Title
Finding Similar Days for Air Traffic Management

Permalink
https://escholarship.org/uc/item/0dn6d6h9

Author
Gorripaty, Sreeta

Publication Date
2017

Peer reviewed|Thesis/dissertation
Finding Similar Days for Air Traffic Management

By

Sreeta Gorripaty

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 Alexey Pozdnukhov
Professor Laurent El Ghaoui

Spring 2017
Abstract

Finding Similar Days for Air Traffic Management

by

Sreeta Gorripaty

Doctor of Philosophy in Civil and Environmental Engineering

University of California, Berkeley

Professor Mark Hansen, Chair

Airports are central to stimulating the growth and development of economies around the world. With the trend of increasing demand for air travel and transportation of goods, there is mounting pressure on the existing airport infrastructure. This leads to an imbalance in the capacity and demand at airports that costs airlines, passengers and the economy at large billions of dollars. A cost-effective solution to this growing imbalance is to develop ways to utilize the existing infrastructure well. Information from historical days that are similar to a day of operations can be used to gain insights to support traffic management decisions on that day. Recent advances in machine learning and computing power have made it possible to mine and analyze sizable historical archives of different variables that characterize and influence airport operations. Finding similar historical days can help better understand the impact of different traffic management initiatives (TMIs) and identify areas of capacity underutilization. Reduction in airport capacity underutilization can lead to reduction in airport delays.

Decision support tools that can identify similar days and the TMIs taken on these past similar days and their resulting outcomes can augment controller experience to guide decision-making on the reference day at an airport. This information can allow air traffic managers to make less conservative decisions and thus improve airport capacity and reduce delays. This dissertation develops similarity measures between days using airport capacity and demand data. We find that dimensionality reduction is feasible for capacity data, and base capacity similarity on the principal components. Dimensionality reduction cannot be efficiently performed on demand data; consequently demand similarity is based on original data in this case. We find that both capacity and demand data lack natural clusters and thus propose that similarity be viewed as a continuous measure. Finally we estimate measures of overall distance based on both capacity and demand similarity. The estimated distances are visualized using Metric Multidimensional Scaling plots and indicate that most days with significant air traffic management activity are similar to certain other days, validating the potential of this approach for decision support.
Accurate demand and capacity estimates are necessary to generate meaningful similarity measures that can be used in decision-support tools. Predicting airport capacity accurately can also help make better tradeoffs between allowing more flights to operate at the airport and minimizing expensive airborne delay. We develop accurate demand estimates from the Aggregate Demand List (ADL), which contains fine-grained flight schedule data of all the flights operating at an airport. Capacity of an airport can be observed only at sufficiently large demand. However, if the throughput of an airport is limited by the demand, we can only conclude that the capacity is larger than or equal to the observed throughput. The inability to directly observe capacity makes capacity prediction a challenging and less explored problem. This dissertation applies machine-learning methods that incorporate observations censored by insufficient demand to develop an airport capacity prediction model. Specifically, we explore Kaplan Meier estimator, Cox Proportional Hazards model and Random Survival Forest model to predict airport capacity. These models predict a capacity distribution rather than a single capacity value for an hour of interest at an airport using its weather, fleet mix and scheduled demand data. The model results also indicate the influence of different variables on the capacity of the airport. Model performance is compared using several validation measures, including Integrated Brier Score (IBS), Concordance Index (C Index), $R^2$ and RMSE of predicted throughput, that account for the presence of censored observations. The RSF model consistently outperforms the KM estimator and Cox model across all the validation measures.

This dissertation also develops capacity based similarity metrics between two days using the predicted hourly capacity distributions. The evaluation of the estimated similarity metric is challenging owing to the lack of ground truth similarity measures. In this dissertation, we propose a framework to validate the estimated similarity metric between two days using predicted capacity CDFs, demand, TMI and operational outcomes data. The assumption for this framework is that days that are similar based on their capacity, demand and TMI features should be similar based on their operational outcomes. We use a Random Forest model to combine the capacity, demand and TMI based similarity metrics, supervised by operational outcomes similarity metric. This combined similarity matrix is evaluated by measuring its correlation with the operational outcomes similarity matrix on test data. The methodology developed in this dissertation to identify similar days can be extended to any airport around the world using their respective weather, demand, TMI and operational outcome data. This framework uses data that can be forecasted and thus can be used to guide decisions on a day-of-operations in order to guide decision making, as well as a post-operations setting to compare decisions and outcomes on similar historical days.
To Amma, Nanna, Annaya and Nikku
# Table of contents

Abstract ........................................................................................................................................... 1
Table of contents ............................................................................................................................. ii
List of figures .................................................................................................................................... iv
List of tables ..................................................................................................................................... vi
Acknowledgments ............................................................................................................................ vii

1. Introduction ................................................................................................................................. 1
   1.1. Background and Motivation .................................................................................................... 1
       1.1.1. Airport Capacity ............................................................................................................. 2
       1.1.2. Air Traffic Management ............................................................................................... 3
       1.1.3. Similar Days .................................................................................................................. 4
   1.2. Research Objective ................................................................................................................ 7
   1.3. Dissertation Overview .......................................................................................................... 7
   1.4. Contributions ....................................................................................................................... 9

2. Similar Days .................................................................................................................................. 10
   2.1. Introduction .......................................................................................................................... 10
   2.2. Data ..................................................................................................................................... 11
   2.3. Correlation Analysis .......................................................................................................... 12
   2.4. Principal Component Analysis: Capacity Data .................................................................... 15
   2.5. Clustering analysis: capacity data ...................................................................................... 19
   2.6. Principal Component Analysis: demand data ..................................................................... 22
   2.7. Clustering analysis: demand data ....................................................................................... 23
   2.8. Continuous Similarity .......................................................................................................... 25
       2.8.1. Maximum Distance Similarity .................................................................................... 25
       2.8.2. Euclidean Distance Similarity ..................................................................................... 27
       2.8.3. Air Traffic Management Case Study .......................................................................... 28
   2.9. Conclusion ............................................................................................................................ 30

3. Capacity Prediction ...................................................................................................................... 33
   3.1. Introduction .......................................................................................................................... 33
   3.2. Censored Data Analysis ....................................................................................................... 35
       3.2.1. Kaplan–Meier estimator .............................................................................................. 39
       3.2.2. Cox Proportional Hazards Model ................................................................................. 39
       3.2.3. Random Survival Forest Model .................................................................................. 40
   3.3. Model Evaluation ................................................................................................................ 42
   3.4. Data Preparation .................................................................................................................. 45
3.5. Results – SAER event status ................................................................................. 49
  3.5.1. Kaplan–Meier estimator .................................................................................... 49
  3.5.2. Cox Proportional Hazards Model ..................................................................... 51
  3.5.3. Random Survival Forest .................................................................................. 53
3.6. Capacity Prediction Case Study ........................................................................... 60
3.7. Results – TAER event status ................................................................................ 62
  3.7.1. Kaplan–Meier estimator .................................................................................... 62
  3.7.2. Cox Proportional Hazards Model ..................................................................... 63
  3.7.3. Random Survival Forest .................................................................................. 65
3.8. Conclusions ........................................................................................................... 68
4. Validation of Similar Days ......................................................................................... 71
  4.1. Introduction ........................................................................................................... 71
  4.2. Data .................................................................................................................... 73
  4.3. Capacity Similarity ............................................................................................... 76
  4.4. Demand Similarity .............................................................................................. 81
  4.5. Operational Outcomes Similarity ....................................................................... 83
  4.6. Validation ............................................................................................................. 84
  4.7. Conclusions ......................................................................................................... 91
5. Conclusions .............................................................................................................. 92
  5.1. Summary .............................................................................................................. 92
  5.2. Contributions ....................................................................................................... 94
  5.3. Future Work ......................................................................................................... 95
  5.4. Concluding Remarks ........................................................................................... 95
Bibliography .................................................................................................................. 97
List of figures

Figure 1.1: Components that determine airport capacity .................................................. 2
Figure 1.2: Air Traffic Control (ATC) in different stages of flight ...................................... 3
Figure 1.3: Similar days architecture .................................................................................. 6
Figure 2.1: Correlation plot for capacity data at EWR airport ........................................... 14
Figure 2.2: Correlation plot for demand data at EWR airport ........................................... 14
Figure 2.3: EWR: Cumulative variance explained vs number of PCs; Loadings of the first 6 PCs on 7-22 hours .................................................................................. 16
Figure 2.4: Cumulative variance explained vs number of PCs; Loadings of the effective number of PCs for SFO, ORD and JFK respectively ........................................ 18
Figure 2.5: PC1 vs PC2 for capacity data of EWR, SFO, ORD and JFK respectively .... 21
Figure 2.6: ASW and WSS for EWR capacity data ............................................................ 21
Figure 2.7: Cumulative variance vs number of PCs for EWR, SFO, ORD & JFK airports ................................................................................................................................. 22
Figure 2.8: PC1 vs PC2 for demand data of EWR, SFO, ORD and JFK respectively .... 24
Figure 2.9: MDS plot for maximum distance for EWR, SFO, ORD and JFK respectively ................................................................................................................................. 27
Figure 2.10: MDS plot for Euclidean distance for EWR, SFO, ORD & JFK respectively ................................................................................................................................. 28
Figure 2.11: Capacity and Demand Profiles for Reference Day and Two Similar Days ................................................................................................................................. 29
Figure 3.1: CDF of arrival capacity, EWR airport ................................................................. 49
Figure 3.2: Comparing capacity CDF between ceiling-filtered data, EWR airport ......... 50
Figure 3.3: KM-predicted vs observed throughput .............................................................. 51
Figure 3.4: Cox-predicted vs observed throughput ............................................................. 53
Figure 3.5: IBS vs tree depth for RSF model ...................................................................... 54
Figure 3.6: Proportion of data above truncation filter ......................................................... 55
Figure 3.7: Example extrapolation of incomplete predicted CDF ....................................... 56
Figure 3.8: C Index computed for truncation-filtered data ............................................... 57
Figure 3.9: C Index vs tree depth for RSF model ............................................................... 57
Figure 3.10: Throughput prediction R² vs tree depth .......................................................... 58
Figure 3.11: RSF-predicted vs observed throughput .......................................................... 59
Figure 3.12: RSF variable importance using minimal depth ............................................... 59
Figure 3.13: OOB arrival capacity CDF for 7-22 hours, 08/13/2014 at EWR airport ................................................................................................................................. 60
Figure 3.14: Weather over EWR airport on 08/13/2014 from 10-22 hours .......... 61
Figure 3.15: Comparing SAER and TAER based arrival capacity CDFs, EWR airport 63
Figure 3.16: KM-predicted vs observed throughput (using TAER data) ................. 63
Figure 3.17: Cox-predicted vs observed throughput (using TAER data) .......... 65
Figure 3.18: IBS vs tree depth for RSF model (using TAER data) ........................................ 66
Figure 3.19: C Index vs tree depth for RSF model (using TAER data) ................................. 66
Figure 3.20: R²-predicted throughput vs tree depth (using TAER data) ............................... 67
Figure 3.21: RSF-predicted vs observed throughput (using TAER data) ............................... 67
Figure 3.22: RSF variable importance using minimal depth (using TAER data) ................. 68
Figure 4.1: Capacity, demand, TMI and operational outcome similarities relationship ........................................................................................................................................................................................................................................ 72
Figure 4.2: MDS plot of Euclidean Distance of OOB Predicted Capacity CDFs ........................... 77
Figure 4.3: MDS plot of Hellinger Distance of OOB Predicted Capacity CDFs .......................... 78
Figure 4.4: MDS plot of Bhattacharya Distance of OOB Predicted Capacity CDFs ............... 79
Figure 4.5: MDS plot of Proximity Distance using RSF model .............................................. 81
Figure 4.6: Cumulative demand-throughput curves for delay estimation .............................. 83
Figure 4.7: MDS plot of Operational Outcomes based Distance ........................................... 84
Figure 4.8: Evaluation of capacity similarity ............................................................................. 85
Figure 4.9: Regression model of the distance metrics ............................................................... 86
Figure 4.10: Correlation vs Tree Depth for Euclidean Capacity Distance for days with (a) TMIs and (b) without TMIs ................................................................................................................................. 87
Figure 4.11: Correlation vs Tree Depth for Hellinger Capacity Distance for days with (a) TMIs and (b) without TMIs ................................................................................................................................. 88
Figure 4.12: Correlation vs Tree Depth for Bhattacharya Capacity Distance for days with (a) TMIs and (b) without TMIs ................................................................................................................................. 88
Figure 4.13: Correlation vs Tree Depth for Proximity based Capacity Distance for days with (a) TMIs and (b) without TMIs ................................................................................................................................. 89
List of tables

