Title
Changing Fuel Loading Behavior to Improve Airline Fuel Efficiency

Author
Kang, Lei

Publication Date
2017

Peer reviewed|Thesis/dissertation

Changing Fuel Loading Behavior to Improve Airline Fuel Efficiency

By

Lei Kang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Civil and Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark Hansen, Chair

Professor Joan Walker

Professor Bryan S. Graham

Spring 2017

Changing Fuel Loading Behavior to Improve Airline Fuel Efficiency

By

Lei Kang

Abstract

Changing Fuel Loading Behavior to Improve Airline Fuel Efficiency

by

Lei Kang

Doctor of Philosophy in Engineering – Civil and Environmental Engineering

University of California, Berkeley

Professor Mark Hansen, Chair

This dissertation investigates how to improve airline fuel efficiency by changing fuel loading behavior of airline dispatchers and reducing unnecessary fuel loading. First, we estimate the potential benefit of fuel loading reduction for six major U.S.-based airlines. We find that the annual monetary savings per airline of avoiding carrying unused fuel in 2012 ranges from $42 million to $605 million, with a total across all six airlines of $1.16 billion. This suggests that there may be significant benefit from reducing unnecessary fuel loading in the U.S. airline industry.

Second, to capture that benefit, we first study the behavior of dispatchers – who make fuel loading decisions – based on comprehensive historical fuel burn data provided by a major U.S.-based airline. Risk-neutral and risk-averse newsvendor models are applied to better understand how dispatchers trade off safety concerns due to possibly insufficient fuel loading and extra fuel burn costs due to excess discretionary fuel loading. We find that dispatchers place extremely high priority on safety relative to excess fuel cost in making discretionary fuel loading decisions. Furthermore, combining our results with a dispatcher survey, we also find that dispatchers who are detail oriented and conservationists are likely to load less discretionary fuel. Our results imply that airlines may want to select for these characteristics during dispatcher interview and seek to cultivate such behavior in dispatcher recurrent training.

Finally, besides behavioral modeling of dispatchers, we propose two novel discretionary fuel estimation approaches that can assist dispatchers with better discretionary fuel loading decisions. Based on the analysis on our study airline, our approaches are found to substantially reduce unnecessary discretionary fuel loading while maintaining the same safety level compared to the current fuel loading practice. The idea is that by providing dispatchers with more accurate information and better recommendations derived from flight records and related contextual data, unnecessary fuel loading and corresponding cost-to-carry could both be reduced. The first approach involves applying ensemble machine learning techniques to improve fuel burn prediction and construct prediction intervals (PI) to capture the uncertainty of model predictions. The upper bound of a PI can be used for discretionary fuel loading. The potential benefit of this approach is estimated to be $61.5 million in fuel savings and 428 million kg of CO_2 reduction per year for our study airline. The second approach in estimating discretionary fuel originates from the idea of statistical contingency fuel (SCF). Due to limitations in the current SCF

estimation method, dispatchers have low confidence in applying SCF values and generally load more discretionary fuel than recommended. Therefore, improved SCF estimation offers another practical approach for reducing discretionary fuel loading. The estimated annual benefit of using this approach is $19 million in fuel savings and 132 million kg CO_2 emissions reductions for our study airline. A similar analysis could be easily generalized to other airlines or the industry as a whole when such detailed airline fuel data becomes available.

Contents

List of Figures

List of Table

Acknowledgements

My academic achievement would be meaningless if I could not celebrate it with my family. I would like to express my deepest gratitude to Lulu Cheng, who makes my life colorful and meaningful. Her accompany is the foundation of my Ph.D. work. I want to thank my parents for all the sacrifices they have made to support me to continue my education. My dad always wants me to become a medical doctor. Wish this doctoral degree is a good substitution for his expectation on me. My mom always wants me to be myself and enjoy my own life. This doctoral degree is my gift for her.

I sincerely thank my advisor Professor Mark Hansen for his constant encouragement and guidance. Mark can always find some smart ideas from my stupid questions and guide me towards the right direction. Without his encouragement, I could not finish this degree. I really appreciate his help in making me a better person and researcher.

I would also like to show my appreciation to Professor Joan Walker and Professor Bryan Graham for their valuable comments and generous support on my dissertation work. Additionally, I want to thank Professor Johann Gagnon-Bartsch for brining me into the world of statistics and showing me the beauty of causal inference. Special thanks to Professor Michael Ball, Professor Alan Hubbard, Professor Michael Anderson, Professor Anil Aswani, Professor Alexey Pozdnukhov, and Professor Haiyan Huang, for sharing thoughts with me and shedding lights on my mind.

It is my honor to join the Nextor family and I would like to thank my fellow Nextors: Yulin Liu, Yi Liu, Sreeta Gorripaty, Frank Ketcham, Ivan Tereshchenko, Megan Ryerson, Lu Hao, Bo Zou, Avijit Mukherjee, Michael Seelhorst, and Bowen Yang. It has been a wonderful journey with all of you. I have been in this family for four years and I am so proud of being a member of it.

Thank you everyone at the Berkeley ITS: Dounan Tang, Mogeng Yin, Matt Reiter, Madeleine Sheehan, Akshay Vij, David Shaw, Yanqiao Wang, Andy Mao, Feras Zarwi, Timothy Brathwaite, Lewis Lehe, Andrew Campbell, Sudatta Mohanty, Sid Feygin, Ziheng Lin, David Kan, Wei Ni, Teddy Forscher, Frank Proulx, Paul Anderson, and all other old and new friends, for helping me complete my research and supporting me throughout this process. I feel so lucky to meet every one of you.

1 Introduction

1.1 Background

Reducing fuel consumption is a goal that unifies the aviation industry, owing to the recent significant increase in fuel cost (see Figure 1.1). The price of jet fuel has a significant impact on airline costs. According to the Bureau of Transportation Statistics (BTS), fuel costs account for about 15.5% of total operating expenses for U.S. passenger airlines (BTS, 2016). Furthermore, air transportation contributes 8% of transportation greenhouse gas (GHG) emissions in the U.S. (Environmental Protection Agency (EPA), 2016) and 10.6% of transportation emissions globally (Intergovernmental Panel on Climate Change (IPCC), 2014). The global GHG emissions by 2020 from aviation are projected to be around 70% higher than in 2005 even if fuel efficiency improves by 2% per year (International Civil Aviation Organization (ICAO), 2014). As the share of global emission from aviation is expected to increase dramatically, there is consequently intense focus on reducing fuel consumption from many stakeholders (e.g. governments, airlines, and aircraft manufacturers), who have undertaken a wide range of efforts and initiatives (see Table 1 for summary of selected initiatives).

On the government side, enhanced air traffic management (ATM) aiming at aviation system efficiency has been estimated to provide from 6% to 12% savings in fuel burn (IPCC 1999). For instance, the benefits of the enhanced capacity of the Next Generation Air Transportation System (NextGen) in the U.S. have been estimated at $132.5 billion from delay savings and environmental emissions reductions over the period of 2013–2030 (Federal Aviation Administration (FAA), 2014). All of the emissions reductions and a large portion of the delay savings value are related to fuel burn. Other ATM initiatives include the Single European Sky (SES) program in Europe (European Commission, 2010) and the Seamless Asia Sky (SAS) program in Asia (ICAO, 2012). Besides ATM, regulatory initiatives such as the European Union Emissions Trading System (ETS) (European Commission, 2008) and fuel and environmental taxes are also means to help reduce aviation emissions.

Another way of reducing fuel consumption is through improved fuel efficiency through new designs of aircraft and engines. Intensive efforts have been made in aircraft and engine improvements by aircraft manufacturers. According to the IPCC (2014), aircrafts being produced today are about 70% more fuel efficient per passenger-kilometer than 40 years ago, and fuel efficiency is projected to improve 2% every year for the foreseeable future. One example of such improvement is that the addition of winglets to the wingtips of aircraft has been shown to improve the aerodynamics of aircraft and hence reduce fuel burn by 2.5-5% (Irrgang et al., 2011).

High fuel prices and the environmental effects of aviation are also motivating strong interest in alternative jet fuels. Adoptions for alternative aviation fuels (e.g. bio-jet fuel and synthetic fuel) may also help improve price stability and provide possible reductions in GHG emissions in the face of dramatic fuel cost fluctuation. Since 2006, the FAA, United States Department of Defense (DoD) and the aviation and fuels industries have aligned their efforts to explore the potential for alternative jet fuels through the Commercial Aviation Alternative Fuels Initiative (CAAFI) (Hileman et al., 2008). Although initiatives in support of aviation alternative fuels are active, their timeline is highly uncertain (Ryerson et al., 2015).

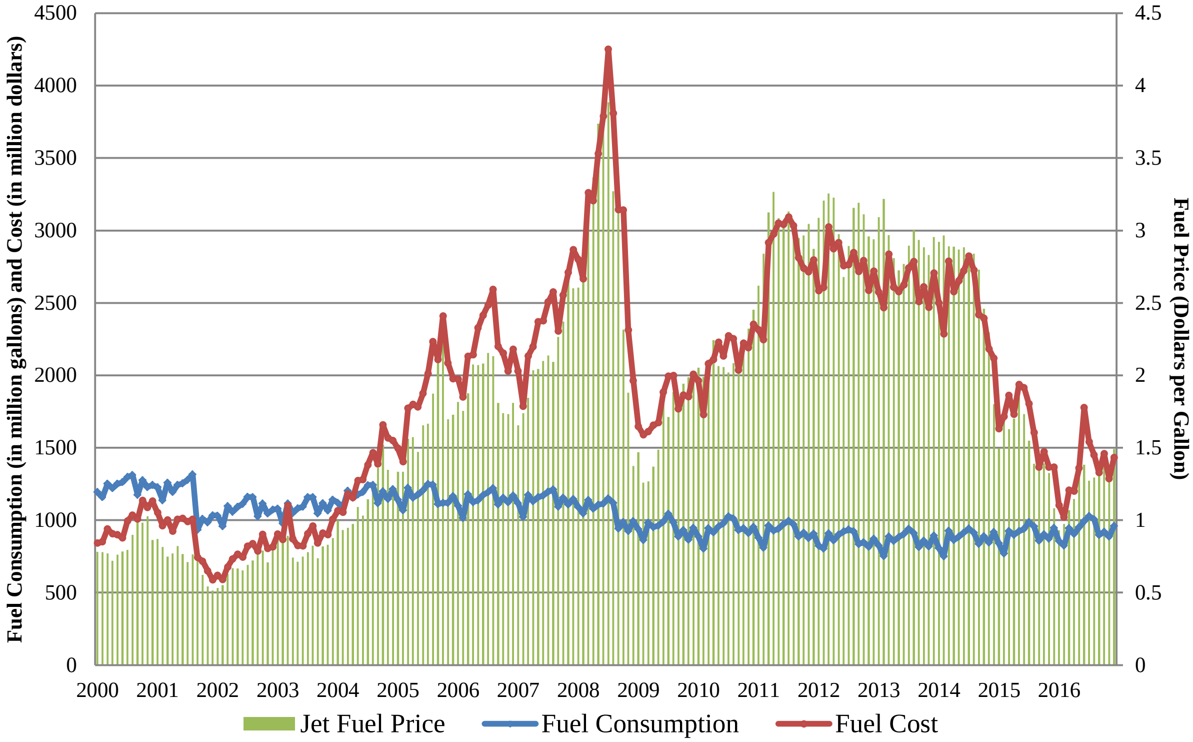


Figure 1.1 Monthly fuel price[1], consumption, and cost (domestic)[2] from 2000 to 2016 in current values

Meanwhile, airlines are also seeking to improve fuel efficiency by improving operations. The fuel consumption of a flight can be influenced by many factors, including flight planning (Schiefer and Samuel, 2011), load factor (International Air Transport Association (IATA), 2004)), ground operations (Hao et al., 2016a), usage of auxiliary power units (APU) (Walker, 2010), single-engine taxi-out and taxi-in (Walker, 2010)), speed and altitude (Lovegren and Hansman, 2011; Roberson, 2010), route selection (Altus, 2009), continuous descent approaches (CDA) (Clarke et al., 2004; Irrgang 2011)), and aircraft maintenance (IATA, 2004; Irrgang 2011). See Table 1.1 for a summary of initiatives for reducing fuel burn by adjusting these various aspects of operations.

Another approach to reducing fuel consumption is to reduce aircraft weight (Ryerson et al., 2015). The lighter the aircraft is, the less thrust is required from the engine for a given speed and altitude, and hence the less fuel is consumed. For this reason, airlines are purchasing aircraft made with lightweight materials (Lee et al., 2009), and charging passengers for luggage (Abeyratne, 2009). Other efforts include accommodating lighter seats and galleys and reducing drinking water loads (European Commission, 2015). However, even though the biggest source of excess weight added to the aircraft is excess fuel (Sadraey, 2012; Irrgang, 2011), there has been little discussion on reducing unnecessary fuel loading. Among very few studies in this field, Ryerson et al. (2015) find that reducing unnecessary fuel loading by dispatchers could result in

---

[1] U.S. Gulf Coast Kerosene-Type Jet Fuel Spot Price:
http://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=pet&s=eer_epjk_pf4_rgc_dpg&f=m
[2] U.S. Carriers Fuel Cost and Consumption: http://www.transtats.bts.gov/fuel.asp

fuel savings on the order of $220 million and 660 million kg CO_2 reduction per year based on the analysis of a major U.S. airline. The fuel saving estimate amounts to 4.48% of total flight fuel consumption for this airline. Motivated by the potential benefit of reducing unnecessary fuel loading demonstrated by Ryerson et al. (2015), based on one major U.S airline, this study focuses on strategies that can help airlines reduce unnecessary fuel loading. First, as part of this study's motivation, we estimate the potential benefit of fuel loading reduction for the U.S. airline industry as a whole. Second, we focus on the behavioral analysis of dispatcher fuel loading decisions. Third, to better assist dispatcher fuel loading, we propose two approaches to improve current airline fuel planning practice. In order to better understand the context of this study, it is useful to describe the general fuel loading practice, before moving on to the problem statement and research questions. In next section, we will focus on the fuel planning process which is a key element of domestic flight planning.

Table 1.1 Summary of fuel consumption reduction initiatives

Stakeholders	Category	Efforts/Initiatives
Government	Enhanced ATM	NextGen (FAA, 2014); SES (European Commission, 2010); SAS (ICAO, 2012)
	Regulation	EU ETS (European Commission, 2008); Fuel and environmental taxes (European Commission, 2015)
Airlines	Operational responses	Optimized flight planning (Schiefer and Samuel, 2011); Single engine taxi (Walker, 2010); Operating at fuel efficient speed and altitude (Lovegren and Hansman, 2011); Minimizing the usage of APU (Walker, 2010); Dynamic route optimization (Altus, 2009); Speed selection through Cost Index (Roberson, 2010); Using CDA (Clarke et al., 2004; Irrgang 2011); Maintaining clean and efficient airframes and engines (IATA, 2004); Maximizing the aircraft's load factor (IATA, 2004)
	Weight reduction	Lightweight materials (Lee et al., 2009); Lighter weight seats and galleys (European Commission, 2015); Luggage charge (Abeyratne, 2009); Fuel Loading (Ryerson et al., 2015)
Aircraft Manufactures	Aircraft and engine improvements	Winglets (Irrgang et al., 2011); Engine evolution (IPCC, 1999; European Commission, 2015); Alternative fuels (IPCC, 1999; European Commission, 2015)

1.2 Domestic Flight Planning Fundamentals

Airlines rely on flight dispatchers to perform the duty of flight planning. Based on a study of a U.S. major airline (which is also the study airline for the following analysis), we summarize the typical flight planning process. For both domestic and international flights, a flight plan is created and submitted from a dispatcher to Air Traffic Control (ATC) for clearance around two hours prior to the scheduled departure time and will be revised if necessary based on updated information. A flight plan[3] is a regulatory requirement, and its contents are developed in order to ensure an airplane meets all of the operational regulations for a specific flight, to give the flight crew information to help them conduct the flight safely, and to coordinate with ATC (Altus, 2009). Given a pair of cities, a flight plan could be characterized by two inter-related parts: route-related decisions (the route of flight, the planned speed, and the planned altitudes along the route) and fuel planning (the quantity of different categories of fuel loaded).

More specifically, the duties of a dispatcher involve carrying out strategic decisions such as checking weather forecasts and operating conditions, selecting routes and flight levels, and determining fuel loads, as well as tactical decisions, such as providing pilots with real-time updates, coordinating between various parties to resolve maintenance issues, and continuously monitoring the flight from takeoff to landing (Ryerson et al., 2015). Assisted by the Flight Planning System (FPS), dispatchers choose a route of flight among several possible routes. A dispatcher will generally prefer the “economical” route, or “econ” route, for short (the route with the lowest fuel consumption). However, the selected route might deviate from econ route due to forecasted en route weather or traffic constraints.

Regarding fuel planning, there are generally four components to the fuel loaded on the airplane: mission fuel, reserve fuel, alternate fuel, and contingency fuel. The amount of mandated fuel needs to satisfy the U.S. Federal Aviation Regulations (FARs) for over-land flights (i.e. within the 48 contiguous states and the District of Columbia), or both the FARs and the International Civil Aviation Organization (ICAO) Annex 6 for over-water flights (outside the 48 contiguous states and the District of Columbia) and international flights. Domestic (over-land) and international flights (over-water plus outside the U.S.) require different considerations in dispatch and fuel planning. In this study, we will focus on domestic operations.

According to the U.S. FARs Part 121.639 (e-CFR: Title 14, 2015)[4], fuel requirements for all domestic operations under Instrument Flight Rules (IFR) include:

(a) enough fuel to fly to the airport to which the airplane is dispatched;

(b) thereafter, to fly to and land at the most distant alternate airport (if required by weather conditions) from the airport to which it is dispatched; and

(c) thereafter, to fly for 45 minutes at normal cruising fuel consumption.

Item (a) is described as mission fuel, which is the amount of fuel required from take-off at the departure airport to landing at the destination airport. This fuel is calculated by the FPS based on factors including the route selection, altitude structure, winds and other weather forecasts, anticipated traffic delays, and aircraft performance.

Item (c) is termed (domestic) reserve fuel. The quantity of reserve fuel is required for the aircraft to continue to fly for 45 minutes at normal cruising speed, presumably to enter a holding

[3] A sample flight plan form could be found at http://www.faa.gov/documentLibrary/media/Form/FAA_7233-1_PRA_revised_12-2013.pdf

[4] http://www.ecfr.gov/cgi-bin/text-idx?SID=a8d3c4800d167b64bbfa2349ec337755&mc=true&node=pt14.3.121&rgn=div5#se14.3.121_1639

pattern above either the destination airport or an alternate airport or to enter a holding pattern en route in the case of reduced airport or airspace capacity (Ryerson et al., 2015). This fuel requirement cannot be changed by the dispatcher, and the amount of fuel is calculated based on aircraft type. According to the interview with dispatchers of the study airline, reserve fuel should be completely unused when the flight lands, except in extraordinary circumstances.

Item (b) is called alternate fuel, which is the quantity that would be needed to fly from the destination airport to the alternate airport. An alternate airport is an airport in the vicinity of the destination airport which can be used in case the destination airport becomes unusable while the flight is in progress (e.g. due to weather conditions). An alternate is required by the FARs when the forecasted visibility is less than 3 mile or the ceiling at the destination airport is less than 2,000 feet at the flight's estimated time of arrival, plus or minus one hour. The policy of our study airline also recommends adding an alternate when thunderstorms are forecast for the same time window. A second alternate might also be added if the forecasted weather conditions at both the destination and first alternate airport do not meet the minimum ceiling and visibility requirement. When a dispatcher enters an alternate airport into the release, the FPS calculates the amount of alternate fuel needed.

The fourth common category of loaded fuel is called contingency fuel, which is usually added to the release by dispatchers to allow for known and unknown airborne contingencies such as arrival delays and in-flight weather changes. This added fuel is based on the dispatcher's experience on a certain route and aircraft (Trujillo, 1996) and reflects a dispatcher's assessment of traffic and weather uncertainties. In addition to adding contingency fuel labeled as such, the alternate airport and the corresponding alternate fuel might also be added by dispatchers as a form of contingency fuel to provide an extra buffer for unexpected events even when an alternate is not required by weather conditions.

To provide consistent and objective fuel planning, some FPSs provide recommended contingency fuel numbers based on a statistical analysis of historical fuel consumption for similar flights. Carriers usually term this Statistical Contingency Fuel (SCF) (Schiefer and Samuel, 2011). In the case of our study airline, the set of similar flights consists of those that took place over the previous year (365 days) and have the same origin, destination, and scheduled hour of departure. For each historically similar flight, the difference between the actual trip fuel consumption and the planned mission fuel consumption is calculated. If this difference is negative, it is then called "under-burn"; otherwise, it is termed "over-burn." We will call this the over/under burn value. The FPS converts the over/under burn value in pounds to minutes and estimates a normal approximation of the distribution of this excess required fuel burn. The 95th and 99th percentiles of the distribution, which are also called the SCF95 and the SCF99, are provided to dispatchers by the FPS as guidelines for contingency fuel loading. The interpretation of SCF95 (SCF 99) is that based on historical fuel consumption, loading the quantity of contingency fuel specified by SCF95 (SCF99) would result in a flight being able to land without dipping into any reserve fuel 95% (99%) of the time, assuming not alternates. Because the SCF values are based on actual fuel consumption, these contingency guidelines implicitly account for weather and other events in history. More details regarding SCF can be found in Karisch et al. (2012). However, in the airline focused upon in this case, these SCF values are rarely trusted, and dispatchers typically load much more contingency fuel on a flight than the SCF numbers recommend (Ryerson et al., 2015).

One final loaded fuel type arises as a result of fuel cost differences; it is sometimes less expensive to carry fuel into particular airports for subsequent use than it is to purchase it at these airports. Therefore, airlines might decide to carry tanker fuel. Other considerations also feature in tanker fuel loading. For instance fuel may be unavailable because the destination airport is in a remote location, or the airline may not be permitted to refuel at certain airports because of fuel contracts (Guerreiro et al., 2013). For some airlines, dispatchers may also need to decide taxi fuel loading (Trujillo, 1996). Similar to contingency fuel, the FPS will provide historically average taxi fuel. The dispatcher can make adjustments to the suggested taxi fuel based on experience, weather, and traffic conditions at the airport.

1.3 Problem Statement

As discussed in Section 1.2, loaded fuel can be generally categorized into mission fuel, reserve fuel, and discretionary fuel. Here, we define discretionary fuel as the summation of contingency, non-weather-required 1st alternate fuel, and all 2nd alternate fuel added by dispatchers. Reserve fuel quantity is mandated by the federal government and mission fuel (as well as required alternate) is calculated by the FPS, making these quantities non-discretionary.

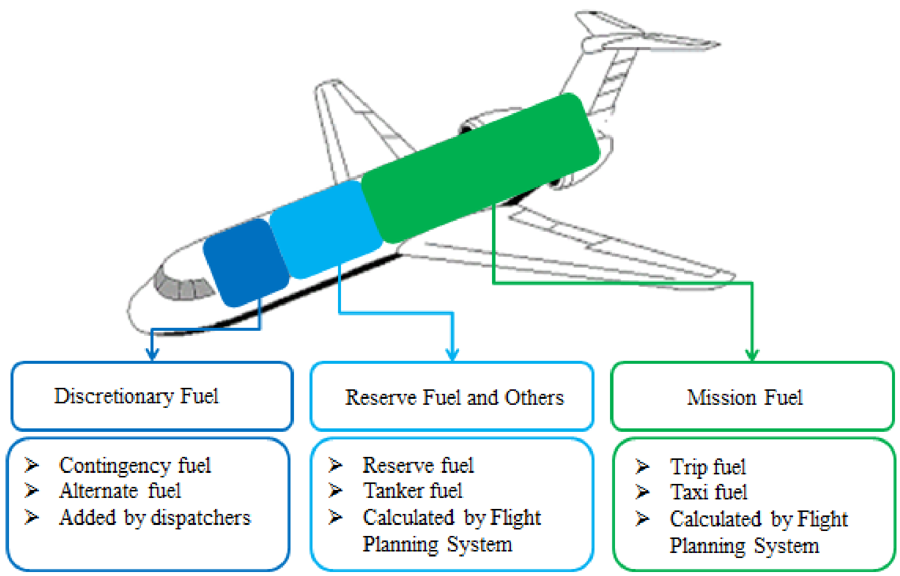


Figure 1.2 Fuel categorization

On top of mission fuel and reserve fuel, airline dispatchers will always load some discretionary fuel on the aircraft to hedge against various uncertainties. In other words, dispatchers load discretionary fuel to cover the uncertainty associated with the over/under burn distribution (and possible need to use an alternate) so that a flight can land without touching reserve fuel. However, dispatchers are found to be conservative in discretionary fuel loading. This behavior has been observed by Trujillo (1996), Ryerson et al. (2015), and Hao et al. (2014). For instance, based on a survey of 50 U.S. pilots and dispatchers about their fuel-loading practices, Trujillo (1996) finds that airline dispatchers and pilots load contingency fuel above the contingency value suggested by the airline. Moreover, the amount of contingency fuel has been found to be related to weather uncertainty and system predictability (Hao et al., 2014). In some cases, this extra fuel is actually needed, but in most cases, the vast majority of it remains in the tanks when the flight pulls into the destination gate.

Therefore, given the significant potential benefit of reducing fuel loading, the objectives of this study are two-fold: (1) estimate the benefit of reducing unnecessary fuel loading at the airline level; (2) seek strategies that can help reduce unnecessary discretionary fuel loading while maintaining the same safety level.

1.4 Overview

To achieve the first goal mentioned above, we estimate the potential benefit of fuel loading reduction for six major U.S.-based airlines. In particular, the cost to carry (CTC) unused fuel when flights push into the gate (termed as gate-in fuel [GiF]) will be evaluated. To the best of our knowledge, this study offers the first comprehensive evaluation of fuel costs due to carrying extra fuel at an airline level. The annual monetary savings of avoiding carrying unused fuel for six airlines in 2012 ranges from $42 million to $605 million, with a total across all six airlines of $1.16 billion. This analysis justifies the importance of considering reducing unnecessary fuel and provides us with a benefit pool. Details can be found in Chapter 3.

To answer the second question, about how to (safely) reduce discretionary fuel loading, we will leverage the detailed flight-level fuel data of one major U.S.-based airline and tackle this problem from two directions. The first direction is to study dispatcher discretionary fuel loading behavior and identify opportunities to reduce fuel loading. In Chapter 4, a newsvendor model is applied to better understand how dispatchers trade off safety concerns due to possibly insufficient fuel loading and extra fuel burn cost due to excess discretionary fuel loading. Under several identification assumptions (see Chapter 4 for details), we find that dispatchers place extremely high priority on safety relative to excess fuel cost in making discretionary fuel loading decisions. Combining our results with a dispatcher survey, we also find that dispatchers who are detail oriented and conservationists are likely to load less discretionary fuel. When airlines interview dispatchers, in addition to skill- and behavior-based performance evaluations, it would be helpful to also test the detail-orientation of dispatchers, as well as their belief in conservation. This might help airlines to select dispatchers who are less likely to overload fuel resulting in potential fuel savings. Airlines may also target dispatchers for recurrent training. Adding training topics on detail-orientation and conservation may also encourage dispatchers to load less unnecessary fuel.

While Chapter 4 focuses on the behavioral modeling of dispatcher discretionary fuel loading, in Chapter 5 and 6 we explore the second direction by proposing two novel discretionary fuel estimation approaches. The idea is that by providing dispatchers with more accurate information and better recommendations, unnecessary fuel loading and CTC unnecessary fuel loading could both be reduced. In Chapter 5, a prediction interval (PI)-based discretionary fuel estimation approach is discussed. It involves applying ensemble machine learning techniques to improve fuel burn prediction and construct PIs to capture the uncertainty of over/under burn distribution. The upper bound of a PI can be used for discretionary fuel loading. The potential benefit of this approach is estimated to be $61.5 million in fuel savings and 428 million kg of CO_2 reduction per year for our study airline.

The second approach in estimating discretionary fuel originates from the idea of SCF. Due to limitations in the current SCF estimation method (see Chapter 6 for details), dispatchers have low confidence in applying SCF values (e.g. SCF95) and generally load more discretionary fuel than recommended. Therefore, improved SCF estimation offers another practical approach

for reducing discretionary fuel loading. The estimated annual benefit of using this approach is \$19 million in fuel savings and 132 million kg CO_2 emissions reductions for our study airline.
In sum, the research questions addressed in this study include the following:
(1) What is the size of the benefit of reducing unnecessary fuel loading? (Chapter 3)
(2) How do dispatchers' make a trade-off between safety and cost in discretionary fuel loading, and how are these trade-offs influenced by dispatcher attitudes? (Chapter 4)
(3) How can reliable discretionary fuel recommendations be provided to dispatchers? (Chapter 5 and 6)

Chapter 2 summarizes several data sources being used in this study. Summary and conclusions can be found in Chapter 7.

2 Data

In this chapter, we will summarize six datasets used in this study. The FAA fuel burn data (Section 2.1) and BTS T-100 data (Section 2.2) are used to estimate potential benefit of reducing unnecessary fuel loading for six major U.S. airlines. However, the FAA fuel burn data doesn't include detailed fuel-loading decisions from the six airlines. In order to better understand airline fuel loading behavior as well as develop fuel efficiency improvement strategies, we also leverage a detailed airline fuel burn dataset (Section 2.3) along with a dispatcher survey (Section 2.4). This offers us an excellent opportunity to study airline dispatchers' behavior in fuel loading decisions. One limitation of this detailed airline fuel burn dataset is the lack of weather and traffic information. To this end, we also incorporate terminal weather forecast (Section 2.5) and historical traffic information (Section 2.6) into this analysis. The applications of detailed airline fuel dataset will be discussed in Chapter 4, 5, and 6.

2.1 FAA Fuel Burn Data

Recent legislation has mandated that airlines provide flight-level fuel data to the FAA (referred to as *FAA data*). The FAA collects detailed flight-level information from six major airlines for 1,956,822 domestic flights (airlines being presented anonymously in the following text). The *FAA data* covers domestic operations between January 2012 and September 2014. Specifically, for a given flight, this dataset contains its origin and destination (OD) airports, departure and arrival time, aircraft type, aircraft pushback weight, aircraft fuel burn for three flight phases (taxi-out, airborne, and taxi-in). However, the *FAA data* does not include detailed fuel-loading information.

2.2 Bureau of Transportation Statistics T-100 Data

The BTS T-100 Domestic Segment data provides monthly aggregate payload information by OD and aircraft type for each airline. This aggregate payload information is expressed in terms of monthly summations of number of passengers transported, freight[5] transported (in pounds), and mail transported (in pounds). It also contains monthly scheduled and actually performed operations for each OD, airline, and aircraft type combination. We combine the BTS payload data with the *FAA data* to estimate the quantity of unused fuel as well as the CTC that unused fuel when a flight lands. Details can be found in Chapter 3.

2.3 Airline Fuel Data

A major U.S.-based airline (different from the six airlines reported to the FAA) provided detailed flight-level information from its domestic and international flights between April 2012 and July 2013 (referred to as *airline data*). This airline operates an extensive domestic network during this period. In addition to basic flight operation characteristics (i.e. aircraft type, OD airports, flight distance, scheduled and actual departure and arrival information), this dataset also contains flight-level fuel loading data in all categories (mission, reserve, tinker, contingency, 1st alternate, and 2nd alternate fuel), target GiF, SCF in units of both pounds and minutes, aircraft pushback weight in units of pounds. It also provides actual fuel burn quantities (in pounds) by flight phase,

[5] Property, other than express and passenger baggage transported by air.

including taxi-out, en route, and taxi-in. For some aircraft, the fuel on arrival is keyed in by the pilot, and as a result is sometimes entered incorrectly. We therefore exclude any flights for which the fuel on arrival is outside of a reasonable range defined by our study airline[6] and flights for which the fuel on arrival is greater than the fuel loaded prior to departure. After excluding outliers with respect to pushback weight, taxi-in fuel burn, taxi-out fuel burn, en route fuel burn, gate-out fuel, and GiF, there were a total of 728,455flights in the airline data within the continental US (with eight major aircraft types[7]).

2.4 Dispatcher Survey Data

Besides fuel data, this major airline also provided results of a dispatcher survey (referred to as *survey data*). This data covers self-reported confidence in decision-support tools, flight planning behaviors, and three sets of personal assessment statements with respect to environment, personal habit, and risk. These questions, designed to capture attitudes and habits of a dispatcher, ask about individuals' degrees of agreement on a number of statements, with answers ranging from 1 (strongly disagree) to 5 (strongly agree). In Section 4.4, we discuss how to leverage these survey responses to better understand dispatcher fuel loading behavior. With survey data and airline data merged, 109 dispatchers are matched with corresponding flights.

2.5 Weather Data

The airline data does not include forecast or actual weather data, and instead we collected weather data from the National Oceanic and Atmospheric Administration (NOAA) database. The weather data included both the actual weather and weather forecast (i.e. terminal aerodrome forecasts [TAFs]) information for major U.S. airports. The actual and forecast weather information contains ceiling, visibility and indicators of the presence of thunderstorms, snow, and visibility conditions by hour, date, and airport.

Dispatchers typically make fuel-loading decisions around the time the flight plan is created. Small deviations in this time do occur from flight to flight, but typically these decisions are made two hours prior to the flight's scheduled departure time. We will refer to this time as the dispatch time. According to an on-site airline interview conducted by Hao et al. (2016b), flight dispatchers also consider actual weather conditions at the dispatch time when making fuel decision. Thus, we merged the weather data with the *airline data* to recreate both the real-time and forecast weather available at the dispatch time for each flight. For the real-time weather, we found the actual weather at the OD airports at the dispatch time. For the forecast weather, we found the most recent forecast issued prior to the dispatch time and refer to the forecast conditions for the origin at the planned departure time and for the destination at the planned arrival time. This matching allowed us to identify the following weather conditions for each flight: (1) the actual weather at the OD airports two hours prior to the scheduled flight departure time (as the actual weather at time of flight planning could influence dispatcher fueling decisions) and (2) the most recent forecasted weather (as forecasted two hours prior to the flight departure) for the OD airports at the scheduled times of arrival and departure. More details about this weather data could be found in Hao et al. (2016b).

