5994
LDA	0.7863	0.0084	0.9933	0.5008	0.0026	0.5976
GBTree	0.8004	0.1088	0.9841	0.5466	0.1342	0.6833
KNN	0.7899	0.0502	0.9852	0.5179	0.0532	0.6232
SVM	0.7898	0.0000	1.0000	0.5000	0.0000	0.5770
RF	0.8021	0.1339	0.9800	0.5570	0.1604	0.7147

To examine the relationship between environmental factors and delay occurrence, a RF model was formulated with the R package *randomForest* (Liaw & Wiener, 2002) using the environmental factors previously described and the binary response variable *late*, which denotes that a sailing was "not delayed" (0) or "delayed" (1). A detailed description of RF models is provided in the previous chapter and will not be repeated here. The model variables are summarized in Table 4.4.

Table 4.4: Variables used in delay occurrence model formulation.

`late`	dependent	categorical
`vessel_cod`	vessel identifier	categorical
`wind speed`	independent	continuous
`wind direction`	independent	categorical
`atmospheric pressure`	independent	continuous
`air temperature`	independent	continuous
`precipitation`	independent	continuous
`ice concentration`	independent	continuous

The model was trained on a training set consisting of 90% of the original data set and tested on a testing set consisting of the remaining 10%. The number of trees *ntrees* was set as 100 and *mtry* was determined by finding the minimal OOB error

for the ensemble over the range of possible values. The plots in Figure 4.12 show that for this model the minimum OOB error occurs when $mtry$=2 and the effect of the number of trees on the OOB error. Values of $ntree$ greater than 50 produce a stable minimum error.

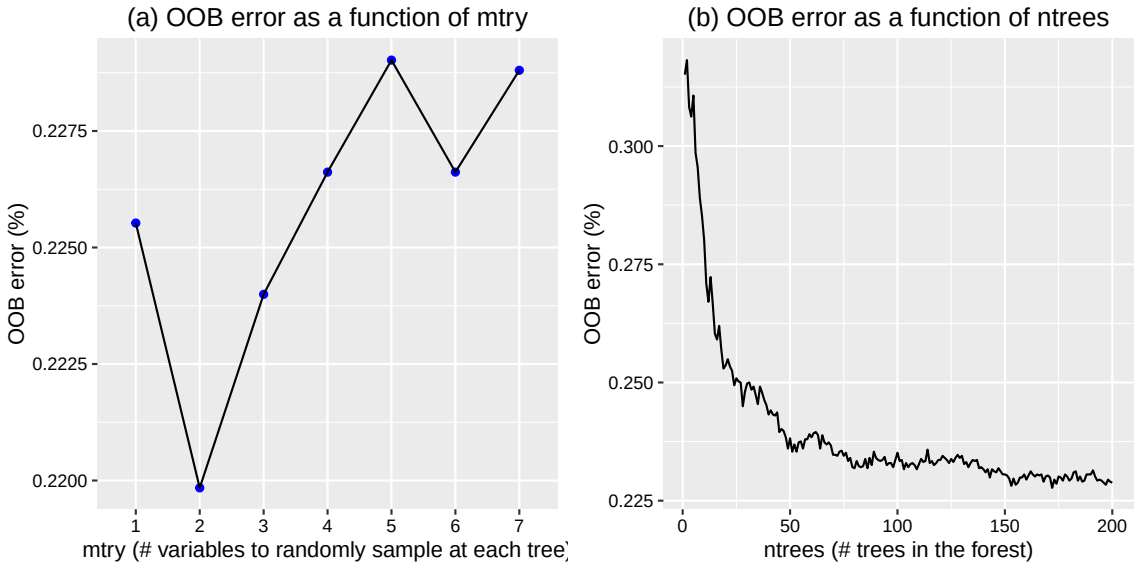

Figure 4.12: Delay occurrence model OOB error as a function of number of variables and of number of trees in the forest.

4.3.2 Model Performance

A detailed description of the model performance metrics was provided in the previous chapter and will not be repeated here.

Table 4.5 shows the key performance metrics of the constructed model. The high measure of specificity is no surprise given the imbalance of the data set. The sensitivity score of 0.1849, however, is very low, which represents the model's inability to correctly predict delayed sailings. This may be an indication of the lack of predictive ability of environmental factors alone. As seen in the exploratory data analysis, there are roughly twice as many delays caused by non-environmental factors as by environmental factors, and these delays are distributed throughout the ranges of the environmental predictors. This makes it very challenging to predict delay occurrence based on environmental factors alone.

Similar to the approach taken in the cancellation prediction model to attempt

Table 4.5: Delay occurrence model performance.

	Accuracy	Sensitivity	Specificity	BalAcc	Kappa	AUC
Model 1	0.7673	0.1849	0.9459	0.5654	0.1679	0.7235

performance improvement, a new feature was added to the data set. The `late` field was examined for cases where it was 1 and recoded to 0 if the environmental conditions during the planned voyage duration were benign. This decision was made based on the improbability of benign environmental conditions being the cause of a delay. In other words, delays that occurred during benign conditions were assumed to be caused by another, non-environmentally-related reason. The same wind speed and ice concentration thresholds as the cancellation model were used (20 kts and 10% ice concentration and 25 kts and 10% ice concentration). Table 4.6 shows that a modest gain in specificity is achieved using the lower wind speed threshold, however the model remains a poor predictor of delayed sailings.

Table 4.6: Delay occurrence model performance using increased wind speed and ice concentration thresholds.

	Accuracy	Sensitivity	Specificity	BalAcc	Kappa	AUC
Model 1	0.7673	0.1849	0.9459	0.5654	0.1679	0.7235
Model 2	0.9250	0.3725	0.9868	0.6797	0.4646	0.9540
Model 3	0.9665	0.3529	0.9878	0.6703	0.3971	0.9878

4.3.3 Results and Discussion

In this section the various aspects of the effects of the independent variables on model outcomes are explored.

Variable importance

As discussed previously, RF have two methods of evaluating variable importance, one that results in a score for the mean decrease in accuracy, and the other a score for the mean decrease of gini. Figure 4.13 shows the variable importance evaluated by both methods. The methods agree that air temperature and Vessel are the most and least important variables, respectively, and that pressure ranks as number three. Wind speed, wind direction, precipitation, and ice concentration rank differently, most

noticeably between ice concentration and wind speed. This disagreement between importance measures may be a symptom of the poor performance of the model in general. Interestingly, air temperature, which can indicated the presence of ice, is more important than ice concentration, which is an actual measure of the presence of ice.

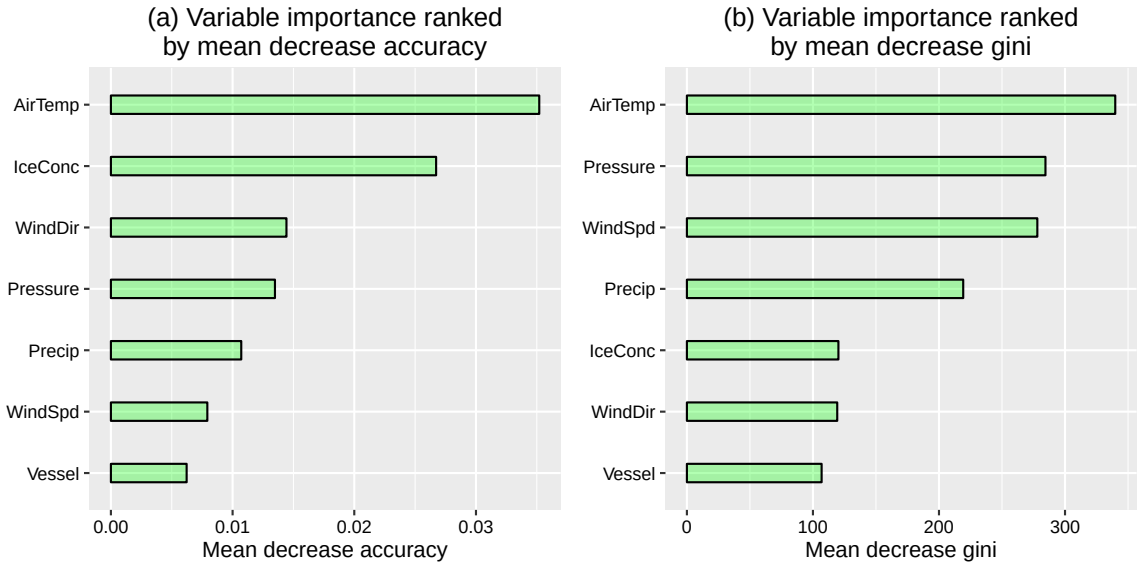

Figure 4.13: Delay occurrence model variable importance by mean decrease accuracy and mean decrease gini.

Variable Responses

Figure 4.14 provides a comparison of the predicted and actual responses over the range of each variable. This provides an estimation of how closely the predicted responses are to the actual responses on a per variable basis. The variables with higher importance (wind speed and pressure) track more closely to the actual predictions, however none of the variables track particularly closely, which is indicative of poor model performance.