Table 2.1: Summary of PCA of capacity data for EWR, SFO, ORD and JFK airports ... 17
Table 2.2: Summary of PCA on demand data for EWR, SFO, ORD & JFK airports...... 23
Table 2.3: TMI information & operational outcomes of similar days from case study
................................................................................................................................. 29
Table 3.1: Applying censored variable analysis to airport capacity.......................... 38
Table 3.2: Important functions in censored variable analysis................................. 38
Table 3.3: Summary statistics of EWR hourly airport data, 2012-2014 (3 years)..... 48
Table 3.4: Cox proportional hazards (PH) model estimation results ...................... 51
Table 3.5: Cox proportional hazards model estimation results (using TAER data) .. 64
Table 3.6: Validation measures summary.................................................................... 70
Table 4.1: Summary statistics for TMI and operational outcome data (2011-2014) 75
Table 4.2: Scheduled arrival demand and IFR capacity profile for EWR, 1/14/2013
........................................................................................................................................................ 82
Table 4.3: Relative importance (days with TMIs) ...................................................... 90
Table 4.4: Relative importance (days without TMIs) .................................................. 90
Acknowledgments

Namaste. I do not know if I can find the words to convey my deepest gratitude to the waypoints that have made my flight towards the PhD possible, but let me try here. There are many more I thank for this day than I can mention here. Graduate school is something of a privilege which cannot be taken for granted for a young women like me, who was born in a village more than 8000 miles away. I have been extremely fortunate to be born to my amma and nanna, who firmly instilled in me that I can rightfully dream to reach the skies (they probably foresaw the dissertation title) and I would like to dedicate my graduation to them. Coming to Cal has meant the world to me and has taught me to be comfortable about being uncomfortable, during those without-prerequisite stats classes and those borderline-crazy ideas presented during qualifying exams. In Telugu (my mother tongue), we use the word krtajnata to convey profound gratitude to people. I will take this section to offer my krtajnata to those who made my PhD possible.

I have grown up reciting verses such as ‘Gurur Brahma Gurur Vishnu… ’, which say that guru (teacher) is a god to an individual. I have been extremely fortunate to share my PhD journey with my guru Dr. Mark Hansen. He has been instrumental in helping me channel my energy to constantly learn and grow through the varied phases of my journey at Berkeley. His enthusiasm and acceptance of my crazy ideas and his wisdom in shaping them into meaningful goals has been a constant source of my energy through the PhD. I am lucky to have found an advisor and a friend in Mark and will never forget our long discussions on education and women empowerment. I offer my krtajnata to Prof. Mark Hansen.

I would like to offer my krtajnata to my committee Dr. Alexey Pozdnukhov and Dr. Laurent El Ghaoui for their guidance and suggestions on my dissertation. Dr. Joan Walker, Dr. Bin Yu, Dr. Mike Ball, Dr. Dave Lovell and Dr. Michael Jordan inspired me through the different phases of my PhD. My journey of learning would not be possible without the lectures, discussions and advice that I got from all of them through my years at Berkeley. The similar days team from ATAC, University of Maryland, NASA and UC Berkeley gave me valuable feedback on my work and helped me immensely. Yi Liu, Yulin Liu, Alex Estes, Dr. Shon Grabbe, Kennis Chan, Corey Warner, John Schade and Danqing Zhang have been very important collaborators for my research.

My journey at Cal would not have been possible without the love and constant support of my Berkeley family. Caroline Le Floch has been one of my first friends here and has definitely added a French flair to my life. Our wine, cheese and cooking sojourns and our hikes together made me feel at home away from home. Feras El Zarwi has been my source of strength through tough PhD times and my source of the cool Arabic slang. Timothy Brathwaite’s tireless efforts to get me on a bike and his support have made the uphill parts of the PhD ride seem pleasant. My swim sessions with Srinidhi, walks with Darren, chats with Andrew Campbell and
Akshay Vij, chai and Indian food sessions with Berkeley friends have kept me going. Lei Kang, Mogeng Yin, Maddie Shaheen, Teddy Forscher, Sudatta Mohanty, Stephen Wong, Sid Feygin, Matthew Reiter, Mustapha R Harb, Nathalie Saade, the 116 McLaughlin crew and the Transoc crew have made all those long nights and groggy mornings enjoyable. Shelley Okimoto, Bernadette Edwards, Helen Bassham and Jeanne Marie have been an immense help during the various phases of the PhD.

My best friends Sonika Malloth and Dhanasree Molugu have made the many years of my life joyous. You guys have stood by my side through all seasons of my life. Aditya Maheshwari, you have been an amazing friend through tough and fun times during my PhD. I want to offer my krtajnata to my brother Praharshit Gorripaty, who has been my pillar of strength and my constant companion for the past 25 years. Annaya, you have showed me what patience and compassion are. I want to give my krtajnata to my husband Nikunj Bajaj for his friendship, unconditional love, support through the ups and downs of the PhD and my associated moods. Nikunj, you have been my light through dark passages, my adventure buddy and you have always reminded me that I have it in me to make this world a better place. I am so excited to be sharing my life with you through many more exciting journeys. My krtajnata to Nikunj’s parents, sisters and brother-in-laws for all their love and excitement regarding my PhD.

The divine energy is this universe has been kind for sending me to my amma and nanna and providing me with the strength and courage to make the most of life. Finally and most importantly, I want to offer my krtajnata to amma and nanna for bringing me to this world, giving me everything beyond what I could imagine. Amma, you have made innumerable sacrifices and provided constant encouragement to reach for the stars. Nanna, without your belief in me, I could never have imagined that I would one day graduate with a PhD from UC Berkeley. You both have truly taught me the meaning of life, love and compassion and I am indebted to you for everything.
1. Introduction

1.1. Background and Motivation

Airports serve as doorways to the world enabling travel and trade between cities and countries at large. In the US, the aviation system contributes to 4.9 to 5.2% of the Gross Domestic Product (GDP) and provides millions of jobs (Eno Center for Transportation, 2013). Rising globalization has resulted in increasing demand for air travel (Button, 2008). Emerging economies are further adding to this increasing demand with new classes of individuals now accessing air travel for leisure and economic opportunities. The surge in E-commerce with short delivery windows is another dimension of the demand faced by airports. Federal Aviation Administration (FAA) predicts a 2.2% average annual passenger growth rate from 2013 to 2033 in the US (Eno Center for Transportation, 2013).

Growing demand puts pressure on the existing airport infrastructure. Many airports worldwide are witnessing the consequences of this surge in demand, juxtaposed with the limited capacity at airports. Ball et al., 2010 estimated that the total cost of air traffic delays in 2007 in the US amounted to $32.9 billion. The component of this cost to airlines is $8 billion, which includes increased cost of crew, fuel and maintenance due to delays. The cost of passenger time lost due to delay accounted for $16.7 billion and $3.9 billion is the cost of lost demand due to passengers avoiding air travel due to delays. Air traffic delays reduced the US GDP by $4 billion in 2007, owing to its indirect effects on businesses in the US economy. Zou and Hansen, 2014, found that delays affect flight frequency and fare in the US, with aggregate fare reduction benefits of over $5 billion annually from delay savings.

One solution to accommodate increasing demand is to expand existing airports by adding runways or building new airports. Airport construction and expansion projects are however expensive, protracted and face challenges such as limited land and environmental pollution. Another alternative is to manage demand using caps on flights and congestion pricing during peak hours. FAA recognizes that these restrictive policies can be useful for short-term management of airport congestion, but are less preferred to improving airport capacity (Ryerson and Woodburn, 2014). It is thus important to make better use of existing airport capacity. In the next few paragraphs, we discuss some important concepts that will help understand ways to accommodate increasing demand by utilizing current airport infrastructure efficiently.
1.1.1. Airport Capacity

Aviation authorities around the world such as the FAA lay out rules governing the airfield and airspace usage. These rules specify aspects such as the minimum separation between different aircraft classes for varying weather conditions. This influences how many operations per unit time a given airport can accommodate, which is referred to as the airport capacity. Factors such as design of the airport, weather conditions, fleet mix and human factors such as the experience of controllers impact the capacity of an airport (Newell, 1979; Venkatakrishnan et al., 1993). The wind conditions at an airport can limit the number of runways available for operations. The visibility and ceiling conditions can impose additional constraints on the minimum separation between aircrafts operating at the airport.

The flights operating at an airport are scheduled taking into account the physical capacity of the airport under ideal conditions. However, the capacity of an airport is highly dynamic, with most of its variability arising from changing weather conditions. Predicting the capacity at an airport is still a challenging problem, with major consequences for allowing more flights to operate than the available airport capacity. At times when more flights intend to takeoff and land at an airport than the capacity permits, it is especially important to manage these flights to ensure safe and efficient airport operations. This motivates the need for air traffic management at airports and is detailed in 1.1.2. The workload of individuals responsible for managing air traffic can also sometimes limit the available airport capacity. Fig. 1.1 summarizes the three main components that govern the capacity of an airport.

Figure 1.1: Components that determine airport capacity
1.1.2. Air Traffic Management

The Air Traffic Management (ATM) system coordinates the flights in the National Airspace System (NAS) to ensure safe and expedited flow of air traffic and is broadly divided into two main components: Air Traffic Control (ATC) and Air Traffic Flow Management (ATFM). The primary mandate of ATC is to avoid collisions between operating aircraft by maintaining the minimum required separation between aircrafts. The Terminal Area Radar Approach Control (TRACON) facility handles the operations near an airport, whereas air route traffic control centers (ARTCCs) handle en route operations. Fig. 1.2 summarizes the ATC implemented during the different stages of flight starting from takeoff to landing. When the capacity of an airport is less than the demand, the ATC may get overloaded managing numerous aircrafts waiting to land. To avoid such risky scenarios, ATFM regulates the demand faced by the airports and airspace to balance demand with their capacity. The strategies used to balance capacity and demand are referred to as Traffic Management Initiatives (TMIs). Two commonly implemented TMIs to manage airports are the Ground Delay Program (GDP) and the Ground Stop (GS). A GDP is implemented to balance the arrival demand and capacity at an airport by delaying flights at their origin airports. When the capacity shortfall is sudden and severe, a GS maybe implemented and all the flights are stopped at their origin airports.