[6] Ratio of target gate-in fuel over actual gate-in fuel within [0.7, 1.3].
[7] A319, A320, B757-300, B757-200, M88, M90, DC9, B737-800

2.6 Aviation System Performance Metrics

The FAA Aviation System Performance Metrics (ASPM) database includes individual flight data for the 77 large airports in the U.S.. The *airline data* does not contain en route traffic information, so we used historical traffic conditions obtained from the ASPM as a proxy. For a given flight in the *airline data*, we considered at its historical flights with same OD pair, scheduled departure hour, and month as in the previous year. Then we calculated the mean and standard deviation (SD) of airborne time based on these historical similar flights. To ensure reliable estimates for SD, only combinations of these arguments with more than 10 flights were kept in our dataset. This historical airborne time information provided a good approximation of possible weather conditions for a current flight. The deviation of historical airborne times and the corresponding flight plan airborne times served as another measure of flight time variability. Hence, we also computed the mean and SD of the difference between actual airborne time and planned airborne time, based on a current flight's previous year's counterparts.

2.7 Summary

The following chapters are based on the datasets described in this chapter. After merging weather data and ASPM data with the *airline data*, we ended up with 368,607 flights with no missing weather information. This merged dataset is used in Chapters 5 and 6. Regarding dispatcher fuel loading behavior modeling, we merged the original dataset with dispatcher survey data. We retained 175,617 flights, corresponding to 109 dispatchers. Data summary is provided in Table 2.1.

Table 2.1 Variable description

Category	Variable	Note	Data Source
Weather	Ceiling condition at origin and destination airports	Current weather at the time of flight planning	NOAA
	Visibility condition at origin and destination airports		
	Thunderstorm condition at origin and destination airports		
	Snow condition at origin and destination airports		
	IFR condition at origin and destination airports		
	Forecasted ceiling condition at origin and destination airports	Forecasted weather at the time of scheduled departure and arrival	
	Forecasted visibility condition at origin and destination airports		
	Forecasted thunderstorm condition at origin and destination airports		
	Forecasted snow condition at origin and destination airports		
Traffic Conditions (in minutes)	Standard deviation of airborne time	Based on flights falling into same OD-hour-month in previous year	ASPM
	Standard deviation of the difference in airborne time from flight plan		
	Average airborne time		
	Average difference in airborne time from flight plan		
Dispatcher Characteristics	Age, working experience	Personal characteristics	Survey
	Personal attributes	Latent variables	
Fuel Loading (in minutes)	Contingency fuel loading	Target variable	Airline Data
	Alternate airport and alternate fuel	Target variable	
	Actual fuel burn	Target variable	
	Planned fuel burn	--	
	SCF95 & SCF99	Target variable	
	Other categories of fuel	--	
Flight Information	Schedule departure/arrival time	--	
	Aircraft type	8 major types	
	Origin-Destination airports	--	
	Flight distance (in nautical miles)	--	
	Planned trip time (in minutes)	--	

3 Benefit Pool from Reducing Unnecessary Fuel Loading

In this chapter, we characterize the extent of the benefit of reducing fuel loading at an airline level. We estimate the potential benefit of fuel-loading reduction for six airlines using the *FAA data*. In particular, CTC GiF is evaluated.

Gate-in fuel is the amount of fuel left in the tanks when a flight pulls into the destination gate. In general, the amount of discretionary fuel loaded could be related to flight predictability, as discussed in Section 1.2. Ideally, in a perfectly predictable system (with no weather or traffic uncertainty, no human error, and no reserve requirements), a flight could be loaded with optimal amount of fuel and land with no fuel left. Under this scenario, the value of GiF is zero. In reality, however, due to the unpredictable nature of the system, the amount of fuel loaded will always be greater than the actual fuel burn. This eventually creates positive GiF. In fact, as we mentioned before, a certain amount of extra fuel is mandated by government regulations. This fuel includes reserve fuel and, in some circumstances, fuel to fly to an alternate airport. However, airline dispatchers may also load extra fuel above these requirements. This behavior has been investigated by Trujillo (1996) and Ryerson et al. (2015). In some cases, this extra fuel is actually needed, but in most cases, the vast majority of it remains in the tanks when the flight pulls into the destination gate. Based on one major U.S. airline data, Ryerson et al. (2015) estimate the CTC GiF would result in around 223 million dollars in annual fuel burn cost. When there is low predictability, the difference between actual fuel burn and fuel loaded will, on average, be greater, resulting in higher GiF. Therefore, assuming consistent fuel-loading behavior of airline dispatchers, changes in GiF over time will also reflect trends in system predictability.

Given the motivation described above, the benefit of reducing unnecessary fuel loading could be evaluated based on the estimation of the CTC GiF. Unfortunately, the *FAA data* does not contain GiF, thus requiring us to estimate GiF for each airline based on aircraft departure weight and fuel consumption. To do so, we use the *airline data*. Our goal is to use the *airline data* to develop a GiF prediction model that can be applied to other airlines reported in the *FAA data*.

3.1 Methodology

One possible GiF modeling strategy is to follow the physical model of aircraft weight (Anderson, 2005) as described in Equation 3.1. Since we are trying to develop a GiF prediction model based on the *airline data* and then apply it to the *FAA data*, this specification can help avoid the impact of airline-specific fuel loading behavior, thus possessing high transferability. For a given flight *i*, its GiF quantity can be obtained by subtracting operating empty weight (OEW)[8], total fuel burn (FB), and payload (PL) from its pushback weight (PB).

$$GIF_i = PB_i - OEW_i - FB_i - PL_i \tag{3.1}$$

Operating empty weights for major aircraft types considered in this paper are presented in Table 3.1.

[8] Operating empty weight (OEW) is the basic weight of an aircraft including the crew, all fluids necessary for operation such as engine oil, engine coolant, water, unusable fuel and all operator items and equipment required for flight but excluding usable fuel and the payload.

Table 3.1 Operating empty weight

Aircraft Type (Engine type)	Operating Empty Weight (lbs)
Airbus A319-100(CFM56)	87930
Airbus A321-200 (V2500s)	105875
Airbus A320-200 (V2500s)	93079
Boeing 737-300 (CFM56)	72490
Boeing 737-400 (CFM56)	76200
Boeing 737-500 (CFM56)	70510
Boeing 737-700 (CFM56)	84100
Boeing 737-800 (CFM56)	90710
Boeing 737-900 (CFM56)	93680
Boeing 757-200 (RB.211 SERIES)	220000
Boeing 757-300 (RB.211 SERIES)	142400
Boeing 767-200 (CF6-80 SERIES)	177500
Boeing 767-300ER (CF6-80 SERIES)	198800
Boeing 767-300 (PW4000 SER)	192100
Boeing 767-400 (CF6-80 SERIES)	227300
Boeing 777-200 (GE90)	304500
McDonnell Douglas MD-82/83 (JT8D SERIES)	78000
McDonnell Douglas MD-88	77976
McDonnell Douglas MD-90-30	88000
Douglas DC-9-50	61880
Embraer 190 (GE CF34-10E)	61910

Source:
http://www.airliners.net/aircraft-data/
http://en.wikipedia.org/

However, flight level payload information is not available in either *airline data* or *FAA data*. We therefore use BTS payload data as a proxy for actual payload. In this regard, BTS T-100 Domestic Segment data (discussed in Section 2.2) which provides monthly aggregate payload information by OD, airline, and aircraft type for the study time period (January 2012 through September 2014), was merged with the *FAA data* and *airline data*. This aggregate payload information is expressed in the form of monthly summations of number of passengers transported, freight[9] transported (in pounds), and mail transported (in pounds). By dividing the total number of performed operations in a specific OD, airline, month, year, and aircraft type combination, we can get its mean payload quantity. In order to develop prediction models for GiF using monthly payload data, we aggregate the airline data based on OD-airline-month-year-aircraft type (denoted by c) and merged that data with the BTS payload data[10]. Thus, a mean GiF is enabled to match with its mean payload and mean fuel burn in a more accurate manner. Equation 3.1 can thus be replaced by Equation 3.2, with mean representation.

[9] Property, other than express and passenger baggage transported by air.

[10] To reduce the error caused by unusual operations, we exclude combinations with 10 operations or less. Moreover, due to outliers removal in the early data cleaning stage, the performed number of operations reported in the BTS data for a given type combination may not match the number of operations left in the *airline data*. To improve approximation accuracy of aggregate level payload statistics obtained from the BTS to actual payload, only combinations with match rate (defined as the ratio between BTS reported operations and airline recorded operations) within [0.8, 1.2] were left. Ultimately, 9608 combinations were generated based on the airline data for later model development.

$$\overline{GIF}_c = \overline{PB}_c - OEW_{type} - \overline{FB}_c - \overline{PL}_c \tag{3.2}$$

Here we use *type* to denote aircraft type. Furthermore, payload can be categorized into freight, mail, and passenger transported and fuel burn can be decomposed into taxi-in, airborne, and taxi-out fuel burns, see Table 3.2 for details. Thus we were able to establish the relationship between the mean GiF ($\overline{GIF}_c$) and its mean pushback weight ($\overline{PB}_c$), mean payload (represented by mean freight ($\overline{PL\_F}_c$), mail ($\overline{PL\_M}_c$), passenger transported ($\overline{PL\_P}_c$), mean fuel burn (represented by mean taxi-out ($\overline{FB\_OUT}_c$), airborne ($\overline{FB\_AR}_c$), and taxi-in fuel burn ($\overline{FB\_IN}_c$).Note that while freight and mail payload data are in pounds, passenger payload is in units of passengers, which needs to be converted to pounds. The means for doing this is discussed below.

Table 3.2 GIF Physical Model Notations

Notations		Explanations
$\overline{GIF}_c$		Mean GIF
$\overline{PB}_c$		Mean pushback weight
$\overline{PL}_c$	$\overline{PL\_F}_c$	Mean freight payload
	$\overline{PL\_M}_c$	Mean mail payload
	$\overline{PL\_P}_c$	Mean number of passenger transported
$\overline{FB}_c$	$\overline{FB\_OUT}_c$	Mean taxi-out fuel burn
	$\overline{FB\_AR}_c$	Mean airborne fuel burn
	$\overline{FB\_IN}_c$	Mean taxi-in fuel burn
OEW_{type}		Operating empty weight

Based on the above-mentioned decomposition, Equation 3.2 can be expressed in a statistical manner:

$$\begin{aligned} \overline{GIF}_c &= \beta_0 + \beta_1 * \overline{PB}_c + \beta_2 * \overline{FB\_IN}_c + \\ &\beta_3 * \overline{FB\_OUT}_c + \beta_4 * \overline{FB\_AR}_c + \beta_5 * \overline{PL\_M}_c + \\ &\beta_6 * \overline{PL\_F}_c + \beta_7 * \overline{PL\_P}_c + \beta_8 * OEW_{type} + \varepsilon_c \end{aligned} \tag{3.3}$$

where $\beta's$ are the parameters associated with each variable. In the physical model setting, intercept β_0 and error term ε should be equal to 0, β_1 should be equal to 1 while other parameters, except for β_7, all equal to -1.

3.2 Estimation Results

We investigate the performance of the physical model prediction (Equation 3.3) in the following section. Mean GiF is treated as the dependent variable and under the physical model, as shown in Equation 3.3, we fix the parameters of mean taxi-in fuel burn, mean taxi-out fuel burn, mean airborne fuel burn, mean mail weight, mean freight weight, and OEW to be -1 and constrain the parameter of mean pushback weight to be 1. However, we also need to estimate average passenger weight (denoted by β_7, see Table 3.3) which is the weight summation of a passenger and that passenger's checked bags and carry-on items. Aircraft-specific fixed effects were also included to capture unobserved mean effects and ensure the unbiased estimation of average passenger weight.

The estimation results are shown in Table 3.3. The average passenger weights vary by month; for example, the estimated winter average passenger weights are greater than summer weights, due to clothing. These values are consistent with the FAA's suggested values[11] (FAA, 2005). The estimated negative aircraft-type-specific effects indicate that there are missing items not being captured, presumably due to differences in OEW between the aircraft used in the observed flights and the values in Table 3-1. These differences could be caused by differences in fixed equipment, repairs and alteration, and aircraft aging (FAA, 2007). As an aircraft ages, its weight usually increases due to trash and dirt collecting in hard-to-reach locations, and moisture absorbed in the cabin insulation (FAA, 2007). Usually weight grows average 0.1% to 0.2% per year, up to about 1% total (Irrgang, 2011) due to aging. In the following prediction analysis, we assume that the estimated aircraft-specific effects do not carry over to other airlines, and GiF is predicted following the physical model using the estimated passenger weights in Table 3.3[12].

[11] The standard average weight for a checked bag is 30 lbs. Standard average weight for adult male is 205 lbs at winter and 200 lbs at summer, whereas for adult female is 184 lbs at winter and 179 lbs at summer. Child weight is 87 lbs at winter and 82 lbs at summer. Thus 213 lbs per passenger is a reasonable estimation.

[12] Another option is to assume the estimated fixed effects are in general aircraft-type specific and we tried to apply the fixed effects estimates to other airlines, but for one airline, we got a large number of negative GIF predictions. So we decide to make an assumption that these estimated fixed effects are airline-specific.

Table 3.3 Estimation results (dependent variable: mean gate-in fuel in lbs)

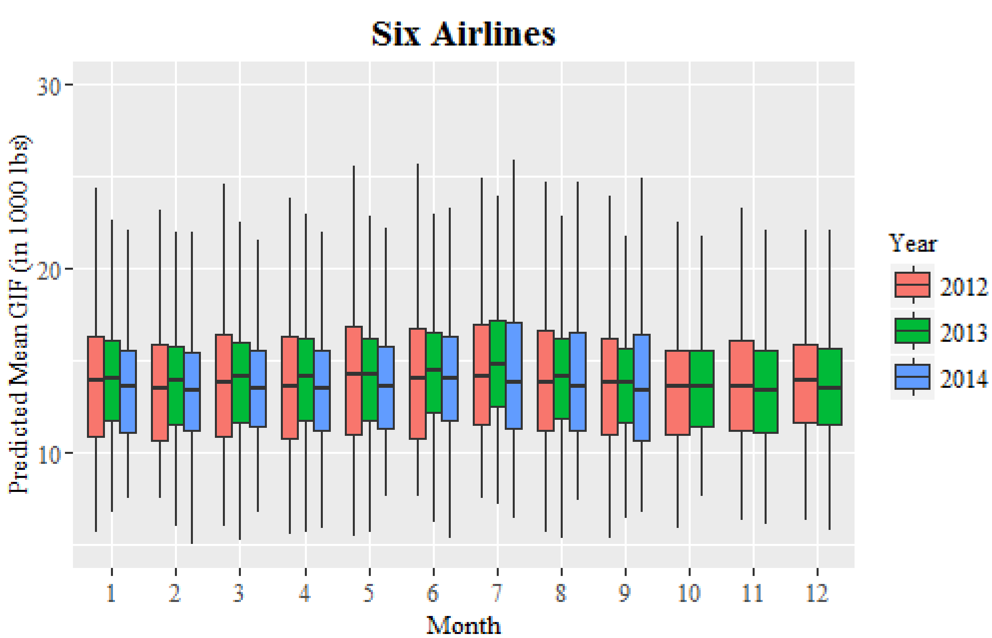
Notation	Variable	Parameter Estimates (T-statistics)
β_0	A319	-4915.65 (-74.24) ***
	A320	-3519.46 (-45.57) ***
	B737-800	-7625.63 (-90.47) ***
	B757-300	-9220.69 (-71.65) ***
	B757-200	-8806.77 (-81.79) ***
	DC9	-9231.69 (-145.12) ***
	M88	-9235.17(-122.00) ***
	M90	-7853.08 (-91.44) ***
β_1	Mean pushback weight (lbs)	1 (fixed)
β_2	Mean taxi-in fuel burn (lbs)	-1 (fixed)
β_3	Mean taxi-out fuel burn (lbs)	-1 (fixed)
β_4	Mean airborne fuel burn (lbs)	-1 (fixed)
β_5	Mean mail weight (lbs)	-1 (fixed)
β_6	Mean freight weight (lbs)	-1 (fixed)
β_7	Average number of passengers in January	-208.62 (-305.97) ***
	Average number of passengers in February	-208.77 (-288.02) ***
	Average number of passengers in March	-206.36 (-298.94) ***
	Average number of passengers in April	-205.06 (-311.01) ***
	Average number of passengers in May	-202.03 (-320.95) ***
	Average number of passengers in June	-199.31 (-320.03) ***
	Average number of passengers in July	-196.47 (-326.94) ***
	Average number of passengers in August	-199.28 (-313.00) ***
	Average number of passengers in September	-203.20 (-315.89) ***
	Average number of passengers in October	-203.05 (-294.01) ***
	Average number of passengers in November	-206.57 (-310.02) ***
	Average number of passengers in December	-207.53 (-300.98) ***
β_8	Operating empty weight (lbs)	-1 (fixed)
Residual standard error		897
10-fold cross validation error		899
Number of observations		9,608
Adjusted R-squared		0.76

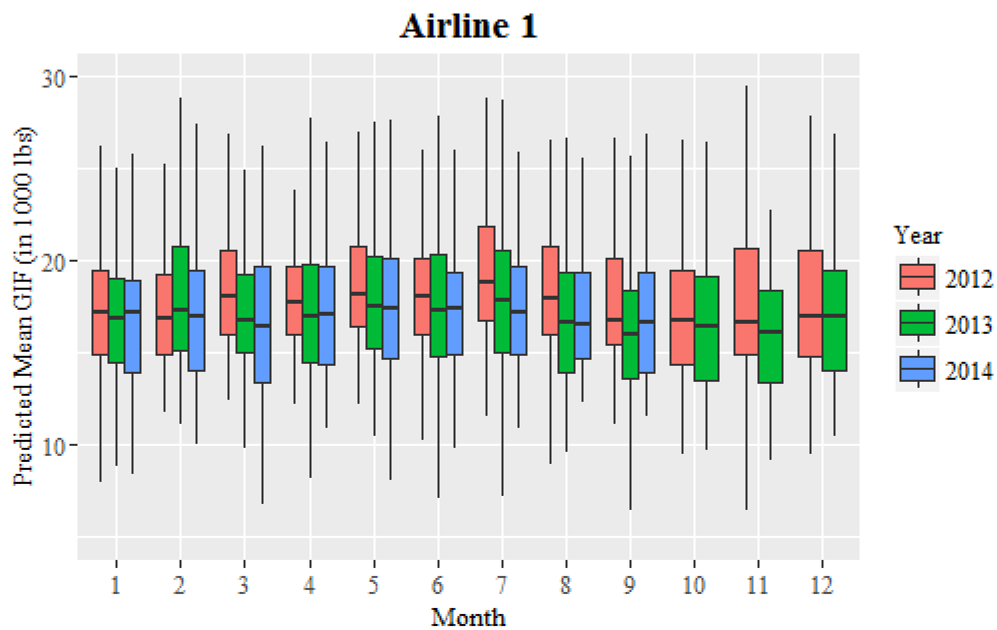
Note: *** indicates 1% significance level.

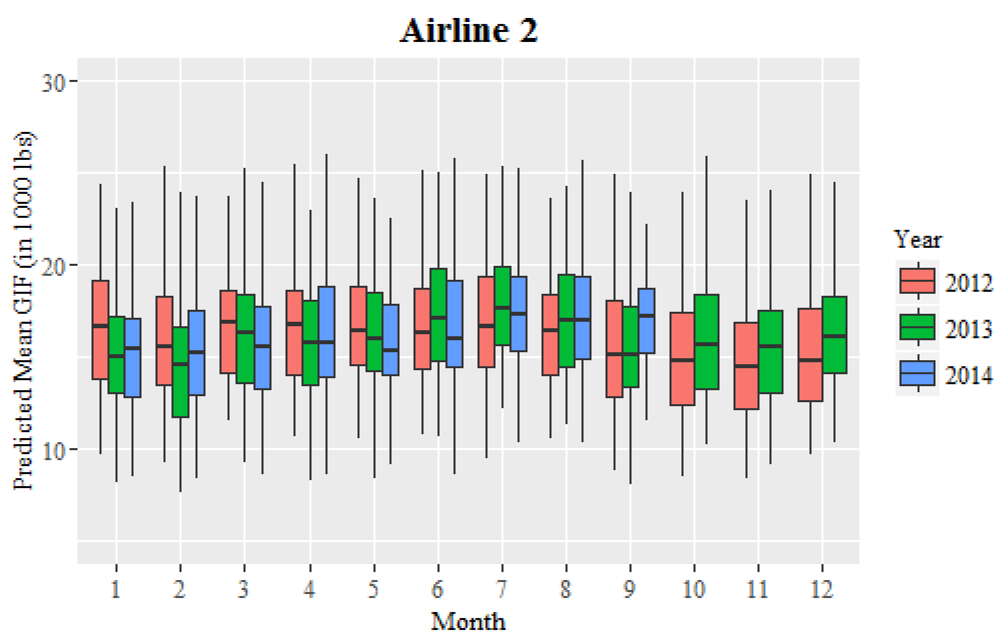
3.3 Gate-in Fuel Prediction and Metrics Trends

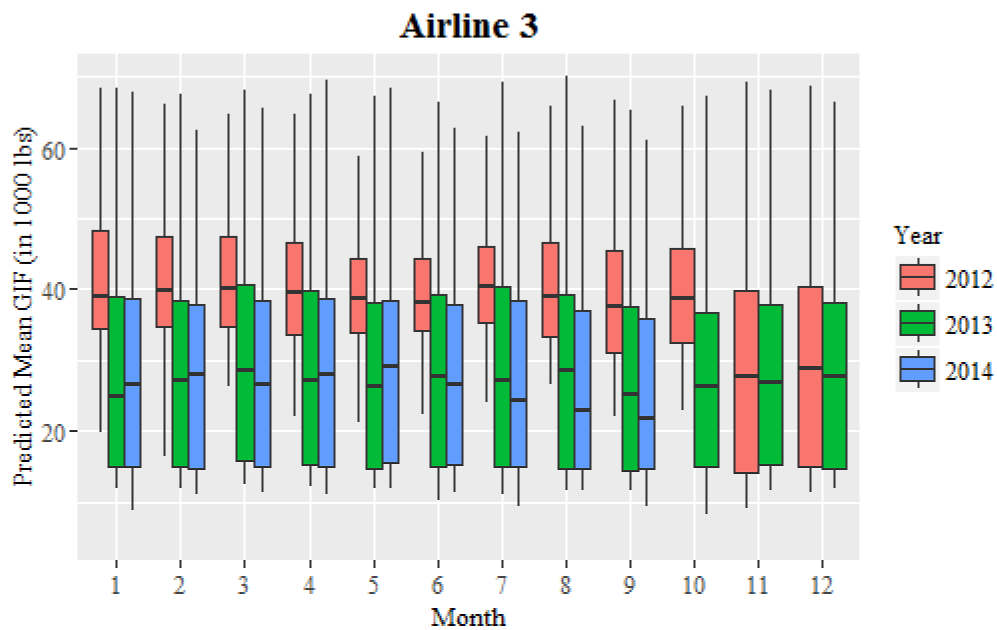
In order to predict gate-in fuel (GiF) for the other airlines in the FAA data using the aircraft weight physical model and BTS monthly average payload information, we needed to aggregate the *FAA data* based on the same grouping criterion discussed in Section 3.1 which is {OD, airline, month, year, and aircraft type}. Then we apply aggregate level prediction on the *FAA data* using the physical model specification and passenger weights estimated from the previous section. Prediction results for six airlines are presented in Figure 3.1. Note that these results assume the operating empty weights in Table 3.1, whereas the results for single airline presented above suggest that, at least for that airline, these weights are several thousand pounds greater. If these weight disparities carry over to the other airlines in the *FAA data*, the results presented here would be overestimates of GiF. On the other hand, if we adjust the results for the six airlines assuming the additional weights implied by the β_0 values in Table 3.3, we obtain negative GIF values for a number of observations. For this reason, we assume the extra weights do not apply to the airlines included in the *FAA data*.

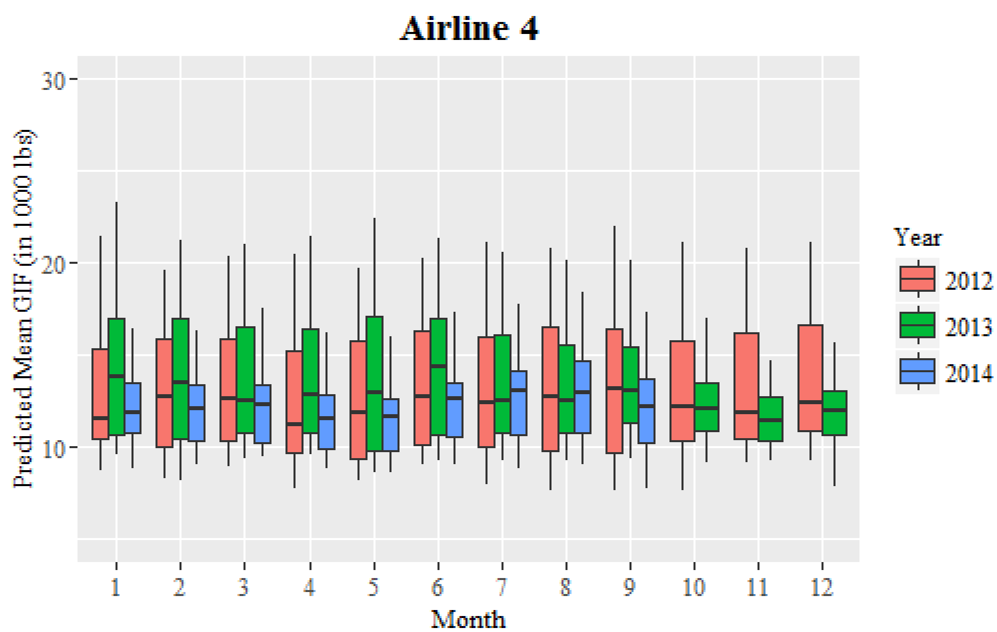
As shown in the boxplots across six airlines, the summer period, especially in July, is found to have higher GiF. This increase is due to the aviation system's usually severe congestion and delay during summer period, which lowers system predictability. Accordingly, as responses to low system predictability, on average, we would expect to observe higher GiFs. It is also noted that airlines have different fuel loading strategies regarding flight predictability. As shown in Figure 3.1, Airline 3 was found to have unusually high GiFs in 2012, suggesting that it was loading excessive fuel in this period. This changed in 2013 and 2014, when Airline 3 GiFs fell within a range comparable to other airlines, but still with high variance. Airline 5 was found to have the lowest GiF level across the six airlines. This result suggests that Airline 5 has less conservative fuel loading strategies than others. It is also possible that Airline 5 operates in a region with higher predictability.




Airline 1
Predicted Mean GIF (in 1000 lbs)
30
20
10
1
2
3
4
5
6
7
8
9
10
11
12
Month
Year
2012
2013
2014


Airline 2
Predicted Mean GIF (in 1000 lbs)
30
20
10
1
2
3
4
5
6
7
8
9
10
11
12
Month
Year
2012
2013
2014


Airline 3
Predicted Mean GIF (in 1000 lbs)
60
40
20
1
2
3
4
5
6
7
8
9
10
11
12
Month
Year
2012
2013
2014


Airline 4
Predicted Mean GIF (in 1000 lbs)
30
20
10
1
2
3
4
5
6
7
8
9
10
11
12
Month
Year
2012
2013
2014

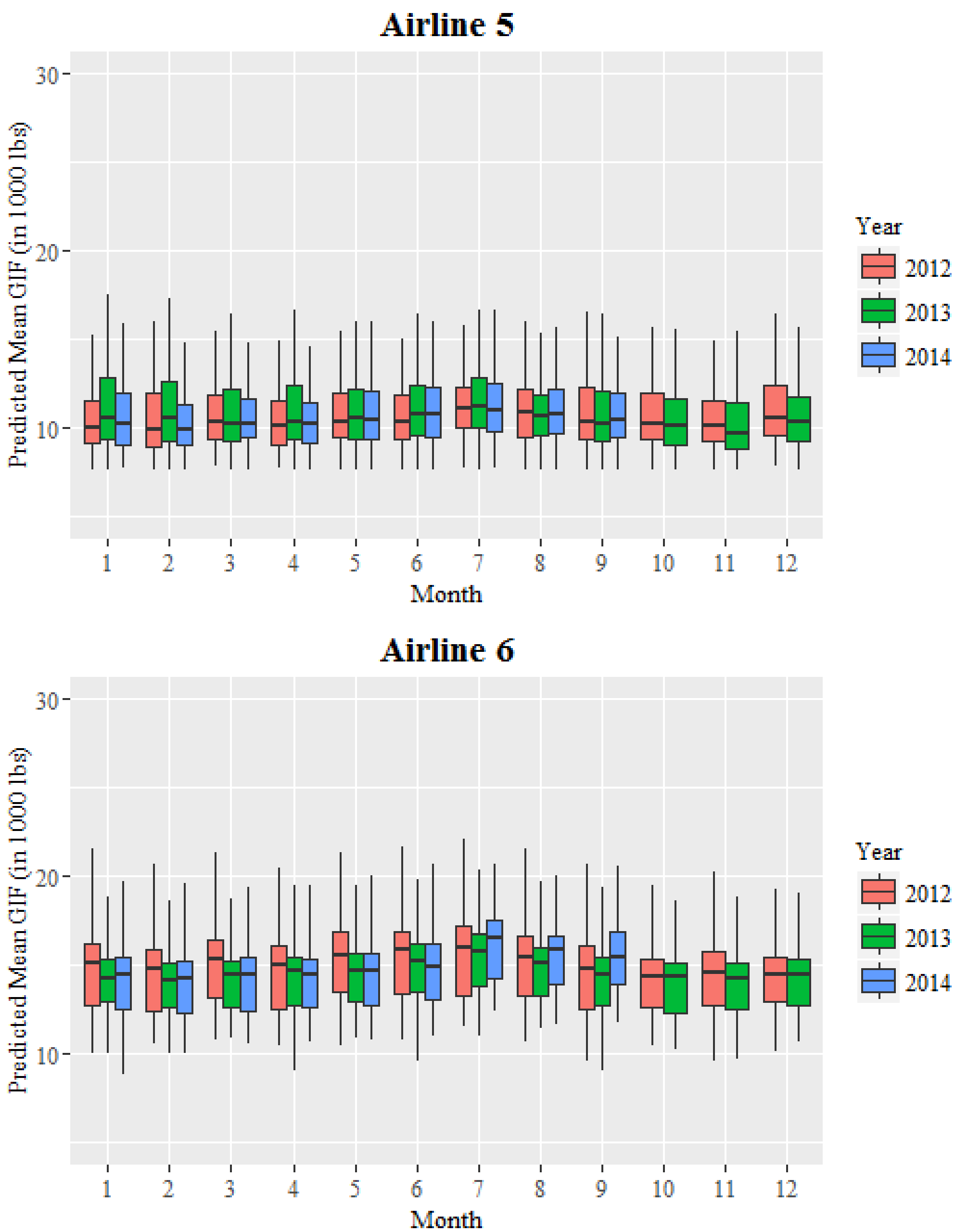


Figure 3.1 Predicted GiF for six airlines

3.4 Cost-to-Carry Analysis of Gate-in Fuel

Based on gate-in fuel (GiF) prediction, we can calculate the fuel burned to carry the GiF using CTC factors. "Cost-to-carry" is defined as the pounds of fuel consumed per pound of fuel carried per mile and it varies across aircraft types and flight distances. We borrowed the estimated CTC factors from Ryerson et al. (2015). Those factors were estimated using Piano-5, state-of-the-practice aircraft performance analysis software by Lissy (Pham et al., 2010). Piano is frequently used in both research and practice for aviation fuel modeling to predict fuel consumed as a function of aircraft dynamics and flight mission characteristics (Ryerson et al., 2015).

Table 3.4 Approximated cost-to-carry factors

Aircraft Types	Mass Parameter (γ_1)	Mass-Distance Parameter (γ_2)	Cost-to-carry Factor ($\frac{\gamma_1}{d}+\gamma_2$)
Airbus A300-600	0.024	3.591× E - 05	0.024/d+ 3.591× E - 05
Airbus A319-100	0.020	4.604× E - 05	0.020/d+ 4.604× E - 05
Airbus A321-200	0.019	4.644× E - 05	0.019/d+ 4.644× E - 05
Boeing 767-200	0.020	4.463× E - 05	0.020/d+ 4.463× E - 05
Boeing 777-300ER	0.028	3.030× E - 05	0.028/d+ 3.030× E - 05
Boeing 737-300	0.020	5.289× E - 05	0.020/d+ 5.289× E - 05
Boeing 737-400	0.020	5.289× E - 05	0.020/d+ 5.289× E - 05
Boeing 737-500	0.020	5.289× E - 05	0.020/d+ 5.289× E - 05
Boeing 737-700	0.021	5.476× E - 05	0.021/d+ 5.476× E - 05
Boeing 737-900	0.021	5.476× E - 05	0.021/d+ 5.476× E - 05
Embraer 190	0.020	5.289× E - 05	0.020/d+ 5.289× E - 05
McDonnell Douglas MD-11	0.024	3.591× E - 05	0.024/d+ 3.591× E - 05

For a given flight *i* with aircraft type *a*, flight distance *d* (in miles), weight *m* (in pounds), and gate-to-gate fuel consumed *b* (in pounds), its CTC factor (in pounds/pounds-mile) is estimated by Ryerson et al. (2015) in the below way. Gate-to-gate fuel consumption is assumed to be a function of distance and weight, as shown in Equation 3.4.

$$b_{i,a} = \gamma_{1,a} m_{i,a} + \gamma_{2,a} m_{i,a} d_{i,a} + \gamma_{3,a} d_{i,a} \tag{3.4}$$

Then, the CTC factor can be expressed as

$$\gamma_{i,a} = \frac{\gamma_{1,a}}{d_{i,a}} + \gamma_{2,a} \tag{3.5}$$

where γ_1 and γ_2 are parameters associated with mass and mass-distance. CTC factors for a majority of aircraft types have been estimated by Ryerson et al. (2015). However, several aircraft types in the *FAA data* that have still not been covered. Thus we approximated CTC factors for those uncovered aircraft types based on similar aircraft types reported in Ryerson et al. (2015) (see Table 3.4). More technical details about the CTC factor estimation can also be found in

Ryerson et al. (2015). The estimated GiF CTC's for six airlines are shown in Table 3.5. In this case, we will assume fuel price is $3/gallon.