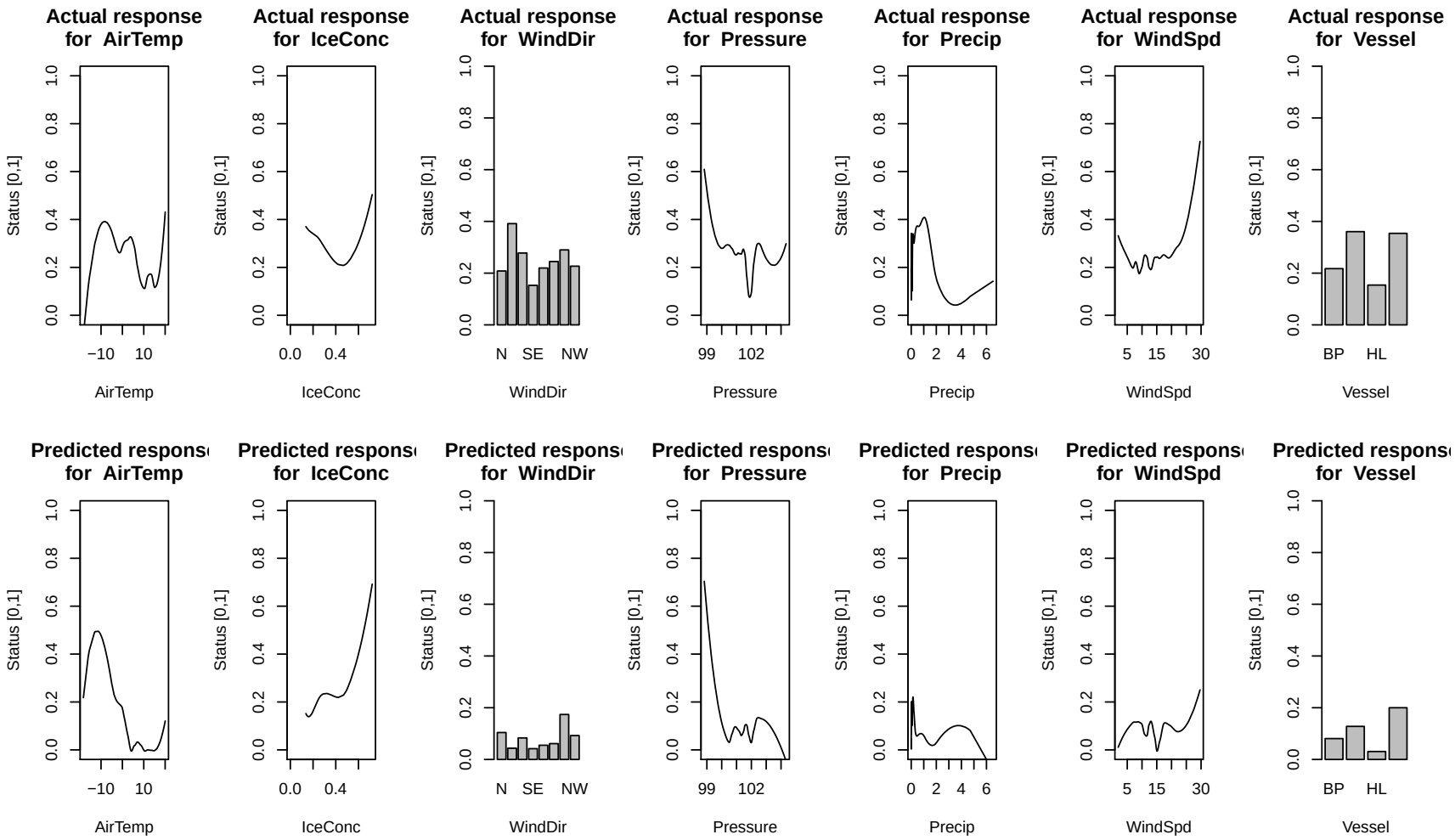

Figure 4.14: Delay occurrence model predicted vs actual responses for each variable.

Variable Partial Dependence

Figure 4.15 shows the partial dependence of each of the independent variables, in order of importance, referenced to the "delayed" class. The plots show the range of each environmental factor for which the probability of predicting the cancelled class are highest, independently (i.e., not accounting for interactions). The probability of the model prediciting the positive class ("delayed" class) is highest when the the air temperature is very low. As expected, ice concentration has very low probability of predicting a delay in the absence of ice, and higher probability in the presence of ice. Pressure has a higher probability at its extremes, indicative of the presence of storms (low pressure) or possibly high traffic levels during the good summer weather (high pressure). Wind speed and precipitation show a general increase in probability as those variables increase. The two categorical variables, wind direction and vessel, do not show strong trends in the probability of predicting a delay, although the lower probability associated with the MV HL ties in with the results of the ANOVA test conducted previously.

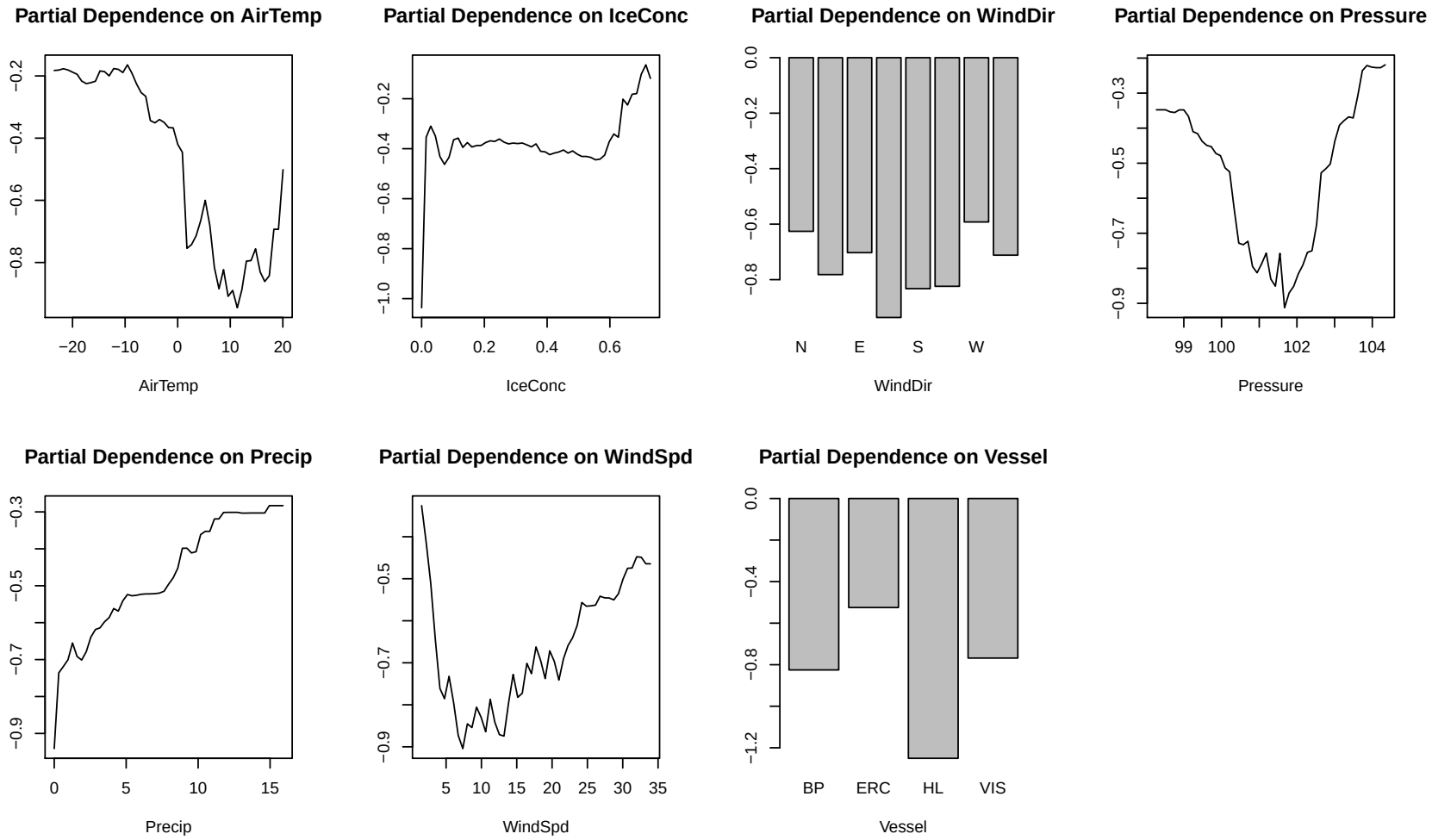

Figure 4.15: Delay occurrence model partial dependence of independent variables.

Figure 4.16 shows the bivariate partial dependence plots, which estimate interactions between two variables, for combinations of air temperature, pressure, and wind speed. The interactions do not show clear trends, however some observations can be made. The correlation between air temperature and ice concentration is evident in the shape of the first bivariate plot; there is a higher probability of predicting a delay when the temperature is lower and the ice concentration is higher. The second plot of air temperature and wind speed emphasizes the importance of air temperature relative to wind speed; little can be assessed about wind speed from this plot. The third plot of air temperature and pressure also shows the relative importance of air temperature over pressure, and the decrease in probability as pressure moderates is also evident.

4.4 Delay Length Modelling

Another way to analyze delayed sailings is by developing a model to predict the delay length. If the delay length is zero, the sailing is exactly on time. If the delay length is positive, it is late arriving, if negative, it is early arriving. This analysis examines the relationship between the dependent variable `delta` (the difference between actual and scheduled arrival times) and the environmental factors described above as independent variables. The exploratory data analysis established trends between the dependent and independent variables as well as limited interactions between variables (which were inconclusive). In order to analyse the response of the entire set of independent variables, regression modelling techniques were employed.

4.4.1 Model Development

Several regression models were investigated in order to determine the most suitable approach to proceed, including logistic regression, classification tree, gradient boosted trees, linear discriminant analysis, k-nearest neighbours, support vector machines, and random forests.

The *caret* package in R (Kuhn et al., 2016) was used to design and implement a standardized test for model selection. Each model was trained on the same training data set with the same random seed. Repeated 10-fold cross-validation was used to improve each model's performance and reduce overfitting. The metrics of RMSE

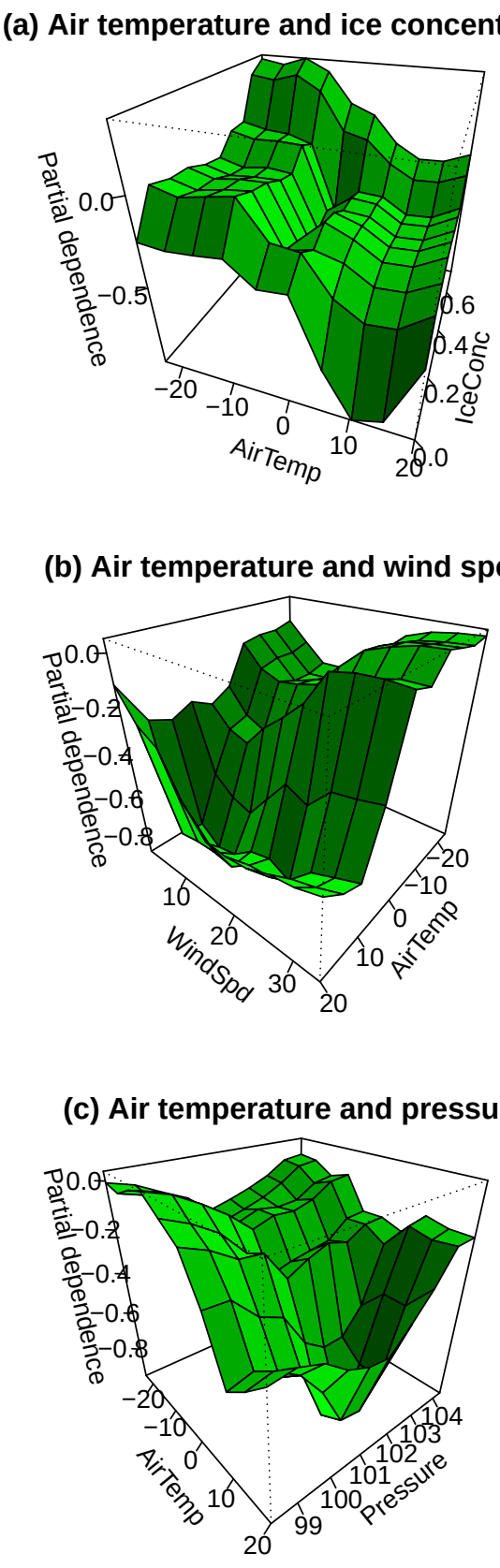

Figure 4.16: Delay occurrence model bivariate partial dependence of wind speed, pressure, and air temperature.

and R^2 were measured for each model and are summarized in Table 4.7. No model performed particularly well; a RMSE of 0.8 equates rougly to a 48 minute error in prediction of delay length. Based on these metrics, however, the Random Forest (RF) model exhibited the best performance and was selected for the analysis.