<table>
<thead>
<tr>
<th>Tower</th>
<th>TRACON</th>
<th>ARTCCs</th>
<th>TRACON</th>
<th>Tower</th>
</tr>
</thead>
</table>

![Figure 1.2: Air Traffic Control (ATC) in different stages of flight](image)

TMIs are implemented during times of capacity-demand imbalance, to manage the airport safely and efficiently. There are substantial penalties such as diversions, airborne delay and excess controller workload for releasing more flights to an airport than its available capacity. Forecasting the capacity and demand for an airport and controlling the release of arriving flights using TMIs in accordance with the predicted airport capacity can mitigate these consequences. The penalties for
releasing fewer flights to an airport include ground delay and cancellations. It is thus important to understand the tradeoffs between using different TMIs in managing operations at an airport. As an example, GDP substitutes less predictable, unsafe and expensive airborne delay with ground delay. However, when a GDP is implemented, airport capacity will be underutilized if the weather clears off earlier than predicted. Traffic management specialist experience plays a vital role in understating these tradeoffs. This motivates the need to develop tools that can augment controller experience and lead to more efficient TMIs, thus enabling the better utilization of existing airport capacity under uncertainty. We introduce some important concepts on such algorithms and their applications in 1.1.3.

1.1.3. Similar Days

Important parameters relating to TMIs such as the start times, stop times and arrival rates are often decided by FAA decision makers. They rely on intuition, experience and sometimes simulation to decide the optimal parameters for a given scenario. The past traffic management actions taken under different scenarios have been recorded in National Traffic Management Log (NTML) since 2000. Historical archives on weather, demand, operational outcome and traffic management actions data can be used to understand underlying patterns and identify similar days in the past. The historical days similar to a day of interest can be used to intelligently augment relevant experience to the decision-maker. Recent advances in machine learning and computing power have made it possible to mine and analyze these sizable historical archives. With decision support tools at their fingertips, traffic management specialists can make better decisions and improve airport operations during times of uncertainty. Supplementing experience with data-driven similar historical days will especially help less experienced traffic management specialists in employing more effective TMIs.

There are two main use cases of finding similar days: day-of-operations and post-operational. In the day-of-operations use case, we identify days similar to the current day at the airport based on capacity and demand data. We visualize relevant TMI and operational outcome data for the similar days. The decision-maker can analyze the operational outcomes of the various decisions taken on the historical similar days and choose an action for the given day that results in the most desirable operational outcomes. In the post-operational use case, the performance of the TMIs on a given day can be evaluated by comparing its operational outcomes with those of similar historical days with different TMIs. This will allow decision-makers to make quantitative comparisons of different strategies.

TMI decision-making informed by the analysis of historical data can lead to more systematic air traffic management. This predictability is especially important for flight operators who plan airline operations. Awareness of the possible actions that might be taken when capacity-demand imbalance occurs, can better equip them
to prepare for these scenarios. The ability to identify similar days is also useful to researchers. The traffic management actions taken on similar historical days and how successful they were based on their operational outcomes can provide useful guidance on ‘what-if’ analysis of different TMIs. This is especially useful for understanding the impact of TMIs that interact with other NAS components, which may not be easily incorporated in simulation studies. Such interactions may be observed in New York metroplex where EWR, JFK & LGA airports are closely located.

An important question in identifying similar days is what an appropriate similarity measure would be. (Mukherjee et al., 2013; Klein et al., 2009) define similar days based on Weather Impacted Traffic Index (WITI) data. WITI (Callaham et al., 2001) captures the interaction between bad weather and air traffic for a region of interest. Further work by Grabbe et al., 2013 used data on the cause and location of GDPs along with weather data to identify similar days. To be useful for TMI decision-making, similarity between days should incorporate capacity and demand information. The first step towards developing algorithms to identify similar days is to generate representative features from weather, capacity and demand data. The next step is to recognize a meaningful distance metric between days in NAS. A machine-learning algorithm to identify similar days based on this distance needs to be trained using historical data. Finally, the TMIs and operational outcomes data on these similar days should be provided. With an important application of similar days being decision support, it is important to visualize key information about the similar days.

Fig. 1.3 shows an architecture that combines the different aspects of finding similar days at an airport that can be used for decision support. The main collections of airport data include weather, capacity, demand, TMI and operational outcome data for all days. This data is used to find a suitable similarity measure, which is further used to estimate the similarity matrix. Each element of the similarity matrix contains the similarity between a pair of days. For a post-operations use case, we can estimate this similarity matrix for all days in a period of interest. To evaluate the strategies on a certain day (reference day), we find the $n$ most similar days to it. We can then investigate a summary table of these $n$ days containing important information about the capacity, demand and weather patterns on those days and compare it to the reference day’s summary. We can also study detailed visualizations of TMIs, convective weather maps and delay patterns on those days.
Figure 1.3: Similar days architecture
1.2. Research Objective

Capacity demand imbalance leads to inefficient airport operations and costs billions of dollars in the US. It is thus important to better manage existing airport capacity during times of imbalance. Finding similar historical days can guide traffic management specialists and flight operators make well-informed decisions. The effectiveness of TMIs can be evaluated by comparing the operational outcomes of different TMIs implemented on similar historical days. The main objective of the research detailed in this dissertation is to find similar days for an airport to aid traffic management decision-making. Finding similar days can be divided into three complementing objectives:

1) **Capacity and Demand estimates**: For similar days to guide TMI decision-making, it is important that they incorporate capacity and demand data. In developing the estimates for capacity and demand, it is important to use data that can be forecasted easily. This will allow for the application of the methodology developed here for the day-of-operations use case. This dissertation aims to develop a methodology to generate accurate estimates of capacity and demand at airports.

2) **Similarity Measure**: An important question in identifying similar days is to develop a method for estimating similarity between days. Another question that arises in this pursuit is if the similarity measure should be discrete or continuous. The objective of this dissertation is to use machine-learning methods to develop a measure of similarity between days that incorporates capacity and demand data.

3) **Validation**: To be usable for decision support at airports, the similarity measure needs to be validated. This dissertation aims to develop a framework that will evaluate the predicted similarity measure between days at an airport. To fulfill this objective, this research will explore questions on how to assess the validity of a given similarity measure.

1.3. Dissertation Overview

This dissertation is organized into five chapters. The second chapter develops similarity measures based on capacity and demand data at four airports – EWR, SFO, ORD and JFK (Gorripaty et al., 2017). We use capacity and demand data at these airports. In this chapter, we explore dimensionality reduction techniques on capacity and demand data to obtain their efficient representations in similarity estimation. We investigate discrete similarity measures using clustering techniques based on a combined capacity-demand distance. This similarity between days is visualized using a Metric Multidimensional Scaling (MDS) plot. This work shows the potential of developing such similarity measures for decision-support.
One of the limitations of the research described in chapter two is that the capacity data used may not be reliable. The third chapter explores the possibility of generating better estimates of capacity and demand data that can be used to develop similarity measures (Gorripaty et al., 2016). We generate scheduled demand data using Aggregate Demand List (ADL), which contains fine-grained flight schedule data of all the flights operating at an airport. The capacity of an airport can be directly observed only at sufficiently large demand. When the throughput at the airport is limited by the demand, we can only observe the lower bound of its capacity. In this chapter we develop capacity prediction models that incorporate observations censored by insufficient demand. We predict the capacity distribution of an hour, instead of a single estimate for capacity. We explore Kaplan Meier estimator, Cox Proportional Hazards model and Random Survival Forest model to predict the capacity distribution for an hour based on the its weather, scheduled demand, fleet mix and throughput data. In this chapter, we also detail the validation measures that incorporate censored observations when evaluating the model. We present a case study that illustrates the agreement between capacity predictions on a test (out-of-bag) sample and the observed data for that sample.

In this dissertation, we mainly focus on similarity estimation application of the capacity prediction models. However, the capacity models developed in this dissertation are not limited to similarity estimation and have numerous other applications. Predicting airport capacity can help make better tradeoffs between allowing more flights to operate at the airport and minimizing expensive airborne delay, thus improving the efficiency of ATM. Accurate capacity predictions can be used to validate controller published Airport Acceptance Rate (AARs) and other existing capacity models. Capacity predictions based on weather, demand and fleet mix data will allow air traffic specialists to analyze the performance of ATFM strategies and quantitatively provide a measure of capacity underutilization of numerous TMIs. This performance analysis is crucial in facilitating the design of better airport management strategies. Capacity predictions can also be used in post-deployment analysis, which assesses how capacity changed as a result of some improvement such as a new runway or a decision support tool.

The fourth chapter explores different ways to estimate similarity using capacity distributions generated from models discussed in chapter 3. We also estimate similarity measures based on scheduled demand data. The capacity and demand distances have different scales and may have a non-linear interaction in determining the distance between two days. In this chapter we develop a methodology to combine the capacity and demand based similarity measures into one similarity measure. We investigate a framework to validate the estimated similarity measure. The assumption for this framework is that days that are similar based on their capacity, demand and TMIs should be similar based on their operational outcomes. The correlation between the combined capacity, demand and TMI distance and the operational outcomes distance thus provides a measure of how well we capture the similarity between two days.
The fifth chapter offers concluding thoughts on the dissertations and discusses directions for future research.

1.4. Contributions

This dissertation presents some new directions in developing similarity measures between days in the NAS. The three main contributions of this dissertation to the literature are summarized below:

1) This dissertation takes a first step towards exploring dimensionality reduction of capacity and demand data. Based on the results, we use the reduced dimensional capacity data and the full dimensional demand data for clustering. We also validate clustering results and develop a similarity measure between two days.

2) The development of capacity prediction models that incorporate observations censored by insufficient demand is a key contribution of this dissertation. In this work, we also incorporate convective weather data into predicting airport capacity. Furthermore, we detail novel validation measures to evaluate the capacity predictions derived from the censored data models. We also develop scheduled demand estimates from the fine-grained ADL data.

3) We make the first effort in developing a methodology for the combination of capacity, demand and TMI similarity supervised using operational outcome similarity. We also evaluate how efficiently our similarity estimates capture reality using test data.
2. Similar Days

2.1. Introduction

Aviation contributes to the development and growth of the global economy, transporting people and enabling trade. Poor airport performance leads to losses to the airlines, passengers and the economy. A study by Zou (2012) predicts the cost savings to US airlines alone of improved operational performance ranges from $7.1–13.5 billion for 2007. As discussed in chapter 1, traffic specialists at the Air Traffic Control System Command Center, when faced with a possible demand-capacity imbalance at some airport, are responsible for planning traffic management initiatives (TMIs) that balance competing priorities of efficient capacity utilization and avoiding the over-delivery of flights. The decisions made to manage air traffic play a significant role in keeping the operations in the NAS safe and efficient.