Based on weighted average CTC estimates, we also seek the annual GiF CTC for six airlines. Total costs were extrapolated based on the number of performed operations obtained from the BTS data. However, such extrapolation might over-estimate the over-loading impact, because the average CTC estimates were based on the airline-reported OD pairs (in the *FAA data*) and these ODs are likely to have more predictability issues than those not reported. Since actual flight distance is not available in the *FAA data*, great circle distance between two OD pairs is used instead in the CTC calculation. The GiF CTC values for the different airlines range from $59 million to $667 million in 2012 and from $63 million to $428 million in 2013. The system level (total across six airlines) GiF CTC values for 2012 and 2013 are estimated to be $1.46 billion and $1.22 billion, respectively.

The GIF quantities estimated above include mandatory reserve fuel - the quantity of fuel required for an aircraft to continue to fly for 45 minutes at normal cruising speed. A more reasonable basis for estimating the costs of extra fuel loading is the difference between the GiF and the reserve fuel. Although detailed reserve fuel was not reported in the *FAA data*, we could use reserve fuel statistics obtained from the *airline data* as a benchmark. To improve estimation precision, average reserve fuel for an OD segment was used if the *FAA data* covered the same OD segment as the airline data. For those segments in the FAA data not being covered in the *airline data*, airline-level average reserve fuel (which is 4,225 lbs) was used.

After the removal of nominal reserve fuel from the predicted GiF, the estimated CTC values for the different airlines in 2012 ranged from $42 million to $605 million, with a total across all six airlines of $1.16 billion. Similarly, in 2013, the total GIF CTC estimate was $0.93 billion. Thus, the vast majority of the GiF and its CTC cannot be attributed the FAA-mandated reserve fuel.

Table 3.5 Cost-to-carry estimates for year 2012 and 2013

Year	Items	**Airline 1**	**Airline 2**	**Airline 3**
2012	Average of predicted GIF (lbs)	17445	16727	39364
	Total CTC extrapolated (lbs)	5.92E+08	2.75E+08	1.48E+09
	Total fuel cost at 3$/gallon	2.66E+08	1.24E+08	6.67E+08
	Second order fuel cost at 3$/gallon	2.02E+08	9.41E+07	6.05E+08
2013	Average of predicted GIF (lbs)	16667	17046	27861
	Total CTC extrapolated (lbs)	5.71E+08	2.87E+08	9.51E+08
	Total fuel cost at 3$/gallon	2.57E+08	1.29E+08	4.28E+08
	Second order fuel cost at 3$/gallon	1.91E+08	9.90E+07	3.72E+08
		Airline 4	**Airline 5**	**Airline 6**
2012	Average of predicted GIF (lbs)	13692	10400	15028
	Total CTC extrapolated (lbs)	1.33E+08	5.14E+08	2.39E+08
	Total fuel cost at 3$/gallon	5.98E+07	2.31E+08	1.08E+08
	Second order fuel cost at 3$/gallon	4.29E+07	1.42E+08	7.69E+07
2013	Average of predicted GIF (lbs)	13617	10283	14597
	Total CTC extrapolated (lbs)	1.40E+08	5.09E+08	2.46E+08
	Total fuel cost at 3$/gallon	6.32E+07	2.29E+08	1.11E+08
	Second order fuel cost at 3$/gallon	4.51E+07	1.40E+08	7.84E+07

3.5 Summary

Due to the unpredictable nature of the aviation system, the amount of fuel loaded per flight will be higher than the amount needed in a perfectly predictable world, where in order to minimize fuel cost, flights could land without any fuel left. Therefore, the GiF metric reflects flight predictability. Furthermore, this metric can also connect predictability with fuel consumption, which offers us a direct monetization scheme through CTC estimation. Therefore, the proposed GiF metric can also be integrated in the benefit assessments of various initiatives and programs, such as NextGen.

In this chapter, a comprehensive dataset for a specific airline was used as a training set to estimate passenger weight and test the performance of the physical model specification. Since we wanted to develop a model that could be transferred to other airlines, the physical model specification allowed us to avoid airline-specific fuel loading behavior. However, due to measurement error, presumably due to OEW, uncertainty remained concerning the prediction performance using the physical model. Therefore, one recommendation for policy makers is to add GiF as another required reporting item in the fuel burn data currently being provided to the FAA.

With the help of CTC factors, we were also able to calculate the extra fuel burn to carry GiF. The extra fuel burn could then be translated into monetary costs. We calculated the CTC values for six U.S. airlines for the year 2012 and 2013, based on the predicted GiF and CTC factors. Per flight GiF CTC and the annual number of performed operations were combined to extrapolate the CTC to an airline level. In 2012, the monetary CTC values were found to range from \$59 million to \$667 million, showing significant benefits toward improving flight predictability, with a total across all six airlines of \$1.46 billion. Even considering removing the impact of reserve fuel, the second order CTC values still ranged from \$42 million to \$605 million, with a total of \$1.16 billion. In 2013, the total benefit after removing the impact of reserve fuel was estimated to be \$0.93 billion.

This chapter suggests that there may be significant benefit from reducing unnecessary fuel loading in the U.S. airline industry. To capture that benefit, we must better understand the behavior of the agents – dispatchers –who make fuel loading decisions, as well as the ways in which historical fuel burn data can be leveraged to inform better fuel loading choices. These are the subjects of the following chapters, which focus on the detailed fuel loading data set available from a single airline.

4 Behavioral Modeling of Discretionary Fuel Loading Decisions

Safety is the number one priority of airline operations. To ensure safety, dispatchers may overload discretionary fuel to account for various uncertainties such as weather uncertainty, traffic congestion uncertainty, traffic control uncertainty, and so forth. As discussed in Section 1.2, discretionary fuel is defined as the summation of contingency fuel, non-weather required 1st alternate fuel, and all 2nd alternate fuel. On the other hand, a flight has to burn extra fuel to carry excess discretionary fuel which raising fuel costs for the airline. Therefore, dispatchers need to make a trade-off between safety and cost in discretionary fuel loading. As shown in Chapter 3, U.S. airlines incur a substantial cost from carrying unused fuel. Moreover, reducing unnecessary fuel loading is shown to be effective in reducing fuel consumption and CO_2 emissions (Ryerson et al., 2015; Kang et al., 2016). Previous studies have found that dispatchers' discretionary fuel decisions are influenced by system predictability and weather forecasts (Hao et al., 2016b). However, from a behavioral perspective, dispatcher fuel-loading decisions under risk[13] have not been fully investigated. In this study, we bridge this gap. Specifically, the objectives of this chapter are as follows:

- Understand how dispatchers trade off perceived safety cost of possible insufficient fuel and CTC excess fuel;
- Gain behavioral insights by relating dispatcher safety-cost trade-offs to dispatcher survey responses.

To meet the stated goals, we leverage the airline fuel data with detailed fuel-loading decisions and dispatcher survey information. To get a rough idea of discretionary fuel loading in our study airline, Figure 4.1 presents the relationship between discretionary fuel loading and over/under burn across all flights. Blue dots, accounting for 99.96% of total operations, indicate flights with discretionary fuel greater than over/under burn value. This is desirable because the goal of loading discretionary fuel is to cover the discrepancy between actual fuel burn and FPS planned fuel burn so that a flight can land without touching reserve fuel. On the other hand, the red dots are non-covered cases, which constitute about 0.04%. To be consistent with dispatcher fuel-loading practice, fuel burn is measured in minutes in this chapter. Using minutes instead of pounds to describe fuel burn helps to avoid the influence of different fuel-burn rates across aircraft types. Two observations can be drawn from Figure 4.1. Looking vertically, we can see that in most situations, discretionary fuel is greater than over burn, sometimes far greater, which suggests that most discretionary fuel loading is unnecessary. If we examine the figure horizontally, we find that there is a huge variability in over/under burn distribution. With such a wide spread of over/under burn distribution presumably due to inaccurate planned fuel burn predictions, it is unsurprising that, for the safety consideration, dispatchers tend to overload discretionary fuel to minimize the risk of encountering fuel exhaustion or flight diversions.

[13] An old idea in the economics literature, dating back at least to Frank Knight, is that a distinction should be drawn between risk (probability distribution of the potential outcomes is known) and uncertainty or ambiguity (i.e. probability distribution of possible outcomes is unknown) (Levin, 2006; de Palma et al., 2008). In the later analysis, I will assume probability distribution is known which fits into the regime of decision making under risk.

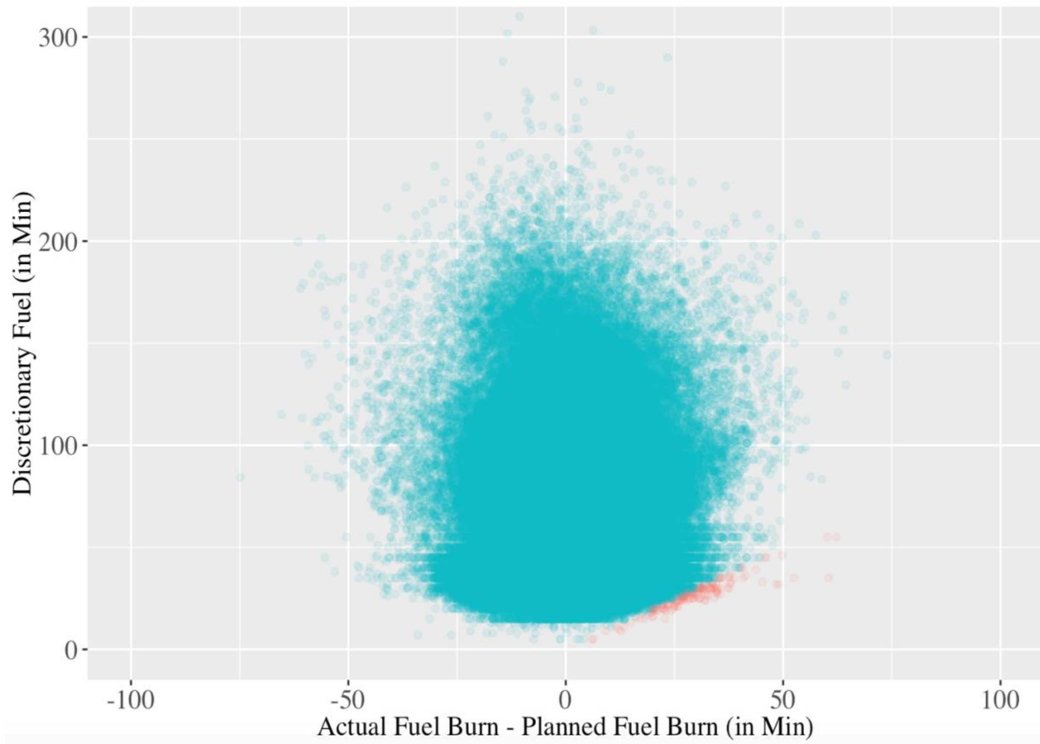


Figure 4.1 Discretionary fuel loading overview

The outline of this chapter is as follows. In Section 4.1, we introduce the basic newsvendor model and the risk-averse newsvendor model used to estimate dispatchers' trade-offs between the perceived safety costs of insufficient fuel and CTC excess fuel. Section 4.2 describes approaches in estimating key elements in the newsvendor model. Section 4.3 and Section 4.4 focus on the estimation results of safety-cost trade-offs and how such trade-offs can be related to dispatcher survey responses, respectively. The summary and conclusions are discussed in Section 4.5.

4.1 Modeling Framework

The newsvendor problem is one of the classical problems of inventory management. The original goal of newsvendor model is to determine optimal inventory levels where the demand is stochastic and the costs are deterministic. The costs arise either from not having enough inventories to meet demand or from holding excess inventory that must be discarded because of the perishable nature of the good. Under the basic newsvendor model set up, the optimal inventory level is jointly determined by stochastic demand distribution and cost ratio (which is also the trade-off between these two cost items). A comprehensive review of newsvendor models can be found in Qin et al. (2011). In the context of fuel loading, dispatchers face two costs when deciding on discretionary fuel (analogy to inventory level): perceived safety cost when discretionary fuel is less than over/under burn value (analogy to not having enough inventory to meet demand) and CTC excess fuel (cost from holding excess inventory). However, instead of determining optimal discretionary fuel, we are interested in inversely estimating the cost ratio, given that we have observed the optimal discretionary fuel loading decisions by dispatchers. By assuming dispatchers make rational fuel loading decisions, the newsvendor model provides us a framework of estimating the safety-cost trade-offs dispatchers make in fuel loading decisions.

4.1.1 Basic Newsvendor Model

We denote over/under burn (in minutes) as a random variable u following a distribution F: $u \sim F$ where F refers to a cumulative distribution function (CDF). Let a denote discretionary fuel loading (in minutes) and $L(a,u)$ denote the cost of a discretionary fuel loading decision. As shown in Figure 4.2, we also use q to represent fuel-burn CTC excess fuel and p to represent perceived safety cost of possibly insufficient fuel loading, where p and q are both positive.

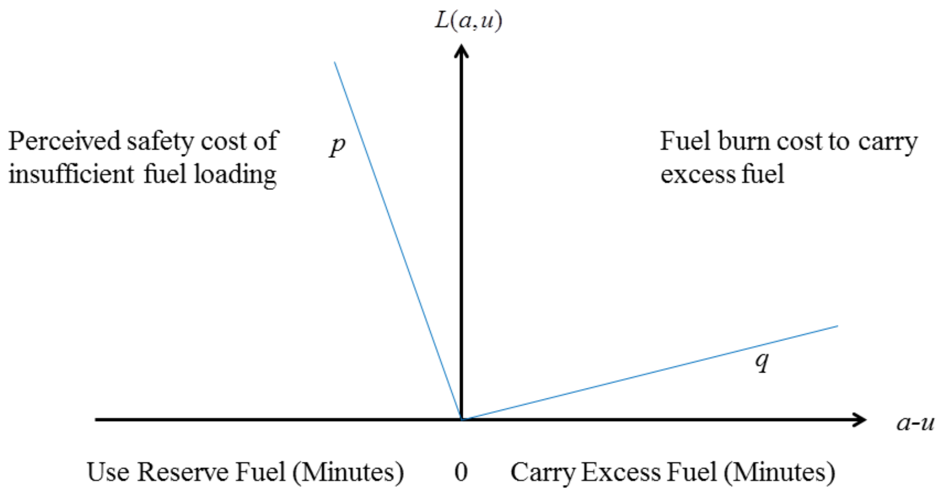


Figure 4.2 Newsvendor model setting

Then the cost can be written as

$$L(a,u) = p(u-a)^+ + q(a-u)^+ \tag{4.1}$$

where $(u-a)^+ = \begin{cases} u-a & (u>a) \\ 0 & o.w. \end{cases}$. Assuming a dispatcher makes discretionary fuel-loading decisions by minimizing the expected cost, the optimal discretionary fuel a can be determined by solving the following minimization problem.

$$\begin{aligned} \min_a E_u(L(a,u)) &= \min_a E_u(p(u-a)^+ + q(a-u)^+) \\ &= \min_a [\int_a^{+\infty} p(u-a)dF(u) + \int_{-\infty}^a q(a-u)dF(u)] \end{aligned} \tag{4.2}$$

where the expectation is taken over over/under burn u. If we take the derivative of the function we are minimizing and set this derivative equal to zero, then we can get a closed form solution of optimal a (Arrow et al., 1951)

$$F(a^*) = \frac{p}{p+q} \tag{4.3}$$

Thus, the optimal discretionary fuel is the $p/(p+q)$ -th quantile of the over/under burn distribution F. Moreover, the ratio between p and q gives us a measure of how a dispatcher trade-offs between cost of insufficient fuel and excess fuel. A higher p/q ratio implies a dispatcher is willing to load more fuel to reduce the risk of having to burn reserve fuel.

4.1.2 Risk-Averse Newsvendor Model

The basic newsvendor model assumes a decision maker is risk-neutral. In the context of flight fuel loading, it is also reasonable to assume dispatchers are risk-averse, given that safety is the top priority in the airline industry. In this section, we consider how to account for risk-aversion explicitly in the newsvendor model. In the setting of newsvendor problems, risk-averse behavior is generally modelled using three approaches (Qin et al., 2011): expected utility maximization with concave utility functions (EU), mean-variance analysis (MV), and conditional value-at-risk (CVaR) minimization.

Within the EU framework, risk-aversion is captured by concave and non-decreasing utility functions. Decision makers place orders by maximizing utility. For example, exponential utility function has been used by Choi and Ruszczynski (2011), Katariya et al. (2014), Bouakiz and Sobel (1992). However, due to the complexity of the first order condition, this approach is not applicable in our context. To be more specific, the utility function can be written in the following way, where λ is a general risk-aware parameter.

$$\min_{a} E_u(L(a,u)) = \min_{a}[\int_a^{+\infty} p(u-a)^{\lambda_1}\, dF(u) + \int_{-\infty}^{a} q(a-u)^{\lambda_2}\, dF(u)] \tag{4.4}$$

However, if we take the derivative of Equation 4.4 with respect to *a*, unlike risk neutral newsvendor model described in Section 4.1.1, here no closed form expression exists for the parameters we are interested in, making behavioral modeling very difficult.

Under MV and CVaR analysis, risk-averse behavior is characterized by downside risk measures. For MV (Lau,1980; Anvari, 1987), the variance or standard deviation of profit is used as a risk measure. Then, decision makers try to maximize the difference between the expected profit and the variance of the profit multiplied by a positive constant *b*, which is used to capture the degree of risk-aversion. When *b*=0, the decision maker is risk-neutral. However, there are criticisms of the MV approach. First, as pointed out by Choi (2009), in theory, the MV criterion is inconsistent with stochastic dominance[14] because the MV criterion is symmetric and treats over-performance and under-performance equally. For instance, an efficient option (in the sense of the MV criterion) may be stochastically dominated by another option. Second, variance as a risk measure is not coherent[15]. Moreover, the expected profit formulation is not applicable in our setting. In our case, only cost is incurred, meaning that an MV formulation only contains a

[14] Some notes on stochastic dominance: The distribution of a random outcome X is preferred to random outcome Y in terms of a stochastic dominance relation if and only if expected utility of X is preferred to expected utility of Y for all utility functions in a certain class (Choi, 2009). For example, a random outcome X first-order stochastically dominates a random outcome Y if and only if every expected utility maximizer with an increasing utility function prefers X over Y.

[15] A coherent measure of risk is a functional that satisfies the following four axioms: convexity, monotonicity, translation equivariance, and positive homogeneity (Artzner et al., 1999). MV risk function does not satisfy convexity and positive homogeneity (Choi, 2009).

variance term for cost. This is equivalent to the concave utility function by setting $\lambda_1 = \lambda_2 = 2$ and we are back to the previous problem.

To overcome the limitations of the MV and concave utility settings in our application, we applied the CVaR formulation to capture the risk-aversion of dispatchers in discretionary fuel loading. The idea of CVaR is that instead of focusing on the entire distribution of cost, a decision maker is more sensitive to the upper tail of the distribution (Rockafellar and Uryasev, 2002). Concretely, let $D(\eta \mid a) := P(L(a,u) \le \eta)$ which is the CDF of cost L. Define $\beta - VaR$ as the β-th quantile of $D(\eta \mid a)$. A β-tail distribution focusing on the upper tail part of the cost distribution can then be written as

$$D^{\beta}(\eta \mid a) := \begin{cases} 0 & \eta < \beta - VaR \\ \dfrac{D(\eta \mid a) - \beta}{1 - \beta} & \eta \ge \beta - VaR \end{cases} \tag{4.5}$$

CVaR model assumes decision makers care about the upper tail part of the loss distribution that they are trying to minimize:

$$L^{\beta}(a) := E_{\beta}(L(a,u)) \tag{4.6}$$

where the expectation is taken under the β-tail distribution $D^{\beta}(\eta \mid a)$.

It has been shown that minimizing $L^{\beta}(a)$ is equivalent to minimizing an auxiliary function (Rockafellar and Uryasev, 2002):

$$\kappa_{\beta}(a, \lambda) := \lambda + \frac{1}{1-\beta} E((L(a,u) - \lambda)^{+}) \tag{4.7}$$

and

$$\min_{a \in A} L^{\beta}(a) \Leftrightarrow \min_{(a,\,\lambda) \in A \times \Re} \kappa_{\beta}(a, \lambda) \tag{4.8}$$

By solving the equivalent minimization problem $\min_{(a,\,\lambda) \in A \times \Re} \kappa_{\beta}(a, \lambda)$, Gotoh and Takano (2007) finds an analytical solution for the above minimization problem under the total cost formulation. In the total cost formulation, a decision maker wants only to minimize the total cost, and no profit is involved. This formulation suits our application perfectly. The optimal discretionary fuel under over/under burn distribution F can be computed as follows:

$$F(a^{*}) = F\left(\frac{q}{p+q} F^{-1}\left(\frac{p(1-\beta)}{p+q}\right) + \frac{p}{p+q} F^{-1}\left(\frac{q\beta + p}{p+q}\right)\right) \tag{4.9}$$

where $F^{-1}(\cdot)$ is the inverse of F and where p and q are defined as before. β is a risk-averse parameter ranging from 0 to 1. If β = 0, then Equation 4.9 reduces to Equation 4.3, the risk-neutral case.

CVaR also has several good theoretical properties. It is coherent (Artzner et al., 1999) and consistent with first (or higher) order stochastic dominance (Shapiro et al., 2009). In particular, the consistency with the stochastic dominance implies that minimizing the CVaR never conflicts with maximizing the expectation of any risk-averse utility function (Ogryczak and Ruszczynski, 2002).

4.2 Estimation Approach

In order to estimate dispatchers' safety-cost trade-offs using the newsvendor modeling framework, we will discuss how to compute different elements in the newsvendor model, namely *F* (over/under burn distribution), *q* (CTC excess fuel), and *p* (perceived safety cost of insufficient fuel).

4.2.1 Estimation of q

The fuel-burn CTC one minute of excess fuel is a function of trip distance and aircraft type (Ryerson et al., 2015). Hence, *q* can be computed for each individual flight using the CTC factor provided by Ryerson et al. (2015). As shown in Section 3.4, for a flight *i* with aircraft type *ac* and flight distance d_i (in statute mile), the additional burn q_i to carry one minute of excess fuel could be expressed in the following form:

$$q_i = \gamma_{1,ac} + \gamma_{2,ac} d_{i,ac} \tag{4.10}$$

where $\gamma_{1,ac}$ and $\gamma_{2,ac}$ are the mass parameter and mass-distance parameter for aircraft type ac respectively (see Section 3.4 for details).

4.2.2 Estimation of F

In order to estimate an over/under burn distribution *F* perceived by a dispatcher during flight planning stage, the following assumptions are made:

- A1: Dispatchers are fully aware of *F* and make discretionary fuel decisions based on *F*.
- A2: The over/under burn distribution *F* varies according to flight distance, terminal weather forecasts, airports, and month.

According to FAA regulation and airline interviews, reserve fuel should be completely unused when the flight lands, except in extraordinary circumstances. Thus, it is reasonable to assume the goal of discretionary fuel loading is to cover possible fuel insufficiency conditions such that dispatchers' discretionary fuel decisions are based on *F* (A1). Note that the *F* refers to perceived over/under burn distribution by dispatchers, so *F* must be conditional on information available to dispatchers. However, since we do not have access to the airline traffic and weather forecast system, the best available information was used: terminal weather forecasts (from NOAA), flight distance, OD airports and month (from airline data).

- A3: For simplicity, we also assume *F* follows a normal distribution.

As shown in Figure 4.1 and Section 5.4 (to be discussed in details later), over/under burn distributions generally exhibit bell shapes, thus it is reasonable to make the normality assumption.

In light of the above assumptions, the original estimation task is transformed into a mean-SD estimation problem where the mean and SD of *F* are functions of the observables described in A2. We can formulate this problem in the following way. The over/under burn u_i of flight *i* can be expressed as

$$u_i = \mu(X_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma(X_i)) \tag{4.11}$$

where $\mu(X_i) = X_i\beta$ and $\log(\sigma(X_i)) = X_i\theta$. β and θ are parameters to be estimated. The SD function is modeled through a log-linear model to avoid any possible negative SD predictions. The conditional probability density of u_i is given by

$$f(u_i \mid X_i) = \frac{1}{\sqrt{2\pi}\sigma(X_i)} \exp\left(-\frac{(\mu(X_i) - u_i)^2}{2\sigma^2(X_i)}\right) \tag{4.12}$$

The negative log-likelihood of *N* observations in the dataset can be written as

$$-LL = \sum_{i=1}^{N}\left[\log\sigma(X_i) + \frac{(\mu(X_i) - u_i)^2}{2\sigma^2(X_i)}\right] \tag{4.13}$$

The estimation goal is to find $\hat{\beta}$ and $\hat{\theta}$ that minimize $-LL$ (or maximize the likelihood). However, $-LL$ is non-convex (Cawley et al., 2004), making reaching the global minimum not guaranteed. But if we decompose $-LL$ into two separate functions, one for the mean function the other for the SD function, then each sub-problem is convex and can be solved using gradient descent or a stochastic gradient descent approach. Although the original minimization problem is non-convex and its solution converges to a local minimum, by experimenting with different random starting values, we can alleviate the influence of non-convexity and obtain a good solution.

This line of reasoning leads to the next alternating estimation algorithm for estimating $\hat{\beta}$ and $\hat{\theta}$.

Step 0: Initialization

Initialize $\sigma(X_i)=1$ for all observations.

Step 1: μ-step

Fix $\sigma(X_i)$, solve for $\hat{\beta} = \arg\min \sum_{i=1}^{N}(w_i(\mu(X_i \mid \beta) - u_i)^2)$ with $w_i = [2\sigma^2(X_i)]^{-1}$ and $\mu(X_i \mid \beta) = X_i\beta$. This is equivalent to solving a weighted least squares problem.

Step 2: σ-step

Fix $\mu(X_i \mid \hat{\beta})$, solve for $\hat{\theta} = \arg\min \sum_{i=1}^{N} (Z_i + \delta_i \exp(-2Z_i))$ with $Z_i(X_i \mid \theta) = \log(\sigma(X_i)) = X_i\theta$ and $\delta_i = \frac{(\mu(X_i \mid \hat{\beta}) - u_i)^2}{2}$. This minimization problem is convex and can be solved using stochastic gradient descent.

Step 3: Compute $-LL$ based on the current estimates of $\hat{\beta}$ and $\hat{\theta}$.

Step 4: Repeat *Step 1*, *2*, and *3* until convergence is achieved.

The standard error estimates can be computed using observed information under the asymptotic normality theorem.

4.2.3 Estimation of p

After obtaining the estimated distribution *F* and CTC excess fuel *q* for flight *i*, we can inversely solve for *p*. In this case, for a dispatcher *j* who has handled N_j flights, under basic Newsvendor model (risk-neutral), *p* can be estimated for each dispatcher using non-linear least squares by minimizing the squared difference between observed quantile and theory-predicted quantile:

$$\hat{p}_j = \arg\min_{p_j} \sum_{i=1}^{N_j} \left(F_i(a_i^*) - \left(\frac{p_j}{p_j + q_i}\right)\right)^2 \tag{4.14}$$

where a_i^* is the discretionary fuel loaded by a dispatcher for flight *i*. Similarly, under a CVaR risk-averse setting, *p* and β can be estimated for each dispatcher using the same minimization criterion:

$$(\hat{p}_j, \hat{\beta}_j) = \arg\min_{(p_j, \beta_j)} \sum_{i=1}^{N_j} \left(F_i(a_i^*) - F_i\left(\frac{q_i}{p_j + q_i} F_i^{-1}\left(\frac{p_j(1-\beta_j)}{p_j + q_i}\right) + \frac{p_j}{p_j + q_i} F_i^{-1}\left(\frac{q_i\beta_j + p_j}{p_j + q_i}\right)\right)\right)^2 \tag{4.15}$$

Since we have only two unknowns in Equation 4.15, grid search can be used to find the best combination of $(\hat{p}_j, \hat{\beta}_j)$.

4.3 Estimation Results

In this section, we focus on the estimation results of *F* and *p* under different newsvendor model settings. A conditional mean-SD model is estimated using weather-matched airline fuel data. For *p* estimation, to ensure robustness, we consider only 96 dispatchers who have handled more than 100 flights in the dataset.

4.3.1 Estimation Results of F

The estimation results of mean-SD model are shown in Table 4.1. Based on the likelihood ratio test, the estimated model is found to provide a statistically superior fit to the constant model with a p-value of less than 0.0001. In this case, the constant model is equivalent to using sample mean and sample standard deviation to make predictions.

Regarding terminal area weather forecast, all adverse weather conditions are found to have a positive effect on both the conditional mean and conditional SD of over/under burn distributions. The construction of low ceiling and low visibility variables accords with the criteria used in FARs, which require a flight to carry enough fuel to travel to an alternate airport if the weather conditions are such that visibility is less than 3 miles and the ceiling at the destination airport is less than 2,000 feet at the flight's estimated time of arrival, plus or minus one hour.

Longer flights are also found to result in higher mean and SD of over-burn distributions, partly because longer flights are more likely to experience unpredicted unfavorable en route weather and traffic conditions. Turning to monthly effects, the winter season is found to have both higher mean and SD of over-burn. In other words, winter season is less predictable than other months. Major destination airports fixed effects are presented in Figures 4.3 and 4.4. Most hub airports are found to have higher actual fuel burns than planned fuel burns. This is because most hub airports of our study airline are busy airports with high delay and fuel burn. Moreover, LGA, JFK, LAX and BOS airports are found to have relative higher variability to over/under burn, which suggests those hub airports are relatively less predictable than others. After estimating the mean-SD model, we can apply them to predict perceived over/under burn distribution *F* for each flight.

Table 4.1 Mean-SD model results (dependent variable: over/under burn in minutes)

Variables		Parameter Estimates for conditional mean	Parameter Estimates for conditional standard deviation (log-linear model)
	Intercept	-4.866 ***	2.201 ***
TAF weather forecast for destination airports	Low visibility indicator (1-if lower than 3 miles, 0-otherwise)	0.557 ***	0.106 ***
	Low ceiling indicator (1-if lower than 2000 feet, 0-otherwise)	2.524 ***	0.099 ***
	Thunderstorm indicator (1-if thunderstorm presents, 0-otherwise)	2.165 ***	0.202 ***
	Snow indicator (1-if snow presents, 0-otherwise)	0.973 ***	0.108 ***
Distance	Flight distance (in nautical miles)	0.001 ***	0.0001 ***
Season (Baseline: January)	February	-0.093	-0.012 ***
	March	-0.803 ***	-0.035 ***
	April	-0.240 ***	-0.031 ***
	May	-0.170 **	-0.003 ***
	June	-0.422 ***	-0.044 ***
	July	-0.604 ***	-0.018 ***
	August	-0.359 ***	-0.035 ***
	September	-0.712 ***	-0.038 ***
	October	-0.686 ***	-0.062 ***
	November	-0.796 ***	-0.085 ***

	December	-0.121	-0.028 ***
Log-likelihood at constant		-1,062,083	
Log-likelihood at convergence		-1,026,468	
Number of observations		368,607	

Note: Hub destination airport fixed effects are presented in separate figures. To save space, hub origin airport fixed effects are not presented.

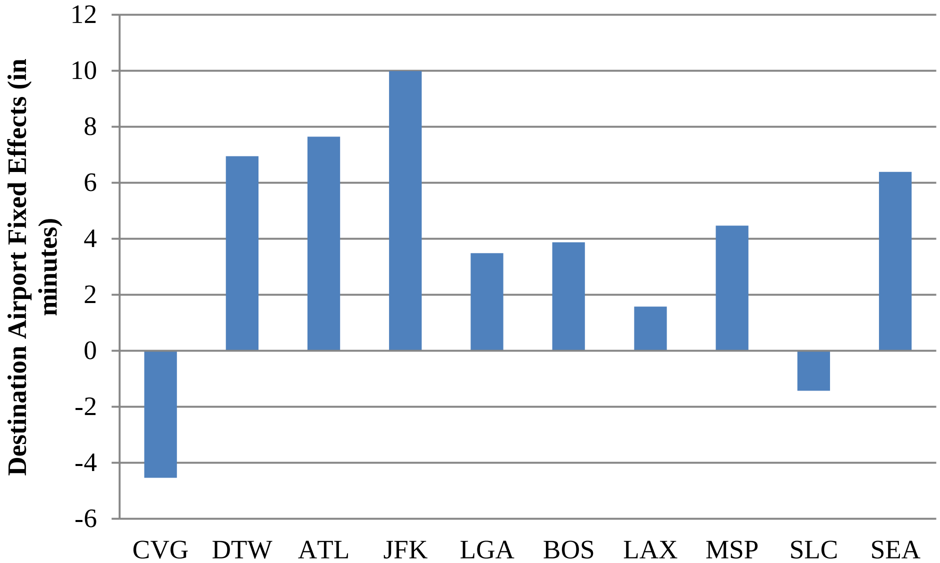


Figure 4.3 Hub destination airport fixed effects for conditional mean (baseline is non-hub airports)

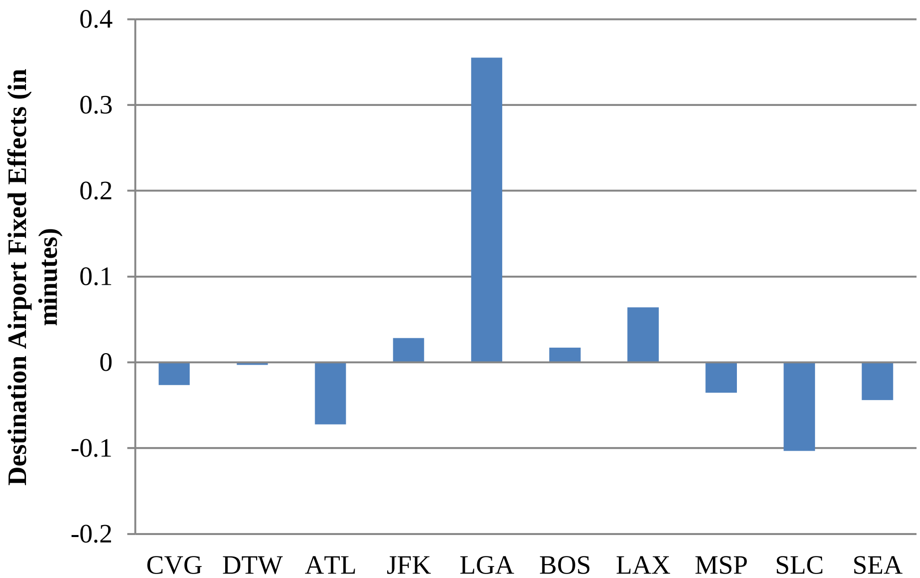


Figure 4.4 Hub destination airport fixed effects for conditional SD (baseline is non-hub airports)[16]

[16] The estimates presented in this figure has been taken exponential transformation and compared against 1 instead of 0.