Table 4.7: Performance metrics used for delay length model selection

Model	RMSE	R2
Rtree	0.8455	0.0556
LinReg	0.8417	0.0441
GBTree	0.8219	0.0947
KNN	0.8388	0.0507
SVM	0.8395	0.0742
RF	0.8050	0.1337

To examine the relationship between environmental factors and delay length, a RF model was formulated with the R package *randomForest* (Liaw & Wiener, 2002) using the environmental factors previously described and the continuous variable of delay length `delta`, measured in hours. The model variables are summarized in Table 4.8.

Table 4.8: Variables used in delay length model formulation.

`delta`	dependent	categorical
`vessel_cod`	vessel identifier	categorical
`wind speed`	independent	continuous
`wind direction`	independent	categorical
`atmospheric pressure`	independent	continuous
`air temperature`	independent	continuous
`precipitation`	independent	continuous
`ice concentration`	independent	continuous

The model was trained on a training set consisting of 90% of the original data set and tested on a testing set consisting of the remaining 10% of the original data set. The number of trees *ntrees* was set as 150 and *mtry* was determined by finding the minimal MSE error for the ensemble over the range of possible values. Figure 4.17 shows that the minimum MSE error was found when *mtry*=4, and that the MSE converges as *ntrees* increases past 50.

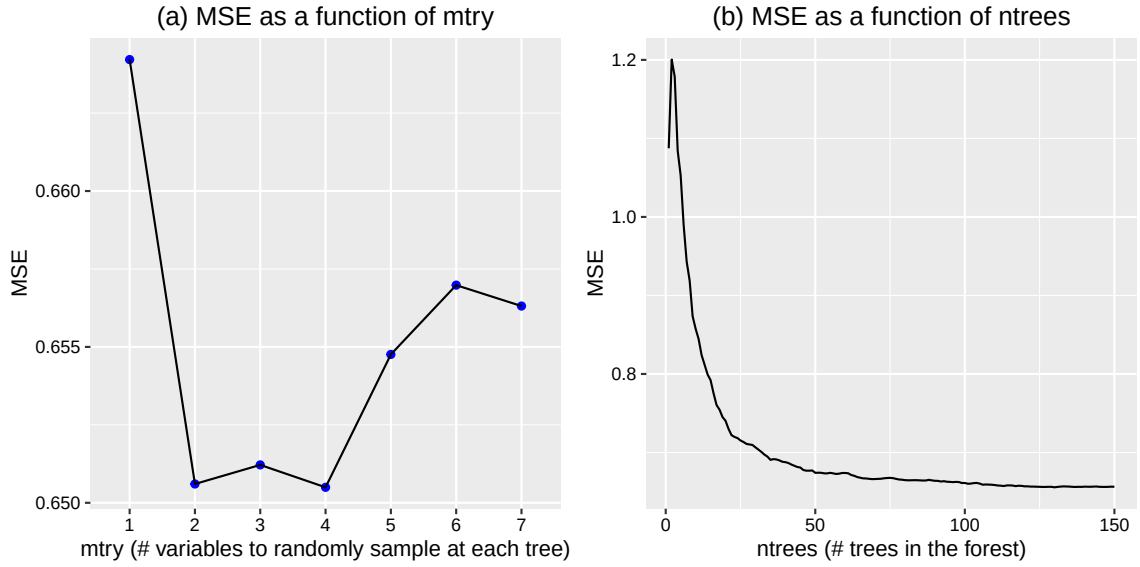

Figure 4.17: Delay length model MSE error as a function of number of variables and number of trees.

4.4.2 Model Performance

Similar to the delay occurrence model, the delay length model has poor predictive performance. After tuning, the model achieved a RMSE of 0.7726 and R^2 of 0.162 on the testing data set. This is attributed to the fact that the model uses only environmental factors as predictors, while there are many other factors that cause delays.

4.4.3 Results and Discussion

In this section the various aspects of the effects of the independent variables on model outcomes are explored.

Variable Importance

Figure 4.18 shows the variable importance evaluated by both methods (mean decrease accuracy and mean decrease gini). The methods agree that air temperature is the variable of greatest importance and vessel is the variable of least importance, but disagree about the relative importance of the remaining factors. These results are similar to the delay occurrence model and may be a symptom of the poor performance of the model in general. Like the delay occurrence model, air temperature, which can

indicate the presence of ice, is more important than ice concentration, which is an actual measure of the presence of ice.

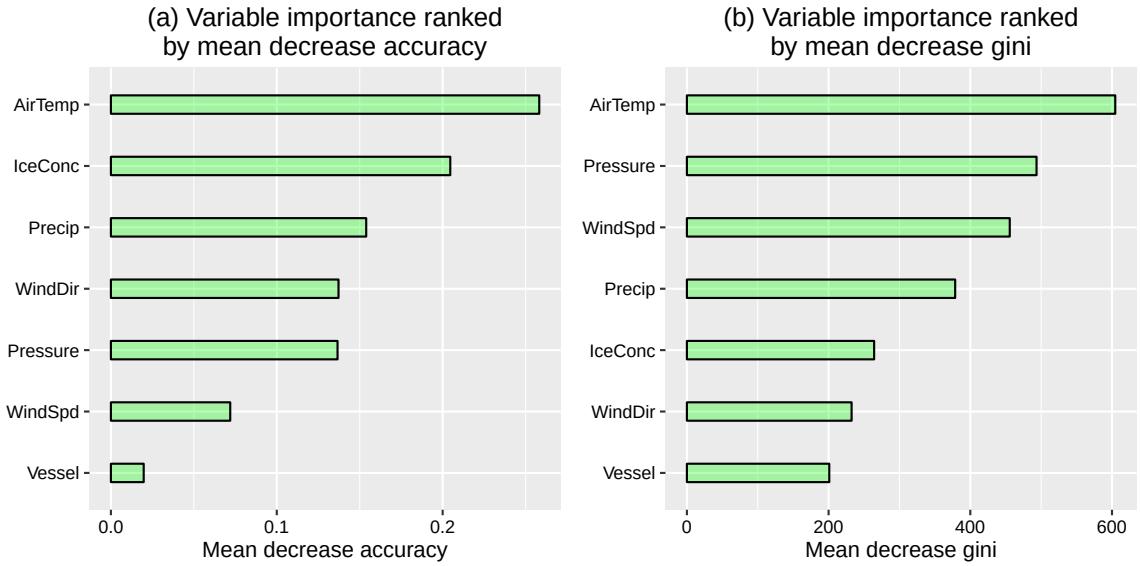

Figure 4.18: Delay length model variable importance by mean decrease accuracy and mean decrease gini.

Variable Responses

Figure 4.19 provides a comparison of the predicted and actual responses over the range of each variable. This provides an estimation of how closely the predicted resonses are to the actual responses on a per variable basis. The variables with higher importance track more closely to the actual predictions, however none of the variables track particularly closely, indicative of weak model performance.

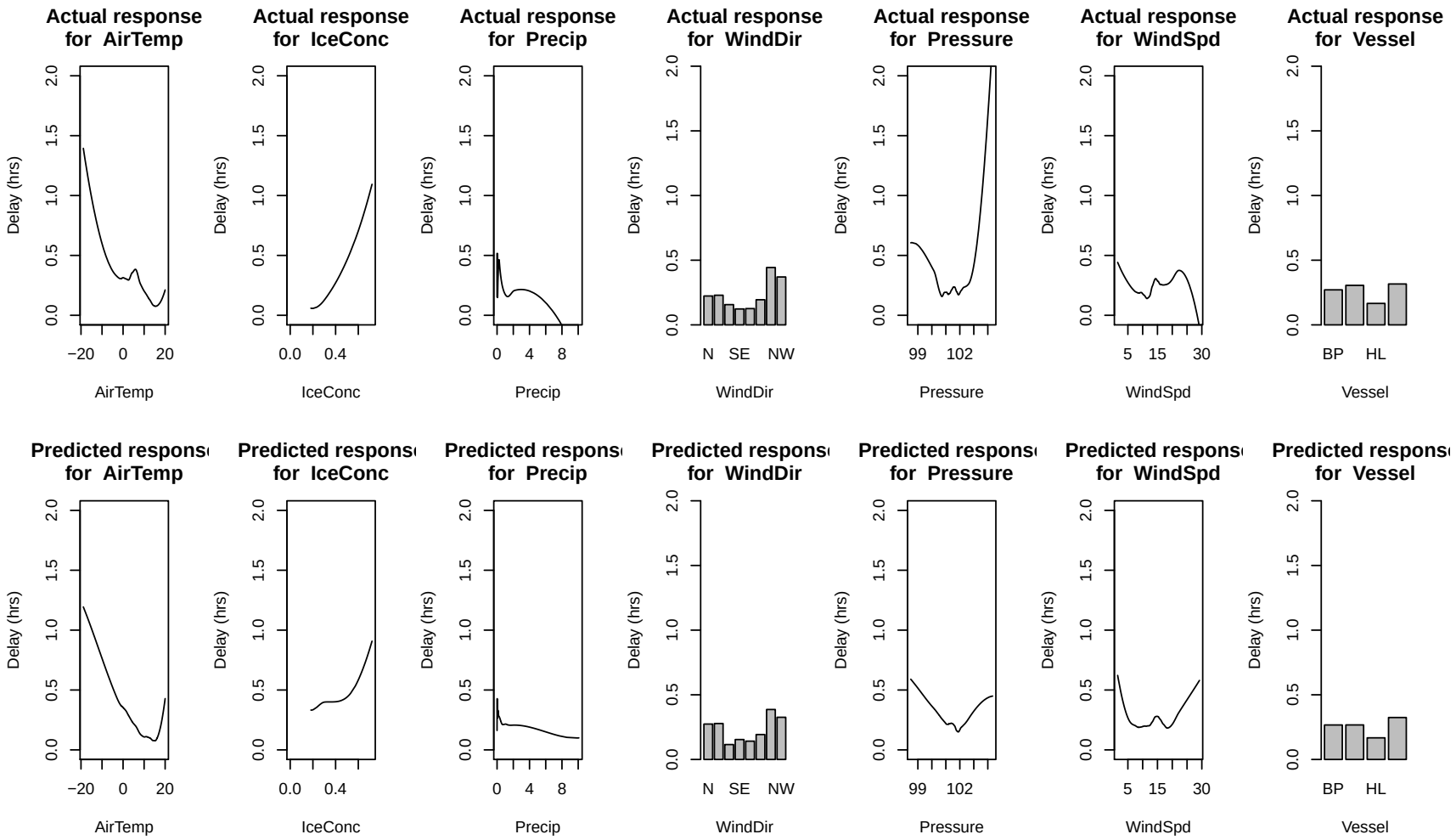

Figure 4.19: Delay length model predicted vs actual responses for each variable.