Numerous studies in the past have developed algorithms and simulation studies to incorporate the stochastic nature of weather and capacity into decision-making (Ball et al., 2010; Cook and Wood, 2010; Dhal et al., 2013; Liu and Hansen, 2015; Mukherjee, 2004; Mukherjee and Hansen 2007; Nilm and El Ghaoui, 2004; Provan et al., 2011; Smith and Sherry, 2008; Wang, 2012; Wang 2011). Much of this literature concerns models and algorithms for making decisions that minimize the expected value of a loss function that takes into account ground delay and more expensive airborne delay. More recent contributions incorporate other performance goals, such as equity and predictability (Liu and Hansen, 2015). It has proven difficult, however, for such methods to gain traction with air traffic specialists, who generally place more stock in their own judgment and experience than recommendations from tools developed by researchers.

Recognizing this, the research community has in recent years turned more attention to developing tools that allow air traffic management specialists to tap their own experience. In essence, the idea is to identify days in the past that are similar to some reference day, and consider the TMI actions taken on these days and the operational outcomes that resulted. This information can be used in post-operational analysis to find if an experience in a recent day has echoes in the past, and for day-of-operations decision-making when the necessary forecasts are available. This chapter contributes to similar days literature by developing methods for identifying similar days based on features of specific relevance to air traffic managers assessing—either prospectively or retrospectively—TMIs for balancing arrival demand and capacity at individual airports.

In this setting, the most salient features are profiles of arrival capacity and demand for a given day (in a day-of-operations application, these profiles would be
derived from forecasts). Similarity between days depends on how closely both the demand and the capacity profiles match. We therefore propose and implement methods for assessing profile similarity and identifying similar profiles. There are three interesting research questions surrounding the investigation of similar days. First, is similarity between two days better viewed as a categorical or continuous variable? In other words, are there natural clusters of similar days, or simply a range of similarity without clear boundaries between similar and dissimilar? Second, how much guidance can similar days provide in situations in which TMI decisions may be required. Are days with TMIs more likely to be “odd balls” dissimilar from most historical days, or are TMI decisions often made in situations similar to those encountered many times before? Third, how can the estimated similarity between pairs of days be used to aid efficient decision-making in air traffic management?

To address these questions we develop capacity and demand similarity metrics for historical days between 2007 and 2015 for four major U.S. airports: Newark (EWR), Kennedy (JFK), Chicago O’Hare (ORD), and San Francisco International (SFO). We then combine the demand and capacity similarity to determine a metric for overall (capacity-demand) similarity between days. We also characterize each day in terms of its level of TMI activity in order to assess the availability of similar days for the subset of days with significant TMIs. Finally, we offer case studies on how the similarity measures can be used to make better decisions in the context of air traffic management.

The remainder of this chapter is organized as follows. Section 2.2 describes the data used and pre-processing done for the analysis. Section 2.3 presents the correlation analysis performed to explore the data. In section 2.4, we perform PCA on capacity data at the four airports and summarize the results. Section 2.5 describes the clustering analysis on capacity data and discusses the possibility of developing a discrete measure of similarity between days. PCA on the demand data is presented in section 2.6. Section 2.7 presents the clustering analysis on demand data with a discussion on developing discrete measures of similarity based on demand data. In section 2.8, we identify similar days using continuous measures of similarity and visually represent the similarity between days with different levels of TMIs. Section 2.9 summarizes the findings of the study and identifies future research needed to make similar days a useful decision support tool.

2.2. Data

We use quarter-hour Aviation System Performance Metrics (ASPM) data for the analysis. We use two variables from the data: arrival demand and capacity. Our data set covers the period from January 2007 through August 2015, and covers four airports: EWR (Newark Liberty International Airport), SFO (San Francisco
International Airport), ORD (O'Hare International Airport) and JFK (John F. Kennedy International Airport).

The arrival demand variable in ASPM dataset reflects the total number of flights that would land in a given quarter-hour time period in the absence of capacity constraints. This includes flights that have been held at their departure airports or while en route in order to avoid queues from building in the terminal area of the arrival airport. We used this data to compute a “new” demand variable for each time period by subtracting flights that were also contributing to demand in the previous time period. This is done to avoid counting the same flights in multiple time periods and estimate the real demand experienced by the airport. The demand data at an airport for an observation, which is one day, is a vector of quarter-hourly demand data for that day.

The capacity variable is the Airport Acceptance Rate (AAR), which reflects FAA estimates of the number of arrival flights that could land in each quarter-hour time period. Discussions with Air Traffic Controllers (ATCs) reveal that the AAR may not be a reliable estimate of airport capacity, particularly when demand is low enough that estimation errors are inconsequential. However, during the hours of the day with a high demand, the controllers are likely to be more attentive to update the AAR regularly due to higher pressure to provide more reliable airport capacity estimates. In light of this, we only use hours that have a high demand-capacity ratio for our analysis.

To determine which hours are appropriate, the median value of the demand-capacity ratio is calculated for each quarter-hour and airport, and the hours for which this value is over 0.75 were identified. (For this purpose, we used the original demand variable reported in ASPM rather than the “new” demand.) Thus, on the majority of days in these hours, demand is at least 75 percent of the announced capacity (AAR). Such periods can be viewed as busy and we can expect more accurate estimates of AAR in them. On this basis, we identify contiguous time periods in the day in which the airport is usually busy (These contiguous periods may contain a few short periods of lower demand). The hours selected as the busy period are: 7 to 22 for EWR; 7 to 22 for JFK, 8 to 22 for SFO; 7 to 20 for ORD. The capacity data at an airport for an observation, which is one day, is a vector of quarter-hourly AAR data for that day. It is important to note here that a “day” refers to the busy hours of the day, which varies from airport to airport.

2.3. Correlation Analysis

To find methods that accurately represent the similarity between different days, we begin with an exploratory analysis of airport capacity and demand data as detailed below. Our analysis focuses on daily capacity and demand observations, where each
observation is a vector containing time series of quarter-hourly capacity (AAR) and “new” demand data respectively over the busy period of a given day and airport. To gain a better understanding of the relationship between different variables (quarter-hours) of an observation (day), we look at their correlation plots. The correlation plots give an overall picture of the pattern of capacity and demand variation across different hours in a day. For illustrative purposes, we show the correlation plots for capacity and demand data for EWR busy periods in Fig. 2.1 and Fig. 2.2 respectively. The axes in both figures correspond to time-of-day on an hourly scale with resolution of quarter-hour. The color at a given x-y coordinate indicates the correlation between the AARs or demand at times x and y; blue indicates positive correlation, red indicates negative correlation, and white indicated the absence of correlation. The correlation plots in Fig. 2.1 and Fig. 2.2 are plotted for quarter-hourly data, but only the full hour is labeled on the axes in the correlation plots to make the plots more legible.

From Fig. 2.1, it is evident that there is strong correlation between the capacities at times that are proximate to each other (i.e. near the diagonal) and that it attenuates—while, however, still remaining positive—as the times become further apart. This indicates that capacity data of contiguous hours is correlated. This is as expected since capacity is greatly dependent on weather, which is similar in contiguous hours. We analyze the correlation of the demand data for busy hours in Fig. 2.2. The white color in the majority of non-diagonal elements indicates very little or no correlation in the demand data across different hours of the day. Fig. 2.2 indicates that demand data is uncorrelated across different hours and suggests that the factors that drive demand—decisions and preferences of flight operators and air travelers—lack inter-temporal persistence. Motivated by the correlation in the capacity data, we explore methods that can provide us with reduced dimension data, thus removing redundancy, while retaining most of the information in the data.
Figure 2.1: Correlation plot for capacity data at EWR airport

Figure 2.2: Correlation plot for demand data at EWR airport
2.4. Principal Component Analysis: Capacity Data

The highly correlated nature of quarter hourly capacity variables in capacity data as discussed in section 2.3 motivates the use of Principal Component Analysis (PCA) to transform these correlated variables into a smaller subset of linearly uncorrelated variables with minimum information loss. The data are centered and scaled prior to the PCA analysis. The intuitive explanation of PCA is to find the projections (principal components) of the data on a new coordinate system (eigenvectors) that is aligned with the directions that capture the most variance (eigenvalues) of the data. It is important to note that the principal components may not map onto any physically interpretable construct and is used to reduce the dimensionality of data containing correlated variables. This is in contrast with factor analysis, which is based on the hypothesis that certain underlying latent factors can predict the observed variables. By construction, the principal components are perpendicular to each other and thus uncorrelated. The PCA finds an equivalent number of components as the number of dimensions of the data. In correlated variables, most of the variance of the data is captured by a smaller subset of the principal components. The number of principal components required to represent the data with minimal information loss is called the effective dimension. We explore four widely used rules (Raîche et al., 2013) to identify the effective dimension of the data in the PCA. The eigenvalue corresponding to an eigenvector gives a measure of the variance captured by that eigenvector. These rules, in brief, are:

- **Parallel analysis**: Finds the eigenvalues on a random dataset (noise of the same dimensions) and checks for eigenvalues that are above noise levels.

- **Kaiser rule**: Drops all eigenvalues under 1.0 since these components contain more noise than signal.

- **Optimal Coordinates**: Checks for the last eigenvalue above the linear extrapolation of the preceding eigenvalue (to the right), by a regression line between the preceding eigenvalue coordinates and the last eigenvalue coordinates in the scree plot (defined in this section). These components contain more information than expected by a linear extrapolation.

- **Acceleration Factor**: Checks for the elbow in the scree plot, where there is a large drop in the eigenvalue between consecutive components. There is a significant gain in information by keeping the component just before the large drop. The number of components until the component just before the elbow is the recommended effective dimensionality from acceleration factor.

A scree plot is a plot of the eigenvalues against the component number (ordered by eigenvalue) and is useful for identifying the reduced dimension of the data and visualizing these rules. The effective dimension of the data is taken to be the dimension that is recommended by the majority of rules. When there is no clear
majority in the effective dimension by these four rules, the dimension that is most commonly recommended (mode) is used. We apply PCA on the airport capacity data of each of the four airports separately and summarize findings from the analysis with the help of informative plots in this section.

First, we discuss the results from PCA analysis for EWR airport in detail. In EWR airport, each observation of the capacity data is a vector of quarter-hourly AAR data from 7 to 22 hours. From the four rules described above, an effective dimension of six is suggested by the dimensionality rules. From cumulative variance plot in Fig. 2.3, we see that a dimension of six (out of 61) captures about 90% of the variance of the data. This means that using just six new variables, we can capture 90% of the variance captured in the 61-dimensional capacity data, where each dimension is a quarter-hourly capacity value between 7-22 hours. The six effective variables, which are called the principal components, are certain linear combinations of the quarter hourly AAR variables. Thus, for further analysis, we use the first six principal components (PC) to represent capacity data at EWR.