4.3.2 Estimation Results of p

The estimated *p* for each dispatcher under risk-neutral newsvendor and risk-averse newsvendor settings are presented in Figure 4.5. It can be seen that the estimation results are consistent across the two models, although the estimated *p* values under the risk-averse model are systematically lower than for risk-neutral estimates. These lower values occur because in the risk-averse (CVaR) model, we also need to estimate a risk-averse parameter. Since risk-averse parameter β and safety-cost parameter *p* both capture a dispatcher's risk attitudes, when we estimate these two competing parameters simultaneously, they are expected to be negatively correlated with each other. This expectation is also confirmed in Figure 4.6. It is also noted in Figure 4.5 that several *p* estimates have extremely high values.

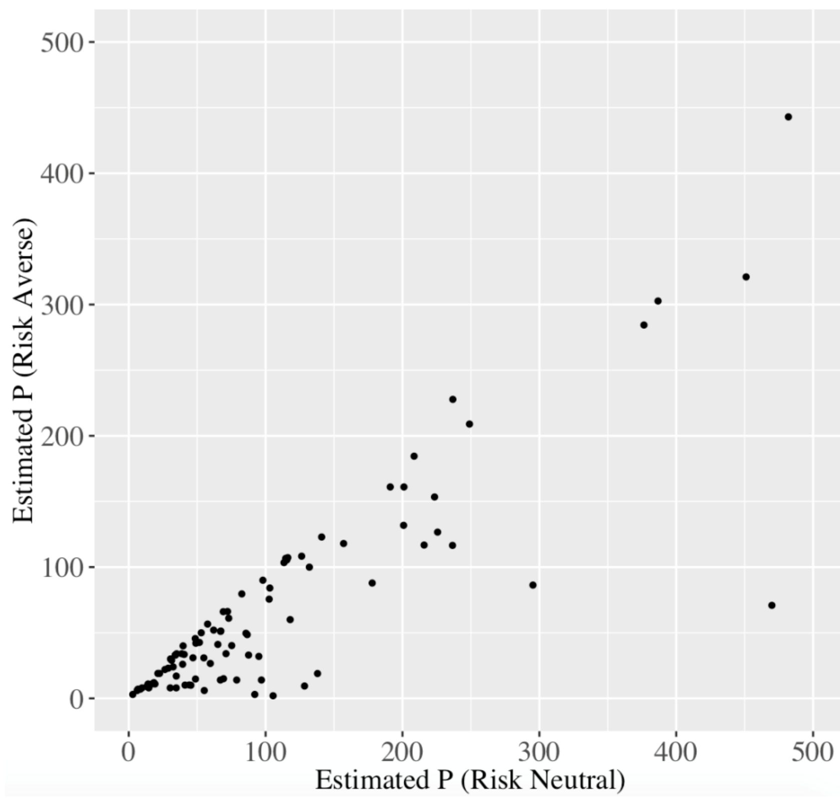


Figure 4.5 Safety cost parameter estimates (unit is in cost/minute reserve fuel)

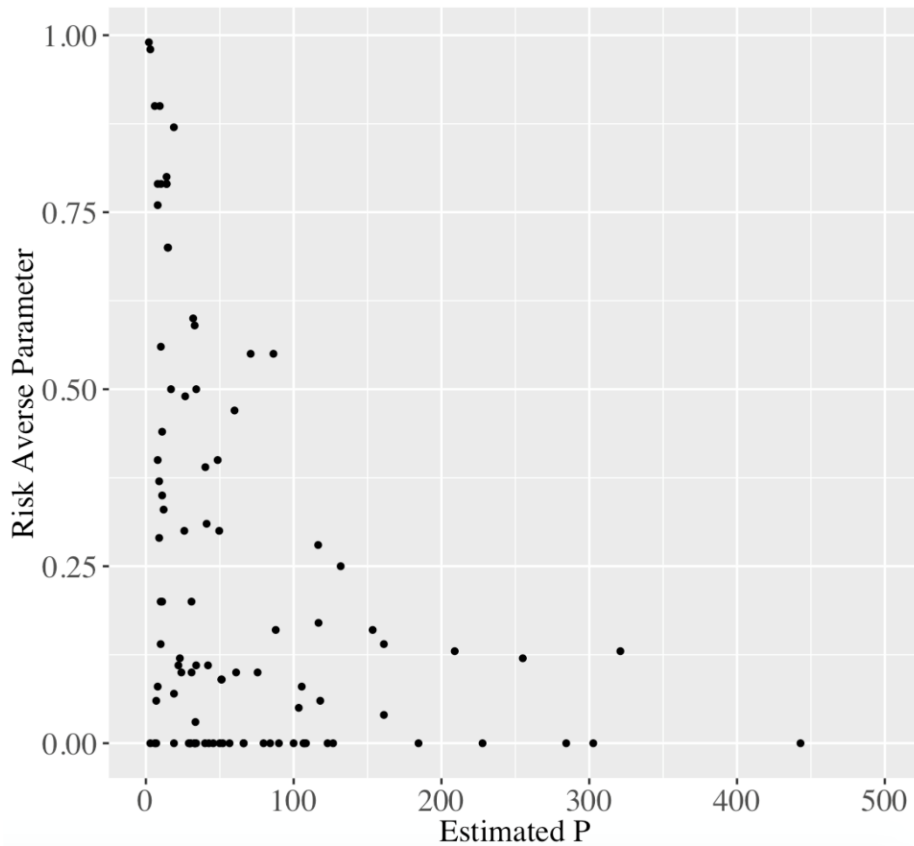


Figure 4.6 Relationship between risk-averse parameter and p parameter

After obtaining the *p* estimates for dispatchers, we can compute the ratio between *p* and *q* for each flight handled by a dispatcher. This safety-cost ratio gives us a measure of how a dispatcher trade-offs between cost of insufficient fuel and excess fuel. As shown in Figure 4.7, for each dispatcher, we compute the mean ratio (red dot), 2.5th percentile ratio and 97.5th percentile ratio (error bar). This 95% interval gives us a sense of the situations faced by dispatchers. We can observe that this ratio varies significantly even with the same dispatcher. Across the 96 dispatchers being considered, the mean ratio is about 1200, indicating that dispatchers assign a much higher cost to burning reserve fuel than to burning non-reserve fuel. In the next section, we demonstrate the use of survey instruments in predicting the mean trade-offs of dispatchers.

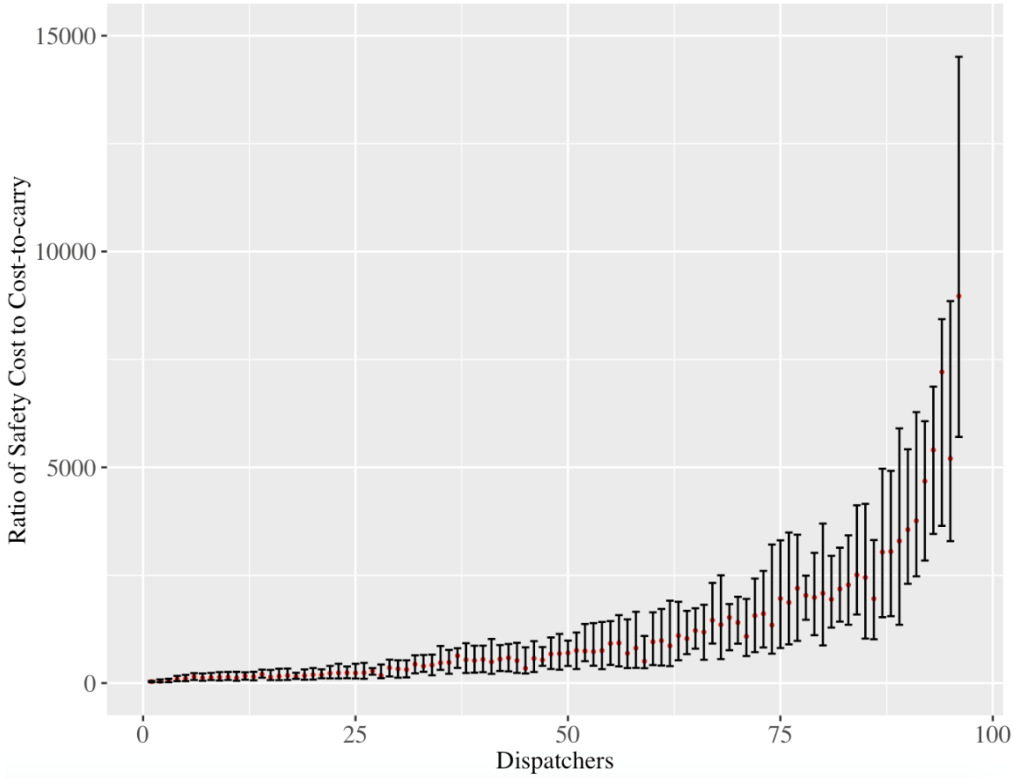


Figure 4.7 95% Safety-cost ratio for dispatchers

4.4 Latent Variable Modeling

In this section, we incorporate dispatcher survey responses (the *survey data* discussed in Section 2.4) to investigate links between dispatcher beliefs and attributes and fuel loading decisions. Specifically, structural equation models are applied to model the relationship between mean safety-cost ratio and various latent variables derived based on the dispatcher survey. The objectives of this section are two-fold. One is to identify relationships between latent variables derived from survey responses and a dispatcher's safety-cost ratio. Another is to propose initiatives based on these relationships that can help reduce discretionary fuel loading.

Based on exploratory factor analysis (EFA), among environmental attitudes statements, two latent variables have been revealed as shown in Table 4.2. The number of latent variables is determined according to eigenvalue decay. In this case, we retain only latent variables with eigenvalues greater than 1 (Kaiser, 1960). Since our goal is to find an optimal interpretable structure capturing dispatchers' attitudes, the oblique factor rotation method which allows for correlation among latent variables is used in the EFA instead of orthogonal rotation (Yong and Pearce, 2013). Based on factor loadings, the two revealed latent variables can be interpreted as *environmentalist* and *conservationist*. *The environmentalist* latent variable refers to the attitude that environmental concerns are important whereas *conservationist* captures one's attitudes in conserving resources. Similar EFA has also been performed on the risk attitudes statements and behavioral habits statements (See Table 4.3 and Table 4.4). Two risk-related latent variables: *prudent* and *risk-taking*, have been identified. In addition, two personal habit-related latent variables: *not detail-oriented* and *detail-oriented*, have also been revealed.

Table 4.2 EFA and summary statistics on environmental questions

Environmental attitudes statements	Summary Statistics		Factor Loadings	
	Mean	Standard Deviation	Environmentalist	Conservationist
Current levels of greenhouse gases do not pose a threat to the environment.	2.83	1.17	-0.712	--
I am willing to change my lifestyle to reduce my environmental impact.	3.43	0.97	0.468	0.463
I take an active role in recycling.	4.13	0.91	--	0.794
Individuals should not be responsible for the environmental impacts caused by others.	2.86	1.06	-0.793	--
I would contribute money to a campaign to strengthen environmental protection.	2.68	1.01	0.520	--
Consideration of environmental impacts rarely affects my lifestyle choices.	2.58	1.00	-0.664	--
We must conserve our resources for future generations.	3.86	0.98	--	0.601
The world's oil reserves are adequate for sustaining current levels of consumption.	2.70	1.22	-0.510	--
I would be prepared to pay higher taxes in order to protect the environment.	2.65	1.16	0.468	--

Note: Factoring method: principal axis factoring; Factor loadings less than 0.3 are not presented.

Table 4.3 EFA and summary statistics on risk questions

Risk attitudes statements	Summary statistics		Factor Loadings	
	Mean	Standard Deviation	Prudent	Risk-taking
People who know me would describe me as a cautious person.	3.41	0.78	0.385	--
I generally look for safer investments, even if that means lower returns.	3.01	0.84	0.709	--
I associate the word “risk” with the idea of “opportunity”.	3.01	0.78	--	0.518
I generally prefer bank deposits to riskier investments.	2.56	0.87	0.807	--
I’d rather take my chances with higher risk investments than increase the amount I’m saving.	2.63	0.64	--	0.569
I think it is more important to have safe investments and guaranteed returns than to take a risk to have a chance to get the	3.01	0.82	0.619	--

highest possible returns.				
I would never consider investments in shares because I find this too risky.	2.12	0.69	0.490	--
If I think an investment will be profitable, I am prepared to borrow money to make this investment.	2.31	0.96	--	0.531
I want to be certain that my investments are safe.	3.40	0.80	0.438	--
I get more and more convinced that I should take greater financial risks to improve my financial position.	2.53	0.75	--	0.728
I am prepared to take the risk to lose money when there is also a chance to gain money.	3.21	0.82	--	0.559

Note: Factoring method: principal axis factoring; Factor loadings less than 0.3 are not presented.

Table 4.4 EFA and summary statistics on habit questions

Behavioral habits statements	**Summary statistics**		**Factor Loadings**	
	Mean	Standard Deviation	Not detail-oriented	Detail-oriented
I do chores right away.	3.84	0.90	-0.557	--
I'll leave my things lying around.	2.10	0.99	0.668	--
I neglect my obligations.	1.31	0.67	0.495	--
I have an eye for details.	4.27	0.83	--	0.653
I am accurate in my work.	4.53	0.71	--	0.957
I forget to put things back where they belong.	1.81	0.94	0.638	--
I am always well prepared.	4.11	0.64	--	0.543
I often make a mess of things.	1.71	0.85	0.758	--
I like order.	4.24	0.69	-0.681	--

Note: Factoring method: principal axis factoring; Factor loadings less than 0.3 are not presented.

These latent variables reveal dispatcher personal attitudes and habits that may help predict the trade-offs between perceived safety cost and fuel cost. For each dispatcher, we are interested in the mean trade-offs measured by the mean ratio between p and q. To ensure the robustness of the following structural equation models, we consider dispatchers with p estimates less than 200. As we can observe in Figure 4.7, the mean safety-cost ratio is highly skewed. Therefore, a logarithm transformation is applied to the mean safety-cost ratio to create a more normal distribution. Model specifications for the risk-neutral model and risk-averse model are presented in Figure 4.8 and 4.9 respectively. To save space, for each category of attitude questions, we use a single item (e.g. habit statements) to represent all statement questions associated with that category. For the risk-averse model, we allow the error terms of log mean ratio and risk-averse parameter to be correlated.

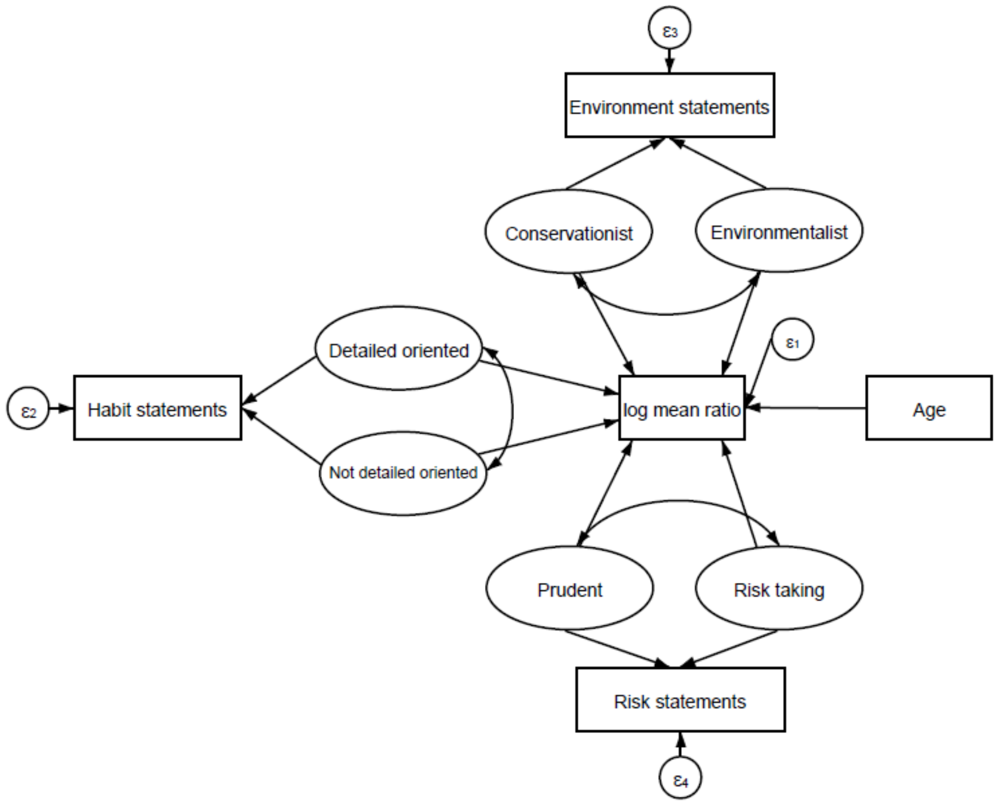


Figure 4.8 Risk-neutral SEM model specification

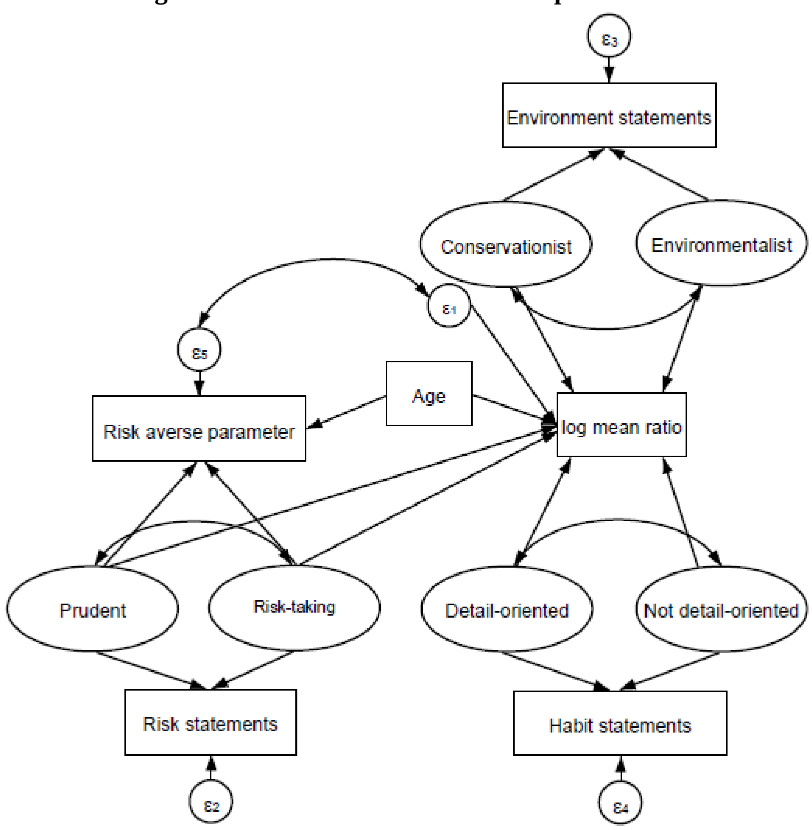


Figure 4.9 Joint modeling specification

The estimation results for the structural equation models are presented in Table 4.5. The second column shows the risk-neutral model results, where the dependent variable is the log mean safety-cost ratio estimated using the basic newsvendor model (risk-neutral). A variable with a positive estimated parameter means an increase in that variable, on average, predicts a higher mean safety-cost ratio, resulting in higher discretionary fuel loading. Detail-oriented dispatchers tend to load less discretionary fuel. This may be because dispatchers who care about details are more likely to make careful evaluation of a fuel plan, given various sources of information. Dispatchers with higher factor loadings on the conservationist latent variable are also found to load less discretionary fuel. Such dispatchers may be responding to the CTC unnecessary discretionary fuel and the resulting harm to the environment. Regarding personal characteristics, it is found that older dispatchers tend to load more discretionary fuel than young dispatchers.

Under CVaR newsvendor model, perceived safety cost p and risk-averse parameter β are estimated simultaneously and both parameters capture the risk-averse attitudes of a dispatcher to some degree. Therefore, we decided to perform a joint estimation of mean safety-cost ratio and risk averse parameter β. The estimation results are presented in columns 3 and 4 in Table 4.5. It is shown that dispatchers who are detail-oriented are more likely to load less discretionary fuel. This result is consistent with the risk-neutral model results. The conservationist latent variable is not significant in the CVaR model in contrast to the risk neutral one. The correlation between the error terms of the two models was found to be negative, which is also in line with Figure 4.6, showing that perceived safety cost p and risk averse β parameter are negatively correlated.

Table 4.5 Estimation results of structural equation models

		Risk-Neutral	**Risk-Averse (Joint Modeling)**	
Variables		Log mean ratio (p/q))	Log mean ratio (p/q) under CVaR	Risk-averse parameter
Latent variables		*Parameter estimates*		
Personal habit	Detailed oriented	-0.272 *	-0.252 *	--
	Not detailed oriented	-0.126	-0.169	--
Environment	Conservationist	-0.521 *	-0.279	--
	Environmentalist	0.367	0.082	--
Risk	Prudent	0.278	0.312	0.074
	Risk taking	0.134	0.014	-0.009
Personal characteristics		*Parameter estimates*		
Age (35 years old and younger is set as baseline)	Age (36-45 years old)	0.421	0.160	0.054
	Age (46-55 years old)	0.896 *	0.464	0.087
	Age (56 years old and above)	0.609	0.340	0.123
Intercept		6.084 *	5.795 *	0.211 *
Correlation estimates				
Correlation between detail oriented and not-well organized		-0.268	-0.269	
Correlation between environmentalist and non-environmentalist		0.772 *	0.774 *	
Correlation between prudent and risk taking		-0.598 *	-0.592 *	
Correlation between log mean ratio and risk-averse parameter		--	-0.154 *	

Note: 1. * denotes significant at 5% significance level.
2. To establish the scale of a latent variable, the variance for the latent variable is fixed to be 1 (Ullman, 2006).

The above estimation results suggest that dispatchers who are detail-orientated are more likely to load less discretionary fuel. There is also some evidence that conservationist dispatchers load less fuel. It is also shown that survey instruments designed to capture dispatchers' attitudes towards personal habits and environment are useful in predicting perceived safety-cost trade-offs. Based on the identified significant latent variables, there could be two small yet powerful strategies for airlines. One is about screening for dispatchers. When airlines interview dispatchers, in addition to skill- and behavior-based performance evaluations, it would be helpful to also test the detail-orientation of dispatchers, as well as their belief in conservation. This might help airlines to select dispatchers who are less likely to overload fuel resulting in potential fuel savings. Another implication is to target dispatchers for recurrent training. Most dispatcher recurrent training programs[17] focus on the knowledge and skills which would improve the operational safety and efficiency of flight dispatching. However, adding training topics on detail-orientation and conservation may also encourage dispatchers to load less unnecessary fuel.

4.5 Summary

In this chapter, we estimate dispatchers' trade-offs between safety and cost in discretionary fuel loading using newsvendor models. We assume dispatchers make discretionary fuel loading decisions by minimizing the expected cost of over-loading (CTC excess fuel) and under-loading (safety cost due to insufficient fuel loading). This problem can then be formulated as a newsvendor problem with closed-form solutions. Both risk-neutral and risk-averse newsvendor model formulations are explored. The ratio between perceived safety cost and CTC excess fuel provides us a measure of how dispatchers trade off safety and fuel cost.

In order to estimate perceived safety cost for each dispatcher, we first estimate the perceived over/under burn distribution for each flight, involving a joint estimation of mean and standard deviation models. An alternating estimation algorithm is used to estimate over/under burn distribution conditional on various covariates such as weather forests, flight distance, month, and airports. After obtaining the estimated over/under burn distributions, based on actual discretionary fuel loading, we can inversely solve for perceived safety cost for each dispatcher. On average, dispatchers are found to value reserve fuel 1,200 times more than other fuel. This underscores the focus on safety in dispatcher culture.

To better understand dispatchers' safety-cost trade-offs, results from a dispatcher survey is incorporated in the mean safety-cost ratio modeling. Six latent variables are identified with respect to attitudes towards personal habits, environment, and risk. It is found that dispatchers who are detail oriented and conservationists are likely to load less discretionary. Airlines may want to select for these characteristics and seek to cultivate them in recurrent training. However, one limitation of this analysis is that the perceived over/under burn distribution by dispatchers might be hard to estimate without imposing strong assumptions. Moreover, it is also difficult to evaluate the monetary savings of the two proposed strategies.

[17] https://atpflightschool.com/aircraft-dispatcher-training/dispatcher-recurrent-training.html

5 Fuel Burn Prediction Model Development

Accurate flight fuel burn predictions are crucial in the aviation industry. The required trip fuel is calculated by an airline's Flight Planning System (FPS) based on factors including the route selection, flight altitude, winds and other weather conditions, anticipated traffic delays, and aircraft performance. However, the FPS trip fuel predictions are not perfectly accurate, because of they are based on forecast conditions and imperfect aircraft performance models, and cannot account for variation pilot technique. If planned trip fuel is higher than actual trip fuel, then a flight will waste fuel by carrying excess fuel weight. On the other hand, if trip fuel is under-estimated, then a flight might be required to make an unscheduled fuel stop, which is operationally disruptive, or, worse, risk fuel exhaustion, as occurred very recently in the tragic crash of LaMia Flight 2933[18]. In practice, in response to prediction uncertainty, airlines carry discretionary fuel (e.g. contingency fuel) to avoid the latter (under-prediction) case. However, carrying unnecessary fuel can prove costly to airlines and harmful to the environment. Some of this wasted fuel burn can be avoided by more accurate fuel predictions and a better understanding of the prediction uncertainty.

Large airlines operation thousands of flights every day, and the records of these flights offer a rich base of empirical data for improving fuel burn predictions. This chapter investigates the potential improvement fuel burn prediction accuracy that can be attained by leveraging such data. In particular, ensemble learning is used to improve fuel burn prediction and construct prediction intervals (PI). Then the upper PI can be provided to dispatchers as recommended discretionary fuel.

This chapter makes the following contributions to the airline fuel loading field: (1) under the ensemble learning framework, we find that ensemble learning can significantly improve fuel burn prediction accuracy compared to airline FPS and offer some improvement compared to the best individual machine learning algorithms; (2) we propose a PI-based discretionary fuel estimation approach which can reduce unnecessary discretionary fuel loading while maintaining a safety performance; (3) we quantify the benefit of improved fuel prediction and the use of PI-based discretionary fuel estimation approach through a CTC analysis.

The remainder of this chapter is organized as follows. In Section 5.1, exploratory data analysis is performed on the airline fuel data and stability-based K-means clustering technique is used to spatially segment flights into three OD clusters. The ensemble learning framework is discussed in Section 5.2 and construction of PI is discussed in Section 5.3. Model performance is presented in Section 5.4. The benefit assessment of improved fuel burn prediction is described in Section 5.5. Section 5.6 provides a summary of this chapter.

5.1 Exploratory Data Analysis

Over/under burn value, defined as the difference between actual trip fuel burn and FPS planned trip fuel burn, is used to gauge the prediction performance of the airline FPS. In this chapter, we use pounds (lbs) to measure fuel burn and fuel loading since it is a more generic unit for describing fuel burn quantity. The FPS of our study airline tends to over-predict trip fuel burn at an average of 132 lbs per flight. Moreover, the SD of over/under burn is 940 lbs. Figure 5.1 presents the distributions of over/under burn for eight selected OD pairs. This figure reveals

[18] https://en.wikipedia.org/wiki/LaMia_Flight_2933

significant heterogeneity across different OD pairs in terms of the performance of FPS fuel burn prediction. For OD segment SFO-JFK, the median over/under burn value is about 1500 lbs, whereas for DEN-LGA the median value is -400 lbs. The large discrepancy across different OD pairs is partly due to route-specific directional effects (e.g. wind and regional congestion effects). This difference also suggests that a global fuel burn prediction model might not perform as well as one that includes some OD specificity. Therefore, we segment flights into different clusters based on OD airports and develop cluster-specific fuel burn prediction models[19]. Clustering based on airports' geographic location has several advantages. One advantage is that it allows us to build a cluster-specific model tuned to capture unique directional patterns for the flights within the same cluster. Another advantage is the clustering results are interpretable and can be applied in airline practice. In this study, we performed K-means clustering based on the coordinates (latitude, longitude) of the OD airports. We refer to these clusters as OD clusters (origin-destination clusters).

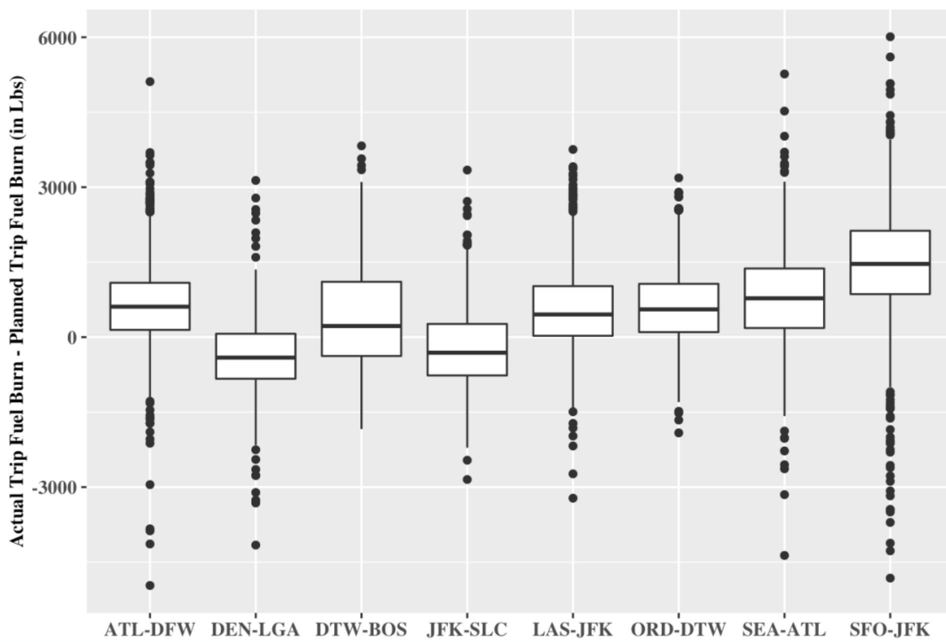


Figure 5.1 Over/under burn distributions across selected OD pairs

One challenge of using K-means clustering is deciding the optimal number of clusters. To tackle this issue, we implemented the stability-based K-means clustering algorithm proposed by Benhur et al. (2002). The basic idea of stability-based clustering is that true clusters should be insensitive to data perturbation. Thus, if we randomly select two subsamples of original data (for instance each accounts for 80% of original sample size) and perform K-means on each subsample, we will be able to assess how stable the clustering assignments are for the observations that are shared by both subsamples. Concretely, we can express the clustering labels L of data X using a matrix C with components:

---

[19] The performance of a global model is also tested compared to cluster-specific models. It is shown that cluster-specific models outperform the global model. The results are shown in Appendix.

$$C_{ij} = \begin{cases} 1, if \ x_i \ and \ x_j \ belong\ to\ the\ same\ cluster\ and\ i \neq j \\ 0, otherwise \end{cases} \quad (5.1)$$

Let L_1 corresponds to the labels from the first subsample of data (accounting for 80% of the original data) with matrix representation C^1, and let L_2 correspond to the second subsample with matrix representation C^2 (another 80% random sample of the original data). The dot product $< C^1, C^2 >= \sum_{i,j} C_{ij}^1 C_{ij}^2$ computes the number of pairs of points clustered together in subsample 1 and subsample 2. Then we can define two similarity measures. The first one is correlation similarity measure:

$$cor(L_1, L_2) = \frac{< C^1, C^2 >}{\sqrt{< C^1, C^1 >< C^2, C^2 >}} \quad (5.2)$$

The second one is called the Jaccard coefficient defined as:

$$jac(L_1, L_2) = \frac{< C^1, C^2 >}{< C^1, C^1 > + < C^2, C^2 > - < C^1, C^2 >} \quad (5.3)$$

Both metrics are based on how well entries of the two matrices agree and the values of metrics vary between 0 and 1. A higher metric value indicates more stable clustering assignments, thus helping us identify “true” clusters. We perform the above mentioned resample and K-means algorithm 100 times, with k, the number of clusters, ranging from 2 to 15. The distribution of corresponding similarity measures is shown in Figure 5.2. The most stable number of clusters is 3 since its stability measures are close to 1 according to both metrics. After determining the optimal number of clusters, we run K-means 100 times and select the cluster assignment that corresponds to the lowest total within cluster sum of squares to reduce the influence of random initialization of K-means algorithm. The resulting OD clusters are shown in Figures 5.3-5.5. For ease of demonstration, we present the origin airports (green dots) and destination airports (red dots) in two separate graphs. The three clusters generally correspond to east-bound, west-bound, and eastern short-haul traffic. This clustering allows us to separate flights into groups with similar directional patterns.