Variable Partial Dependence

Figure 4.20 shows the partial dependence of each of the independent variables, in order of importance. The plots show the marginal effect of each variable on the predicted response. Air temperature shows a strong marginal effect at the lowest temperatures, dropping off as the temperature warms above -10°C. Ice concentration shows a spike in the marginal effect as the value increases from zero and as concentrations exceed 60%. Logicially speaking, however, it does not make sense that 10% ice concentration has more of an effect than 50%, so the spike at the lower end is expected to be due to other factors. Lower pressure also has a larger marginal effect, while moderate pressure has minimal effect. The increase in effect as pressure increases is probably due to other factors (like traffic density), similar to the effects observed in the delay occurrence model. Except for a small increase in marginal effect as wind speed increases, slightly larger effect when the wind is from the west, and slightly smaller effect for the HL vessel, there is no further compelling evidence of strong marginal effects in the remaining variables.

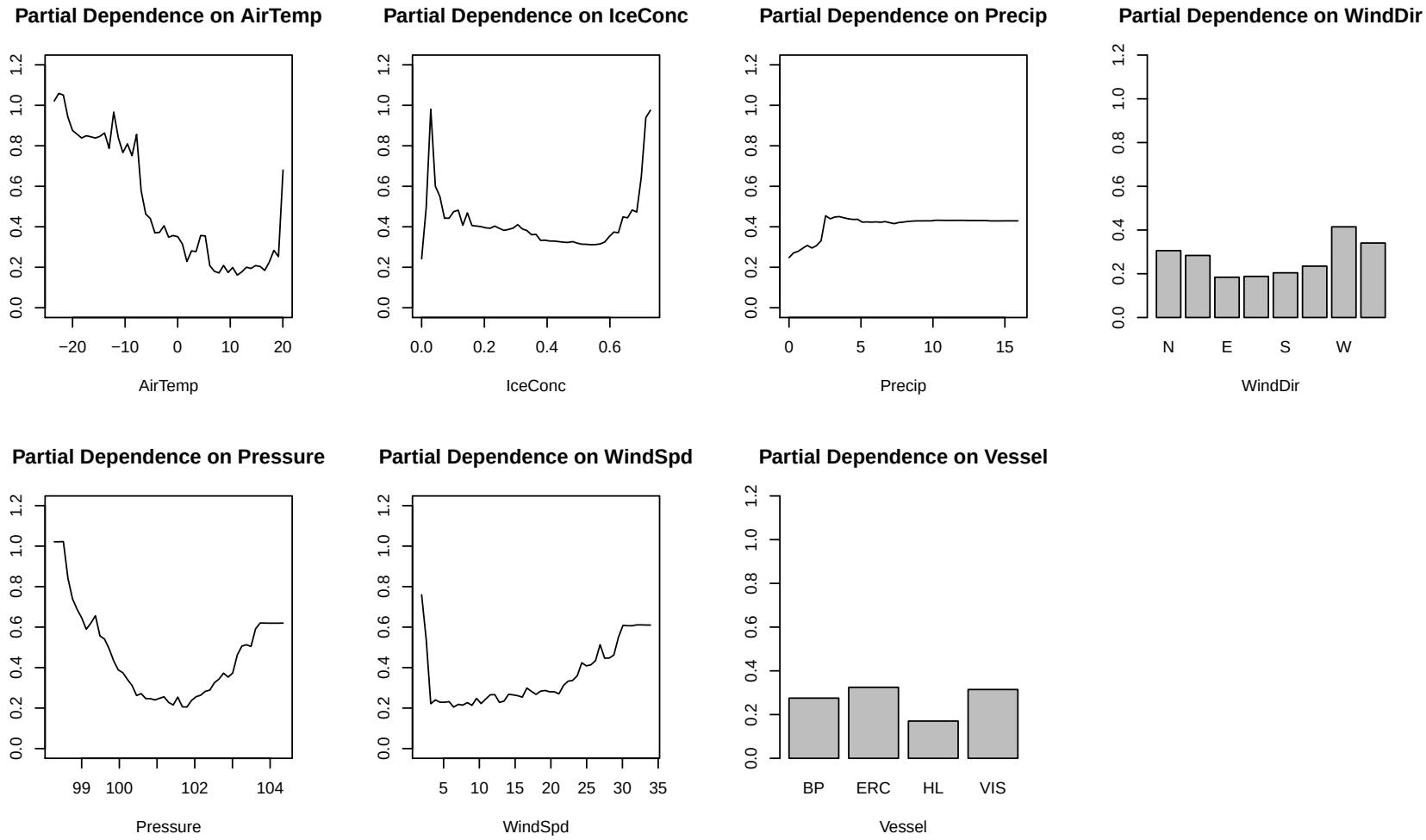

Figure 4.20: Delay length model partial dependence of independent variables.

Figure 4.21 shows the bivariate partial dependence plot, which estimates interactions between two variables, of air temperature and pressure. Bivariate dependence for other combinations of variables were not calculated due to the low marginal effects of those variables, or due to a high correlation (as is the case for air temperature and ice concentration). The air temperature by pressure plot shows largest marginal effect when pressure is lowest, and for moderate pressures air temperatures of less than -10°C increase the marginal effect. Moderate temperatures and pressures have little effect, while higher temperatures show an increase in effect over the range of pressure.

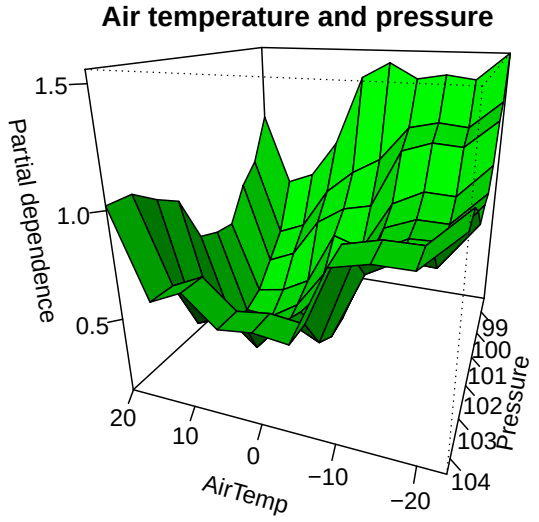

Figure 4.21: Delay length model bivariate partial dependence of air temperature and pressure.

4.5 Model Run-Times

The RF delay length and delay occurrence models were both run on an Apple Macbook Air with a 1.7 GHz Intel Core i7 processor and 8 GB of RAM, which was more than capable of running the models. Model training times of less than 20 seconds were typical, and the next longest processing times were due to the bivariate partial dependence plot computations, which each took approximately 8 seconds.

Chapter 5

Discussion

5.1 Introduction

The aim of this research is to analyse the effects of environmental factors on MAI ferry operations. This was approached through a statistical analysis of ferry sailing cancellations and delays with respect to environmental factors. The analysis of sailing delays was further broken down into an analysis of delay occurrence and delay length. Through exploratory data analysis and statistical modelling, the likelihood and impact of each environmental factor on operations was examined. Results demonstrated that environmental factors alone are good predictors of sailing cancellation but poor predictors of sailing delay.

This chapter is divided into three parts. The first is a discussion of the interrelationships between environmental factors, cancellations, and delays, and their connections to decision-making. The second part explores expected trends based on on likely climate change scenarios and their potential impacts on company operations. The third part provides overall conclusions, recommendations for further research, and final thoughts.

5.2 Relationship Between Cancellations and Delays

The dominant environmental factor for cancellation prediction is wind speed, followed second by pressure and third by air temperature. Precipitation, wind direction, and ice concentration followed these three with considerably less influence. The dominant factor in delay prediction was air temperature, however there was disagreement between the variable importance measures for the remaining variables. There was some agreement that pressure and wind speed factored moderately high, however wind direction, precipitation, and ice concentration factored lower and with differing priority.

The delay prediction models did not have good performance, so the variable importance is harder to reconcile and draw conclusions from, however some observations can be made.

The fact that wind speed factors much more prominently in cancellation prediction than delay prediction reflects current company policy and decision-making. MAI staff stated that in the past (prior to the period of this study) the decision to cancel was made less frequently. The vessels were very seaworthy so the goal was to send the ferries even if the wind was high and attempt to enter the destination harbour as soon as the wind speed dropped to a safe level. Although this would allow ferries to arrive at the earliest possible time given the weather, it would often cause ferries to be significantly delayed, waiting outside of the harbour for hours until the wind dropped. More recently the company found that costs could be reduced by cancelling sailings based on the forecast, and customer satisfaction also increased because customers in general preferred to wait for the next sailing on shore than endure bad weather at sea for extended periods. This change in approach to handling bad weather caused an increase in cancellations due to wind speed and a related reduction in delays due to wind speed, reflected in Figure 5.1, which shows this relationship during the 2010-2011 period when this change in policy was made, and also through a comparison of Figures 3.3 and 4.7, wherein cancellations increase drastically with wind speed, but delays increase only slightly. The decision to cancel more frequently due to wind speed effectively removes wind speed as a dominant factor in delay prediction. Conversely, one could reasonably assume that cancelling less frequently in the presence of high winds would increase the overall frequency and length of delays.