![Figure 2.3: EWR: Cumulative variance explained vs number of PCs; Loadings of the first 6 PCs on 7-22 hours](image)

The loadings of a principal component associate different weights to the capacity at different hours of the day in defining a PC. In Fig. 2.3, we plot the loadings of different hours for the first six PCs of the EWR capacity data. From Fig 2.3, we see that the first PC captures the average behavior of the day, the second PC captures the contrast between morning and evening hours, the third PC captures the contrast between the middle of the day and other times, and so on. The oscillatory fifth PC is an artifact of the way in which the AAR data are reported at the quarterly hour level: to maintain integer quarter-hourly values, rates may oscillate up and down, for example 7-8-7-8. For the remaining airports, we have summarized the results of the PCA in Fig 2.4 and Table 2.1.
As explained earlier, the PCA finds an equivalent number of principal components (PC) as the number of dimensions of the data. The number of dimensions for the capacity data is the number of quarter-hours in the busy hours of a day. Based on the busy hours considered for each airport, the dimensionality of the capacity data in SFO, ORD and JFK airports can range between 53-61 dimensions. Fig. 2.4 shows the plots of percentage cumulative variance plotted against PC number and the plots of loadings of the important PCs (as found from the four rules) plotted against the corresponding hour for SFO, ORD and JFK airport respectively. In Table 2.1, we summarize the results from PCA on the capacity data at the four airports. Note that the percentage cumulative variance is rounded up to the nearest multiple of five when reported throughout the paper.

| Table 2.1: Summary of PCA of capacity data for EWR, SFO, ORD and JFK airports |
|--------------------------------|-----|-----|-----|-----|
| Airport | EWR | SFO | ORD | JFK |
| Effective dimension | 6   | 5   | 5   | 8   |
| Cumulative variance explained | 90% | 90% | 90% | 90% |
| Variance explained by 1<sup>st</sup> PC (Capturing the average capacity) | 60% | 65% | 65% | 45% |
| Variance explained by 2<sup>nd</sup> PC (Capturing the contrast between morning and evening capacity) | 15% | 15% | 15% | 15% |
| Variance explained by 3<sup>rd</sup> PC (Capturing the contrast between noon and evening capacity) | 10% | 5%  | 5%  | 15% |

In estimating the effective dimensionality of the capacity data, three rules were consistent in their dimensionality recommendation. The fourth rule of Acceleration Factor (AF), always recommended a dimension of one, since the drop in information captured from the first to the second PC is large. This means that the elbow in scree plot is between 1 and 2, meaning that a dimension of 1 is recommended by the AF rule. It is interesting to note that across the airports, the amount of variance explained by the effective number of PCs is about 90%, making it a very efficient dimensionality reduction. The larger number of PCs required for capturing the JFK capacity data may indicate a higher variability in capacity at the JFK airport. The loadings of the PCs on the hours depict strikingly similar stories across the different airports. For each airport, the first PC captures the overall level of capacity in the day, the second PC contrasts the morning and the evening capacity, and the third PC contrasts the midday and evening capacity. There is also
consistency in the percentage of variance captured by each of the PCs, with the 1st (overall capacity) PC capturing 45-65% of the variance, the 2nd (morning-evening contrast) PC capturing about 15% of the variance and the 3rd (midday-evening contrast PC) capturing 5-15% variance. As expected, in a highly variable airport such as JFK, the variance of data is distributed more evenly across different components. This consistency in effective dimensionality, cumulative variance, and temporal patterns of factor loading for airports spanning different layouts, weather conditions and traffic patterns is somewhat surprising.

Figure 2.4: Cumulative variance explained vs number of PCs; Loadings of the effective number of PCs for SFO, ORD and JFK respectively
2.5. Clustering analysis: capacity data

PCA allows the similarity between capacities on two different days at a given airport to be measured more accurately. The capacity vector consists of correlated variables before PCA transformation. Commonly, components that capture the most variance are important for clustering purposes and render robustness to clustering results (Ben-Hur and Guyon, 2003). In light of this, we use the $m$ uncorrelated PCs with the highest eigenvalues, where $m$ is the effective dimension, instead of the raw correlated data as the input to the clustering analysis. We employ a weighted Euclidean distance metric to measure the distance between two observations (vectors of PCs) to account for the varying importance of different PCs, based on the variance explained as shown in equation 2.1.

\[ d_{ij} = \sqrt{\sum_k \alpha_k (f_{ik} - f_{jk})^2} \]  

(2.1)

where $f_{ik}$ is the value of factor $k$ on day $i$ and $\alpha_k$ is the proportion of variance explained by factor $k$. Days that are closer to each other based on the weighted Euclidean distance measure can be considered to be more similar based on capacity. The next question is whether similar days form natural clusters. We therefore search for inherent clusters in the capacity data, using the distance metric defined in equation 2.1. We check for clusters qualitatively by inspecting graphical representations of the data and also use quantitative measures. Specifically, the clustering criteria employed are:

- **Visual inspection:** The scatterplot between the first and second principal components, which capture a large proportion of the variance of the data (60-80% for capacity data), can reveal clusters.

- **Within sum of squares (WSS):** We generated different numbers of clusters in the data using k-means. The WSS naturally decreases with increasing number of clusters. However, if we plot WSS against the number of clusters and observe an elbow in the plot between the values $n - 1$ and $n$, this suggests that the data inherently has $n$ clusters. This elbow suggests that there is a large decrease in the variance of the data when $n$ clusters are formed as compared to other number of clusters.

- **Average Silhouette Width (ASW):** The ASW estimate gives a measure of the normalized separation and cohesion properties of clusters. The separation of clusters is a measure of how separated the points in one cluster are from the points in the next nearest cluster and cohesion is a measure of how close the points in one cluster are to the points in the same cluster. For a point $i$, let
$b(i)$ being the lowest average dissimilarity (distance) with other clusters to which it does not belong (separation measure). Let $a(i)$ be the average dissimilarity for point $i$ with all other points in the cluster it belongs to (cohesion measure). The definition of silhouette for a point, $s(i)$ is given in equation 2.2. Points with $s(i)$ close to 1 are clustered well and those with $s(i)$ close to 0 are in between the assigned cluster and the nearest neighboring cluster and those with negative $s(i)$ values may be assigned to the wrong cluster and are a better fit in their nearest neighboring cluster. The ASW is the average of the silhouette over the entire dataset and is a measure of how well the data is clustered. The ASW has a value between -1 and 1, with values closer to 1 indicating a higher quality of clustering and values away from 1 indicating poor clustering.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2.2}$$

In employing the clustering criterion described above for capacity data, we use the important PCs as the input data as decided in section 2.4 and use the weighted Euclidean distance measure as described in equation 2.1. We applied the above methods to look for evidence of clustering for the capacity data of the four airports. For each of the four airports, the first and second PCs are plotted against each other, as shown in Fig. 2.5.

Each point in the plots represents a day, while the color of the point reflects the average capacity on that day. The plots in Fig. 2.5 for SFO, ORD and JFK show that days with a higher average capacity have a higher PC1 value and for EWR, higher average capacity have a lower PC1. PC1 captures the overall capacity behavior over the day. Also, there seems to exist symmetry in all the plots about a value of 0 for PC2, for a given value of PC1. This symmetry represents days with the same average capacity, with some days having a positive PC2 indicating a certain contrast between the morning and evening capacities and the other symmetric days with negative PC2 indicate the opposite contrast between the morning and evening capacities. With opposite contrasts, these symmetric days, with the same value of PC1, have the same overall average capacity. We also observe that PC2 is close to zero on the days with the highest (or lowest) PC1; on these days capacity is high (or low) throughout the day, so there can be little contrast.

These plots do not suggest natural clusters at any of the four airports. To verify this finding, we consider the quantitative clustering criteria. The ASW estimates for all the airports are less than 0.5, which provides no evidence of the presence of clusters in the data. Also, the WSS plots for all the four airports do not provide sufficient evidence for inherent clusters in the capacity data of the airports.
For illustrative purposes, the ASW and WSS plots for varying number of clusters plotted for EWR airport are shown in Fig. 2.6.

Figure 2.5: PC1 vs PC2 for capacity data of EWR, SFO, ORD and JFK respectively

Figure 2.6: ASW and WSS for EWR capacity data

To summarize, we find no clusters in the capacity data in the four airports considered, based on both qualitative and quantitative measures. The plots in Fig.
suggest that though there are no clusters in the data, days with higher average capacities are closer to each other and days with lower average capacities are closer to each other. Given the interpretation of PC1, it is not surprising to observe the higher average capacity days bunched together in one end of the plot, with the lower average capacity days trailing on the other end. The lack of clusters motivates us to consider more continuous notions of similarity between days, which we will explore in more detail in later sections in the paper. First, however, we turn to the analysis of the demand data.

2.6. Principal Component Analysis: demand data

The PCA methodology described in detail for capacity analysis in section 2.4 is also applied to the demand data. The variables for an observation (a day) are the quarter-hourly demand data in the busy hours. For the four airports analysed, the original dimension of the data is between 53-61 dimensions, based on the busy hours considered at the airport. The same four rules that are used to investigate the effective dimension of the capacity data are also applied to the demand data of each of the four airports. The cumulative variance explained is plotted against the number of PCs for each of the four airports in Fig. 2.7. Table 2.2 summarizes the PCA results on the demand data at the four airports.

Figure 2.7: Cumulative variance vs number of PCs for EWR, SFO, ORD & JFK airports
Table 2.2: Summary of PCA on demand data for EWR, SFO, ORD & JFK airports

<table>
<thead>
<tr>
<th>Airport</th>
<th>EWR</th>
<th>SFO</th>
<th>ORD</th>
<th>JFK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective dimension</td>
<td>12</td>
<td>18</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>Cumulative variance explained</td>
<td>40%</td>
<td>50%</td>
<td>50%</td>
<td>40%</td>
</tr>
</tbody>
</table>

The amount of variance explained by the effective number of PCs is rather small. This is in marked contrast to the results of PCA on capacity data, in terms of the effectiveness of the dimensionality reduction. The plots of the cumulative variance explained by different number of PCs in Fig. 2.7 show that the additional variance captured by all PCs beyond the first PC is small and is fairly constant across all the remaining PCs. Owing to the low percentage of variance captured by the effective number of PCs across all airports, we do not use the PCs to represent demand data in the next steps analyzing the inherent clusters in the data and estimating the dissimilarity matrix. Since the variance explained by each of the PCs for the demand data is small, the information captured by the PC1, PC2 and other PCs is not discussed in detail.

2.7. Clustering analysis: demand data

Having found that the dimensionality reduction technique using PCA is ineffective in capturing the information present in the demand data, we use the entire vector of quarter-hourly demand data (raw demand data) for the busy hours at the airports for a given day. It is important to note that this change in strategy changes the dimensionality of the distance estimation problem, with the entire demand data consisting of more than 60 components, as opposed to less than 20 components that are recommended by the PCA. In this light, it is useful to discuss the implications of the higher dimensionality data.

In higher dimensions, the volume of the space increases at such a pace, that the data in practice tends to become sparser at higher dimensions. The reasons for such sparsity is that there are numerous dimensions defining the data point, making it unlikely that two data points will be similar across all dimensions. Thus, in higher dimensions, points are generally more dissimilar from each other. While we would be more likely to find clustering if we lowered the dimension of the demand data through PCA, the clusters would be of questionable validity since the lower dimension representation does not effectively capture the true variation in demand. We therefore perform the cluster analysis on the full demand profile over the busy hours. We use PCA only to create the plots for visual inspection, using the first two principal components.
The plots are shown in Fig. 2.8 and each point represents a day, with the coordinates representing the scores on the first two demand PCs and the color of the point indicating the average demand of that day. It is difficult to gain much information from the scatter plots of the first two PCs of the demand data due to the small proportion of variance explained by each of these components. Thus, these scatterplots may not represent the behavior of the demand data across all the variables and are less reliable in revealing the number of clusters as compared to the capacity data analysis, where a large proportion of the variance of the data was captured by the first two PCs. Even so, these plots do not provide any evidence of clusters in the demand data for all the four airports.