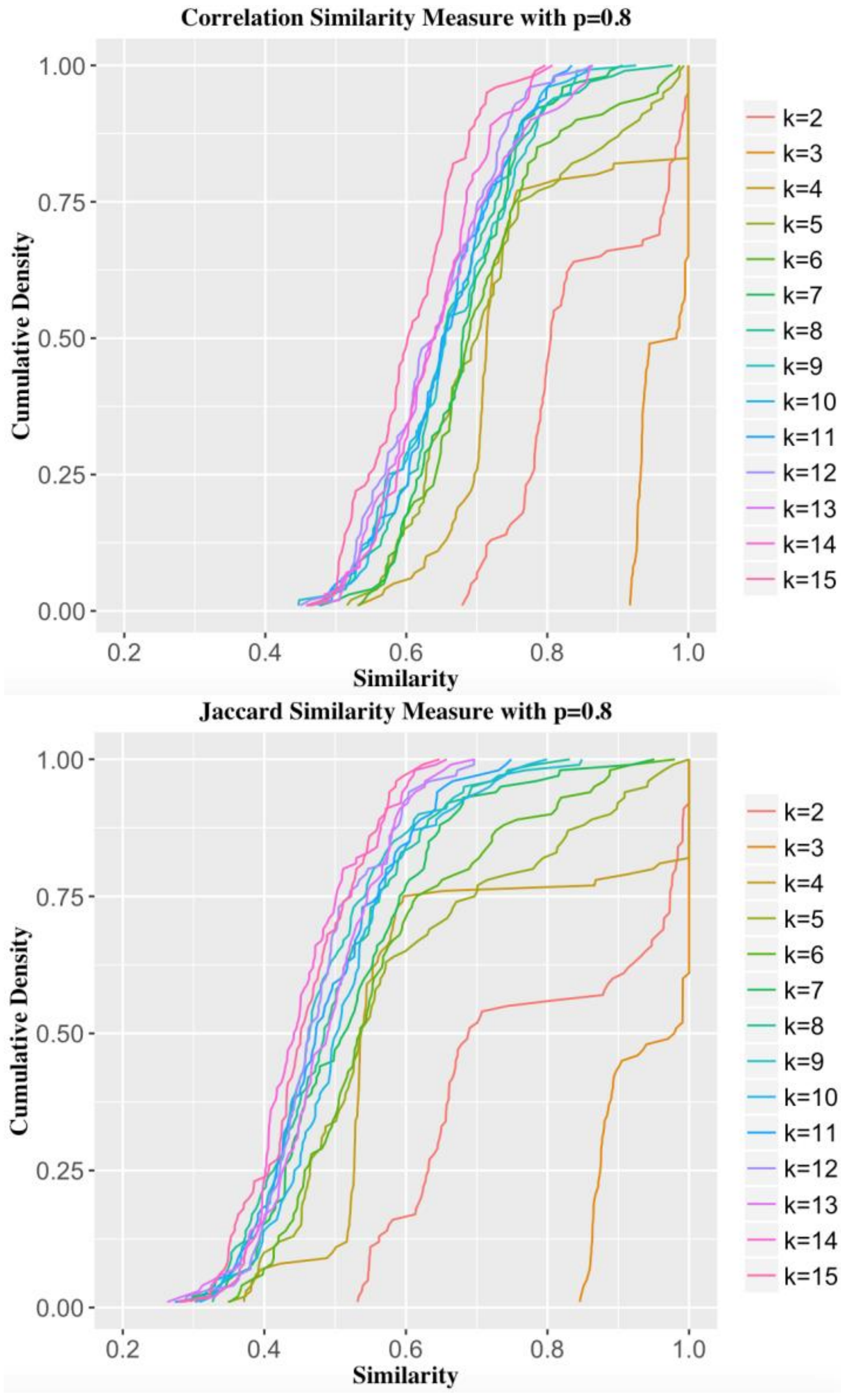


Figure 5.2 Stability-based clustering results

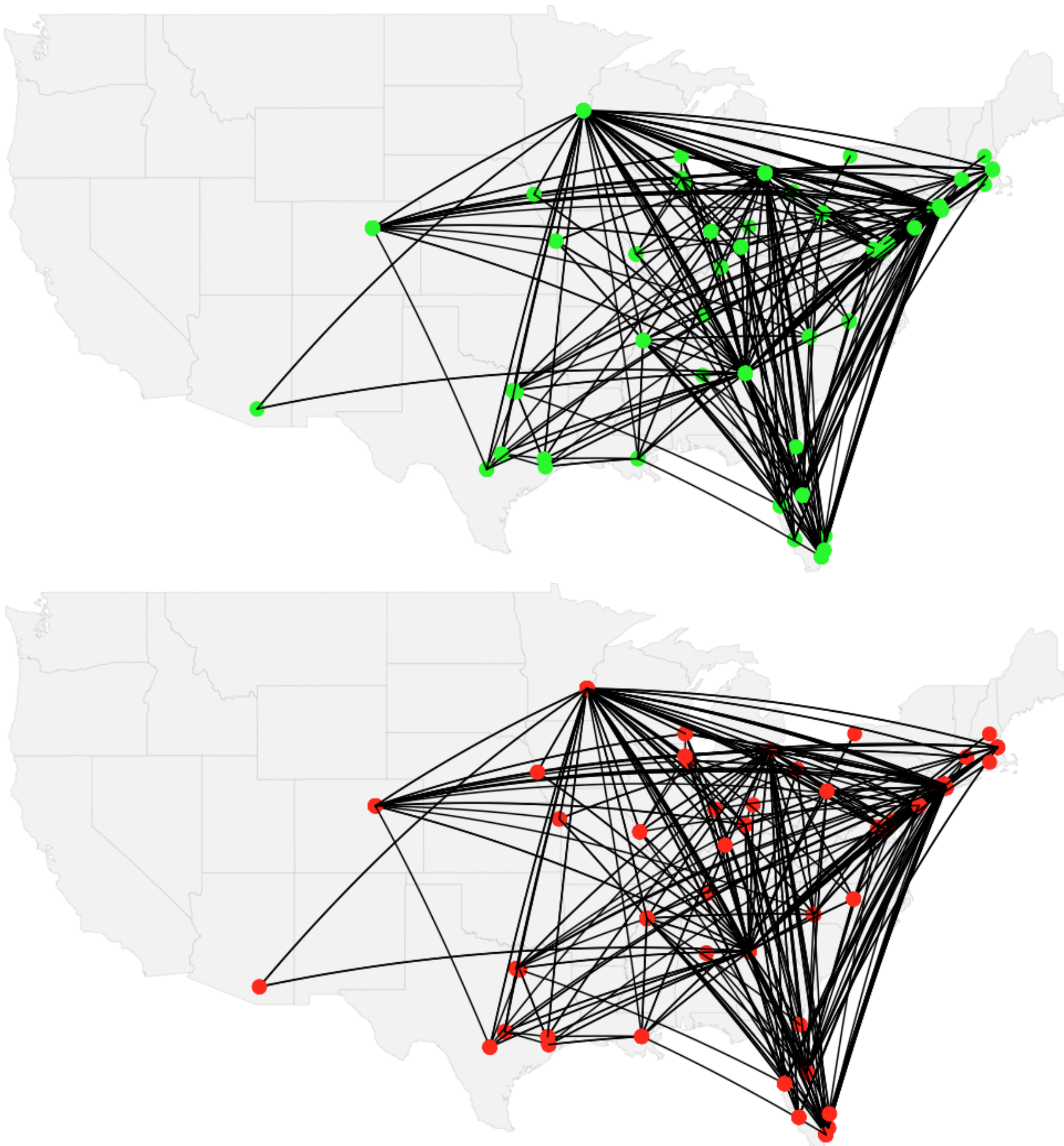


Figure 5.3 OD cluster 1 (Green dots denote origin airports, red dots denote destination airports)

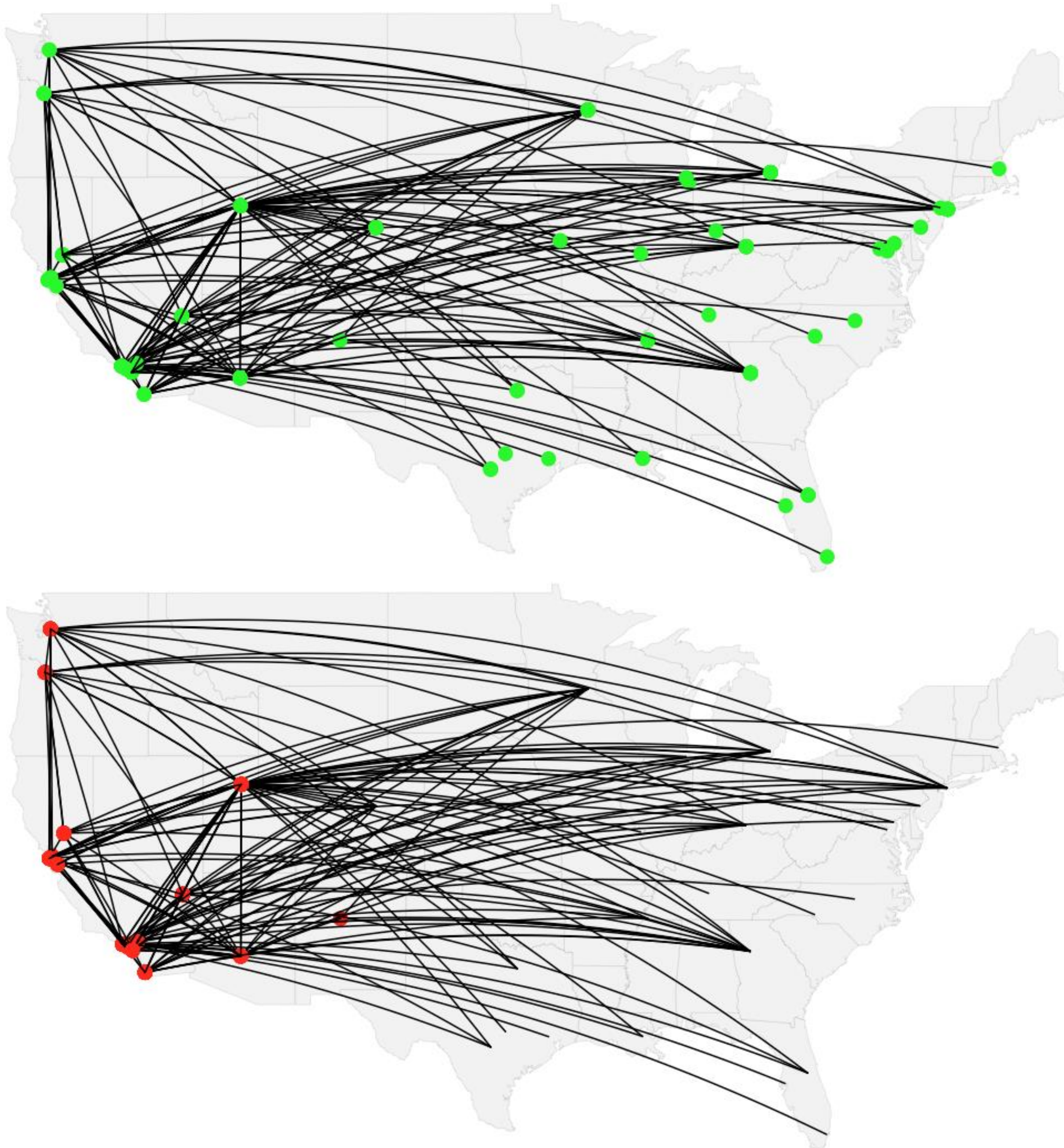


Figure 5.4 OD cluster 2 (Green dots denote origin airports, red dots denote destination airports)

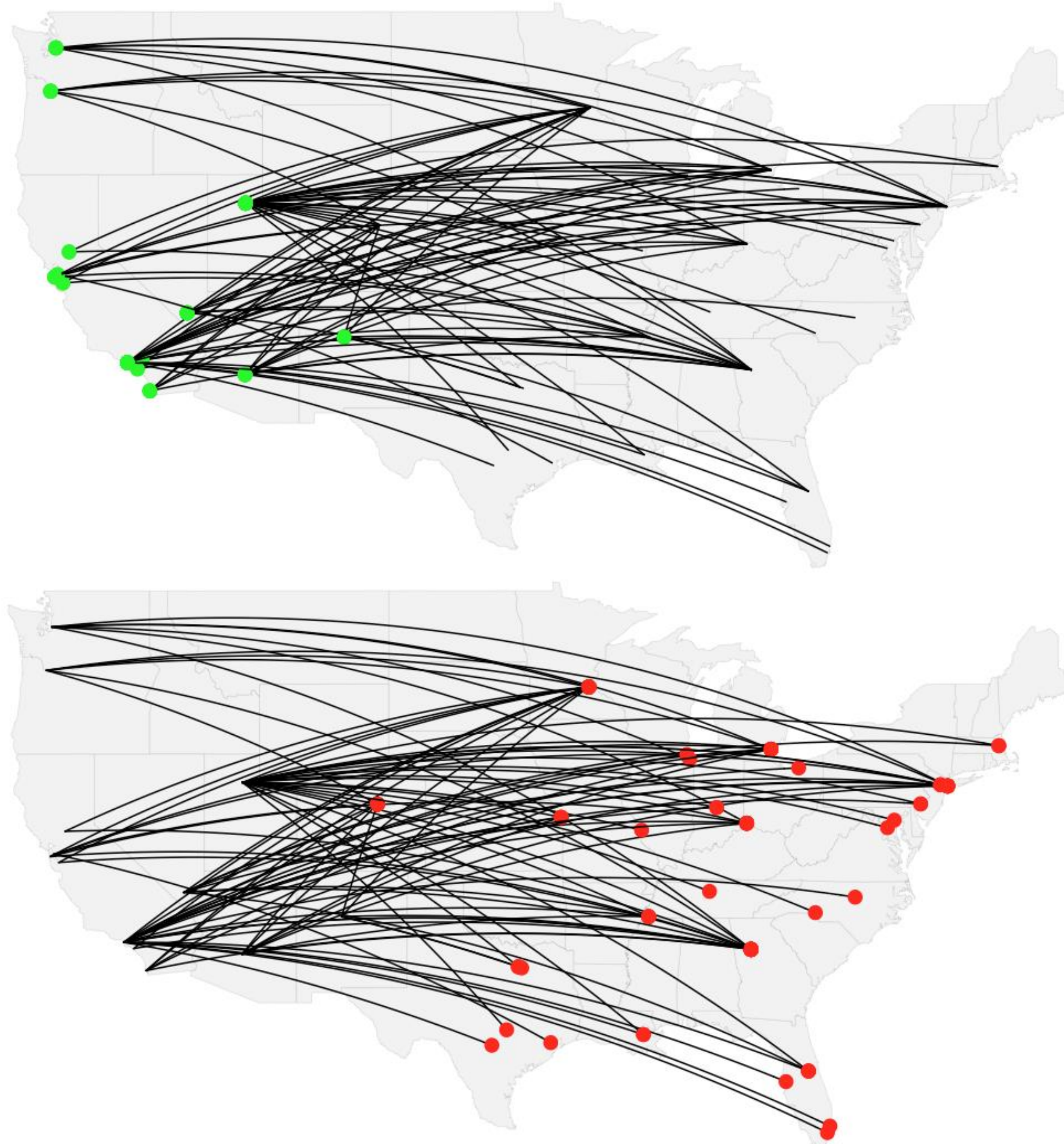


Figure 5.5 OD cluster 3 (Green dots denote origin airports, red dots denote destination airports)

Next, cluster-specific fuel burn prediction models are developed. In Section 5.2, we describe an ensemble learning framework that has been shown to significantly improve upon current popular machine learning algorithms on flight fuel burn prediction.

5.2 Ensemble Learning

Besides safety concerns, fuel burn prediction also effects the quality and validity of many other aircraft parameter predictions (such as gross weight, center of gravity position, maximal altitude, flight distance) as well as trajectory computations and optimizations (including altitudes, flight times, and costs) (Dancila et al., 2013). The equations describing the instantaneous fuel burn rate are complex and dependent on the particular aircraft-engine configuration. Therefore, traditionally due to limited computing power, many airlines' FPS (including that of our study airline) use a simplified fuel burn performance description model based on linear interpolation tables (Liden, 1992). With the enhanced computing power and emerging machine learning algorithms, many researchers have demonstrated promising results for applying machine learning algorithms for flight fuel burn prediction. For instance, popular algorithms applied in this field include decision tree (Chati and Balakrishnan, 2016), random forests, boosting tree, and neural network (Horiguchi et al., 2017).

A natural question to ask is: How much can we improve prediction accuracy by using an ensemble of machine learning methods compared to using a single method? In this section, we discuss an ensemble learning framework that can be used to combine the prediction results of different machine learning algorithms and yield more accurate prediction results.

Ensemble learning methods are often used when the true prediction function is not easily approximated by a single algorithm (Hastie et al., 2009; LeDell, 2015). The idea is to build a prediction model by combining the strengths of several base learning algorithms or base learners. Commonly used ensemble machine learning methods include Bagging (Breiman, 1996a) and Boosting (Freund and Schapire, 1997). Bagging (also known as Bootstrap aggregation) averages prediction results over a collection of bootstrap samples. One popular variant of Bagging is the Random Forest (Breiman, 2001). Boosting, on the other hand, starts with building weak base learners and iteratively adds them together.

Another powerful ensemble learning method proposed by Wolpert (1992) is called stacking. Stacking combines multiple base learners into a single prediction function through a second layer of training called meta-learning. It seeks an optimal linear combination of different base learners (not necessarily weak learners as required in Boosting) to give improved prediction accuracy (Wolpert 1992; Brieman 1996b; Leblanc and Tibshirani, 1996; van der Laan et al., 2007). Van der Laan et al. (2007) provide the theory for stacking and show that stacking (they call it "Super Learner", but it is essentially stacking) performs asymptotically as well as the best possible weighted combination of the base learners. In some sense, stacking provides a more general framework of ensemble learning in which it can also incorporate Boosting and Bagging as its base learners. From a Bayesian perspective, one can also average a large number of base learners with respect to the posterior probabilities of models. This algorithm is called Bayesian Model Averaging (Raftery, 1995). It has been shown that if the true prediction function is contained within the base learner library, BMA is never worse than stacking. However, if the true prediction function is not well approximated by the base learner library, then stacking will significantly outperform BMA (Clarke and Yu, 2003). In this chapter, we will focus on the stacking algorithm and designing meta-learning algorithms that work well for the airline fuel burn prediction.

5.2.1 Stacking

By leveraging the strengths of different base learners (such as linear regression, boosting, random forests, support vector machine, neural network), one can achieve prediction accuracy improvement. The idea is to perform K-fold cross-validation on the original dataset (also known as level zero data) using each base learner, then stack the cross-validated predictions from each learner to form a new set of features (also known as level-one data) and perform another layer of training based on level-one data.

Suppose we have selected L learners $f_1(X), \dots f_L(X)$ of a response variable y in terms of a vector of **X** on the same training set $S = \{(y_i, X_i), i = 1, \dots N\}$. Instead of following the winner-takes-all strategy by selecting a single learner based on lowest cross validation error (e.g. squared loss error), we can stack the prediction output from different learners to form a new design matrix and combine them in the following form (Wolpert 1992; Brieman 1996b):

$$ST(X) = \sum_{l=1}^{L} \alpha_l f_l(X) \tag{5.4}$$

where ST stands for stacking and α_l are the weights associated with each base learner. Given the training set $\{(y_i, X_i), i = 1, \dots N\}$, we can choose $\{\alpha_l\}$ to minimize:

$$\sum_i (y_i - \sum_{l=1}^{L} \alpha_l f_l(X_i))^2 \tag{5.5}$$

$\{\alpha_l\}$ can be estimated through ordinary least squares. However, as Brieman (1996) pointed out, if the level-one data is constructed using training set S, and the $\{\alpha_l\}$ are also selected by minimizing squared error over S, then the resulting $\{\alpha_l\}$ will overfit the data and suffer from high prediction inaccuracy on a test set. As a remedy to this problem, Wolpert (1992) suggests using leave-one-out cross-validation (CV) to generate leave-one-out prediction $f_l^{LOOCV}(X)$ and Breiman (1996) proposed to the use K-fold CV to generate cross-validated prediction $f_l^{CV}(X)$ to save computational costs. The K-fold CV algorithm works in the following way (LeDell, 2015):

- Randomly divide the original training set (level-zero data) into K folds with roughly equal size: $X_{(1)}, \dots, X_{(K)}$;
- For each base learner f_l in the library, train a model using K-1 folds data and make prediction on the hold-out fold data. Repeat the process K times, where in each a different fold of the data is used as the hold-out data

For each base learner, this algorithm will generate an n-dimensional vector of cross-validated predictions with the same size as the original training data. Then we can stack the cross-validated prediction output of L base learners to form a level-one design matrix Z with dimension $n \times L$. Accordingly, the original training data is often referred to as the level-zero data with dimension $n \times p$, where n is the number of observations and p is the number of predictors. The $\{\alpha_l\}$ are then chosen to minimize:

$$\sum_i (y_i - \sum_{l=1}^{L} \alpha_l f_l^{CV}(X_i))^2 \tag{5.6}$$

Along this line, Leblanc and Tibshirani (1996) also compare CV-based level-one data generation with a bootstrap-based method. In some cases, bootstrap-based level-one data generation works slightly better than the CV-based method, but it is computationally more expensive.

5.2.2 Meta-learning

After obtaining level-one data, the question remaining is how to estimate $\{\alpha_l\}$. This process is called meta-learning with the objective of finding the optimal linear combination of L base learners. Breiman (1996) found that when the coefficients in equation 5.6 were constrained to be nonnegative and $\sum \alpha_l = 1$, stacking showed better prediction error than any of the base learner $f_l(\mathbf{x})$. However, the solution for Equation 5.6 obtained using nonnegative least squares (Lawson and Hanson, 1987) might be too restrictive in finding optimal solutions. Hence, a data-adaptive meta-learning approach is needed. Since all the base learners are predicting the same *y*, their CV-predictions will be highly correlated, thus regularization-based meta-learning via lasso (Friedman and Popescu, 2003) is preferable. Lasso can help shrink the weights of “unnecessary” learners to zero and select “useful” learners. Now the minimization problem becomes the following:

$$\begin{aligned} &\min_\alpha \sum_i (y_i - \sum_{l=1}^{L} \alpha_l f_l^{CV}(X_i))^2 \\ &s.t. \|\alpha\|_1 \le t \end{aligned} \tag{5.7}$$

or equivalently

$$\min_\alpha \sum_i (y_i - \sum_{l=1}^{L} \alpha_l f_l^{CV}(X_i))^2 + \lambda \|\alpha\|_1 \tag{5.8}$$

The regularization parameter λ can be selected using validation set or K-fold cross-validation in the training set. Note that there will be no intercept term in the meta-learning process.

Broadly speaking, meta-learning can also be viewed as a typical function approximation or learning task in which one can apply any type of parametric or nonparametric algorithm to level-one data and learn a meta-learner *g* using a bounded loss function (LeDell, 2015). For instance, one can train random forests on level-one data and use K-fold cross-validation or validation to tune model parameters.

$$\arg\min_g \sum_i (y_i - g(Z_i))^2 \tag{5.9}$$

However, in general, complicated meta-learning algorithms suffer from overfitting issue. Thus, it is common to estimate an optimal linear combination of the base learners.

Note that the stacking framework can be applied to any user-defined loss function. In the case of fuel burn prediction, a squared error loss is used in both base learners training and meta-learning. As shown in Chapter 6, we can also apply stacking to quantile error loss in the setting of quantile regression. Details can be found later.

It must be mentioned that from a practical point of view, the stacking framework also enjoys another advantage. Recall that we use the CV-prediction results of L base learners to form level-one data for meta-learning. These L base learners are not restricted to the best-tuned base learners. We can incorporate all the training efforts into the level-one data. For instance, in the typical model training of gradient boosting, one may specify a different number of trees for boosting (say 1,000, 2,000, or 3,000). In this case, L is 3 instead of 1.

5.3 Prediction Uncertainty

From a dispatcher's perspective, a point prediction of trip fuel burn from a machine learning model might not be adequate. It is desired to construct prediction intervals (PIs) associated with point prediction so that dispatchers can evaluate model results and make fuel loading decisions with a high level of confidence. Furthermore, the development of PI can also be used to replace discretionary fuel loading.

In linear regression models, one can derive closed-form expression for PIs assuming model errors follow independent normal distributions (Wonnacott and Wonnacott, 1996). However, for many non-linear models, there are no closed form expressions for model errors and one has to analyze the statistical properties of the model error which is a non-trivial task for most machine learning algorithms (Chryssoloiuris et al., 1996). Another method is to use simulation and re-sampling. However, this method is often computationally intensive. Other non-probabilistic approaches such as the fuzzy set approach (Maskey et al., 2004), require better understanding of the uncertainty associated with model predictors.

Since historical residuals (defined as the observed differences between observed data and predicted values) are the best indicators of future prediction errors, in this study we follow the procedure described by Shrestha and Solomatine (2006) to construct PIs for test set flights based on the empirical distributions of residuals in the validation set. The main idea is to partition the input space (model predictors) into different clusters having similar model errors (relaxing the constant error assumption) using clustering techniques. Then PI is constructed for each cluster. In the case of hard-clustering (e.g. K-means), for a flight in the test set belonging to cluster A, its corresponding $(1-\alpha)\times 100$ percent PI can be computed by finding the $(\alpha/2)\times 100$ and $(1-\alpha/2)\times 100$ percentile values from the empirical distribution of residuals in validation set with the same cluster assignment. However, in the case of soft-clustering (e.g. Gaussian mixture) where the clustering assignment is not deterministic, but rather a flight belongs to more than one cluster with corresponding membership probabilities, we need to consider membership probabilities in the PI construction as well.

To be more specific, we first need to sort flights in the validation set with respect to the residuals (prediction errors) in ascending order. Let p_{ij} denote the membership probability of flight j belonging to cluster i, e_j to indicate the residual of flight j (sorted), and n to denote the total number of observations in the validation set. Then, PIC_i^{Lo}, the lower PI for cluster i is e_j where j is the maximum value that satisfies the following inequality:

$$\sum_{k=1}^{j} p_{ik} < \frac{\alpha}{2} \sum_{j=1}^{n} p_{ij} \tag{5.12}$$

Similarly, PIC_i^{Up}, the upper PI for cluster *i* is e_h where *h* is the maximum value of it that satisfies the following inequality:

$$\sum_{k=1}^{h} p_{ik} < (1-\frac{\alpha}{2}) \sum_{h=1}^{n} p_{ih} \tag{5.13}$$

Once we have computed upper and lower PI for each cluster, the PI for a flight *j* is the weighted average of the PI of each cluster:

$$PI_j^L = \sum_{i=1}^{c} p_{ij} PIC_i^{Lo} \tag{5.14}$$

$$PI_j^U = \sum_{i=1}^{c} p_{ij} PIC_i^{Up} \tag{5.15}$$

where PI_j^L and PI_j^U are the lower and upper PI for flight *j*, respectively. The constructed PI can be gauged using the coverage probability which is the percentage of flights in the test set with actual fuel burn values falling into the constructed PI plus fuel prediction.

5.4 Model Performance

In this section, we evaluate the performance of stacking-ensemble learning and clustering-based PI construction using the airline data. The airline data described in Chapter 2 (368,607 flights with no missing weather information) is randomly divided into three categories: 60% training set, 20% validation set, and 20% test set. Prediction performance was evaluated based on test set, and PIs were constructed using validation set. Table 5.4 shows the number of flights in each OD cluster

Table 5.1 Sample summary

Number of flights	Training set	Validation set	Test set
OD Cluster 1	158,733	52,911	52,911
OD Cluster 2	36,519	12,174	12,174
OD Cluster 3	25,911	8,637	8,637

5.4.1 Prediction Performance

In terms of base learners, we want a library that covers the function space well in places where they are needed and base learners sufficiently different from each other for the meta-learning to be effective (Hastie et al., 2009). To this end, we used the following base learners: lasso, ridge, random forests (RF), gradient boosting machine (GBM), k-nearest neighbor (KNN), and multivariate adaptive polynomial spline regression (MARS) to capture both linear and non-linear functional forms. Table 5.5 presents the features used in the prediction model development. The selected features include terminal weather forecasts, historical airborne time distribution, and flight characteristics such as aircraft type. Since we do not have en route weather and traffic forest information, we incorporate FPS planned fuel burn and planned trip time as proxies for the missing information.

Table 5.2 Summary of predictors

Weather-related predictors			
Forecasted visibility for destination airports	Forecasted ceiling for destination airports	Forecasted thunder storm conditions for destination airports	Forecasted snow conditions for destination airports
Requirement for 1st alternative airports	Requirement for 1st alternative airports	Planned alternate fuel	Month (seasonality)
Predictability-related predictors			
Mean airborne time	Median airborne time	75th percentile of airborne time	80th percentile of airborne time
85th percentile of airborne time	90th percentile of airborne time	95th percentile of airborne time	Standard deviation of airborne time
Max airborne time	Median deviation from the median of airborne time	Mean deviation from flight plan	Median deviation from flight plan
75th percentile of deviation from flight plan	80th percentile of deviation from flight plan	85th percentile of deviation from flight plan	90th percentile of deviation from flight plan
95th percentile of deviation from flight plan	Standard deviation of deviation from flight plan	Max deviation from flight plan	Median deviation from the median of deviation from flight plan
Flight characteristics			
Coordinates of OD	Planned trip time	Aircraft type	Flight distance
Fuel capacity	Hub airport indicators	FPS planned trip fuel	

The base learners were tuned using 5-fold cross validation to find optimal hyper-parameter settings, such as maximum tree depth in random forest, number of trees in gradient boosting, and so forth. The cross-validation error and test set error using various algorithms can be found in the Appendix. Since KNN performed unexpectedly worse than the FPS, its test set prediction results are not presented in the following figures (see Appendix for details). Figures 5.6 to 5.8 display the model prediction results (measured in mean-squared prediction error) on the test set for each OD cluster. The best base learner was found to be gradient boosting, followed closely by random forests. It is shown that by applying machine learning techniques, the test error can be reduced by about 50% compared to our study airline's FPS. Using ensemble learning, we can achieve about 2–5% error reduction compared to the best base learner. The lasso-based meta-learner is also found to yield better prediction performance than non-negative-least-squares (NNLS)-based meta-learner. This comparative superiority results from lasso providing flexibility in finding the optimal weights of different base learners.