Ice concentration factored relatively highly in variable importance of the delay prediction models in terms of mean decrease accuracy, but quite low in the cancellation prediction model. This again reflects company policy in that sailings are rarely cancelled due to ice. In general the vessels are capable of completing their voyage in the majority of ice conditions found in the Cabot Strait, but delays may be encountered along the way if the ice concentration is high. The approach to ice concentration is different than for wind speed, however, in that delays due to ice are generally more acceptable. This is likely for two reasons. One, sailing through ice does not typically cause an increase in discomfort for passengers because the presence of ice reduces the

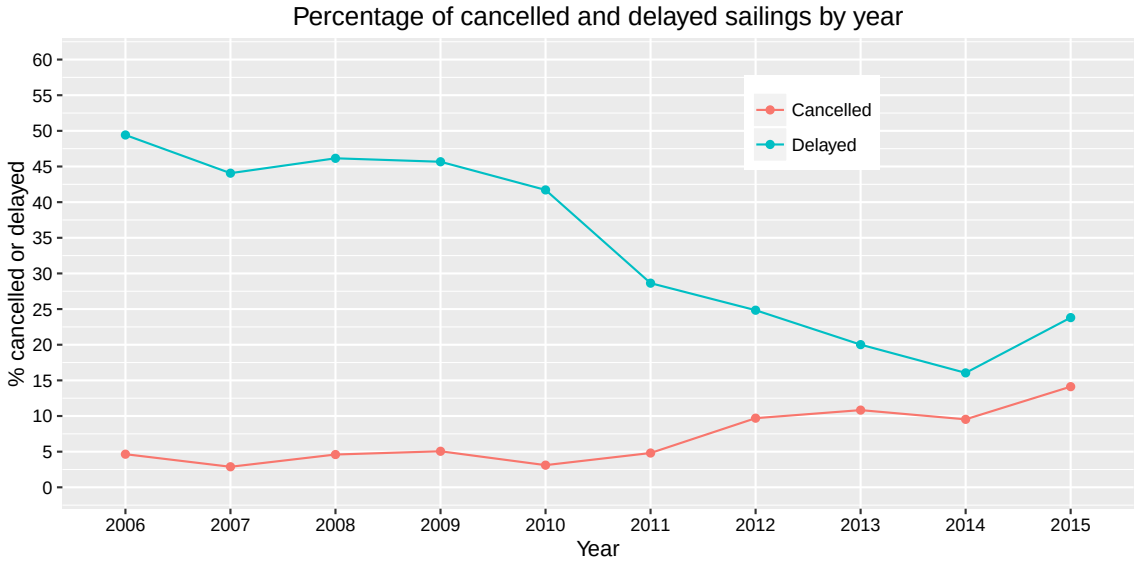

Figure 5.1: Percentage of cancelled and delayed sailings, 2006-2015.

effects of waves and wind on the vessel. Two, the logic behind cancelling a sailing due to high wind does not hold for ice because high wind typically passes within a small time window, so it is reasonable to expect that another sailing can occur within a reasonable amount of time. Ice, however, does not pass quickly and could remain for several days or weeks. Cancelling for these longer periods is not an option, so the risk of delay is more acceptable.

Interestingly, air temperature proved to be a more important predictor of delays than ice concentration. This may be due in part to the procedures for collecting, interpolating, storing, and displaying these data sets, and also to the fact that ice concentration is only one factor of concern with respect to the presence of ice (other factors such as thickness and age were not considered in this study). However, it is clear that air temperature is a good predictor of delay and a moderate predictor of cancellations. Figures 3.3 and 4.7 demonstrate that at the lowest temperatures delays are quite frequent, which is probably an indication of the presence of ice and a reflection of the company's decision-making, but also that cancellations increase significantly as temperature moderates, while delays decrease. This again reflects the decision making with respect to wind speed, because moderate temperatures reflect less ice, but also the presence of storms that bring warmer, moist air along with high winds.

At the beginning of this study wind direction was expected to have more importance in predicting cancellations and delays based on the statements of MAI staff with respect to the difficulty in entering harbour in higher winds from various directions. Although this is almost certainly a practical reality of harbour navigation, this study found that wind direction was not a prominent factor. The statistical significance of cardinal wind direction on delay length found in Chapter 4 by ANOVA determined that only a westerly wind was significant in producing longer delays, however neither the cancellation model nor the delay models found wind direction to be of great importance.

As stated earlier, one fundamental difference between modelling cancellations and delays is the fact that delays are generally not a result of a decision, but an effect realized due to some external factor, and that cancellations are always the result of a decision. This highlights a limitation in the modelling of cancelletions: the environmental data used for this study is based on observed conditions, but decisions are made based on forecasts. By using observed conditions it must be acknowledged that error may be introduced, because actual conditions often have some degree of difference from their associated forecast. For example, a prediction of high winds may cause a cancellation, but if the high winds do not actually occur the model will learn that the cancellation occured during otherwise acceptable wind speeds. Data were not available on the reliability of forecasts, however the error associated with this is assumed to be minimal because the decision to cancel is typically made 12-48 hours prior to sailing, when forecasts have a higher reliability. Furthermore, environmental factors typically affect operations in their extremes, and the relative frequency of forecasts being incorrect to the degree that a decision would be changed is assumed to be low, so the statistical significance of these "incorrect" values of environmental factors would be low and have little effect on the model.

5.3 Future Trends

The preceding chapters encompass a thorough investigation of the effects of environmental factors on MAI ferry operations, using a variety of statistical exploration and modelling techniques, and from a historical perspective. The findings obtained

through rigorous statistical analysis reinforce what MAI ferry captains and operations managers have known for years: that the likelihood of cancellation increases as wind speed increases, and the likelihood and impact of delays increases as ice concentration increases (or, as previously demonstrated, as air temperature decreases). Less obvious in practice, perhaps, are some of the subtle relationships of and between environmental factors and their combined effects on cancellations and delays, i.e., the high likelihood of cancellation if the atmospherice pressure is very low even if wind speeds are low, and the fact air temperature is actually a better predictor of delays than ice concentration. It is hoped that these and other findings from Chapters 3 and 4 provide a straightforward explanation of the effects of environmental factors on current and recent MAI operations.

This knowledge may provide insight into the decision-making for current operations, however the lack of discovery of a significant and previously unknown environmental effect limits the degree to which changes in company policy or decision-making would be required. The key to mobilizing this gained knowledge for the benefit of future decisions lies in attempting to predict how circumstances may change over the coming years and decades, and what effect that will have on the company. A determination of how operations may be affected by future environmental scenarios and what impacts these changes may have can inform longer-term decision making and initiate further investigation into areas of concern.

5.3.1 Predicting Future Cancellations

In order to better understand how future operations will be affected by cancellations caused by environmental factors, the RF cancellation prediction model constructed in Chapter 3 was used to project future cancellation likelihoods using data from select Climate Model Intercomparison Project Phase 5 (CMIP5) climate change models. Due to the poor performance of the delay models using environmental factors alone, the projection study was limited to projecting cancellations. CMIP5 is based on the latest agreements of the World Climate Research Programme's Working Group on Coupled Modelling to promote coordinated atmosphere-ocean general circulation climate model experiments. The efforts of this working group and the twenty climate

modelling groups that comprise it provide for freely available state-of-the-art multi-modal datasets to allow for the wider advancement of climate variability and change (Taylor, Stouffer, & Meehl, 2012).

The CMIP5 models attempt to project the effects of forcing due to the internal interactions of the complex and non-linear climate system itself (such as El Nino and the North Atlantic Oscillation), as well as externally-forced responses due to natural causes (such as large volcanic eruptions) and anthropogenic activities (such as the burning of fossil fuels). External forcing is standardized through the "representative concentration pathways" (RCP) protocol, which establishes radiative forcing scenarios that provide model inputs. The RCP used in the projection of cancellation probability was RCP8.5, which is based on radiative forcing of 8.5 W/m^2 in 2100, and represenative of a high emissions scenario (Moss et al., 2010).

Five climate model data sets were used for the projections: CMCC-CM (Scoccimarro et al., 2011), CNRM-CM5 (Voldoire et al., 2013), INM-CM4 (Volodin, Dianskii, & Gusev, 2010), IPSL-CM5A-LR (Dufresne et al., 2013), and IPSL-CM5A-MR (Dufresne et al., 2013). These were selected based on their availability of data for the timeframe in question, total size (to remain within download and storage limitations), variables represented in the data set (i.e., wind speed, wind direction, atmospheric pressure, air temperature, precipitation, and ice concentration), and observation frequency to fit the cancellation projection model (eight times daily).

The timeframe with which to make cancellation projections was selected as the 20-year period spanning 2026 to 2045. This period represents a medium- to long-term planning horizon that supports strategic planning, deicision-making, and initiatives, such as fleet-recapitalization and infrastructure projects, but is not so far in the future as to be meaningless to current MAI decision-makers. The intent is to determine if cancellations are likely to increase, decrease, or remain the same, so that the impacts of these outcomes may be considered in future planning.

The geographical area of interest is the same as the statistical analysis of cancellations and delays in Chapters 3 and 4, the Cabot Strait between North Sydney, Nova Scotia, and Port aux Basques, Newfoundland. The climate model data sets are provided in NetCDF format (similar to the historical data sets used in Chapters 3 and 4) and the process for extracting and formatting the the data into a data set usable

by the projection model was the same as outlined in previous chapters. The resulting data sets (one for each model) consisted of values for each environmental factor for every three hours from January 2026 to December 2045. To simulate ferry sailings a list of fictitious scheduled departure times was generated that closely resembles the operational schedule currently in use, i.e., two sailings in each direction each day. The final step in data matching involved pairing each scheduled sailing with the most adverse environmental factors spanning the duration of the voyage.

5.3.2 Results

Each of the five resulting input data sets were fed into the RF cancellation model developed in Chapter 3, yielding a binary prediction of either "not cancelled" or "cancelled" for every scheduled sailing from 2026 to 2045. These predictions were then aggregated by year and month to determine the projected ratio of cancelled sailings to total sailings and compared to historical data from 2006 to 2015.

Figure 5.2 shows the results of the five climate change data set model runs and the comparison to previous years. Aggregated annually, the results of all models demonstrate an increase in the ratio of cancelled sailings to total sailings aggregated annually. Figure 5.3 shows the mean of the five model results for each projected year, as well as the ranges between the maximum and minimum values for each year.

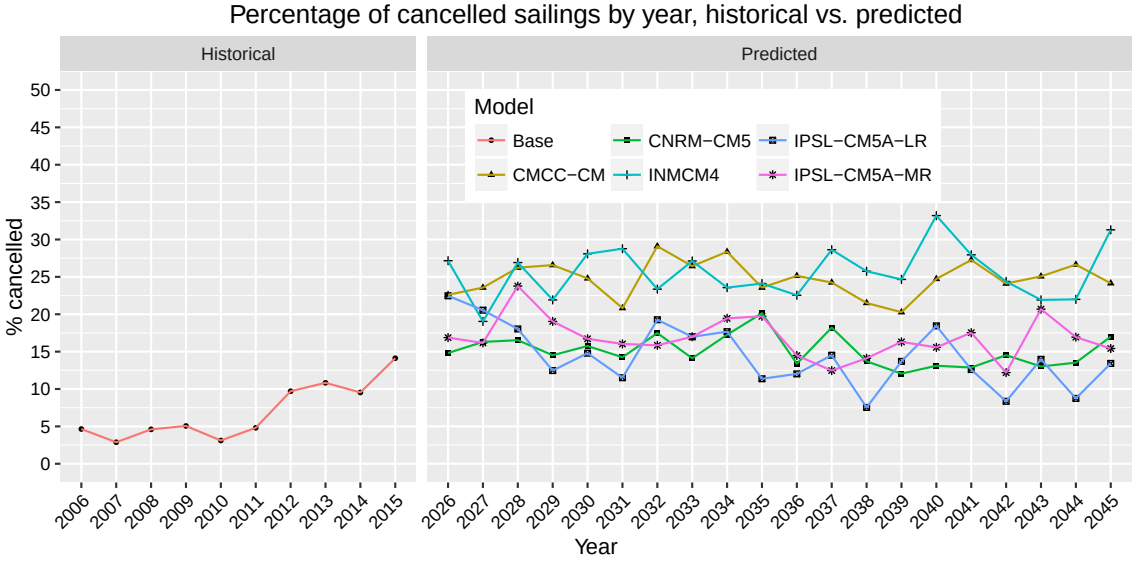

Figure 5.2: Comparison of projected and historical annual cancellation ratios for each climate model.