![Figure 2.8: PC1 vs PC2 for demand data of EWR, SFO, ORD and JFK respectively](image)

Turning to the quantitative clustering criteria applied to the full daily time series demand data for busy hours, we find that the ASW estimates for all airports’ demand data are lesser than 0.2, which suggests the lack of clusters. Also, the WSS plots for the four airports suggest the absence of clusters. Although capacity and demand data differ with regard to dimensionality, they are consistent in lacking natural clustering. This suggests that discrete similarity is not suitable for measuring similarity between days and we explore continuous measures of similarity in section 2.8.
2.8. Continuous Similarity

In a continuous measure of similarity between days, two days are relatively similar if they have a smaller measure of distance between them and relatively dissimilar if they have a larger measure of distance between them. Based on the results from PCA, we based our continuous similarity metric on two components: the capacity distance based on the effective number of PCs and the demand distance based on the entire data for busy hours. There are numerous ways to combine the capacity and demand based distances to develop an overall measure of distance between two days. We explore two ways of combining these distances, one using the maximum of these distances and the other using the sum of squares (Euclidean) and they are discussed in detail in 2.8.1 and 2.8.2 respectively.

The distance matrix for a dataset can be estimated by calculating the distance between every pair of days in the dataset. The estimated distance matrix can be used to aid decision-support for air traffic controllers both in reviewing historical actions and for future decision-making. The controller can identify the closest days to a given day and gain situation-specific guidance on what actions were taken in the past and how different reactions to weather and demand conditions resulted in different outcomes. Thus, applying this concept of similarity into decision-support can help the controller learn from the past actions and experience of many other controllers, helping to identify relevant experience at a glance. In 2.8.3, we illustrate how similarity can be used for decision-making in air traffic management.

2.8.1. Maximum Distance Similarity

In the plots below, we determine distance based on the maximum of capacity and demand distance. Specifically, we calculate the distance as shown in equation 2.3:

\[ d_{ij} = \max (d_{\text{cap}}, d_{\text{dmd}}) \]  

(2.3)

where \(d_{\text{cap}}\) is the distance based on PCs of capacity vectors and \(d_{\text{dmd}}\) is the distance based on demand vectors. For two days to be similar based on this metric, both their capacity and demand profiles must be quite similar. It is not possible for higher similarity in one dimension to offset dissimilarity in the other. The PCs are estimated using centered and scaled capacity data, which are then used to estimate the capacity distance. The demand distance is estimated using centered and scaled demand data. Since the capacity and demand data have the same units and are centered and scaled, the resulting distances can be combined together meaningfully.
To visualize the distances obtained, Metric Multidimensional Scaling (MDS) is used to plot high-dimensional data in a low-dimensional space based on the pairwise dissimilarities/distances between two observations (Liu et al., 2014). The distances are preserved in a 2-dimensional visualization plot in the mean squared sense. Similar days appear nearby each other and dissimilar days appear farther away from each other in the plot, therefore allowing for an informative visualization of the distance matrix. We incorporate additional information into the MDS plots by varying the color of the points that are used to represent each day, based on the number of flights impacted by TMIs. To determine if a flight is so impacted, we use a flag available in ASPM that indicates if a flight was issued an estimated departure clearance time (EDCT). On a day with substantial TMI activity involving a given airport, a large proportion of flights arriving at the airport will have an EDCT issued for them. Therefore, we develop MDS plots using the distance matrix for airports and coloring the points in the plot using the fraction of arrival flights that were given an EDCT hold on that day.

The axes of the MDS plots do not have any physical meaning and should only be interpreted for the relative positions of the points (days). The MDS plots are an approximate visualization of the relative positions and the distances between the days. The MDS algorithm, given the constraint of a two-dimensional plot, tries to preserve the distances between two days as estimated using the similarity metric. The MDS plots in Fig. 2.9 visualize the distances between days based on the maximum distance metric as shown in equation 2.3. Points colored in lighter shades of blue have higher proportions of arriving flights with EDCT. In other words, light blue days have more TMI activity. In all four airports, the days with high TMI activity are intermingled with days with low TMI activity.

For all airports except JFK, there is also discernable separation based on the amount of TMI activity. Most high TMI days are not isolated points, but rather have a considerable number of days with varying level of TMIs near them. On the other hand, the relative separation suggests that we are capturing some of the inherent dissimilarity between high TMI days from low TMI days. The only exception is JFK, where it appears that other factors besides airport demand and capacity may determine TMI activity. These are likely to be convective weather and terminal airspace constraints.
Figure 2.9: MDS plot for maximum distance for EWR, SFO, ORD and JFK respectively

2.8.2. Euclidean Distance Similarity

Distance based on Euclidean norm of the capacity distance and the demand distance is calculated in equation 2.4.

\[
d_{ij} = \sqrt{d_{cap}^2 + d_{dmd}^2} \tag{2.4}
\]

MDS plots based on this metric are shown in Fig. 2.10. We again color days with the proportion of arriving flights given an EDCT. As before, the plots in Fig. 2.10 show intermingling of high and low TMI days but, with the exception of JFK, some separation between high and low TMI days. As discussed in section 2.8.1, the MDS plots in Fig. 2.10 are useful to visualize the relative distances of the days based on the similarity metric and the axes do not have any physical meaning.
2.8.3. Air Traffic Management Case Study

To help readers appreciate how similarity information between two days can be used to make better decisions, we present a case study. We use Euclidean distance metric as detailed in section 2.8.2, to identify similar days at EWR airport. The reference day in this case study is chosen to be 13 August 2014. Using the Euclidean distance metric, we estimate the distance of reference day from all days in the past starting from Jan 2007. For this case study, we pick the two closest days to reference day and analyze how this information might inform decision-making or post-operational analysis on reference day. The two closest days to the reference day are 31 October 2013 and 25 January 2012. Below in Fig. 2.11, we plot the capacity and demand profiles of the reference day and the similar days.

In Fig. 2.11, we can observe that the two similar days, which are chosen based on the demand and capacity similarity, have similar demand and capacity profiles. Demand peaks are slightly higher on the reference day, but there are many periods where demand is higher on the similar days. Capacity wise, the quarter-hourly capacity is 10 for most of the reference day and 25 January 2012, while it
oscillates between 9 and 10 on 31 October 2013. (This oscillatory pattern is how FAA represents non-integer capacity values; in effect it means that the called hourly rate was usually 38, as compared to 40 on the other two days.) We tabulate the TMI and operational outcome data for the reference and similar days in table 2.3

Figure 2.11: Capacity and Demand Profiles for Reference Day and Two Similar Days

Table 2.3: TMI information & operational outcomes of similar days from case study

<table>
<thead>
<tr>
<th>Variable</th>
<th>08/13/2014 (reference day)</th>
<th>10/31/2013 (similar day 1)</th>
<th>01/25/2012 (similar day 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMI start</td>
<td>15:15:00</td>
<td>16:24:00</td>
<td>N/A</td>
</tr>
<tr>
<td>TMI duration (min)</td>
<td>719</td>
<td>395</td>
<td>0</td>
</tr>
<tr>
<td>TMI type</td>
<td>GS/GDP</td>
<td>GDP</td>
<td>N/A</td>
</tr>
<tr>
<td>Average delay (min/flight)</td>
<td>35</td>
<td>29</td>
<td>6</td>
</tr>
<tr>
<td>Total holding time (min)</td>
<td>449</td>
<td>563</td>
<td>10.7</td>
</tr>
<tr>
<td>Total diversions</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Total cancellations</td>
<td>33</td>
<td>5</td>
<td>12</td>
</tr>
</tbody>
</table>

Next, we consider the traffic management initiatives (TMI) that occurred in these three days and how these actions may have influenced operational outcomes in these days. On the reference day, traffic was managed rather aggressively, beginning with a ground stop at 15:15. The combined ground stop and GDP lasted almost 12 hours, and resulted in an average delay of 35 min along with 33 flight cancellations. Control on 31 Oct 2013 was not as tight, with a GDP lasting 6.5 hours. This resulted in less delay and fewer cancellations, but more airborne holding and
diversions. The other similar day, 25 January 2012, did not have a TMI. Apparently, this was not required, since there was little delay or airborne holding and no diversions. Reviewing Figure 2.11, this is probably because the peak demands on 25 January 2012 were considerable less than those on the other day. This points to a possible shortcoming in the demand similarity metric, which gives the same weight to peak and non-peak demand levels.

Supposing that, based on forecasts on demand and capacity, an air traffic management specialist on 8 June 2014 had access to the information in Figure 2.11 as well as the TMI and operational outcome information for the two similar days. They might conclude that 25 January was not sufficiently similar because of its lower peak demands. Regarding 31 October 2013, they might decide to implement a similar TMI with the expectation that this could result in considerable airborne holding, or impose tighter controls resulting in more delay and cancellations but less holding and diversions. Based on the input from flight operators and other considerations, they could make the appropriate tradeoff between these different operational performance possibilities. The information about similar days enables them to make these choices based upon the most relevant past experience.

2.9. Conclusion

In this chapter, we explore the identification of similar days in air traffic management. The contributions of this chapter to the literature are: (1) we identify ways to reduce dimensionality of capacity data by minimizing the information loss using principal component analysis. We find little correlation between demand data and therefore do not recommend dimensionality reduction for demand data; (2) we investigate the existence of inherent clusters in the capacity and demand data and find that there are no inherent clusters that could be used to define a discrete measure of similarity based on cluster membership; (3) given the absence of clusters, we develop a continuous measure of similarity using the capacity and demand data. The results of this study can be summarized into three main components as follows:

**Capacity:** This study presents a very efficient dimensionality reduction technique for capacity data, with the effective PCs capturing about 90% of the variance of the data, consistently for all the airports. Also, the loadings of the effective PCs capture consistent patterns in the capacity data, with PCs capturing capacity behavior such as the average capacity over the day and the contrast between morning and evening capacity, consistently across all airports. The capacity data is observed to have no inherent clusters, as validated by quantitative and qualitative measures. The high-eigenvalue PCs for each airport are used in further analysis to represent capacity data. Thus for studies where AAR is used to represent airport capacity, we can replace the high-dimensional quarter-hourly AAR data (with 53-61 dimensions for busy hours) with eight or lesser number of PCs capturing more than 90% of the
variance of the AAR data. Also, in future studies that are interested in capturing specific aspects of capacity such as the average capacity or the contrast between morning and evening capacities, the specific PCs capturing these aspects can be used.