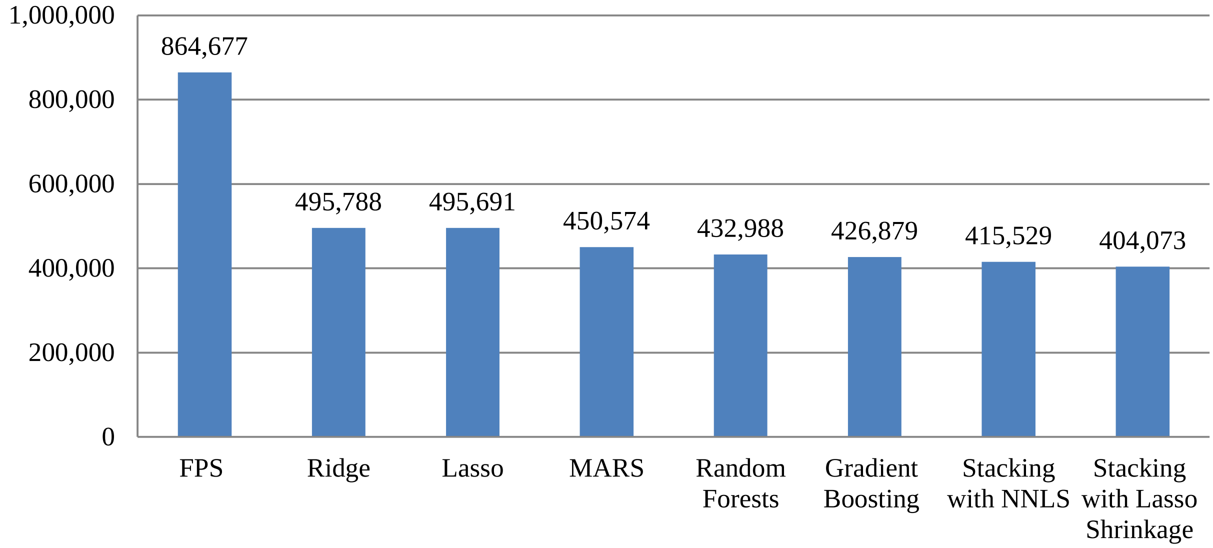


Figure 5.6 Mean squared error on test set (in lbs^2) for OD cluster 1

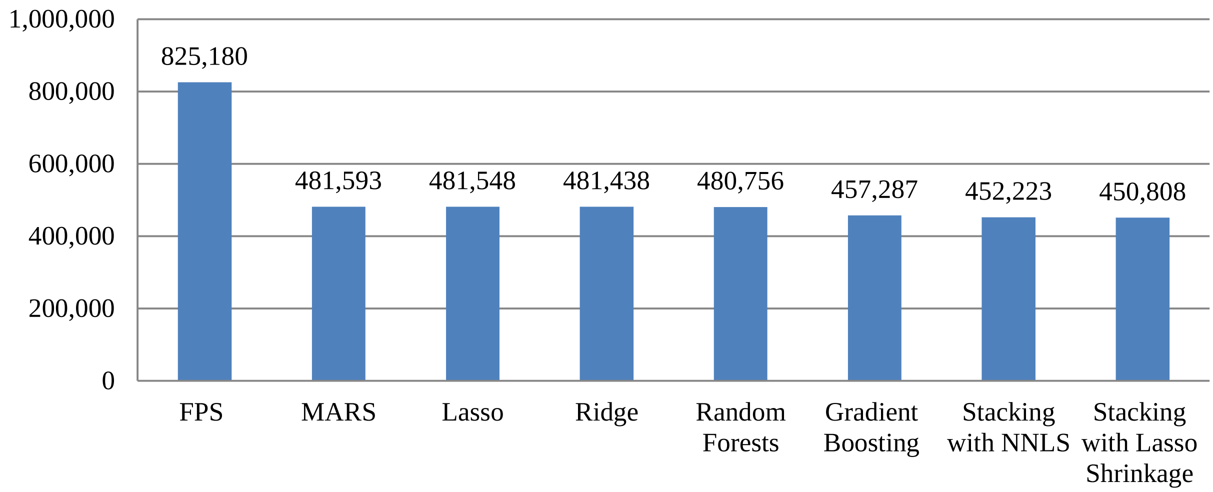


Figure 5.7 Mean squared error on test set (in lbs^2) for OD cluster 2

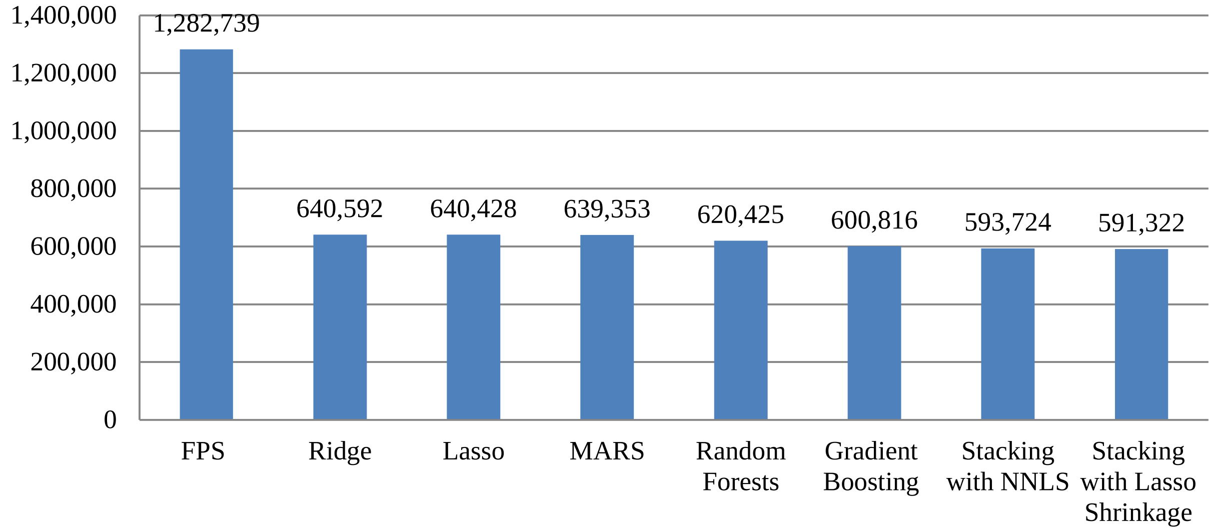


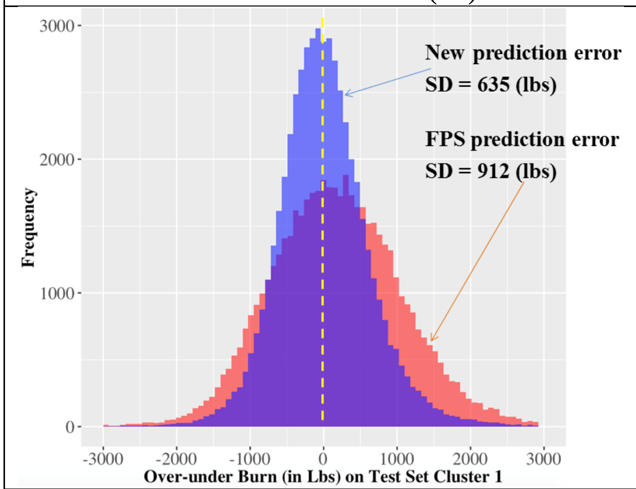
Figure 5.8 Mean squared error on test set (in lbs^2) for OD cluster 3

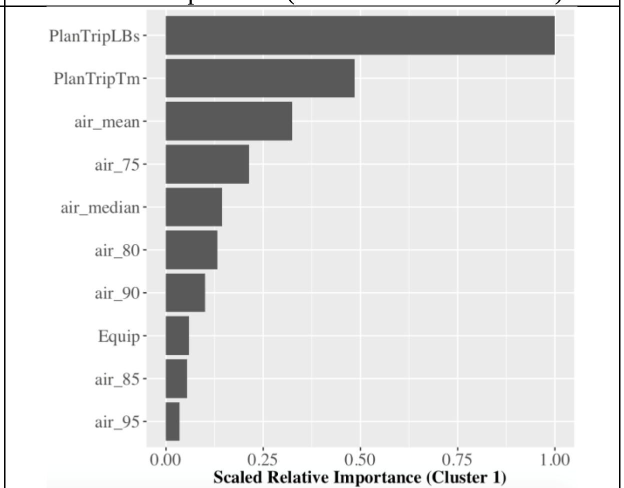
Prediction errors on test sets based on stacking with lasso shrinkage for each OD cluster are presented in Table 5.6. It can be seen that ensemble learning yields more reliable prediction results compared to the airline FPS in the sense of reduced variability and bias (zero-centered means). This bias-variance reduction results in smaller mean squared error shown in Figures 5.6 to 5.8. Since ensemble learning does not provide variable importance measures, to get a rough idea of which predictors are more important, we use the scaled importance measure generated by the random forests algorithm. Variable importance is determined by whether that variable was selected during splitting in the tree building process in random forests training and how much the squared error (over all trees) improved as a result. Then a variable's importance measure is scaled with respect to the most important feature. In other words, the most important feature has a scaled relative importance of 1, and all the other variables are measured relative to the most important feature. The top 10 features are presented for each OD cluster. It can be found that FPS planned fuel burn and planned trip times are the top two important features. Different quantiles of historical airborne time distribution and aircraft type are also found to be important in improving prediction performance.

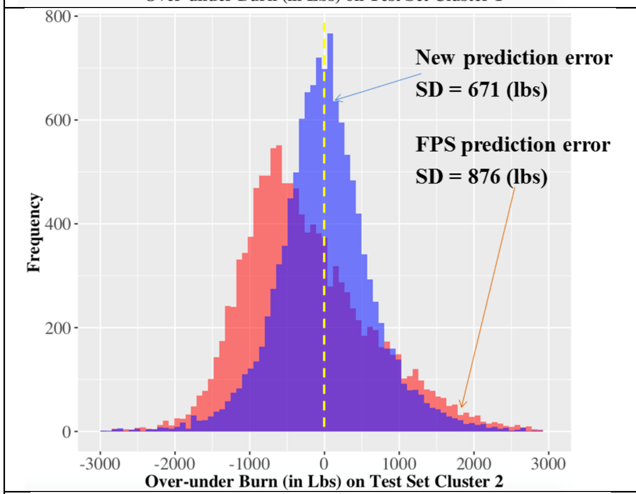
It is not surprising that airline FPS planned fuel burn and trip time are the most important features. They are the best from the FPS based on airline aircraft performance model and en route weather and traffic forecasts that are only available to the study airline. Here, we use them as best the proxies for the missing en route forecast information. The feature importance result also suggests that system predictability (measured by historical flight time distribution) plays an important role in flight fuel burn prediction. This indicates that besides airlines' efforts, the FAA can help achieve fuel cost reduction by improving system predictability. This finding is consistent with the results from Hao et al. (2016b).

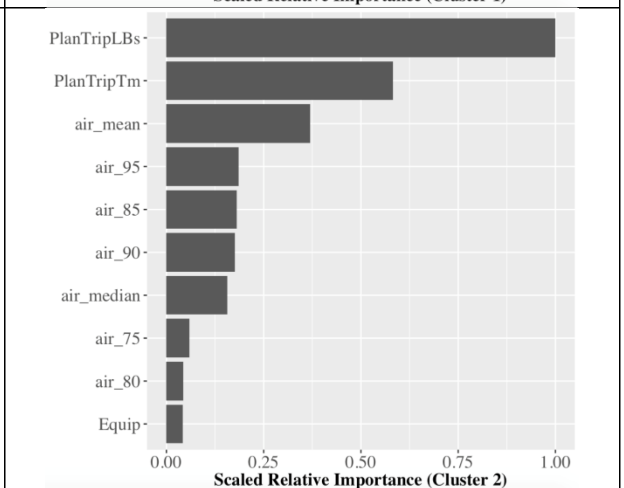
Table 5.3 Prediction error and variable importance

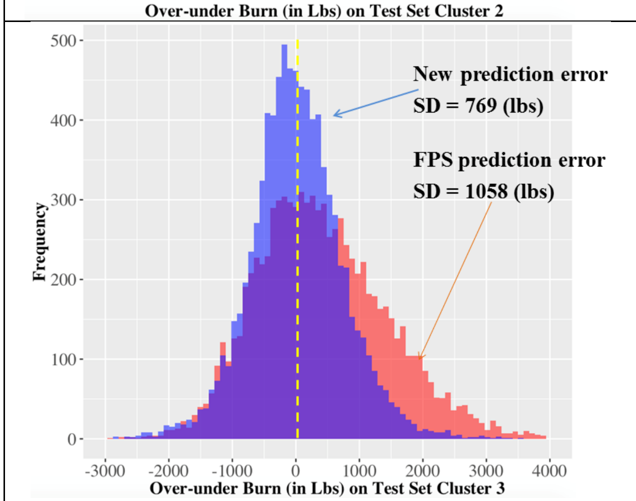
Test set Prediction error (lbs)	Variable Importance (based on random forests)

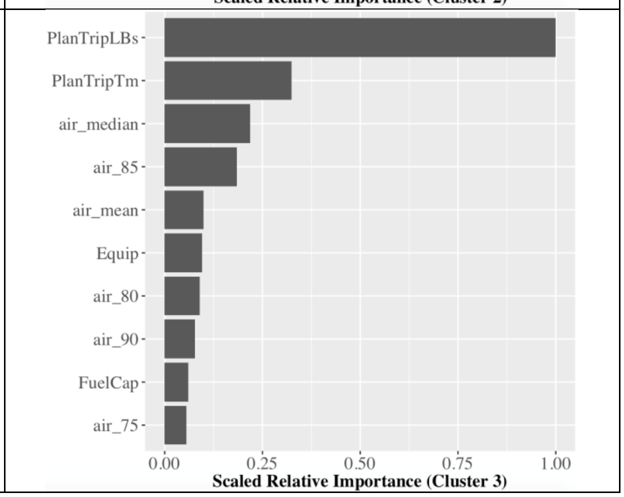












5.4.2 Prediction Interval

In order to achieve more accurate PI construction, within each OD cluster (Layer 1), we perform another layer of clustering based on individual flight characteristics, including historical distribution of airborne times, planned trip time and trip fuel burn, and forecasted weather conditions at the destination airports. For Layer 2 clustering, we apply Gaussian mixture models by assuming that our input features follow multivariate Gaussian distributions with full covariance structure. Compared to K-means, Gaussian mixture clustering is a soft-version clustering technique as it estimates the probabilities of each flight belonging to a Gaussian distribution. This soft version clustering allows us to utilize information from all sub-clusters in constructing PI (see details later).

The optimal number of Gaussian distributions (sub-clusters) per OD cluster is determined based on Bayesian Information Criterion (BIC), an indicator of model fit, where lower BIC indicate a better fit. Shown in the BIC plots (Figure 5.9), as we increase the number of sub-clusters, the BIC decreases. However, after certain numbers, the marginal decrease becomes negligible. Therefore we stop at the position where subsequent improvement in model fit is insignificant and pick the corresponding number as our optimal number of sub-clusters. One reason for adopting this early stopping is that we generally prefer a simple model to complicated models. In the end, we found two sub-clusters for OD cluster 1, two sub-clusters for OD cluster 2, and three sub-clusters for OD cluster 3. For each of the 10 sub-clusters, we constructed a PI based on the prediction residuals in the validation set. Then, for each OD cluster, the PI was a weighted average of its sub-clusters' PIs.

As expected, the PIs developed based on the validation set provide consistent coverage probability in the test set. Take OD cluster 1 as an example, as shown in Table 5.7, using a 95% PI developed based on OD cluster 1's validation set and applying to OD cluster 1's test set, for 95.1% of flights, the differences between actual fuel burn and predicted fuel burn are within the constructed PIs.

Table 5.4 Coverage performance on test set

Test set performance	Coverage probability using 95% PI	Coverage probability using 99% PI
OD cluster 1	95.1%	99.0%
OD cluster 2	95.5%	99.0%
OD cluster 3	94.8%	98.9%

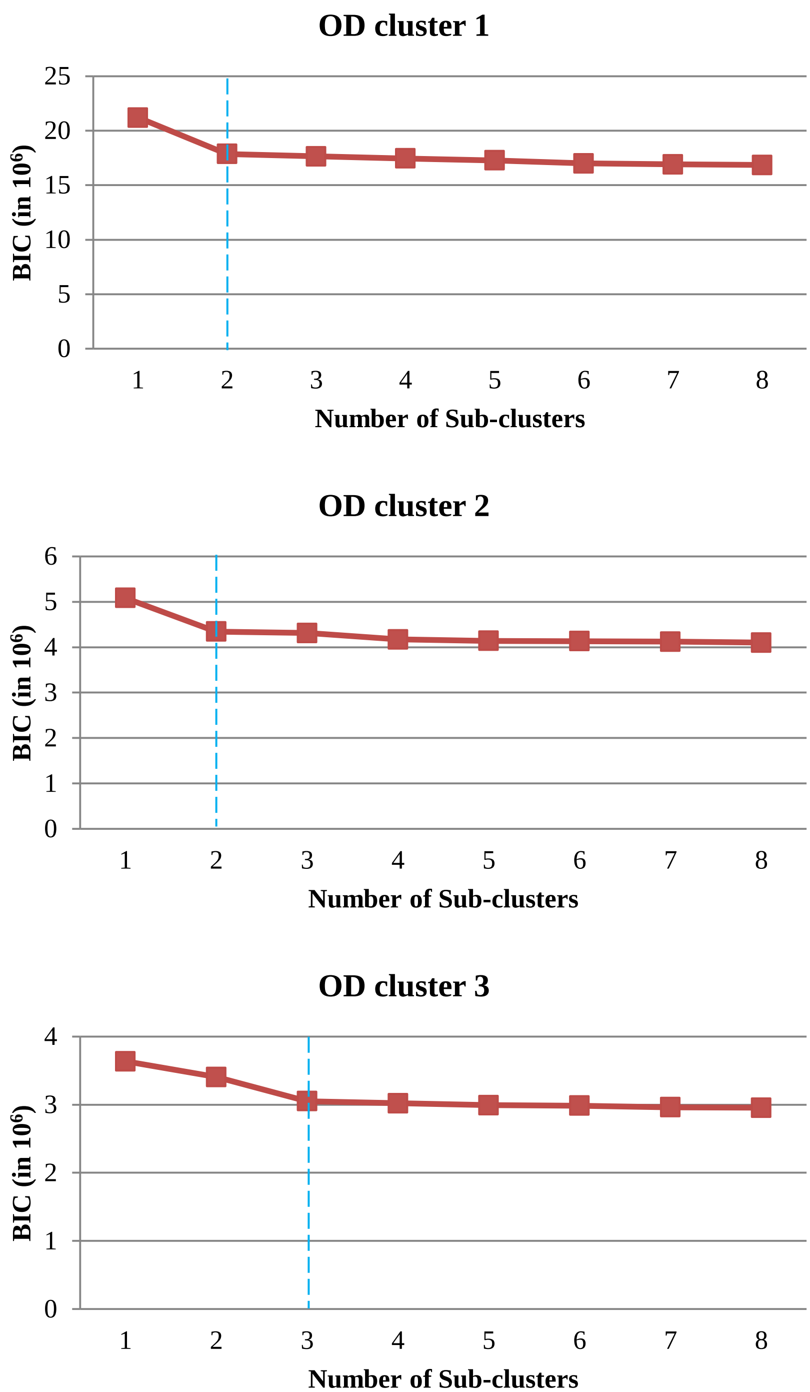


Figure 5.9 Sub-clustering results

5.5 Safety Check and Cost-to-Carry Analysis

In this analysis, we assume dispatchers trust proposed discretionary fuel and perform discretionary fuel loading accordingly. The upper PI (calculated by Equation 5.15) serves as a good candidate for discretionary fuel loading quantity. Next, we need to evaluate whether the proposed discretionary fuel can reduce fuel cost while maintain a same safety level as before? In light of airline practice, we test both the upper 95% PI and upper 99% PI. The selection of these two quantiles are consistent with airline's practice of using SCF95 and SCF99 (refer to Section 1.2 and Chapter 6 for more details about SCF).

As shown in Table 5.9, in the test set 0.03% of flights in OD cluster 1 landed with reserve fuel being used, and 0 flights in OD cluster 2 and 3 used reserve fuel based on original dispatcher discretionary fuel loading. If we use the upper 95% PI and upper 99% PI as discretionary fuel, we can achieve similar safety performance (though slightly worse) in terms of number of flights using reserve fuel. This is because dispatcher discretionary fuel loading is always higher than our proposed discretionary fuel (e.g. upper 95% and 99% PI) which results in safe, but over-conservative fuel loading.

The difference between dispatcher discretionary fuel and our proposed discretionary fuel defines the opportunity for fuel saving. As we mentioned before, there is a CTC extra weight on a flight. Hence, if we can reduce discretionary fuel loading, then the CTC discretionary fuel would also decrease. Recall cost-to-carry factor is defined as the pounds of fuel consumed per pound of fuel carried per mile, and it varies across aircraft types and flight distance. Then, using Equation 3.4, we can estimate fuel savings due to carrying less discretionary fuel.

Column 2, in both Table 5.11 and Table 5.12, presents average CTC fuel saving per test set flight if we use the upper 95% and 99% PI as discretionary fuel. We also consider the breakdown of terminal weather conditions. A weather-impacted flight is defined as a flight with the following TAF weather forecast at destination airport: forecasted ceiling below 2,000 feet, or visibility below 3 miles, or forecasted thunderstorm presence. After obtaining per-flight-basis fuel saving, by extrapolating back to the original size of the data, we can obtain an estimate of airline-wide total savings. The extrapolation is performed in the following way: First, compute fuel savings in the test set. Since test set is a 20% random sample of the airline weather-matched data, we can multiply test set savings by 5. As mentioned in Chapter 2, weather-matched flights represent about half of the original dataset over 14 months of operation, and we further multiply the benefit by 2 to find airline-wide fuel savings. More careful extrapolation differentiating aircraft types or airports, would make for an interesting future research topic; however, such specificity is not the focus of this study, which concentrates on the flight-level impact.

In addition to monetary savings, we also utilize the U.S. Environmental Protection Agency conversion factor for jet fuel (Environmental Protection Agency, 2013) to translate fuel savings into reduction in CO_2 emission in kilograms. Our study shows that airline can save \$76.1 million and 530 million kg of CO_2 emissions if applying the upper 95% PI as discretionary fuel and \$60.4 million and 421 million kg of CO_2 emissions if applying upper 99% PI as discretionary fuel (assuming \$3/gallon as jet fuel price). The upper 99% PI is, of course, always higher than upper 95% PI, resulting in somewhat smaller savings, but the difference is not very large.

The above results suggest significant benefits of using improved fuel burn prediction in discretionary fuel calculation. However, from airline's perspective, the percentage of flights landing with reserve fuel being used based on proposed discretionary fuel still seems high (see

Table 5.9 column 3 and 5). To address this safety concern, we propose to add a safety buffer to PI-based discretionary fuel, which can help achieve the same safety performance as the current practice for our study airline. In this case, we estimate a scaling factor η and the new discretionary fuel quantity for flight *j* becomes $\eta \times PI_j^U$. Scaling factor η can be learned for each OD cluster based on its validation set.

$$\begin{aligned} &\min \eta \\ &s.t. \frac{1}{N}\sum_{j=1}^{N} I(AF_j - PF_j - \eta PI_j^U < 45) < \gamma \end{aligned} \tag{5.16}$$

Using a constrained optimization formulation, we can find a minimum scaling factor η that achieves a given threshold safety γ. In formula 5.16, I is an indicator function. AF_j denotes actual fuel burn for flight *j*, and PF_j denotes planned fuel burn.

Take OD cluster 1 as an example, as shown in Table 5.10, for good weather flights in the validation set, 0.03% of flights landed with reserve fuel being used according to dispatcher discretionary fuel loading, whereas using the upper 99% PI as discretionary fuel would result in 0.3%. Using line search, we find that if we multiple PI-based discretionary fuel for each flight by 1.09, we can also achieve 0.03% in the validation set. Similar scaling factor estimation is carried out for other OD clusters. By applying the learned scaling factor to the test set, we can achieve \$71.7 million fuel saving over 14 months (or \$61.5 million per year) if using the scaled upper 95% PI and \$56.4 million fuel saving over 14 months (or \$48.3 million per year) if using the scaled upper 99% PI with the same safety level as the current practice of our study airline (see column 5 of Table 5.11 and Table 5.12).

Table 5.5 Safety check on test set

	Percentage of flights landing using reserve fuel				
Test set performance	Use original discretionary fuel	Use upper 95% PI as discretionary fuel	Use upper 95% PI with scaling factor as discretionary fuel	Use upper 99% PI as discretionary fuel	Use upper 99% PI with scaling factor as discretionary fuel
OD cluster 1 (52911 flights)	0.03%	1.7%	0.02%	0.3%	0.02%
OD cluster 2 (12174 flights)	0%	2.0%	0%	0.4%	0%
OD cluster 3 (8637 flights)	0%	1.7%	0.01%	0.3%	0.01%

Table 5.6 Safety check on validation set

Validation set performance		Percentage of flights landing using reserve fuel	Percentage of flights landing using reserve fuel if use upper 99% PI as discretionary fuel	Scaling factor η for SCF99	Percentage of flights landing using reserve fuel if use upper 95% PI as discretionary fuel	Scaling factor η for SCF95
OD cluster 1	Good WX	0.03%	0.3%	1.12	2.1%	1.20
	Bad WX	0.01%	0%	1.00	0.05%	1.02
OD cluster 2	Good WX	0%	0.5%	1.11	2.4%	1.18
	Bad WX	0%	0%	1.00	0.2%	1.02
OD cluster 3	Good WX	0%	0.5%	1.03	2.3%	1.07
	Bad WX	0%	0.08%	1.05	0.08%	1.05

Table 5.7 Cost-to-carry saving using 95% PI as discretionary fuel

<table>
<tr><td colspan="2"></td><td colspan="3">Use upper 95% PI as discretionary fuel</td><td colspan="3">Use upper 95% PI with scaling factor as discretionary fuel</td></tr>
<tr><td colspan="2">Test set performance</td><td>Mean CTC fuel savings (in lbs) per flight</td><td>Airline-wide monetary savings ($) (assuming $3/gallon)</td><td>Airline-wide CO_2 reduction (kg)</td><td>Mean CTC fuel savings (in lbs) per flight</td><td>Airline-wide monetary savings ($) (assuming $3/gallon)</td><td>Airline-wide CO_2 reduction (kg)</td></tr>
<tr><td rowspan="2">OD cluster 1 (52911 flights)</td><td>Weather impacted flights</td><td>232</td><td rowspan="6">7.61×10^7</td><td rowspan="6">5.30×10^8</td><td>230</td><td rowspan="6">7.17×10^7</td><td rowspan="6">4.99×10^8</td></tr>
<tr><td>Non-weather impacted flights</td><td>192</td><td>177</td></tr>
<tr><td rowspan="2">OD cluster 2 (12174 flights)</td><td>Weather impacted flights</td><td>172</td><td>169</td></tr>
<tr><td>Non-weather impacted flights</td><td>271</td><td>249</td></tr>
<tr><td rowspan="2">OD cluster 3 (8637 flights)</td><td>Weather impacted flights</td><td>431</td><td>422</td></tr>
<tr><td>Non-weather impacted flights</td><td>404</td><td>394</td></tr>
</table>

Table 5.8 Cost-to-carry saving using 99% PI as discretionary fuel

<table>
<tr><td colspan="2"></td><td colspan="3">Use upper 99% PI as discretionary fuel</td><td colspan="3">Use upper 99% PI with scaling factor as discretionary fuel</td></tr>
<tr><td colspan="2">Test set performance</td><td>Mean CTC fuel savings (in lbs) per flight</td><td>Airline-wide monetary savings ($) (assuming $3/gallon)</td><td>Airline-wide CO_2 reduction (kg)</td><td>Mean CTC fuel savings (in lbs) per flight</td><td>Airline-wide monetary savings ($) (assuming $3/gallon)</td><td>Airline-wide CO_2 reduction (kg)</td></tr>
<tr><td rowspan="2">OD cluster 1 (52911 flights)</td><td>Weather impacted flights</td><td>177</td><td rowspan="6">6.04×10^7</td><td rowspan="6">4.21×10^8</td><td>177</td><td rowspan="6">5.64×10^7</td><td rowspan="6">3.92×10^8</td></tr>
<tr><td>Non-weather impacted flights</td><td>152</td><td>138</td></tr>
<tr><td rowspan="2">OD cluster 2 (12174 flights)</td><td>Weather impacted flights</td><td>95</td><td>95</td></tr>
<tr><td>Non-weather impacted flights</td><td>213</td><td>193</td></tr>
<tr><td rowspan="2">OD cluster 3 (8637 flights)</td><td>Weather impacted flights</td><td>344</td><td>330</td></tr>
<tr><td>Non-weather impacted flights</td><td>341</td><td>334</td></tr>
</table>

5.6 Summary

Having accurate fuel burn predictions is critical for airlines. Loading too much fuel will result in additional CTC. Loading too little fuel, on the other hand, can lead to in-air fuel emergencies. Both cases are harmful and costly to airlines. The current FPS planned trip fuel predictions are subject to errors, the size of which affects the amount of discretionary fuel that dispatchers load. In this study, we have shown how ensemble learning techniques can be used to build more accurate fuel prediction models. The lasso-based stacking is found to reduce the mean squared prediction error by 50% over the current FPS and by 2-5% over the best base learners. So far, we have considered only six base learners. However as we increase library size by incorporating other powerful algorithms in base learners (e.g. neural network, SVM, etc.), we would expect to generate an even more powerful predictions.

Besides more accurate predictions, we also construct PI with 95% and 99% coverage probabilities using a Gaussian mixture clustering idea. We propose to use upper 95% PI and 99% PI as discretionary fuel. By assuming dispatchers trust the proposed discretionary fuel and load discretionary accordingly, we find that our study airline can save roughly $61.5 million and 428 million kg annually with the same safety performance, as measured by the risk of using reserve fuel, is realized under current practice.

To summarize, this chapter provides a good example of applying ensemble learning in improving flight trip fuel burn prediction. Moreover, PI-based discretionary fuel is found to be a good alternative to replace dispatcher discretionary fuel with significant fuel savings and CO_2 reduction. In the next chapter, we provide another approach to tackling the question of how to reduce unnecessary discretionary fuel loading while maintain the same safety level.

6 SCF Estimation Improvement

The underlying goal of Chapters 5 and 6 is to provide airline dispatchers with reliable and accurate recommendations on discretionary fuel loading so that unnecessary fuel loading can be reduced. Unlike Chapter 5, which is established on improved fuel burn prediction models, this chapter studies another approach in estimating discretionary fuel loading. If airlines do not want to improve the FPSs fuel prediction (due to system upgrading costs, etc.), we can still rely on current FPS to estimate discretionary fuel and achieve significant fuel savings. This supposition leads to a discussion of SCF. Recall that in airline fuel-loading practice, SCF is commonly used to provide dispatchers with discretionary fuel recommendations. However, due to limitations in the current SCF estimation procedure, dispatchers have low confidence in applying SCF values (e.g. SCF95) and generally load more discretionary fuel than the recommended numbers. Therefore, in this chapter, we propose a new approach to improve the estimation of SCF. We show that the improved SCF outperforms the current FPS SCF. We also estimate the benefit of using new SCF as discretionary fuel compared to current discretionary fuel loading by dispatchers using a similar CTC analysis as the one described in Section 5.5.

6.1 Motivation

To provide consistent and objective fuel planning, some FPSs provide recommended contingency fuel numbers based on a statistical analysis of historical fuel consumption for similar flights. Carriers usually term this Statistical Contingency Fuel (SCF) (Karisch et al., 2012). When a dispatcher plans a flight (about two hours prior to departure), the FPS pulls historical data (for some number of years as chosen by the airline) of all flights with the same OD pair that were scheduled to depart in the same “hour bank” or time window specified by the airline. As we discussed in Section 1.2, for each historical (non-diverted) flight, the difference between the actual trip fuel consumption and the planned mission fuel consumption is calculated (over/under-burn value). The FPS converts the over/under burn value in pounds to minutes and estimates a normal approximation of the distribution of this excess required fuel burn. The 95th and 99th percentiles of the distribution, which are also called the SCF95 and the SCF99, are provided to dispatchers by the FPS as guidelines for contingency fuel loading. The interpretation of SCF95 (SCF 99) is that based on historical fuel consumption, loading the quantity of contingency fuel specified by SCF95 (SCF99) would result in a flight being able to land without dipping into any reserve fuel (or alternate) 95% (99%) of the time.

SCF has been widely used in the airline industry. According to one survey (Schiefer and Samuel, 2011), many airlines have SCF estimation functionalities (as described above) embedded in their FPSs. These include Air India, British Midland International, United Airlines, Virgin America, Virgin Atlantic, SAS Group of Airlines, to name a few. In the case of our study airline (a major U.S.-based network carrier), the SCF values are calculated based on the set of historical flights took place over the previous year with the same OD and scheduled hour of departure.

However, there are several limitations to the current SCF estimation procedure. First of all, from a statistical perspective, the procedure assumes that under/over-burn is normally distributed, which may not be the case. Second, unless the sample is quite large, the estimate of a 95^{th} or 99^{th} percentile based on the sample mean and SD is subject to considerable sampling error. Likewise, it is of course impossible to calculate SCF values in the case of serving a new OD market with no similar historical flights. Thirdly, although the SCF calculation has implicitly accounted for weather and other events in history by using actual fuel consumption information, dispatchers might still have low confidence in applying those numbers due to oversimplified grouping criterion (e.g. OD-hour). For example, aircraft type is missing from the grouping criterion, even though the aircraft performance models used for different aircraft types may have varying predictive performance. Additionally, in order to increase the confidence level of dispatchers in SCF values, weather forecasts should also be explicitly taken into account. While the high percentiles used to calculate SCF are intended to account for adverse weather, dispatchers are reluctant to trust SCF values in such conditions.

A previous analysis based on the same study airline reveals that dispatchers would almost always load extra fuel above recommended SCF values (Ryerson et al., 2015). As a result, it has been found that 4% of the fuel consumed by an average flight is due to carrying unused fuel. Similar behavior has also been observed in other airlines. For instance, based on a survey of 50 U.S. pilots and dispatchers about their fuel-loading practices, Trujillo (1996) finds that airline dispatchers and pilots always load contingency fuel above the suggested contingency value by the airline. A recent study (Hao et al., 2016b) finds that discretionary fuel loading is related to weather uncertainty and aviation-system predictability. These results suggest that since contingency fuel reflects a dispatcher's assessment of flight uncertainty, improving system predictability can lead to reduction in discretionary fuel loading.

The objective of this chapter is to provide reliable SCF values that dispatchers are more likely to believe. Ideally, their faith in these values would be such that they would generally adhere to them in setting discretionary fuel. To overcome the limitations of the widely used SCF estimation method described above, we propose a new SCF estimation procedure that relies on quantile regression models, focusing on SCF95 and SCF99. A quantile regression model has several desirable properties: (1) it models a given quantile of over/under-burn value directly rather than employing simplified grouping criterion and assuming a normal distribution; (2) it allows covariates to be added into the estimation function so that characteristics such as weather and traffic can be explicitly controlled for; (3) this method also allows us to estimate SCF values for flights where the old method cannot be used because there is not an adequate sample of similar flights.

To get a sense of the relationship between discretionary fuel loading and the SCF value for our study airline, we plot the distribution of discretionary fuel over its corresponding SCF value. In Figure 6.1 and 6.2, the diagonal line represents SCF value. For each SCF category (SCF 95 and SCF99), its corresponding discretionary fuel in general varies considerably and is found to be systematically higher than SCF (boxplot is in general above the diagonal line). This is consistent with the findings from Ryerson et al. (2015) that dispatchers seldom trust SCF values and always load more discretionary fuel than recommended. Table 6.1 presents the mean, SD, 95^{th} and 99^{th} percentile of over/under-burn statistics for eight aircraft types. It can be observed that fuel performance differs across aircraft types. This variance also suggests the need to incorporate aircraft type into SCF estimation, which is missed in the airline's SCF estimation. It is also noted that the SD of the over/under-burn distributions for different aircraft types remain relatively constant.