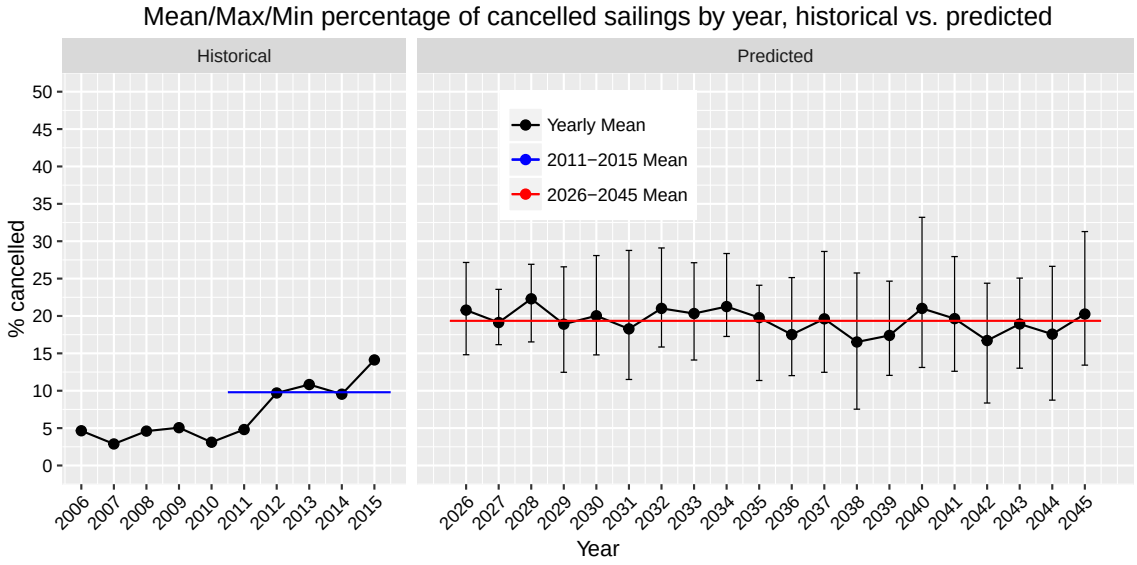

Figure 5.3: Comparison of projected and historical annual cancellation ratios by mean, maximum, and minimum of all five climate models.

The average of the cancellation ratios between 2006 and 2015 is 6.93% and between 2011 and 2015 is 9.8%. Although it is informative to observe the trend from 2006, the latter of these ratios is more useful for future comparison for two reasons: (1) it encompasses the period that only all currently operating vessels were in service (and the older vessels retired), and (2) it encompasses the period of the company's modified policy towards delays and cancellations (i.e., ferries would not sail in adverse conditions and wait for a "weather window" to enter the destination port (thus risking a substantial delay), but would be cancelled prior to sailing).

The mean cancellation ratios for all models in 2026 is projected to be 14.81%, which indicates an increase of 5.02% by 2026. The mean cancellation ratio for all models over the 2026 to 2045 period is projected to be 19.35%, which indicates an increase of 9.55% over the 2011 to 2015 period.

Figure 5.4 aggregates the cancellation ratios by month, showing a bar graph comparison of the monthly means and standard deviations of historical and predicted cancellation ratios (2011-2015 and 2026-2045, respectively). This plot shows that the months from October to June are projected to have significant increases in cancellation rates (almost doubling in some cases and more than doubling in at least three months), while July, August, and September are projected to remain relatively stable. Note that the standard deviation of the historical cancellation ratios is of limited use

because there are only 5 data points for each month (2011, 2012, 2013, 2014, and 2015).

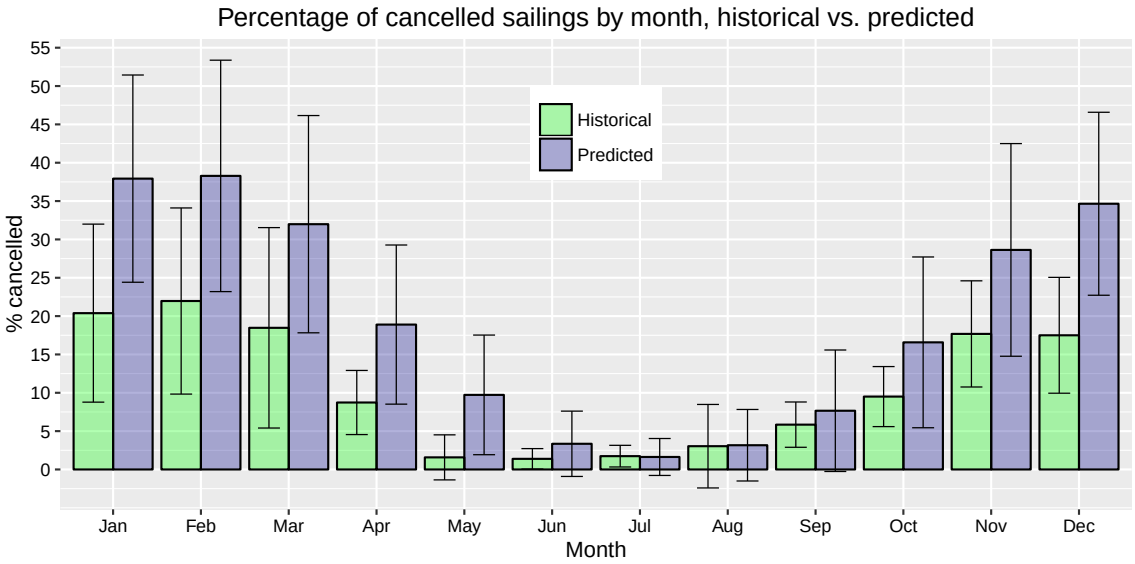

Figure 5.4: Comparison of projected (2026-2045) and historical (2011-2015) mean and standard deviation of monthly cancellation ratios.

5.3.3 Study Limitations

These results should be read with the acknowledgement of various model limitations. First, both the cancellation and the delay prediction models are based on environmental predictors alone. Although the acknowledged poor performance of the delay model makes clear the requirement for additional inputs or alternate approaches, it should be remembered that cancellations also occur for non-environmental reasons, and despite the very good performance of the cancellation model, a degree of error will always be present for this reason. Second, the historical environmental data in the study are subject to measurement errors and errors resulting from the modelling and interpolation methods used in the NARR. Third, the traffic data were found to have what are suspected to be data-entry errors, and although steps were taken to reduce these errors (see the rules in Appendix A), it is unlikely that all of the errors were found. Fourth, only five CMIP5 climate change models were used in this study. Despite the coordination and standardization of climate change modelling provided

by the World Climate Research Programme Working Group, variation between models is expected due to the different model approaches, internal and external forcing parameters, mathematical structures on which the models are based, and the inherent error of each model. A better distribution of cancellation predictions would result from using as many CMIP5 models as possible.

5.4 Knowledge Mobilization

An investigation into the driving forces behind these increases is not the intent of this study, although one hypothesis is that climate change may in general be causing an increase in the strength and frequency of storms over the next several decades. In order to better understand specific drivers, a focused study of the impacts of a larger set of climate models on MAI operations should be undertaken, inlcuding a targeted analysis of projected environmental patterns within Atlantic Canada and the associated variations in specific weather and ice conditions.

A detailed analysis of the decision to cancel sailings could also be undertaken to investigate potential opportunities for mitigating the losses caused by cancelled sailings. This could include a risk analysis to identify the risks and opportunities associated with the decision to cancel, which would provide a framework for understanding the priorities and tradeoffs. Cancelled sailings generate no revenue but also have reduced operating costs because the vessel is not waiting at sea for a chance to enter harbour. This implies an optimal point in the decision between cancelling allowing the sailing to proceed with a certain likelihood of delay. The problem becomes more complex when other company priorities are incorporated, such as customer satisfaction and safety, and would require data on other aspects of MAI operations. However when combined with the impacts of various environmental factors from this study a framework for informed decisions amidst these tradeoffs could be developed. One example that would benefit is the study of queue length as cancellations occur, and how quickly the backlog is dealt with, to determine if more optimal policies are possible.

One of the principal reasons that the vessels are so dominantly affected by high winds is the geogrpahy of Port aux Basques harbour, which is relatively constrained for large vessels that risk being pushed off course by the wind when travelling at slower speeds. The current fleet of vessels are modern, seaworthy, ice-class vessels, but are

limited in their manoeverability due to their fixed shaft lines. Bow thrusters and the retro-fitted Becker rudders significantly improve manoeverability at low speeds, however in constrained harbours without the availability of tug boat assistance, manoeverability remains an issue. Harbour improvement initiatives may improve the situtation but would be very costly.

Fleet recapitalization is typically planned over decades and normally begins with an analysis of requirements based on projected future realities. If the company is concerned about the projected rise in cancellations, a more detailed study of the changing environmental conditions can help shape the requirements for future fleet recapitalization. Ship propulsion and manoevering technology has advanced significantly in recent years and will continue to do so. Examples applicable to this scenario include the introduction of azipods, which can significantly increase the manoeverability and autonomy of vessels (but are limited in ice), as well as advances in engine efficiency to reduce operating costs. Furthermore, potential vessel designs could be simulated in the operating environment to verify the best-performing option to meet requirements, as well as to provide an analysis of tradeoffs between costs and requirements.

5.5 Conclusions and Future Work

This thesis set out to analyze the effects that various environmental factors have on MAI ferry operations and to determine how variations in these factors in the coming decades may change these effects. Various statistical analysis tools were employed and RF was selected to model the occurrence of cancellations, the occurrence and extent of delays, as well as the relative importance of environmental factors on each of these individually and in selected pairs. The cancellation models was then run using five climate change model data sets to project the extent to which cancellations may increase or decrease over the next three decades.