Demand: The results in this study indicate that the dimensionality reduction of demand data is inefficient, with only about 50% variance explained by the effective number of PCs. The cumulative variance plotted against the number of PCs shows that the additional information added by each PC beyond the first PC is low and fairly constant. Similar to the capacity data, no clusters were observed in the demand data, as verified by both qualitative and quantitative criteria. Owing to the inefficient dimensionality reduction by PCA, we use the original demand data for busy hours to represent demand data for all airports. Thus, for future studies, where ASPM demand data is used to represent airport demand, the high-dimensional quarter-hourly demand data needs to be used for analysis. The results of the PCA analysis and dimensionality reduction are very robust and are validated by four widely used effective dimensionality rules of Parallel Analysis, Kaiser rule, Optimal Coordinates and Acceleration Factor.

Similar Days: We explored discrete and continuous measures of similarity between days at four different airports. The absence of clusters found in the study indicates the lack of natural discrete measure of similarity such as membership in a specific cluster of the data. While clusters can arbitrarily be forced onto any dataset, our results suggest that no real clustering exists for airport demand and capacity data. The clustering analysis is validated using the within sum of squares (WSS) and the Average Silhouette Width (ASW) to form robust conclusions on the clustering behaviour of the demand and capacity data.

Two continuous measures of distance were estimated for identifying similar days, with one being the maximum and the other being the Euclidean distance, with the two components being the capacity and demand based distances. Similarities were found between the Euclidean and maximum measure of distance between days, such as the mix of distances of high TMI and low TMI days. In three of the airports, there is also evidence of some separation between high and low TMI days, with the exception being JFK. On the other hand, we also found that for either distance metric, the days that are very similar with respect to demand and capacity can have very different amounts of TMI activity. This is an important area for future research.

The results of this study also indicate that the identification of similar historical days using continuous measures of similarity may be a viable approach to guiding TMI decision making on a given day of operation. However, much more work is required before this approach can be implemented. First, while demand can be known beforehand on a day-of-operation, capacity can only be predicted. Thus it will be necessary to base capacity similarity on a forecast of capacity as opposed to actual capacity values. Further, we need to identify not just what TMI actions were
taken on historical days, but also how well they worked. Finally, our results suggest that other features of the day besides demand and capacity drive TMI activity. These features must be considered in assessing similarity between days in future work.
3. Capacity Prediction

3.1. Introduction

The previous chapters in this dissertation have discussed how the economic contribution of the aviation system depends on the capacities of aviation infrastructure, with airports—in particular airfields—often being the weakest links. As discussed in the previous chapters, growing demand juxtaposed with limited airport capacity is resulting in delays and billions of dollars of economic loss (Eno Center for Transportation, 2013). Numerous traffic management initiatives (TMI) are aimed at creating a balance between the demand and the capacity at an airport. The NextGen (Next Generation Air Transportation System) initiative by FAA (Federal Aviation Administration) strives to increase capacity, save time and fuel, obviate traffic delays and assist controllers in managing aircraft with improved safety standards.

Chapter 2 discusses the estimation of similarity between two days at an airport based on their capacity and demand data. Such endeavors require reliable estimates of demand and capacity variables. The Aggregate Demand List (ADL) dataset, which contains fine-grained information on the flight schedules of all the flights operating at an airport, has made it possible to compute very accurate demand information for airports. One of the limitations discussed in chapter 2 is that the AARs, which are used to represent airport capacity, may be unreliable. It is thus important to measure, model, and predict airport capacity for developing accurate capacity based similarities. Modeling airport capacity will also better enable maximizing the utilization of existing capacity and assessing investments and technologies for improving it.

Analyses of airport operations (Newell, 1979; Venkatakrishnan et al., 1993) indicate that the interaction of factors such as separation rules, fleet mix, weather conditions and human performance largely determine the capacity of an airport. Numerous models with varying degrees of complexity have been developed to estimate airport capacity. Some commercially used capacity models include the FAA Airfield Capacity Model, Airport and Airspace Simulation Model (SIMMOD) and Total Airspace and Airport Modeler (TAAM) (LeighFisher (Firm), et al., 2012), all of which are based on rules that specify minimum time and distance separations of the airport. Several empirical studies of airport capacity have focused on the acceptance rate (AAR), which specifies the number of arrivals that controllers believe an airport can accept in given time period (i.e. the airport arrival capacity) based on operating conditions and past experience (FAA, 2010). FAA archives AAR data for major airports in the US at hourly or quarter-hourly time scales. Recent work (Wang, 2011) has used AAR data to build machine-learning models such as Support Vector
Machine (SVM) and Bagging Decision Tree (BDT). Other studies (Provan et al., 2011; Cox and Kochenderfer 2016) used Markov chain, Hidden Markov Model and regression trees to estimate AAR distributions using observed and forecasted weather and AAR data. Other studies (Liu et al., 2008; Buxi and Hansen, 2011) used clustering technique and scenario trees to estimate capacity profiles using AAR and weather data.

The AARs may not be updated regularly to reflect changing conditions and is based on human judgment, even if it is informed by considerable experience. This human subjectivity, combined with the higher priority controllers give to working traffic compared to recording data, may sometimes make the recorded AARs unreliable, reducing the quality of the models derived from them. Further, there is considerable evidence that, even controlling for operating conditions, airport capacity is quite variable, making it problematic to characterize it with a single AAR value (Hansen, 2004; Kim and Hansen, 2013). These factors limit the value of capacity prediction methods calibrated on AAR data archives.

This study develops airport capacity estimates without relying on AAR data. Our starting point is that airfield capacity is, by definition, the maximum number of operations—arrivals, departures, or the sum of the two—that an airport can handle in a given time period. Therefore, capacity of an airport can be directly observed only at sufficiently large demand. If the throughput of an airport is limited by the demand, we can only conclude that the capacity is larger than or equal to the observed throughput. This inability to directly observe capacity makes it difficult to develop capacity prediction models derived from historical data. In this study we explore methods that incorporate observations censored by insufficient demand to develop capacity prediction models. Specifically we estimate Kaplan Meier estimator, Cox Proportional Hazards and Random Survival Forest models to predict the capacity distribution of an hour (rather than a single capacity value) using historical weather, demand, and throughput data. This approach ensures that the capacity predictions are free from the limitations of the studies based on AAR data as discussed before.

A primary motivation for this research is the need for capacity predictions in strategic air traffic flow management. In this setting, traffic management specialists attempt to balance airport and airspace demand and capacity by implementing traffic manage initiatives (TMIs). The most common TMI is the ground delay program (GDP), in which flights are delayed at their origin airports in order to meter flow into a destination airport (Liu and Hansen, 2013). Inaccurate capacity predictions will result in either airborne holding (or even diversions to other airports), or unnecessary ground delay. With the capacity predictions from this study, specialists can set metering rates that make efficient use of available capacity while minimizing the risk of airborne holding and diversions. Accurate and data driven capacity estimates from this work would also better facilitate numerous decision-support endeavors (Bloem and Bambos, 2015; Kulkarni et al., 2013) which
use machine learning methods to guide TMI decision making, or even help flight operators to predict these decisions.

In this study, we develop an airport capacity prediction model that incorporates demand-censored observations using weather, demand and throughput data. The capacity model is estimated using EWR airport data from 2012-2014 (3 years) and predicts the capacity distribution for an hour of interest. We also discuss validation measures that incorporate censored observations to evaluate the capacity model. The two main contributions of this study are 1) using machine-learning techniques to train and predict airport capacity based on demand-censored observations rather than AAR data 2) developing an evaluation framework to validate the capacity model despite our inability to directly observe capacity distributions. This chapter is organized as follows: Section 3.2 reviews past research on airport capacity and details the censored models proposed in this research. Section 3.3 describes the model evaluation framework for the models discussed in section 3.2. Section 3.4 describes the source and processing of different variables used in model estimation. Sections 3.5 and 3.7 apply the methodologies detailed in sections 3.2 and 3.3 on EWR airport data to develop and validate three capacity prediction models. Section 3.6 discusses a case study where the capacity distributions are predicted for test data. Section 3.8 summarizes conclusions from the study and opportunities for further research.

3.2. Censored Data Analysis

As noted above, for purposes of this research we define capacity as the maximum number of flight arrivals, departures, or operations that can occur at an airport in a given time period. In other words, if we assume a sufficient demand of flights wanting to arrive or depart at an airport over some time period, capacity is the number of flights that can actually operate at the airport in that time period. The throughput is defined as the number of flights that operate at an airport in a given interval of time, with throughput being the minimum of the demand and the capacity for that interval. In this study, unless otherwise specified, capacity refers to arrival capacity and the interval of interest is an hour. The relationship between the capacity, demand and throughput is stated in equation 3.1 below:

\[ q_a(h) = \min\left(c_a(h), d_a(h)\right) \]  

where:

- \( c_a(h) \) is arrival capacity in hour \( h \)
- \( q_a(h) \) is arrival throughput in hour \( h \)
- \( d_a(h) \) is arrival demand in hour \( h \)
Equations of the form of 3.1 have been modeled using censored regression. In these models, in addition to 3.1, we assume:

\[
c_a(h) = \alpha_a + X_a(h)\beta_a + \varepsilon_{a,h} \sim N(0, \sigma_a^2)
\]

where in addition to the previously defined variables,

- \(\alpha_a\) is capacity intercept
- \(\beta_a\) is column vector of coefficients for variables in \(X_a(h)\)
- \(X_a(h)\) is row vector containing the variables of interest for a given hour \(h\)
- \(\varepsilon_{a,h}\) is a random variable with a truncated normal distribution with mode 0 and variance \(\sigma_a^2\). The random variable is truncated so that the capacity variable is always non-negative

For example, Hansen, 2004 employed censored regression to study the impact of a new runway on airport capacity at Detroit. Kim and Hansen, 2010 shows how capacity predictions from conventional models based on simulation, separation requirements, and physical laws vary in range from empirically observed capacities using a censored regression framework. The censored regression model assumes that capacity is normally distributed, which may not be true for capacity at all airports. The linear specification in equation 3.2 cannot readily capture threshold, non-linear, and interaction effects.

Censored data has been extensively analyzed in bioinformatics, studying time duration until the occurrence of an event and in this context is broadly referred to as ‘survival analysis’. In bioinformatics, the ‘event’ varies from incidence of a disease to occurrence of cancer or death. The time till occurrence of these events provides important measures such as mortality rate of a disease or success rates of new drugs. Some methodologies used for censored data analysis include the Kaplan Meier estimator (Kaplan and Meier, 1958), Cox Regression (Cox, 1972) and Random Survival Forest (Ishwaran et al., 2008). In most applications of these survival analysis models, the time till event is the censored observation, which is recorded only if the duration of the study is sufficiently large.

Censored data analysis is not limited to bioinformatics and is also widely used in areas such as infrastructure maintenance (Prozzi and Madanat, 2000) where one application is analyzing the life of a pavement. This literature is commonly referred to as reliability analysis and the models analyzing the time to infrastructure failure are called duration models. There has also been some recent work (Billon et al., 2005; Li and Laurence, 2015) on freeway capacity estimation where the Kaplan Meier estimator is used to study traffic breakdown and capacity distributions.
However, advances in censored data analysis have not yet been implemented into airport capacity prediction where a ‘breakdown’ cannot be directly observed due to safety constrains. The parallel problem setup in survival and reliability analyses motivate us to use censored data analysis to model airport capacity.