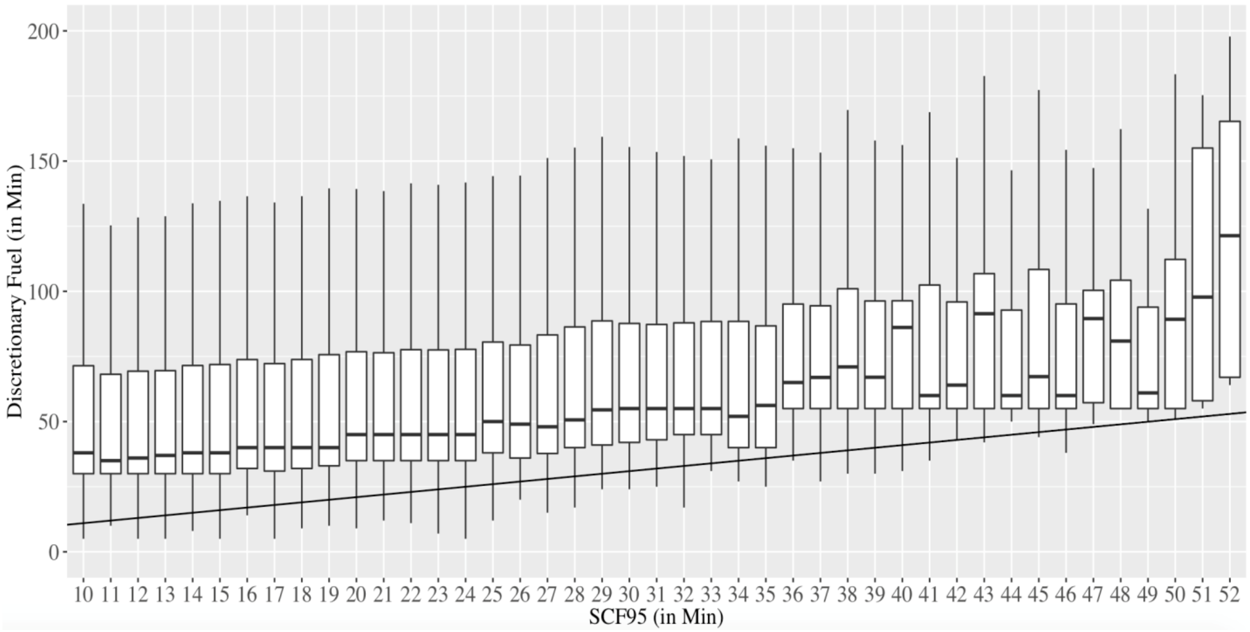


Figure 6.1 Relationship between discretionary fuel and SCF95

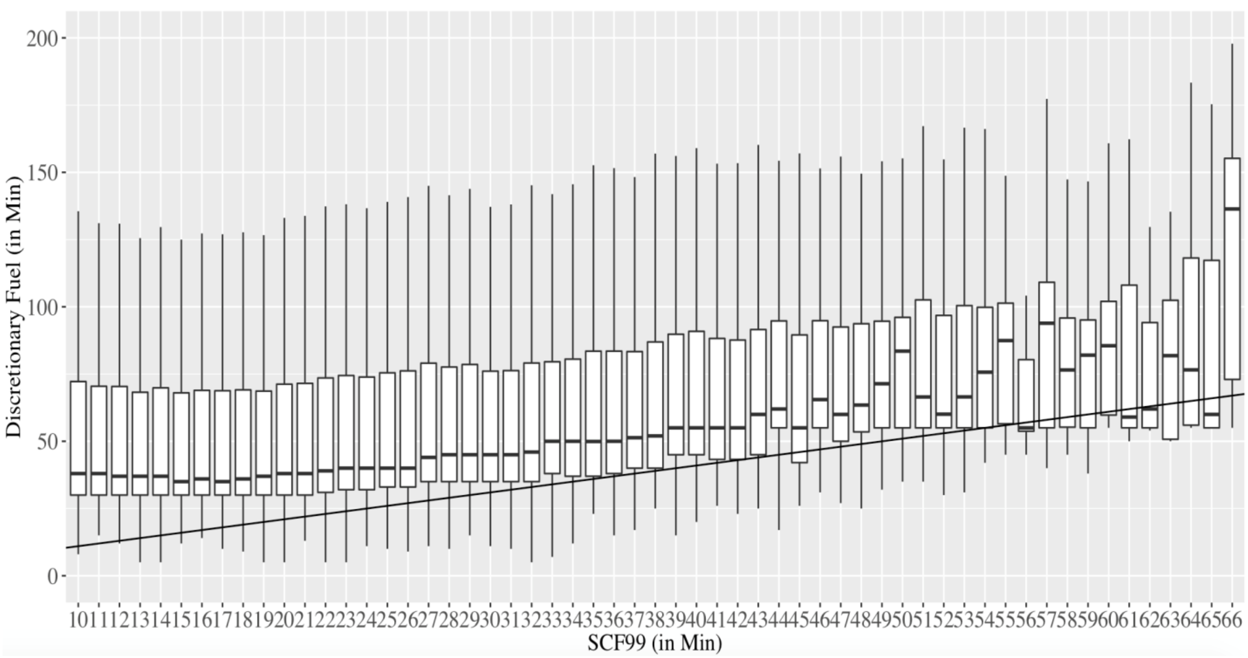


Figure 6.2 Relationship between discretionary fuel and SCF99

Table 6.1 Under/over-burn summary (in minutes)

Aircraft type	Mean	Standard Deviation	95th percentile	99th percentile
A319	-9.8	9.3	5.7	16.0
A320	-7.8	9.5	7.7	18.9
B737-800	-4.2	9.6	12.4	21.9
B757-300	3.2	8.4	16.9	24.2
B757-200	4.6	9.3	20.1	29.6
DC9	6.5	7.4	18.8	26.1
MD88	6.9	8.5	21.2	29.2
MD90	1.3	7.5	13.7	21.3

6.2 Methodological Approach

In this section, we introduce our SCF estimation procedure based on quantile regression method. The response variable Y is the under/over-burn value (in minutes). The covariates X include terminal area weather forests, historical traffic conditions, aircraft types, departure hour window, departure month, and dummies for major airports.

We focus on three techniques: (1) parametric quantile regression (QR); (2) gradient boosting machine (GBM) based quantile regression tree; (3) random quantile forests (RQF). Just as a standard linear regression models conditional mean functions $E(Y \mid X)$, a quantile regression is used to model conditional quantiles of Y: $Q_q(Y \mid X)$, where q denotes a specific quantile. It can be shown that the quantile regression estimator for q-th quantile minimizes the following loss function (Koenker, 2005):

$$J_q(\beta) = \sum_{i=1}^{N} \rho_q(y_i - f(x_i, \beta)) \tag{6.1}$$

where $\rho_q(t) = t(q - \mathrm{I}_{(t<0)})$ and I is an indicator function. In our case, we choose q to be 0.95 and 0.99. If we specify $f(x_i, \beta) = x_i'\beta$, minimizing Equation 6.1 with respect to β would produce parameter estimates of the q-th conditional quantile function. Sometimes, in order to achieve better prediction performance, we could also leverage machine learning algorithms like gradient boosting machine (GBM) to estimate $f(x_i, \beta)$ in a non-parametric manner (Friedman, 2001). GBM is an ensemble learning algorithm which combines different weak learners in a sequential fashion and gradually improves model fit (see Equation 6.2).

$$\hat{f}_t \leftarrow \hat{f}_{t-1} + \zeta h_t(x^k) \tag{6.2}$$

The main idea of applying GBM in the quantile regression setting is to iteratively add simple quantile regression tree models $h_t(x^k)$ to existing models $\hat{f}_{t-1}$, so that the updated model $\hat{f}_t$ can further reduce the quantile loss function specified in Equation 6.1. Here, $h_t(x^k)$ is constructed based on a splitting variable x^k, which reduces the loss function the most at iteration t. ζ is called the learning rate and it is usually set at a small value. We use ζ = 0.005 in this paper. The number of iterations is a tuning parameter that we need to determine. If we set iteration to be a large number, we will be likely to overfit the data. Algorithm details could be found in Ridgeway (2007).

Another powerful learning algorithm comprises the random forests (RF) which has been widely used in the area of conditional mean prediction. The idea of RF is to average the prediction outputs from a large number of decision trees (Breiman, 2001). For conditional mean, the prediction of a single decision tree for a new data point X = x is the mean response of Y in a particular leaf that contains X = x. Then, RF computes a final prediction by averaging predictions from all trees. Drawing an analogy with conditional mean, when it comes to predicting conditional quantiles, instead of using mean response of Y in a leaf, we can report q-th empirical quantiles of Y in that leaf and then average the obtained quantiles across all trees. This algorithm is called random quantile forests (RQF) which has been proposed by Meinshausen (2006). The tuning parameter for RQF is the minimum node size, which is related to how deep we should grow a decision tree. If we set the minimum node size to be a small number, then we obtain a deep tree which is very likely to overfit the data.

In order to improve upon the above suggested individual learners, stacking is applied in the quantile loss setting. Recall that in the meta-learning of stacking, after obtaining validated or cross-validated-prediction results, we can perform another layer of training in seeking the optimal linear combinations of all base learners. Similarly, lasso-based meta-learning can be applied with a quantile loss function.

$$\min_{\alpha} \sum_{i} \rho_q (y_i - \sum_{l=1}^{L} \alpha_l f_l^{CV}(X_i)) + \lambda \| \alpha \|_1 \tag{6.3}$$

Using K-fold CV, we can tune regularization parameter λ and find the best set of $\{ \alpha_l \}$ that gives better prediction performance compared to the best individual learner.

To test the prediction performance of the four proposed quantile regression-based models (QR, GBM, RQF, lasso-based stacking), we look at a loss function-based goodness-of-fit measure (Koenker and Machado, 1999):

$$R(\beta) = 1 - \frac{J_q(f(\beta))}{J_q(\tilde{f}(\beta))} \tag{6.4}$$

where $J_q(f(\beta))$ is the value of loss function on test set using SCF95 estimation function $f(\beta)$. $\tilde{f}(\beta)$ is used to denote a model with the constant term only which is equivalent to using the 95th (or 99th) quantile of Y in the training set as prediction for every flight. Four proposed SCF estimation models plus airline FPS SCF will be measured against a constant-only model developed using the training set. Regarding tuning parameters, in order to achieve the best prediction performance on test set, we should avoid overfitting the training data. Therefore, we need to find the optimal number of iterations in GBM and the minimum node size in RQF using the validation set. Parameter tuning results are provided in later sections. We used the same training, validation, and test sets described in Chapter 5, allowing us to compare the fuel-saving benefits across two approaches.

6.3 Model Performance

In this section, we present the estimation results and prediction performance of four proposed methods: QR, GBM, RQF, and stacking. The parametric QR model results afford easier interpretation compared to the machine learning models; thus, we focus on parameter estimates in the QR model. Prediction performance on the test set across four models will also be discussed.

The estimation results for the parametric QR models with respect to SCF95 and SCF99 are presented in Table 6.2. A positive parameter estimate means an increase in the variable results in higher SCF prediction and vice versa. Regarding aircraft type, the A319 is treated as baseline. The relative magnitudes of parameter estimates are consistent with Table 6.1 which suggests that heavy aircraft in general have higher SCF than smaller aircraft. Longer flights are also found to result in higher SCF, partly because longer flights are more subject to disturbances due to en route weather and traffic. The signs of parameter estimates for historical traffic condition variables are all positive across two SCF models except for the SD of historical airborne time in the SCF95 model which is statistically insignificant. Since we did not have en route weather forecast information, historical traffic predictability measures can serve as proxies for variability in weather and operations conditions that are related to OD, month, and hour of day. The estimation results suggest that if historical traffic conditions are less predictable than represented by large standard deviation of airborne time and large deviation between actual and planned airborne time, then SCF is higher. This also indicates that SCF could be reduced through enhanced ATM targeting on improving system predictability. Forecasted weather conditions for destination airports are found to have bigger impact than origin airports. Among terminal area weather forecasts, forecasted thunderstorm is found to have the biggest impact on SCF, followed by forecasted low-ceiling and low-visibility conditions. The construction of low-ceiling and low-visibility variables is based on the adverse weather definition used in the FARs, which require a flight to carry enough fuel to travel to an alternate airport if the weather conditions are such that visibility is less than 3 miles and the ceiling at the destination airport is less than 2,000 feet at the flight's estimated time of arrival plus/minus one hour. Forecasted thunderstorm indicators at origin airports also have positive impact on SCF. The monthly fixed effects are included in the models to capture seasonality influences not captured by the other variables. It can be found that SCF95 is higher in the winter season, possibly as the results of wind effects. The estimation results of the parametric QR model for SCF99 are also consistent with the SCF95 model with consistent signs of parameter estimates. The interpretation is also similar.

Table 6.2 Estimation results of quantile regression models

Category	Variable	SCF99 Model		SCF95 Model	
		Estimates	T-stat	Estimates	T-stat
--	Intercept	0.518	0.20	-7.829 *	-19.29
Aircraft type (Baseline is A319)	A320	3.485 *	3.68	2.238 *	10.99
	B737-800	2.253 *	2.04	2.042 *	8.58
	B757-300	10.319 *	7.50	11.310 *	42.59
	B757-200	12.306 *	12.32	13.359 *	65.39
	DC9	13.777 *	8.85	15.869 *	59.19
	MD88	15.000 *	15.94	16.547 *	85.59
	MD90	7.538 *	7.13	9.322 *	45.15
Distance	Flight distance (in nautical miles)	0.004 *	1.96	0.003 *	2.94
Historical traffic condition	Median of historical airborne time	0.015	0.97	0.026 *	3.84
	Standard deviation of historical airborne time	0.012	0.06	-0.027	-1.05
	Median of difference between historical actual and planned airborne time	0.171 *	5.08	0.137 *	10.85
	Standard deviation of difference between historical actual and planned airborne time	0.328 *	5.01	0.211 *	8.59
TAF weather forecast for destination airports	Low visibility indicator (1-if lower than 3 miles, 0-otherwise)	3.439 *	3.94	2.444 *	7.62
	Low ceiling indicator (1-if lower than 2000 feet, 0-otherwise)	6.405 *	16.77	5.052 *	34.60
	Thunderstorm indicator (1-if thunderstorm presents, 0-otherwise)	13.268 *	17.14	6.485 *	17.46
	Snow indicator (1-if snow presents, 0-otherwise)	4.937 *	11.82	3.147 *	6.27
TAF weather forecast for origin airports	Low visibility indicator (1-if lower than 3 miles, 0-otherwise)	0.303	0.69	0.151	0.93
	Low ceiling indicator (1-if lower than 2000 feet, 0-otherwise)	0.693	0.71	0.058	0.10
	Thunderstorm indicator (1-if thunderstorm presents, 0-otherwise)	2.582 *	3.12	0.963 *	4.18
	Snow indicator (1-if snow presents, 0-otherwise)	0.202	0.41	-0.217	-0.85
Month (Baseline is January)	February	-0.415	-0.69	0.221	1.04
	March	-2.485 *	-4.03	-1.328 *	-6.96
	April	-1.240 *	-2.18	-0.493 *	-2.75
	May	-0.616	-1.04	0.006	0.03
	June	-1.889 *	-2.86	-1.034 *	-4.89
	July	-1.107	-1.72	-0.781 *	-3.77
	August	-0.971	-1.36	-0.672 *	-3.12
	September	-1.970 *	-2.99	-1.290 *	-6.64
	October	-2.332 *	-3.79	-0.981 *	-5.02
	November	-2.782 *	-4.72	-1.284 *	-6.59
	December	-0.619	-0.99	-0.261	-1.18
Number of observations	221,163				

Note: 1) To save space, airport fixed effects and departure hour fixed effects estimates are not presented in this table. 2) * denotes significant at 5% significance level.

Turning to GBM training results, we have tested the number of iterations ranging from 5,000 to 20,000, and 19,000 iterations were found to produce the smallest loss value on the validation set for the SCF95 model, and 9,000 iterations for SCF99 model. Regarding the node size selection in RQF, 50 was found to achieve the smallest validation set error for SCF95 model, and 100 was selected based on the SCF99 model. In addition, 200 decision trees were trained in an RQF; $1e^{-5}$ was selected as the best regularization parameter for lass-based stacking for the SCF95 model. Similarly, $1e^{-4}$ was found to be the best regularization parameter for the lasso-based stacking SCF99 model.

Using the best model-tuning parameters, the goodness-of-fit measures of four proposed models on the test set are presented in Table 6.3. For the SCF95 estimation, among three individual learners - QR, GBM and RQF, RQF was found to perform slightly better than the other two. However, RQF performs the worst among three individual learners in SCF99 estimation. Lasso-based stacking was found to improve model prediction in both SCF95 and SCF99 estimation. On the other hand, we observed that SCF estimated by the airline FPS was found to provide a poor fit on the test set. Thus, lasso-based stacking was found to be the best prediction model and is thus used in the following benefit assessment.

Table 6.3 Goodness-of-fit measures

Test set performance	Goodness-of-fit measure	
	SCF95	SCF99
FPS[20]	0.076	-0.003 *
QR	0.231	0.196
GBM	0.237	0.200
RQF	0.250	0.187
Stacking	0.268	0.219

Note *: we obtain a negative goodness-of-fit measure for SCF99 model using airline FPS as prediction. This indicates airline FPS SCF99 is performing worse than simply using sample 99th quantile as prediction with respect to the quantile loss we are measuring.

[20] Based on flights with SCF values.

From Table 6.4 to Table 6.7, we plot predicted SCF95 (SCF99) values against airline's FPS SCF95 (SCF99) values. We also consider the breakdown of terminal weather conditions. A weather impacted flight is defined as a flight for with the TAF weather forecast at the destination meets the following criteria: forecasted ceiling below 2,000 feet, or visibility below 3 miles, or forecasted thunderstorm presence. For weather-impacted flights, the quantile regression-based models tend to predict higher SCF values than the FPS, since terminal weather forecast have been explicitly taken into account in the SCF estimation process. This property is desirable for dispatchers because more confidence will be gained in making discretionary fuel decisions by controlling for weather and traffic variables. For non-weather-impacted flights, the proposed methods tend to predict lower SCF values than FPS. Again, by adding weather, traffic, aircraft type information into SCF estimation, dispatchers can also potentially load less discretionary fuel, which would lead to less fuel consumption. It is also noted that some SCF predictions are negative. In these few cases, the fuel burn predicted by the FPS will be higher than the actual fuel burns more than 95% of the time. In other words, for those flights, no discretionary fuel will be needed because FPS over-estimates fuel-burn predictions. However, in order to evaluate the potential fuel-savings of applying the new SCF95, we will follow our study airline's SCF practice and set SCF95 values to be exactly 10 minutes if the corresponding prediction is less than 10. Similar patterns can also be found in SCF99 estimation.

Table 6.4 SCF95 prediction results for weather impacted flights

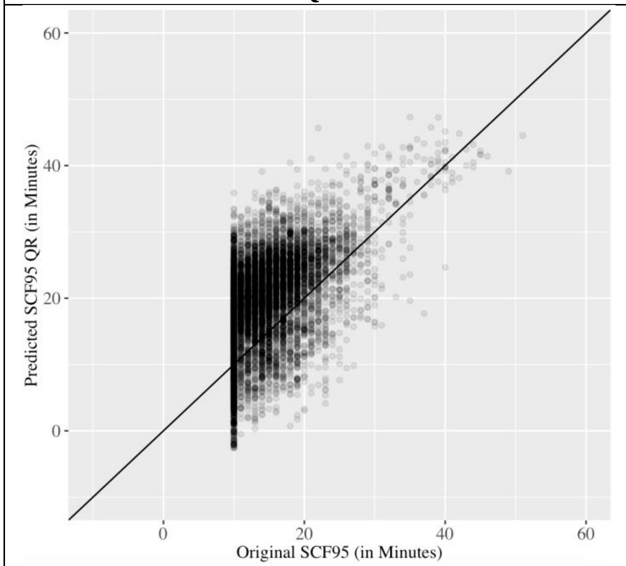
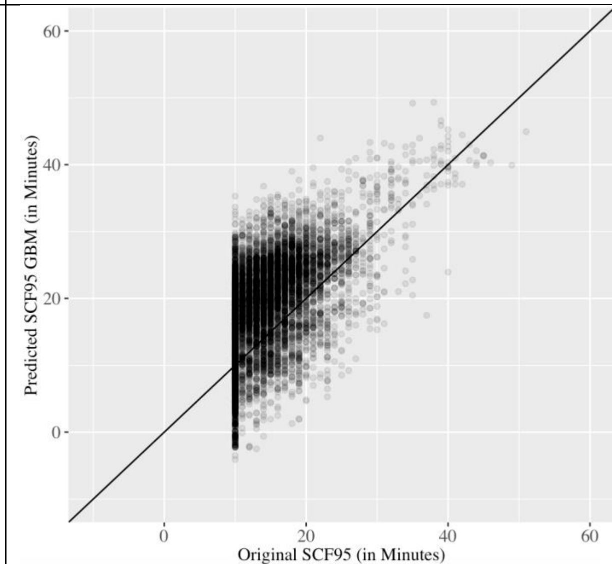
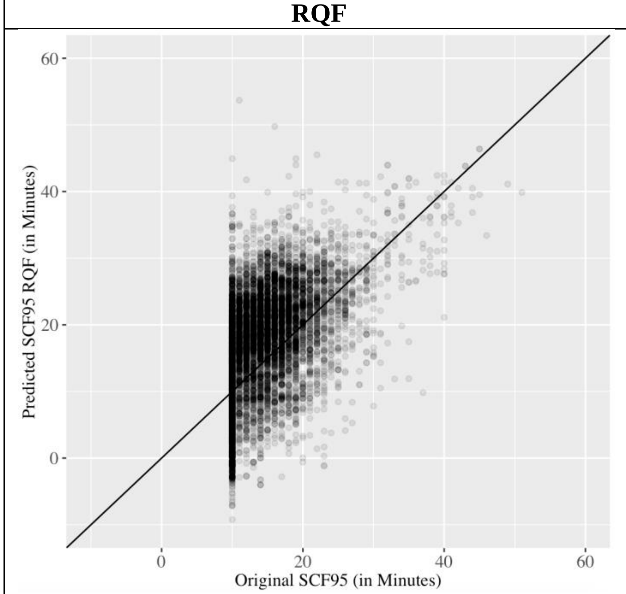
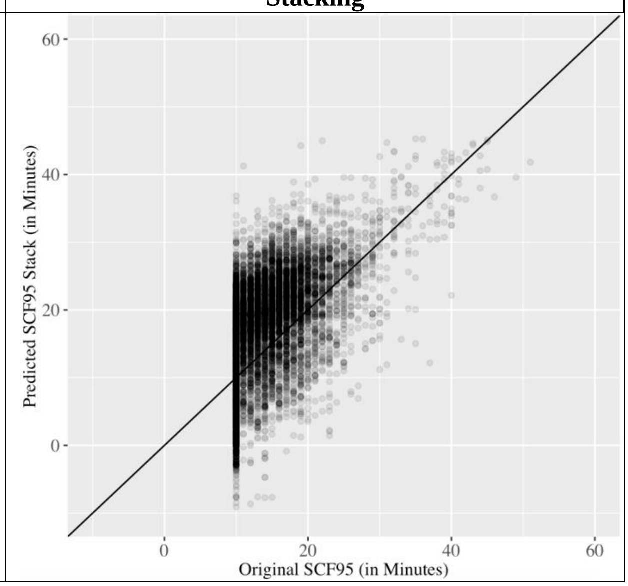
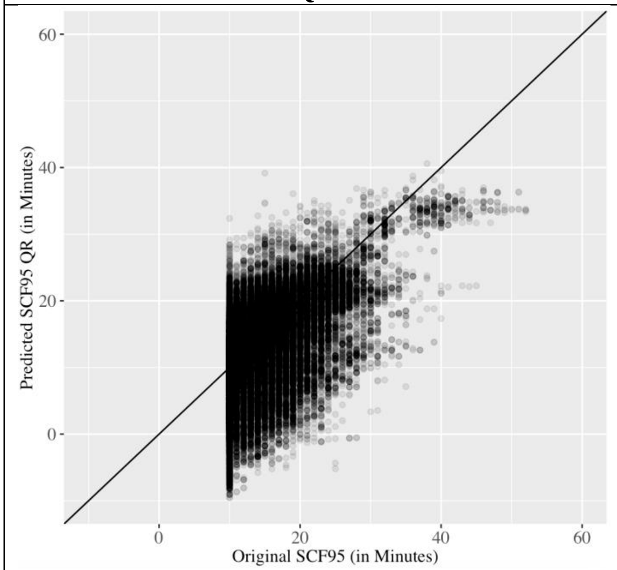
<table>
<tr><th colspan="2">Weather impacted flights in the test set</th></tr>
<tr><th>QR</th><th>GBM</th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><th>RQF</th><th>Stacking</th></tr>
<tr><td>
</td><td>
</td></tr>
</table>

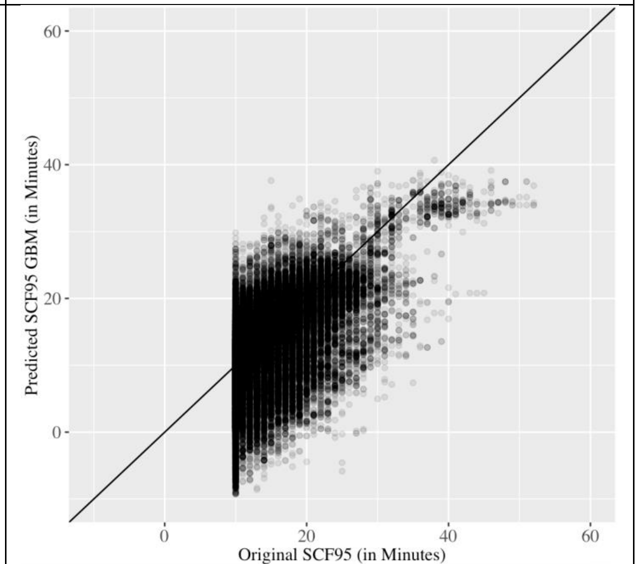
Table 6.5 SCF95 prediction results for non-weather impacted flights

Non-weather impacted flights in the test set

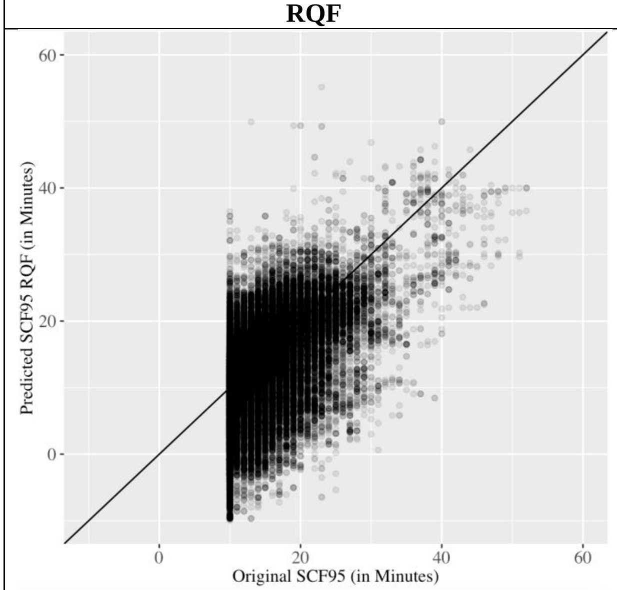
QR



GBM



RQF



Stacking

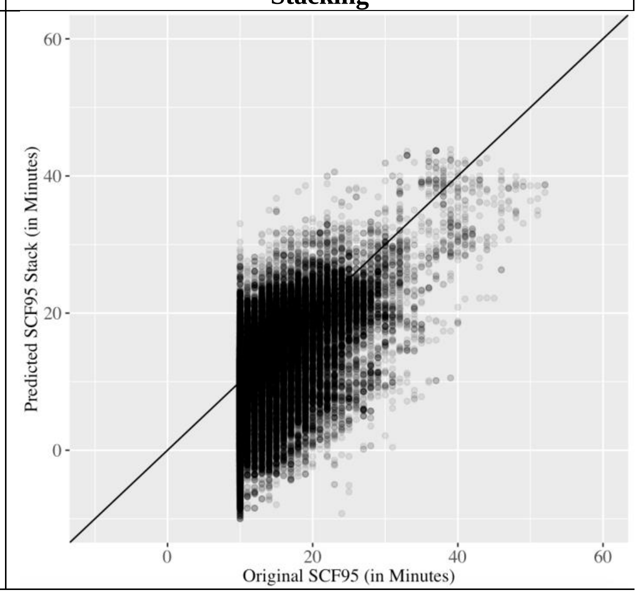


Table 6.6 SCF99 prediction results for weather impacted flights

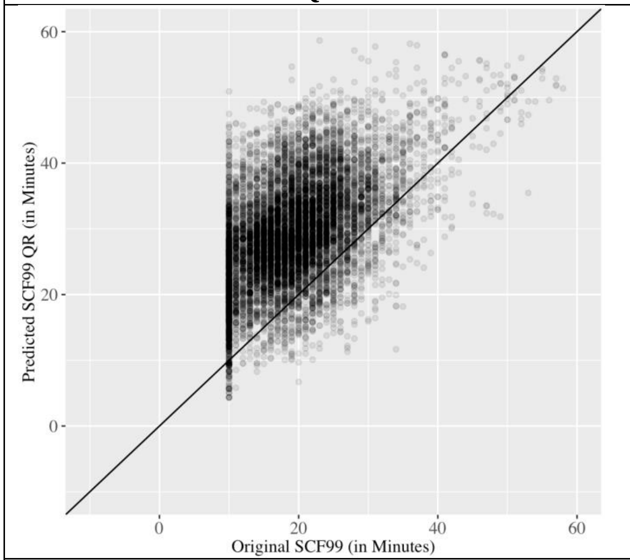
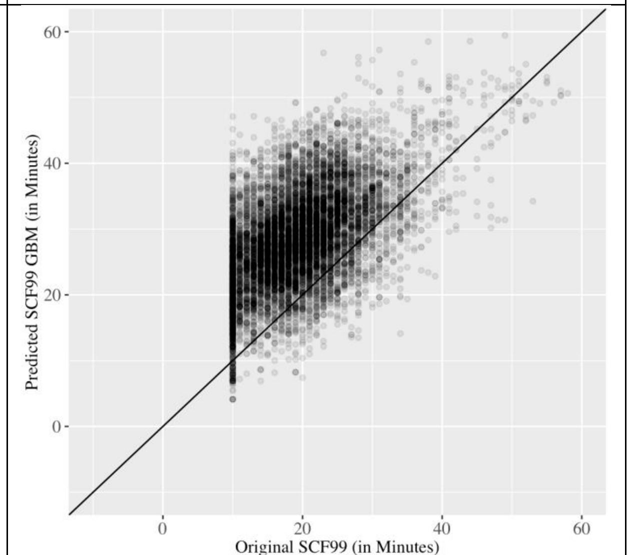
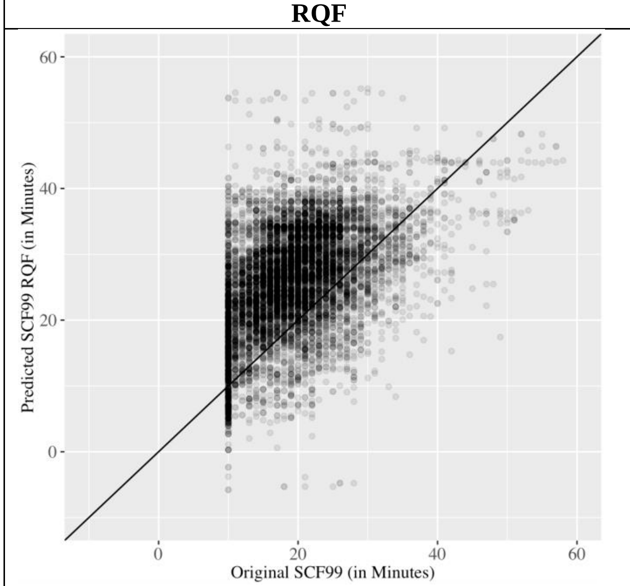
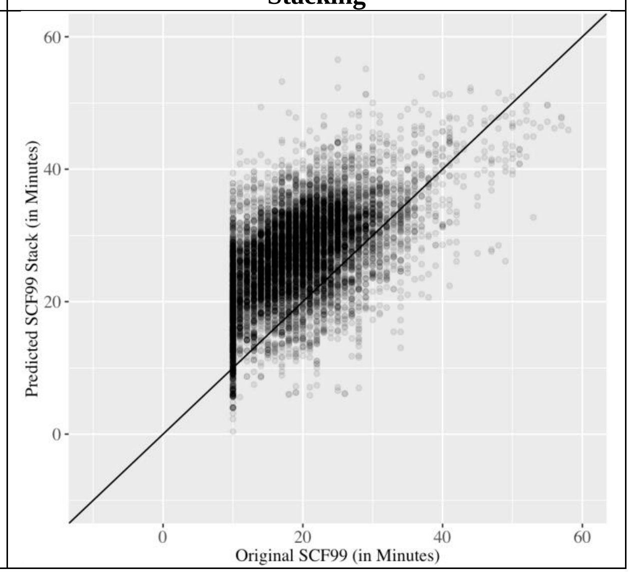
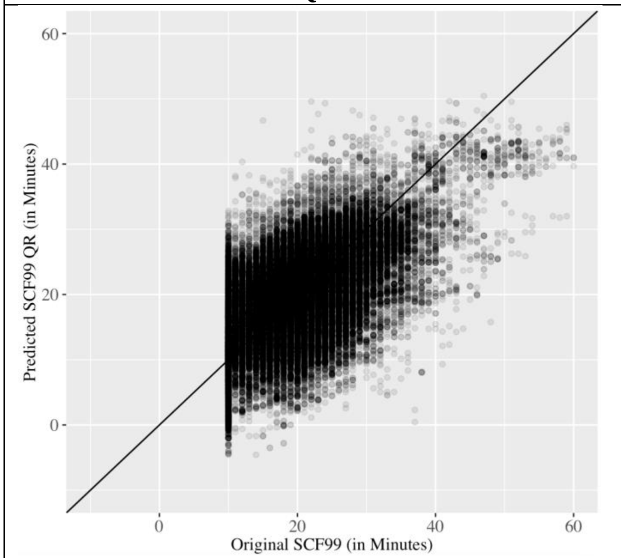
<table>
<tr><th colspan="2">Weather impacted flights in the test set</th></tr>
<tr><th>QR</th><th>GBM</th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><th>RQF</th><th>Stacking</th></tr>
<tr><td>
</td><td>
</td></tr>
</table>

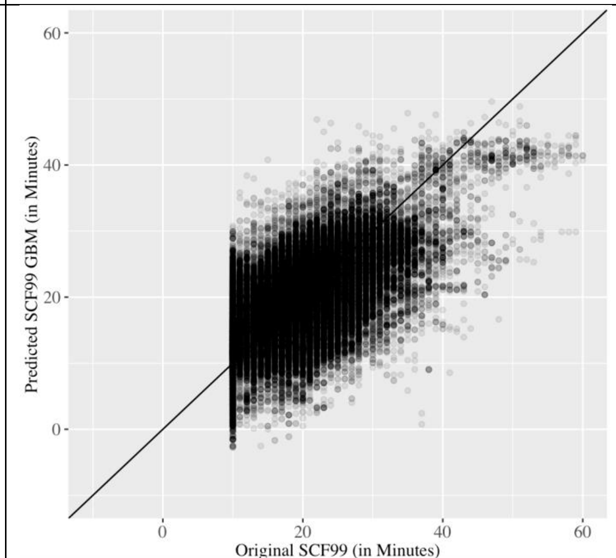
Table 6.7 SCF99 prediction results for non-weather impacted flights

Non-weather impacted flights in the test set

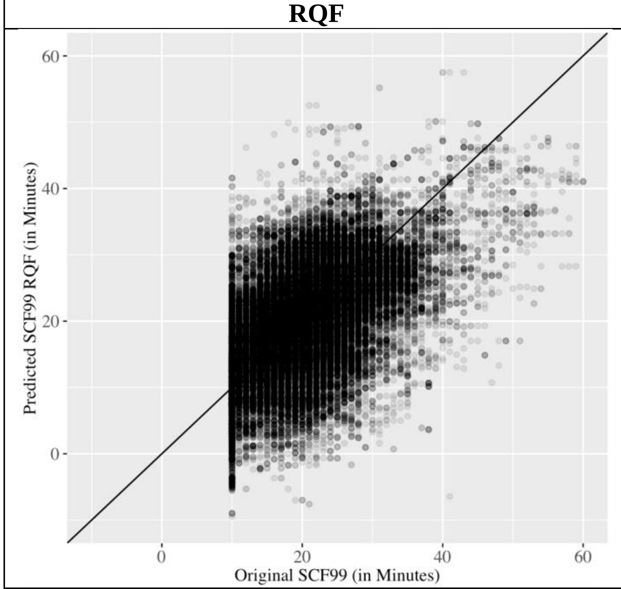
QR



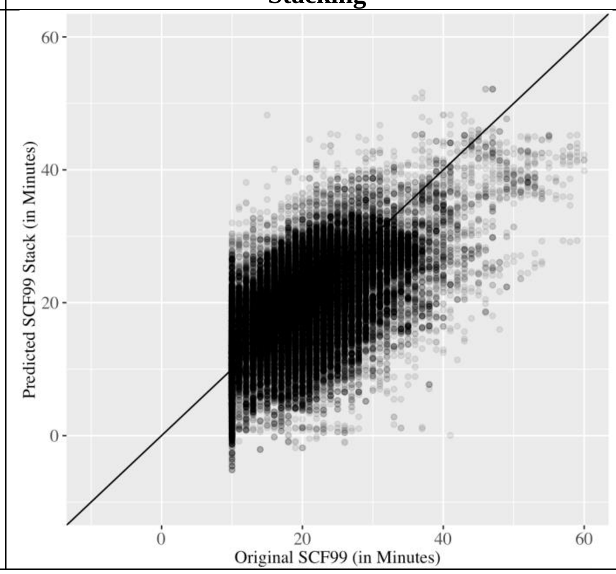
GBM



RQF



Stacking



6.4 Safety Check and Benefit Assessment

In this section, we consider the possible fuel savings associated with implementing a new SCF-estimation procedure. As suggested in Figures 6.1 and 6.2, dispatchers in general load more discretionary fuel than the SCF recommendation. Here we assume that because the SCF values obtained from our models are more believable (for example, by being higher when there is adverse weather), dispatchers will follow them. Thus, the difference between discretionary fuel loading and new SCF value defines our opportunity for fuel saving. If we reduce discretionary fuel loading, then the CTC discretionary fuel would also decrease. By assuming dispatchers follow new SCF recommendations perfectly in loading discretionary fuel, we can compute fuel savings in terms of CTC discretionary fuel reduction using the same CTC analysis mechanism described in Section 5.5.