The RF cancellation model had good performance and demonstrated that environmental factors alone are good predictors of cancelled sailings. Results showed that cancellations increase with wind speed and have an inverse relationship with atmospheric pressure, but are affected to a lesser degree by wind direction, air temperature, precipitation, and ice concentration. Both delay models demonstrated poor performance, which is attributed to the lack of predictive power of environmental factors

alone in this context. Delays are caused by many reasons that are not related to the environment, and further study is recommended in this area. Results, however, did show that delays increase in frequency and length as the air temperature decreases and the ice concentration increases, but show lesser relationships with wind speed, wind direction, atmospheric pressure and precipitation. For the period 2026-2046 there is consensus among the five climate models used for projections that cancellations will increase significantly over the next three decades in all but the summer months.

These conlusions highlight areas in which further study would provide better understanding of MAI operations, both current and in the future, and address some of the study limitations (some previously mentioned but all summarized here):

- Analysis of non-environmental factors that affect the occurrence and length of delays to better understand the nature and impacts of delays in general.
- Explore non-machine-learning modelling techniques for the delay problem, such as hazard-based duration models.
- Analyse trade-offs between environmental and non-environmental factors, and the presence of any covariance that may affect model results.
- Investigate the impacts of cancelled sailings, including loss of revenue, customer satisfaction and experience, cancelled/rescheduled bookings, queue length build-up and recovery, etc., to better understand the cancellation decision and mitigation strategies.
- Conduct a study of a larger set of climate change models for the Atlantic Canada region to better understand how weather and ice conditions are projected to change and the associated impacts on operations.
- Conduct a detailed risk analysis and cost-benefit analysis on the decision to cancel to determine if an optimal policy that balances priorities exists and can be implemented in practice.
- Study the impacts of climate change projections on operations in the context of long-term decision-making, including fleet recapitalization and infrastructure upgrades, to ensure future requirements are identified.

Marine Atlantic Incorporated has a long history of providing essential transportation and logistics links through its ferry operations in harsh environmental conditions year-round. It is hoped that this modest contribution can provide some measure of practical benefit to the already considerable body of knowledge, expertise, seamanship, policies, and decision-making held by the company.

References

Albers, S. J., Dery, S. J., & Petticrew, E. L. (2015). Flooding in the Nechako River Basin of Canada: A random forest modeling approach to flood analysis in a regulated reservoir system. *Canadian Water Resources Journal / Revue canadienne des ressources hydriques*, 1–11.

Ayers, T. (2015). Marine Atlantic ferry finally closing in on North Sydney port. Retrieved from http://thechronicleherald.ca/novascotia/1275679-marine-atlantic-ferry-finally-closing-in-on-north-sydney-port

Bennion, M. D. (2010). *A comparison of operational performance: Washington State Ferries to ferry operators worldwide*. Washington State Department of Transportation. Retrieved from http://www.wsdot.wa.gov/research/reports/fullreports/750.1.pdf

Bhat, C. R. (1996). A generalized multiple durations proportional hazard model with an application to activity behavior during the evening work-to-home commute. *Transportation Research Part B: Methodological*, *30*(6), 465–480.

Boyles, S., Fajardo, D., & Waller, S. T. (2007). A naive Bayesian classifier for incident duration prediction. In *86th Annual Meeting of the Transportation Research Board, Washington, DC*. The National Academies of Sciences, Engineering, and Medicine. Transportation Research Board.

Breiman, L. (1984). *Classification and regression trees*. New York, N.Y.: Chapman and Hall.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Canadian Ferry Operators Association. (2015). *Keeping Canada Moving: A Survey of the Ferry Sector in Canada*. Canadian Ferry Operators Association. Retrieved from http://www.cfoa.ca

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random Forests for classification in ecology. *Ecology*, *88*(11), 2783–2792.

de Osés, F. X. M. & la Castells, M. (2008). Heavy weather in European short sea shipping: Its influence on selected routes. *The Journal of Navigation*, *61*(01), 165–176. Retrieved from http://dx.doi.org/10.1017/S0373463307004468

Dorp, J. R. V., Merrick, J. R. W., Harrald, J. R., Mazzuchi, T. A., & Grabowski, M. (2001). A risk management procedure for the Washington state ferries. *Risk Analysis*, *21*(1), 127–142. Retrieved from http://dx.doi.org/10.1111/0272-4332.211096

Dufresne, J. L., Foujols, M. A., Denvil, S., Caubel, A., Marti, O., Aumont, O., ... Vuichard, N. (2013). Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. *Climate Dynamics; Observational, Theoretical and Computational Research on the Climate System*, *40*(9), 2123–2165.

Fang, M.-C. & Lin, Y.-H. (2015). The optimization of ship weather-routing algorithm based on the composite influence of multi-dynamic elements (ii): optimized routings. *Applied Ocean Research*, *50*, 130–140.

Garib, A., Radwan, A., & Al-Deek, H. (1997). Estimating magnitude and duration of incident delays. *Journal of Transportation Engineering*, *123*(6), 459–466.

Grabowski, M., Ayyalasomayajula, P., Merrick, J., & Mccafferty, D. (2007). Accident precursors and safety nets: leading indicators of tanker operations safety. *Maritime Policy and Management; The flagship journal of international shipping and port research*, *34*(5), 405–425.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *Elements of Statistical Learning, The: Data Mining, Inference, and Prediction*. Springer.

Hayter, A. J. (2012). *Probability and statistics for engineers and scientists* (4th ed..). Boston, MA: Boston, MA : Brooks/Cole, Cengage Learning.

Hensher, D. A. & Mannering, F. L. (1994). Hazard-based duration models and their application to transport analysis. *Transport Reviews*, *14*(1), 63–82. Retrieved from http://dx.doi.org/10.1080/01441649408716866

Jarrah, A. I. Z. (1993). Decision support framework for airline flight cancellations and delays. *Transportation Science*, *27*(3), 266–280.

Ji, Y. (2014). Traffic incident clearance time and arrival time prediction based on hazard models. *Mathematical Problems in Engineering*, *2014*, 1–11. doi:10.1155/2014/508039

Kelman, J. B. (2008). Hazards in the maritime transport of bulk materials and containerized products. *Loss Prevention Bulletin*, (203), 28–36.

Klein, A. (2010). Airport delay prediction using weather-impacted traffic index (WITI) model. In *2010 IEEE/AIAA 29th Digital Avionics Systems Conference (DASC)*. doi:10.1109/DASC.2010.5655493

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., ... Candan, C. (2016). *caret: Classification and Regression Training*. Retrieved from https://CRAN.R-project.org/package=caret

Lee, S., Park, W., & Jung, S. (2014). Fault detection of aircraft system with Random Forest algorithm and similarity measure. *The Scientific World Journal*, *2014*.

Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3).

Lin, L., Wang, Q., & Sadek, A. W. (2015). A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. *Transportation Research Part C*, *55*, 444–459.

Marine Atlantic Inc. (2015). *The View From Here: 2014-15 Annual Report*. Marine Atlantic Inc. Retrieved from http://www.marineatlantic.ca/uploadedFiles/Content/About_Us/Corporate_Information/MAI_AR_2015_ENG_WEB%20(2).pdf

Marine Atlantic Incorporated. (2015). *Traffic Data Set*.

Merrick, J. R. W. (2005). Assessing uncertainty in simulation-based maritime risk assessment. *Risk analysis*, *25*(3), 731–743.

Merrick, J. R. W., Dorp, J. R. V., Mazzuchi, T. A., & Harrald, J. R. (2001). Modeling risk in the dynamic environment of maritime transportation. In *Proceedings of the 2001 winter simulation conference, december 9-12, 2001* (Vol. 2, pp. 1090–1098). Virginia Commonwealth University. Arlington, VA, United states: Institute of Electrical and Electronics Engineers Inc. Retrieved from http://dx.doi.org/10.1109/WSC.2001.977419

Mesinger, F., Dimego, G., Kalnay, E., & Mitchell, K. (2006). North American Regional Reanalysis. *Bulletin of the American Meteorological Society*, *87*(3).

Millard, K. (2015). On the importance of training data sample selection in Random Forest image classification: A case study in peatland ecosystem mapping. *Remote sensing*, *7*(7), 8489–8515.

Moss, R. H., Edmonds, J. A., Hibbard, K. A., Manning, M. R., Rose, S. K., Van, D. P., ... Wilbanks, T. J. (2010). The next generation of scenarios for climate change research and assessment. *Nature*, *463*(7282), 747.

Nam, D. & Mannering, F. (2000). An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A: Policy and Practice*, *34*(2), 85–102.

O'Connor, P. J. & O'Connor, N. (2006). Work-related maritime fatalities. *Accident Analysis and Prevention*, *38*(4), 737–741.

Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A., & Pereira, J. M. C. (2012). Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. *Forest Ecology and Management*, *275*, 117–129.

Pierce, A. D., Farris, C. A., & Taylor, A. H. (2012). Use of random forests for modeling and mapping forest canopy fuels for fire behavior analysis in Lassen Volcanic National Park, California, USA. *Forest Ecology and Management*, *279*, 77–89.

Polishchuk, P. G., Muratov, E. N., Artemenko, A. G., Kolumbin, O. G., Muratov, N. N., & Kuz'min, V. E. (2009). Application of random forest approach to QSAR prediction of aquatic toxicity. *Journal of Chemical Information and Modeling*, *49*(11), 2481–2488.

Provan, C. A., Cook, L., & Cunningham, J. (2011). A probabilistic airport capacity model for improved ground delay program planning. In *2011 IEEE/AIAA 30th Digital Avionics Systems Conference (DASC)*. doi:10.1109/DASC.2011.6095990

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org

Reynolds, R., Smith, T., Liu, C., Chelton, D., Casey, K., & Schlax, M. (2007). Daily High-Resolution-Blended Analyses for Sea Surface Temperature. *Journal of Climate*, *20*(22), 5473–5489, 5491–5496.

Rezaee, S., Pelot, R., & Finnis, J. (2016). The effect of extratropical cyclone weather conditions on fishing vessel incidents' severity level in atlantic canada. *Safety Science*, *85*, 33–40.

Robichaud, B. & Mullock, J. (2001). The weather of Atlantic Canada and Eastern Quebec. Retrieved from http://www.navcanada.ca/EN/media/Publications/Local%20Area%20Weather%20Manuals/LAWM-Atlantic-1-EN.pdf

Scoccimarro, E., Gualdi, S., Bellucci, A., Sanna, A., Fogli, P., Manzini, E., ... Navarra, A. (2011). Effects of tropical cyclones on ocean heat transport in a high-resolution Coupled General Circulation Model. *Journal of Climate*, *24*(16), 4368–4384.

Sen, D. & Padhy, C. P. (2015). An approach for development of a ship routing algorithm for application in the North Indian Ocean region. *Applied Ocean Research*, *50*, 173–191.

Shao, W., Zhou, P., & Thong, S. K. (2012). Development of a novel forward dynamic programming method for weather routing. *Journal of Marine Science and Technology (Japan)*, *17*(2), 239–251. Retrieved from http://dx.doi.org/10.1007/s00773-011-0152-z

Siddiqui, C., Abdel-Aty, M., & Huang, H. (2012). Aggregate nonparametric safety analysis of traffic zones. *Accident Analysis and Prevention*, *45*, 317–325.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation. *AI 2006: Advances in Artificial Intelligence, Proceedings*, *4304*, 1015–1021.

Sokolova, M. & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 307–307.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1947–1958.

Takashima, K., Mezaoui, B., & Shoji, R. (2009). On the fuel saving operation for coastal merchant ships using weather routing. In *8th International Navigational Symposium on Marine Navigation and Safety of Sea Transportation, TransNav 2009, June 17, 2009 - June 19* (pp. 431–436). Tokyo University of Marine Science and Technology. Gdynia, Poland: CRC Press.

Taylor, K., Stouffer, R., & Meehl, G. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498.

Transport Canada. (2015). *Transportation in Canada 2014* (tech. rep. No. 15296 E). Government of Canada.

Treasury Board Secretariat. (2016). Corporate Profiles - Crown Corporations. Retrieved from http://www.tbs-sct.gc.ca/hgw-cgf/finances/rgs-erdg/cc-se/corporate-societe/ccp-pse-eng.asp

Veritas, D. N. (2016). Rules for classification of ships: Part 5, Chapter 1 - Ships for navigation in ice. Retrieved from https://www.dnvgl.com/about/index.html

Voldoire, A., Sanchez-Gomez, E., Salas, Y. M., Decharme, B., Cassou, C., Sénési, S., ... Chauvin, F. (2013). The CNRM-CM5.1 global climate model: Description and basic evaluation. *Climate Dynamics; Observational, Theoretical and Computational Research on the Climate System*, *40*(9), 2091–2121.

Volodin, E., Dianskii, N., & Gusev, A. (2010). Simulating present-day climate with the INMCM4.0 coupled model of the atmospheric and oceanic general circulations. *Izvestiya, Atmospheric and Oceanic Physics*, *46*(4), 414–431.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, *527*, 1130–1141.

Wu, Y., Pelot, R. P., & Hilliard, C. (2009). The influence of weather conditions on the relative incident rate of fishing vessels. *Risk Analysis*, *29*(7), 985–999.

Zhang, Y. (2008). Real-time intermodal substitution: strategy for airline recovery from schedule perturbation and for mitigation of airport congestion. *Transportation Research Record*, (2052), 90–99.

Appendix A

Traffic Data Set Filtering Rules

A.1 Traffic Data Set Filtering Rules

All of the data in the traffic data set was input manually by MAI staff and is therefore prone to data input errors. Some of the errors are observable due to their not complying with the logic of ferry operations (i.e. an arrival time that was earlier than the departure time). Upon inspection it was found that some of these errors could be identified and corrected manually based on the logic of ferry operations, however the size of the data set made this impractical. These errors were therefore addressed by filtering the data through a set of rules based on the logic of ferry operations and deleting the records that were found to be in error. The following paragraphs list the rules used to identify errors and any supplemental notes for increasing the understanding of the traffic data set.

Rule 1: The scheduled departure time must be later than the scheduled arrival time. If the difference between the scheduled arrival time and the scheduled departure time is negative, an error exists. Note that records are from the perspective of the port, not the vessel or the individual sailing, and thus the arrival time must be before the departure time.

Rule 2: The actual departure time must be later than the actual arrival time. If the difference between the actual arrival time and the actual departure time is negative, an error exists. Note that records are from the perspective of the port, not the vessel or the individual sailing, and thus the arrival time must be before the departure time.

Rule 3: The maximum allowable arrival delay is 3 days (72 hours). If the difference between the actual arrival time and the scheduled arrival time is greater than 72 hours, an error exists. This is based on the longest delay in recent years being slightly longer than 2.5 days. Many records had arrival delays of weeks, months, or even years, which are most likely due data entry errors. (Note that while modelling

delays in chapter five this was further reduced to 6 hours and any longer delays were considered as outliers and removed, vastly improving model performance).

Rule 4: The maximum allowable early arrival is 1 hour. If the difference between the actual arrival time and the scheduled arrival time is less than -1 hours, an error exists. This is based on consultation with MAI staff wherein it was learned that vessels almost never arrived more than 1 hour early. Many records had early arrivals of many hours, days, weeks, or even years, which are most likely due to data entry errors.

Rule 5: The maximum allowable departure delay is 12 hours. If the difference between the actual departure time and the scheduled departure time is greater than 12 hours, an error exists. This is based on the MAI schedule, which typically has departures every 12 hours in each direction. Many records had departure delays of days, weeks, months, or even years, which are most likely due data entry errors.

Rule 6: The maximum allowable early departure is 1 hour. If the difference between the actual departure time and the departure arrival time is less than -1 hours, an error exists. This is based on consultation with MAI staff wherein it was learned vessels only depart more than 1 hour early in the event of a schedule change, which is a rare event.

Appendix B

Derivation of Equations for Wind Speed and Direction

B.1 Wind Direction

NetCDF files provide wind data using two vectors, **u** and **v**. **u** is the east-west component of the wind speed and direction vector (the component on the x-axis). It is positive when blowing to the east, and negative when blowing to the west. **v** is the north-south component of the wind speed and direction vector (the component on the y-axis). It is postive when blowing to the north, and negative when blowing to the south.

In order to obtain wind direction from the vectors, the two-argument *arctangent* function is used. $Arctan(y, x)$ determines the angle between the x-axis and the vector from the *origin* to the point (x, y). Thus, in this case, $arctan(v, u)$ determines the angle between the x-axis and the vector from the *origin* to the point (u, v). However, wind direction is in reference to the y-axis (north is at the top), so the agruments are reversed to yield the angle from the y-axis. The direction the wind is blowing *to*, in radians, is then

$$wind\ direction = atan2(\mathbf{u}, \mathbf{v}) \tag{B.1}$$

where $atan2$ is the nomenclature for the two-argument *arctangent* function in most mathematical computing languages.

In order to convert from radians to degrees, the result is multiplied by $180/\pi$, becoming

$$wind\ direction = \frac{180}{\pi} atan2(\mathbf{u}, \mathbf{v}) \tag{B.2}$$

Finally, in order to convert from the direction the the wind is blowing *to* to the direction the wind is blowing *from* (the format used in marine weather forecasts), 180° is added to the result, giving

$$wind\ direction = \frac{180}{\pi} atan2(\mathbf{u}, \mathbf{v}) + 180 \tag{B.3}$$

B.2 Wind Speed

Wind speed is determined simply by using the Pythagorean theorem to determine the length of the vector from the *origin* to (u, v). The wind speed in m/s is then

$$wind\ speed = \sqrt{\mathbf{u}^2 + \mathbf{v}^2} \tag{B.4}$$

There are 1852 m in one nautical mile and 3600 seconds in one hour, so to convert from m/s to nautical miles per hour (kts), the result is multiplied by 3600/1852, giving

$$wind\ speed = \frac{3600}{1852}\sqrt{\mathbf{u}^2 + \mathbf{v}^2} \tag{B.5}$$

Appendix C

Results of ANOVA Tests and Tukey HSD

C.1 Statistical Significance of Cardinal Wind Direction on Delay Length

Table C.1: ANOVA test of wind direction on delay length.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
wdb.w	7	60.1	8.6	6.7	5.6210E-08
Residuals	5075	6467.9	1.3		

Table C.2: Results of Tukey HSD for wind direction on delay length.

	diff	lwr	upr	p adj
N-E	0.0487	-0.2041	0.3016	9.9906E-01
NE-E	0.0040	-0.3098	0.3179	1.0000E+00
NW-E	0.2101	-0.0060	0.4261	6.3647E-02
S-E	0.0306	-0.1904	0.2516	9.9990E-01
SE-E	0.0417	-0.1822	0.2655	9.9925E-01
SW-E	0.0266	-0.2193	0.2725	9.9998E-01
W-E	0.3095	0.0784	0.5406	1.2884E-03
NE-N	-0.0447	-0.3438	0.2544	9.9983E-01
NW-N	0.1613	-0.0326	0.3553	1.8614E-01
S-N	-0.0181	-0.2175	0.1813	9.9999E-01
SE-N	-0.0070	-0.2096	0.1956	1.0000E+00
SW-N	-0.0221	-0.2488	0.2046	9.9999E-01
W-N	0.2607	0.0502	0.4713	4.3596E-03
NW-NE	0.2060	-0.0626	0.4747	2.7982E-01
S-NE	0.0266	-0.2461	0.2992	9.9999E-01
SE-NE	0.0377	-0.2373	0.3126	9.9990E-01
SW-NE	0.0226	-0.2706	0.3158	1.0000E+00
W-NE	0.3054	0.0245	0.5863	2.2015E-02
S-NW	-0.1795	-0.3294	-0.0295	6.9808E-03
SE-NW	-0.1684	-0.3226	-0.0142	2.1061E-02

Continued on next page

Table C.2: Results of Tukey HSD for wind direction on delay length.

	diff	lwr	upr	p adj
SW-NW	-0.1834	-0.3682	0.0013	5.3225E-02
W-NW	0.0994	-0.0651	0.2640	5.9817E-01
SE-S	0.0111	-0.1499	0.1721	1.0000E+00
SW-S	-0.0040	-0.1945	0.1865	1.0000E+00
W-S	0.2789	0.1079	0.4498	2.1486E-05
SW-SE	-0.0151	-0.2089	0.1788	1.0000E+00
W-SE	0.2678	0.0931	0.4425	9.3293E-05
W-SW	0.2829	0.0807	0.4850	5.9364E-04

C.2 Statistical Significance of Vessel on Delay Length

Table C.3: ANOVA of vessel on delay length.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
vessel_cod	3	67.5	22.5	17.7	2.0322E-11
Residuals	5079	6460.5	1.3		

Table C.4: Results of Tukey HSD for vessel on delay length.

	diff	lwr	upr	p adj
ERC-BP	0.1051	-0.0166	0.2267	1.1805E-01
HL-BP	-0.1478	-0.2446	-0.0509	5.1550E-04
VIS-BP	0.1713	0.0391	0.3036	4.8701E-03
HL-ERC	-0.2528	-0.3739	-0.1318	4.9432E-07
VIS-ERC	0.0662	-0.0846	0.2171	6.7207E-01
VIS-HL	0.3191	0.1874	0.4507	3.0688E-09