As shown in Tables 6.10 and 6.11, for weather-impacted flights in the test set, the per-flight fuel saving is 243 lbs for SCF95 and 194 lbs for SCF99, using a lasso-based stacking model results as discretionary fuel. If we extrapolate back to the original dataset with operations over 14 months of operation for our study airline, we can further estimate an airline-wide total saving. By using \$3/gallon as jet fuel price, we can estimate airline-wide monetary savings by flight count. The airline-wide benefits are about \$75.6 million in fuel savings and 526 million kg in CO_2 emissions reduction using the new SCF95 for discretionary fuel (or \$64.8 million and 451 million kg per year) and \$64.1 million fuel saving and 446 million kg CO_2 emissions reduction using new SCF99 as discretionary fuel (or \$54.9 million and 382 million kg per year). Consistent with the findings in Section 5.5, using a lower percentile (SCF95 v.s. SCF99) can help to achieve higher fuel saving. Furthermore, weather-impacted flights are found to have higher fuel-savings opportunities. Dispatchers tend to load more discretionary fuel than necessary for weather-impacted flights to overcome weather and traffic uncertainties. However, in the proposed SCF estimation procedure, weather and traffic-related information has already been taken into account explicitly. Hence, using improved SCF as discretionary fuel for weather-impacted flights would generate more fuel savings compared to dispatchers' overloaded discretionary fuel.

Safety is a dispatcher's major consideration in discretionary fuel loading. As shown in Table 6.8, if we load discretionary fuel exactly as the proposed SCF values (either SCF95 or SCF99), we would still encounter a small proportion of flights using reserve fuel which might be undesirable to airlines. To better apply our proposed SCF estimation method in practice, we try to find a safe buffer on top of the proposed SCF to guarantee a similar safety margin for our study airline. The scaling factor idea, described in Section 5.5, is also applied here. We can learn scaling factors from the validation set and apply them to the test set. Again, we learn scaling factors for weather-impacted and non-weather-impacted flights separately. As shown in Table 6.8 and 6.9, by using the learned scaling factors, we can achieve the same safety performance as the current practice of our study airline. Moreover, we can still achieve an airline-wide benefit of \$19.6 million and \$22.2 million in fuel savings over 14 months, respectively, by using proposed SCF95 and SCF99 for discretionary fuel. These benefits are roughly \$16.8 million and \$19.0 million per year. Note that unlike fuel-saving estimates using unscaled SCF, we obtain higher fuel-savings using the scaled SCF99 model results. This is because in order to achieve a similar safety level, the learned scaling factor is 3.20 for SCF95 and 1.09 for SCF99 among non-weather impacted flights. The big scaling factor of SCF95 offsets the benefit, which results in slightly lower fuel savings comparing to scaled SCF99 model results.

Table 6.8 Safety check on test and validation sets using new SCF95

	Percentage of flights landing using reserve fuel					
	Original discretionary fuel	Use QR SCF95 as discretionary fuel	Use RQF SCF95 as discretionary fuel	Use GBM SCF95 as discretionary fuel	Use Stacking SCF95 as discretionary fuel	Use Stacking SCF95 with scaling as discretionary fuel
Validation set performance	0.02%	3.3%	3.0%	3.3%	3.2%	0.02%
Test set performance	0.02%	3.4%	3.2%	3.4%	3.2%	0.03%

Table 6.9 Safety check on test and validation sets using new SCF99

	Percentage of flights landing using reserve fuel					
	Original discretionary fuel	Use QR SCF99 as discretionary fuel	Use RQF SCF99 as discretionary fuel	Use GBM SCF99 as discretionary fuel	Use Stacking SCF99 as discretionary fuel	Use Stacking SCF99 with scaling as discretionary fuel
Validation set performance	0.02%	0.85%	0.87%	0.84%	0.81%	0.03%
Test set performance	0.02%	0.85%	0.83%	0.85%	0.78%	0.03%

Table 6.10 Cost-to-carry fuel saving using new SCF95

	Use stack SCF95 as discretionary fuel			Use stack SCF95 with scaling factor as discretionary fuel		
Test set performance	Mean CTC fuel savings (in lbs) per flight	Airline-wide monetary savings ($) (assuming $3/gallon)	Airline-wide CO_2 reduction (kg)	Mean CTC fuel savings (in lbs) per flight	Airline-wide monetary savings ($) (assuming $3/gallon)	Airline-wide CO_2 reduction (kg)
Weather impacted flights(10589 flights)	243	7.56×10^7	5.26×10^8	233	1.96×10^7	1.36×10^8
Non-weather impacted flights (63133 flights)	230			31		

Table 6.11 Cost-to-carry fuel saving using new SCF99

	Use stack SCF99 as discretionary fuel			Use stack SCF99 with scaling factor as discretionary fuel		
Test set performance	Mean CTC fuel savings (in lbs) per flight	Airline-wide monetary savings ($) (assuming $3/gallon)	Airline-wide CO_2 reduction (kg)	Mean CTC fuel savings (in lbs) per flight	Airline-wide monetary savings ($) (assuming $3/gallon)	Airline-wide CO_2 reduction (kg)
Weather impacted flights(10589 flights)	194	6.41×10^7	4.46×10^8	194	2.22×10^7	1.54×10^8
Non-weather impacted flights (63133 flights)	197			47		

6.5 Summary

In this chapter, we have shown the possibility of reducing fuel consumption through an improved SCF estimation procedure. A quantile regression-based SCF estimation procedure has been proposed. Three estimation models including parametric quantile regression, gradient boosting machine, and random quantile forests have been found to substantially outperform the airline's FPS in SCF estimation. Ensemble learning is also found to outperform three proposed models by combining them together. The new SCF-estimation procedure overcomes the limitations of the widely used SCF-estimation method, which relies on simplified grouping criterion and normal approximation. The proposed method can also incorporate terminal weather forecast and historical traffic conditions into the SCF estimation.

With the help of CTC factors, we are also able to calculate the extra fuel burn to carry the difference between actual contingency fuel and new SCF value based on model prediction. The extra fuel burn then can be translated into monetary costs and CO_2 emission. The estimated benefit pool for our study airline is in the magnitude of \$75.6 million in fuel savings and 526 million kg of CO_2 emissions reduction over 14 months of operation (or roughly \$64.8 million and 451 million kilogram per year). We further investigate the impact of adding a practical safety buffer (multiplying by a scaling factor), which can help to achieve a similar safety level, as measured by the fraction of flights landing without their full 45 min fuel reserve, as is the current practice for our study airline. Even after applying scaling factors, the estimated benefits are still significant: \$22.2 million in fuel savings and 154 million kg of CO_2 emissions reduction using the SCF99 model (or roughly \$19 million and 132 million kilogram per year). In addition, this study also builds a link between SCF95 estimation and aviation system predictability enabling the proposed models can also be used to predict benefits from reduced fuel-loading enabled by improved ATM.

Based on the fuel-burn data obtained from a major U.S.-based airline, a significant benefit has been estimated by improving SCF estimation, assuming that dispatchers could be persuaded to follow the improved guidance. It is also likely that this benefit would scale if applied throughout the airline industry, both in the U.S. and abroad. Moreover, given the link between system predictability and SCF estimation, a system-wide fuel savings benefit due to improved ATM could also be assessed.

7 Conclusions

In this dissertation, we have investigated dispatcher discretionary fuel loading behavior and proposed two discretionary fuel estimation approaches to help dispatchers make better fuel loading decisions. The contributions of this dissertation are the following:

1. We estimate fuel-burn cost due to carrying unnecessary fuel at an airline level. This broad view provides us a fuel-saving benefit benchmark and draws attention to a simple and well-tested fuel-saving approach: reducing unnecessary fuel loading. In Chapter 3, conservative fuel loading behavior is observed in six major U.S. airlines. This identifies opportunities in achieving significant fuel burn reduction by reducing unnecessary fuel loading in a system level. The benefit in 2012 (after removing the effect of reserve fuel) is estimated to range from \$42 million to \$605 million for six airlines with a total of \$1.16 billion. While it is not possible to attain all of these savings without causing unacceptable risk of fuel emergencies, they at least raise the possibility that an improved fuel loading practice could result in substantial savings

without compromising safety. The results also motivate us to further explore dispatcher fuel loading behavior.

2. Dispatchers' fuel loading behavior is modeled as decision under risk. Starting in Chapter 4, we begin to focus on fuel loading decisions from a major U.S.-based airline. With the help of this detailed flight level fuel burn and fuel loading data, we are able to estimate how dispatchers trade off safety concerns due to insufficient fuel loading and fuel costs due to carrying excess fuel. Using the newsvendor model formulation, we assume dispatchers make discretionary fuel loading decisions by minimizing the expected cost of under-loading (insufficient fuel loading) and over-loading (carry unnecessary fuel). The ratio between perceived safety costs and CTC excess fuel provides us with a quantified measure of the trade-offs between safety and fuel cost. Across 96 dispatchers, on average, dispatchers are found to value reserve fuel 1,200 times more than other fuel. This underscores the focus on safety in dispatcher culture. By integrating a dispatcher survey in modeling dispatchers' safety-cost trade-offs, we find that dispatchers who are detail oriented and conservationists are likely to load less discretionary fuel. These findings provide two important implications for airlines. One is about screening for dispatchers. When airlines interview dispatchers, in addition to skill- and behavior-based performance evaluations, it would be helpful to also test the detail-orientation of dispatchers, as well as their belief in conservation. This might help airlines to select dispatchers who are less likely to overload fuel resulting in potential fuel savings. Another implication is to target dispatchers for recurrent training. Add training topics on detail-orientation and conservation may also encourage dispatchers to load less unnecessary fuel.

3. Two novel discretionary fuel estimation approaches are proposed to assist dispatchers with better discretionary fuel loading decisions. These two approaches are shown to generate reliable discretionary fuel recommendations for dispatchers. In Chapter 5, we propose to use a PI-based discretionary fuel estimation approach in which the idea is to improve flight fuel burn prediction using machine learning techniques and then use PI to capture the prediction uncertainty. Using upper 95% PI as discretionary fuel loading is found to achieve the highest benefit with \$61.5 million in savings and 428 million kilogram of CO_2 in emissions reduction per year. Another property of using this approach is that we can also maintain the same safety performance as in current practice.

In Chapter 6, we explore another discretionary fuel estimation approach which is based on improving SCF estimation. Airlines can still use their FPS for fuel burn prediction without any dramatic change to the current FPS. The only change required is in the SCF estimation module. The proposed quantile regression-based SCF estimation approach is found to outperform current FPS. The estimated annual benefit of using the new SCF99 as discretionary fuel is found to be \$19 million in fuel savings and 132 million kilogram of CO_2 in emissions reduction. Although this approach provides less benefit than the PI-based approach, it still yields significant savings, given we are adhering to the airline (unimproved) FPS. A natural extension of this analysis is to combine improved fuel burn prediction models with quantile regression based on the SCF estimation procedure. This addition may lead to more accurate discretionary fuel estimation results and even greater fuel savings.

Reducing fuel consumption is a unifying goal across the aviation industry. To this end, a simple, well-tested, but commonly overlooked approach is to reduce unnecessary discretionary fuel loading. With small yet powerful changes in dispatchers' discretionary fuel loading behavior, such as influencing their attitudes and providing them with more accurate information, we have shown that significant benefits cane be seized in terms of monetary savings and CO_2 reduction for our study airline. A similar analysis could be easily generalized to other airlines when such detailed airline fuel data becomes available.

Bibliography

Abdellaoui M, Baillon A., Placido L.,, Wakker P., 2011. The Rich Domain of Uncertainty: Source Functions and Their Experimental Implementation. American Economic Review 101, 695-723.

Abeyratne R., 2009. Air Carrier Responses to the Fuel Price Crisis: Some Issues. Global Trade and Customs Journal 4, 31-43.

Ahn D., Choi S., Gale D., Kariv S., 2014. Estimating Ambiguity Aversion in a Portfolio Choice Experiment. Quantitative Economics 5, 195-223.

AirSafe.com, 2015. Air Canada plane crashes since 1970. Retrieved from http://airsafe.com/events/airlines/aircan.htm

Altus S., 2009. Effective Flight Plans Can Help Airlines Economize. AERO Online Special Issue. http://www.boeing.com/commercial/aeromagazine/articles/qtr_03_09/article_08_1.html

Anderson D., 2005. Determination of Dispatch Takeoff Weight. Available at: http://www.smartcockpit.com/download.php?path=docs/&file=Determination_Of_Dispatch_Takeoff_Weight.pdf

Anvari M., 1987. Optimality Criteria and Risk in Inventory Models: the Case of the Newsboy Problem. Journal of the Operational Research Society 38, 625-632.

Artzner P., Delbaen F., Eber J., Heath D., 1999. Coherent Measures of Risk. Mathematical Finance 9, 203-228.

Arrow K., Harris T., Marshak J., 1951. Optimal Inventory Policy, Econometrica 19, 250-272.

Avineri E., and Prashker J., 2005. Sensitivity to travel time variability: Travelers' learning perspective. Transportation Research Part C 13, 157-183.

Barberis N., 2013. Thirty Years of Prospect Theory in Economics: A Review and Assessment. Journal of Economic Perspectives 27, 173-196.

Ben-hur A., Elisseeff A., Guyon I., 2002. A stability based method for discovering structure in clustered data. Pacific Symposium on Biocomputing 7:6-17.

Bouakiz M., and Sobel M., 1992. Inventory Control with an Exponential Utility Criterion. Operations Research 40, 603-608.

Breiman L., 1996a. Bagging Predictors. Machine Learning 24, 123-140.

Breiman L., 1996b. Stacked Regressions. Machine Learning, 24, 49-64.

Breiman L., 2001. Random Forests. Machine Learning 45, 5-32.

Bureau of Transportation Statistics. Airline Financial Data, 2016. Available at: https://www.rita.dot.gov/bts/press_releases/bts061_16 [Accessed 25 January 2017]

Camerer C., and Weber M., 1992. Recent Developments in Modeling Preferences: Uncertainty and Ambiguity. Journal of Risk and Uncertainty 5, 325-370.

Cawley G., Talbot N., Foxall R., Dorling S., Mandic D., 2004. Heteroscedastic Kernel Ridge Regression. Neurocomputing 57, 105-124.

Chati Y., and Balakrishnan H., 2016. Statistical Modeling of Aircraft Fuel Flow Rate. International Council of Aeronautical Sciences (ICAS) Congress.

Chen X., Sim M., Simchi-Levi D., Sun P., 2007. Risk Aversion in Inventory Management. Operations Research 55, 828-842.

Choi S., 2009. Risk-averse Newsvendor Models. (Doctoral Dissertation). Rutgers University. https://rucore.libraries.rutgers.edu/rutgers-lib/26627/

Choi S., and Ruszczynski A., 2011. A Multi-product Risk-averse Newsvendor with Exponential Utility Function. European Journal of Operational Research 214, 78-84.

Chryssoloiuris G., Lee M., and Ramsey A., 1996. Confidence Interval Prediction for Neural Network Models. IEEE Transactions of Neural Networks 7, 229–232.

Clarke J., Ho N., Ren L., Brown J., Elmer K., Tong K., Wat J., 2004. Continuous Descent Approach: Design and Flight Test for Louisville International Airport. Journal of Aircraft 41, 1054-1066.

Clarke B., and Yu B., 2003. Comparing Bayes Model Averaging and Stacking When Model Approximation Error Cannot be Ignored. Journal of Machine Learning Research 4, 683-712.

Dancila B., Botez R., Labour D., 2013. Fuel Burn Prediction Algorithm for Cruise, Constant Speed and Level Flight Segments. The Aeronautical Journal 117, 491-504.

e-CFR: Title 14: Aeronautics and Space Part 121, 2015. Operating Requirements: Domestic, Flag, and Supplemental Operations. Retrieved from http://www.ecfr.gov/cgi-bin/text-idx?SID=a8d3c4800d167b64bbfa2349ec337755&mc=true&node=pt14.3.121&rgn=div5

Environmental Protection Agency, 2013. Emission Factors for Greenhouse Gas Inventories, US Environmental Protection Agency. http://www.epa.gov/climateleadership/documents/emission-factors.pdf. Last accessed: January 2015.

Environmental Protection Agency, 2016. Fast Facts: U.S. Transportation Sector Greenhouse Gas Emissions 1990-2014. Available at: https://nepis.epa.gov/Exe/ZyPDF.cgi?Dockey=P100ONBL.pdf

European Commission, 2010. Report on the SES Legislation Implementation. Available at: http://ec.europa.eu/transport/modes/air/single_european_sky/doc/2011_10_03_ses_legislation_report_2010.pdf

European Commission, 2015. Fuel and Air Transport. Available at: http://ec.europa.eu/transport/modes/air/doc/fuel_report_final.pdf

Federal Aviation Administration, 2005. Aircraft Weight and Balance Control. Federal Aviation Administration, U.S. Department of Transportation, Washington, D.C. Available at: http://www.faa.gov/documentLibrary/media/Advisory_Circular/AC120-27E.pdf

Federal Aviation Administration, 2007. Aircraft Weight and Balance Handbook. Federal Aviation Administration, U.S. Department of Transportation, Washington, D.C. Available at: https://www.faa.gov/regulations_policies/handbooks_manuals/aircraft/media/FAA-H-8083-1A.pdf

Federal Aviation Administration, 2014. Next Gen: the Business Case for the Next Generation Air Transportation System. Available at: https://www.faa.gov/nextgen/media/BusinessCaseForNextGen-2014.pdf

Freund Y., and Schapire R.E., 1997. A Decision-Theoretic Generalization of on-line Learning and an Application to Boosting. Journal of Computer and System Sciences 55, 119-139.

Friedman J., 2001. Greedy Function Approximation: A Gradient Boosting Machine, Annals of Statistics 29(5), 1189-1232.

Friedman J., and Popescu B., 2003. Importance sampled learning ensembles, Technical report. Stanford University, Department of Statistics.

Gotoh J., and Takano Y., 2007. Newsvendor Solutions via Conditional Value-at-risk Minimization. European Journal of Operational Research 179, 80-96.
Guerreiro Fregnani, J. A. T., Müller, C., Correia, A. R., 2013. A Fuel Tankering Model Applied to a Domestic Airline Network. Journal of Advanced Transportation, 47(4), 386–398.
Hao L., Hansen M., Lovell D., Vlachou K., Ball M., 2014. Flight Predictability: Concepts, Metrics and Impacts. NEXTOR Final Report, Available at: http://www.nextor.org/pubs/NEXTOR-II-Predictability-2014.pdf
Hao L., Ryerson M., Kang L., Hansen M., 2016a. Estimating Fuel Burn Impacts of Taxi-out Delay With Implications for Gate-hold Benefits. Transportation Research Part C: Emerging Technologies, DOI: doi:10.1016/j.trc.2016.05.015
Hao L., Hansen M., Ryerson M., 2016b. Fueling for Contingencies: the Hidden Cost of Unpredictability in the Air Transportation System. Transportation Research Part D: Transport and Environment 44, 199-210.
Hastie T., Tibshirani R., Friedman J., 2009. The Elements of Statistical Learning. Second Edition. Springer Science+Business Media LLC.
Hileman J., Wong H., Ortiz D., Brown N., Maurice L., Rumizen M., 2008. The Feasibility and Potential Environmental Benefits of Alternative Fuels for Commercial Aivation. Proceedings of the 26th Congress of the Aeronautical Sciences, Anchorage Alaska.
Horiguchi Y., Baba Y., Kashima H., Suzuki M., Kayahara H., Maeno J., 2017. Predicting Fuel Consumption and Flight Delays for Low-Cost Airlines. Proceedings of the Twenty-Ninth AAAI Conference on Innovative Applications (IAAI-17)
International Air Transport Association, 2004. Environmental Review. Available at: http://web.docuticker.com/go/docubase/5452
International Civil Aviation Organization, 2012. A Planning Framework for Seamless ATM. Available at: http://www.icao.int/APAC/Meetings/2012_APSAPG1/WP15%20IATA_A%20Planning%20Framework%20for%20Seamless%20ATM.pdf
International Civil Aviation Organization, 2014. 2013 Environment Report. Available at: http://cfapp.icao.int/Environmental-Report-2013/files/assets/common/downloads/ICAO_2013_Environmental_Report.pdf
Intergovernmental Panel on Climate Change, 2014. Climate Change 2014: Mitigation of Climate Change, Cambridge University Press.
Irrgang M., 2011. A Look at the State of Airline Fuel Conservation. Available at: http://www.agifors.org/document.go?documentId=2131&action=download
James G., Witten D., Hastie T., Tibshirani R., 2013. An Introduction to Statistical Learning with Applications in R, Springer, Available at: http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf
van de Kaa E., 2010. Prospect Theory and Choice Behaviour Strategies: Review and Synthesis of Concepts from Social and Transport sciences. EJTIR Special Issue 10, 299-329.
Kahneman, D., and Tversky, A., 1979. Prospect Theory: An Analysis of Decision under Risk. Econometrica 47, 263-291.
Kaiser, H. F., 1960. The Application of Electronic Computers to Factor Analysis. Educational and Psychological Measurement 20, 141-151.
Karisch S.E., Altus S.S., Stojković G., and Stojković M., 2012. Operations Quantitative Problem Solving Methods in the Airline Industry, A Modeling Methodology Handbook Series

(International Series in Operations Research & Management Science), New York, Springer.
Katariya A., Cetinkaya S., Tekin E., 2014. On the Comparison of Risk-neutral and Risk-averse Newsvendor Problems. Journal of the Operational Research Society 65, 1090-1107.
Koenker R., and Machado J., 1999. Goodness of Fit and Related Inference Processes for Quantile Regression 94, 1296-1310.
Koenker R., 2005. Quantile Regression. Cambridge University Press. ISBN 0-521-60827-9.
Kolev N., and Paiva D., 2009. Copula-based Regression Models: A Survey. Journal of Statistical Planning and Inference 139, 3847-3856.
Lau H., 1980. The Newsboy Problem under Alternative Optimization Objectives. Journal of the Operational Research Society 31, 525-535.
Lawson C., Hanson R.J., 1987. Solving Least Squares Problems, SIAM
Leblanc M. and Tibshirani R., 1996. Combining Estimates in Regression and Classification. Journal of the American Statistical Association 91, 1641-1650.
LeDell E.E., 2015. Scalable Ensemble Learning and Computationally Efficient Variance Estimation. (Doctoral Dissertation). University of California, Berkeley, USA. http://www.stat.berkeley.edu/~ledell/papers/ledell-phd-thesis.pdf
Levin J., 2006. Choice under Uncertainty. Available at: http://web.stanford.edu/~jdlevin/Econ%20202/Uncertainty.pdf
Li Z., and Hensher D., 2011. Prospect Theoretic Contributions in Understanding Traveller Behaviour: A Review and Some Comments. Transport Reviews 31, 97-115.
Liden S., 1992. Optimum Cruise Profiles in the Presence of Winds, Digital Avionics Systems Conference, Proceedings IEEE/AIAA 11th, available at: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=282147&isnumber=6983&tag=1
Liu Y., Hansen M., Gupta G., Malik W., Jung Y., 2014. Predictability Impacts of Airport Surface Automation. Transportation Research Part C 44, 128–145.
Lovegren J., and Hansman R., 2011. Quantification of Fuel Burn Reduction in Cruise Via Speed and Altitue Optimization Strategies. Report No. ICAT-2011-03. MIT International Center for Air Transportation. Available at: http://dspace.mit.edu/bitstream/handle/1721.1/62196/Lovegren_ICAT-2011.pdf?sequence=1
Maskey S., Guinot V., Price R., 2004. Treatment of Precipitation Uncertainty in Rainfall-runoff Modeling: a Fuzzy Set Approach. Advances in Water Resources 27, 889-898.
McFadden D., 1981. Chapter 5: Econometric models of probabilistic choices. In C. F. Manski, & D. McFadden (Eds.), Structural analysis of discrete data with economic applications (pp. 198–272). Chicago: University of Chicago Press.
McFadden, D., 2001. Economic choices. Nobel lecture, December 2000. American Economic Review, 91(3), 351–378.
Meinshausen N., 2006. Quantile Regression Forests, Journal of Machine Learning Research 7, 983–999.
von Neumann J. and Morgenstern O., 1947. Theory of Games and Economic Behavior, 2nd ed. Princeton: Princeton University Press.
Ogryczak W., and Ruszczynski A., 2002. Dual Stochastic Dominance and Related Mean-risk Models. SIAM Journal on Optimization 13, 60-78.

de Palma A., M. Ben-Akiva, D. Brownstone, C. Holt, T. Magnac, D. McFadden, P. Moffatt, N. Picard, K. Train, P. Wakker and J.L. Walker, 2008. Risk, Uncertainty and Discrete Choice Models, Marketing Letters 19, 269-285.

Pham V., Tang J, Alam S, Lokan C and Abbass H., 2010. Aviation Emission Inventory Development and Analysis. Environmental Modelling & Software 25, 1738–1753.

Qin Y., Wang R., Vakharia A., Chen Y., Seref M., 2011. The Newsvendor Problem: Review and Directions for Future Research, European Journal of Operational Research 213, 361-374.

Quiggin J., 1982. A Theory of Anticipated Utility. Journal of Economic Behavior and Organization 3, 323-343.

Raftery A.E., 1995. Bayesian Model Selection in Social Research. Sociological Methodology 25, 111-163.

Rockafellar T., and Uryasev S., 2002. Conditional Value-at-risk for General Loss Distributions. Journal of Banking and Finance 26, 1443-1471.

Ridgeway G., 2007. Generalized Boosted Models: A guide to the gbm package, Available at: http://www.saedsayad.com/docs/gbm2.pdf

Roberson B., 2010. Fuel Conservation Strategies. Available at: http://www.boeing.com/commercial/aeromagazine/articles/qtr_02_10/pdfs/AERO_FuelConsSeries.pdf

Ryerson M.S., Hansen M., Hao L., Seelhorst M., 2016. Landing on Empty: Estimating the Benefits from Reducing Fuel Uplift in U.S. Civil Aviation. Environmental Research Letters 10, 2015, 094002.

Ryerson, M.S., Hansen M., Bonn J., 2011. Fuel Consumption and Operational Performance. Ninth USA/Europe Air Traffic Management Research and Development Seminar, Berlin, Germany.

Sadraey M.H., 2012. Aircraft Design: A Systems Engineering Approach. John Wiley & Sons.

Savage L.,1954. The Foundations of Statistics. New York: Wiley.

Schiefer G., and Samuel, T., 2011. Managing Flight Operational Costs. SABRE. Retrieved from http://www.sabreairlinesolutions.com/pdfs/ManagingFlightOperationalCosts.pdf

Shapiro A., Dentcheva D., Ruszczynski A., 2009. Lectures on Stochastic Programming: Modeling and Theory. SIAM (Society for Industrial and Applied Mathematics), Philadelphia.

Shrestha D., and Solomatine D., 2006. Machine Learning Approaches for Estimation of Prediction Interval for the Model Output. Neural Networks 19, 225–235.

Trujillo A., 1996. Uncertainties that Flight Crews and Dispatchers Must Consider When Calculating the Fuel Needed for a Flight. NASA Technical Memorandum 110240. Retrieved from http://ntrs.nasa.gov/search.jsp?R=19960042496

Tversky, A., and Kahneman, D., 1992. Advances in Prospect Theory: Cumulative Representation of Uncertainty. Journal of Risk and Uncertainty, 5(4), 297–323.

Ullma J., 2006. Structural Equation Modeling: Reviewing the Basics and Moving Forward. Journal of Personality Assessment 87, 35-50.

van der Laan M., Polley E., Hubbard A., 2007. Super Learner. U.C. Berkeley Division of Biostatistics Working Paper Series, paper 222.

Walker M., 2010. NBAA Flight Planning Optimization and Fuel Uplift Requirements to Reduce Operational Costs. Available at: http://www.srs.aero/wordpress/wp-content/uploads/2010/11/NBAA-2010-Optimized-Planning.pdf

Wolpert, D., "Stacked Generalization," Neural Networks, Vol. 5, 1992, pp. 241-259.

Wonnacott, T. H., and Wonnacott, R. J., 1996. Introductory Statistics. New York: Wiley.
Yong A. and Pearce S., 2013. A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. Tutorials in Quantitative Methods for Psychology 9, 79-94.

Appendix

Table A.1 Cluster1 model performance

	5-fold MSE on training	MSE on validation	MSE on test
FPS	866,229.2	853,987.4	864,676.7
KNN (k=4)	996,344.3	944,664.4	925,123.4
Lasso (lambda=0.01)	504,899.4	494,201.0	495,690.8
Ridge (lambda=0.001)	504,899.9	494,230.9	495,787.7
MARs	459,735.1	448,515.3	450,573.7
Random Forests (depth=15)	446,960.3	432,587.4	432,988.4
Gradient Boosting (N=3500)	438,284.6	426,686.8	426,878.6
Stacking with NNLS	--	415,379.7	415,529.0
Stacking with Lasso Shrinkage (lambda = 10)	--	404,091.5	404,072.5

Table A.2 Cluster2 model performance

	5-fold MSE on training	MSE on validation	MSE on test
FPS	815,295.9	827,276.5	825,179.9
KNN (k=3)	2,034,886.4	1,943,307.9	1,832,740.2
Lasso (lambda=0.01)	469,923.5	490,394.4	481,547.7
Ridge (lambda=0.0001)	469,996.9	490,321.3	481,437.7
MARs	466281.9	489,369.4	481,593.1
Random Forests (depth=15)	473,368.7	485,726.1	480,755.6
Gradient Boosting (N=3500)	454,910.6	471,026.9	457,287.1
Stacking with NNLS	--	462,216.0	452,223.1
Stacking with Lasso Shrinkage (lambda = 1000)	--	459,627.3	450,808.1

Table A.3 Cluster3 model performance

	5-fold MSE on training	MSE on validation	MSE on test
FPS	1,238,647.9	1,239,857.2	1,282,738.9
KNN (k=4)	2,071,680.9	1,959,904.3	2,054,702.4
Lasso (lambda=0.1)	621,848.1	603,089.0	640,428.2
Ridge (lambda=0.0001)	621,813.8	603,103.9	640,591.8
MARs	617,125.2	599,522.6	639,353.0
Random Forests (depth=15)	615,483.1	583,157.6	620,425.0
Gradient Boosting (N=3000)	595,935.4	570,399.8	600,816.2
Stacking with NNLS	--	557,961.2	593,723.8
Stacking with Lasso Shrinkage (lambda = 100)	--	555,620.5	591,321.6

Table A.4 Full model performance

	5-fold MSE on training	MSE on validation	MSE on test
FPS	901,450.8	894,783.6	907,133.1
KNN (k=3)	1,431,021	1,341,977.8	1,339,982.2
Lasso (lambda=0.01)	531,421.6	524,703.2	529,007.0
Ridge (lambda=0.00001)	531,429.5	524,705.2	529,003.1
MARs	573,902.9	484,335.9	491,195.4
Random Forests (depth=15)	471,456.0	460,510.5	463,678.5
Gradient Boosting (N=4000)	465,757.6	457,898.3	460,252.0
Stacking with NNLS	--	446,726.9	449,572.3
Stacking with Lasso Shrinkage (lambda = 100)	--	438,485.6	441,442.7