In this section, we detail the framework of some important censored data models applied to the airport capacity context. The relationship between capacity (censored data), demand (censoring variable) and throughput data is captured in equation 3.1. It is important to first define what an observation is and when it is censored. In airport capacity analysis, the event of interest is the airport reaching capacity. An observation corresponds to an hour and is censored when the event is not observed, i.e., there is not sufficient demand to push the airport to its capacity. Mathematically, the event status for arrivals ($\delta_a(h)$) for an observation $h$ is defined in equation 3.3 using the capacity variable and then in terms of observable variables (demand and throughput).

$$\delta_a(h) = 1(c_a(h) < d_a(h)) = 1(q_a(h) < d_a(h))$$  \hspace{1cm} (3.3)

where in addition to the previously defined variables:

- $\delta_a(h)$ is event status for arrivals at hour $h$
- $1(\cdot)$ is the indicator function (=1 is its argument is true and 0 otherwise)

A value of 1 for $\delta_a(h)$ indicates that the event (reaching capacity) occurred at hour $h$ and the observation is uncensored. In this case the observed throughput is the capacity. A value of 0 indicates that at hour $h$ the airport is not at capacity and the observation is censored. Thus, when $\delta_a(h) = 0$ the observed throughput is only the lower bound of the capacity. In the above equations and throughout this paper, the subscript $a$ indicates that our models and estimates are for airport arrivals.

The application of bioinformatics models to airport capacity analysis requires translation from the time domain to the count domain. While this is straightforward mathematically, the terminology can be confusing as a result of model origins. Before we proceed to discussing individual models in detail, Table 3.1 clarifies the analogues between survival and reliability models based on censored data on the one hand, and airport (or some other transportation facility) capacity models on the other. Table 3.2 defines some important functions that will be used in detailing the censored model frameworks. Here, $C$ refers to the capacity random variable and $q$ refers to a throughput value and is a scalar.
### Table 3.1: Applying censored variable analysis to airport capacity

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Survival Analysis</th>
<th>Reliability Analysis</th>
<th>Capacity Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit of observation</td>
<td>Time to event</td>
<td>Throughput to event</td>
<td></td>
</tr>
<tr>
<td>Event</td>
<td>Death/disease</td>
<td>Infrastructure failure</td>
<td>Reaching Capacity</td>
</tr>
<tr>
<td>Censoring variable</td>
<td>Study duration</td>
<td>Demand</td>
<td></td>
</tr>
<tr>
<td>Observable variables</td>
<td>Time, event status</td>
<td>Throughput, event status</td>
<td></td>
</tr>
<tr>
<td>Event = 1</td>
<td>Time to event &lt; study duration</td>
<td>Capacity &lt; demand</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Observed time &lt; study duration</td>
<td>Throughput &lt; demand</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3.2: Important functions in censored variable analysis

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulative distribution</td>
<td>( F_C(q) = P(C \leq q) )</td>
</tr>
<tr>
<td>Survival function</td>
<td>( S_C(q) = P[C &gt; q] = 1 - F_C(q) )</td>
</tr>
<tr>
<td>Hazard function</td>
<td>( z(q) = \lim_{\Delta \to 0} \frac{P[q &lt; C &lt; q + \Delta</td>
</tr>
<tr>
<td>Cumulative Hazard function</td>
<td>( H(q) = \int_0^q z(u)du = - \ln[1 - F_C(q)] = - \ln S_C(q) )</td>
</tr>
</tbody>
</table>

The survival function estimates the probability that the capacity for that observation is more than some specified value \( q \). We can directly see from the definition that \( S(0) = 1 \) and \( S(\infty) = 0 \). The hazard function estimates the probability of the observation reaching capacity between \( q \) and \( q + \Delta \), given its capacity is not saturated at a throughput less than \( q \). Intuitively, the cumulative hazard function is the expected number of times an hourly capacity is saturated as throughput increases from 0 to \( q \), if saturation were assumed to be a repeatable occurrence. Since saturation can only occur once in a given hour, the cumulative hazard is not very interpretable in a practical sense. Its relation with the hazard function and the cumulative probability distribution function makes it a very useful in survival modeling.
3.2.1. Kaplan–Meier estimator

The Kaplan Meier (KM) estimator, also known as the product limit estimator, is used to estimate the airport capacity distribution from throughput and event status data. The KM estimator is non-parametric and provides an empirical estimate of the overall capacity distribution at an airport. The KM estimator is derived from the following empirical functions of observed throughput $q$.

$s(q)$ Number of observations at capacity for a throughput value of $q$ (i.e. $q(h) < q$ and $d(h) \geq q$).

$r(q)$ Number of observations for which throughput is greater than or equal to $q$ (i.e. which satisfy $q(h) \geq q$). This includes observations for which both demand and capacity are greater than or equal to $q$ (i.e. $d(h) \geq q$ and $c(h) \geq q$).

We sort observed distinct throughput in increasing order ($q_1, q_2, ..., q_n$), where $n$ is the number of distinct throughput values and $q_j$ is the largest throughput value $\leq q$. The Kaplan-Meier estimate for arrival capacity distribution, $\widehat{F}_c(q)$ is given by:

$$\widehat{F}_c(q) = 1 - \prod_{i: q_i \leq q} \left(1 - \frac{s(q_i)}{r(q_i)} \right) = 1 - \left(1 - \frac{s(q_1)}{r(q_1)} \right) \cdots \left(1 - \frac{s(q_j)}{r(q_j)} \right)$$ (3.4)

See Kaplan and Meier, 1958 for more details on the theory and discussion around the KM estimator. In our context, the KM estimator yields the unconditional distribution of capacity at an airport for all of the hours included in the data set. One major drawback of the KM estimator is that it doesn’t directly account for covariates in its estimation. Covariates can be incorporated into the estimation by portioning data based on covariates and separately calculating the KM estimate for these groups. Thus, we can use the KM estimator to compare the capacity distribution of groups of observations. With a large number of covariates however, such grouping based on variables becomes tedious and impractical, especially given that numerous factors influence airport capacity. This motivates us to explore other models that incorporate covariates in capacity estimation.

3.2.2. Cox Proportional Hazards Model

The Cox model studies the impact of different covariates on airport capacity by assuming that the hazard for any observation is proportional to the hazard for any other observation. The Cox model specification is similar to logistic regression
models, with hazard ratio being used instead of odds ratio. Let \((X(1), X(2))\) represent the covariates of hours 1 and 2. The hazard ratio is given by \(\frac{z(q, X(1))}{z(q, X(2))}\). The hazard function can be specified as required by the data or prior knowledge. The commonly used specification for the hazard function in Cox model is shown below:

\[
    z(q, X(h)) = z_b(q) \exp(X(h)\beta) \tag{3.5}
\]

where \(z_b(q)\) is the baseline hazard function and \(z(q, X(h))\) is the hazard function. The baseline hazard function is the hazard for \(X(h) = 0\), without any specific assumptions made about its form. The baseline hazard function serves as a reference of the hazard, with \(\exp(X(h)\beta)\) being the relative risk of observation \(X(h)\) with respect to the baseline. Further details on the estimation of a Cox model can be found in Cox, 1972.

A partial likelihood estimation method is used to estimate the parameters of interest \(\beta\) without estimating the baseline hazard function in the Cox model. For this reason, the Cox model is often referred to as a semi-parametric method. The major limitation of the Cox model is that, like other parametric models, the fit depends heavily on the accuracy of its specification. Also, the assumption that the hazard for any observation is proportional to the hazard for any other observation may not be valid for airport capacity modeling. Thus, we explore non-parametric models where we can capture the effect of covariates without a pre-determined specification.

### 3.2.3. Random Survival Forest Model

The Random Survival Forest (RSF) model for the analysis of right-censored data allows covariates but does not require assumptions such as proportional hazards and does not necessitate a linear model specification. The RSF model addresses the limitations of the KM estimator and Cox model and is robust to data perturbations with its bootstrapped ensemble methodology. The feature selection incorporated into the RSF algorithm helps separate signal from noise when estimating capacity. Ishwaran et al., 2008 discusses the framework of RSF model and details its estimation and implementation. The algorithm for estimating the Random Survival Forest model is briefly explained below:

1) Draw \(B\) bootstrap samples (sample with replacement), where \(B\) is the number of trees in the model. Each sample on average consists of approximately 63% of the data, and the remaining 37% data is used for out-of-bag or OOB estimates. On each sample, grow a survival tree where the node is split to maximize difference between the capacity distributions of observations in the daughter nodes using the log-rank statistic.
2) Grow each tree until specified maximum depth is reached.

3) Let $q_{1,t} < q_{2,t} < \ldots < q_{n(t),t}$ be the $n(t)$ distinct throughput values for observations at capacity in terminal node $t$. The Nelson–Aalen CHF estimate for node $t$, $\hat{H}_t(q) = \sum_{q_{lt,t}} s_l q (\frac{n(\ell)}{r_{lt,t}})$, where $l$ is the observation number in the sorted list of throughput values and $t$ is the terminal node number into which the observation falls. $s_{l,t}$ is the number of observations in node $t$ that are at capacity for a throughput of $q_{lt,t}$. $r_{lt,t}$ is the number of observations that are not yet saturated or censored for a throughput less than $q_{lt,t}$. All the observations within the node $t$ are assumed to have the same CHF.

4) Calculate the cumulative hazard function (CHF) for each tree and obtain the ensemble CHF.

Once the forest is built, for observation $h$, $H(q|X(h)) = \hat{H}_h(q)$, if $X(h) \in t$ (falls into node $t$). Thus, we can find the CHF for any observation $h$, by running $X(h)$ down the tree and using the CHF estimate for terminal node into which the observation falls. Let $H_b(q|X(h))$ denote the CHF for observation $h$ estimated using a tree grown from $b^{th}$ sample out of $B$ bootstrap samples. The ensemble CHF, $H_e(q|X(h))$ is estimated by averaging over all $B$ bootstrap samples: $H_e(q|X(h)) = \frac{1}{B} \sum_{b=1}^B H_b(q|X(h))$. We can obtain the capacity distribution from the RSF model using the relation between the ensemble CHF and the CDF, $F_C(q|X(h)) = 1 - e^{-H_e(q|X(h))}$.

Step 1 of the RSF algorithm discussed above is performed using the Log-Rank splitting rule (LeBlanc and Crowley, 1993), which is the basis for splitting the observations at a node in the tree into two groups (daughter nodes). The log-rank test statistic, also known the Mantel-Cox test statistic, is a non-parametric test that compares capacity distribution of two groups. The estimation of the log-rank statistic between two groups of observations is explained in the steps below:

1) Let $q_1, \ldots, q_k$ be $k$ distinct throughput values where the observation was at capacity in either of the two groups. For each throughput value $q_j$, let $r_{1j}, r_{2j}$ be the number of observations in each group at risk, i.e., neither at capacity nor censored for a throughput less than $q_j$. Let $s_{1j}, s_{2j}$ be the number of observations in each group respectively that reach capacity at throughput $q_j$.

2) Under the null hypothesis that the two groups have identical capacity distribution, the variable $r_{ij}$ follows hypergeometric distribution with parameters $r_j, r_{1j}, s_j$, where $r_j = r_{1j} + r_{2j}$ and $s_j = s_{1j} + s_{2j}$.
3) For this distribution, the expectations is, \( E_{1j} = \frac{s_j}{r_j} r_{1j} \) and variance, \( V_j = \frac{s_j (r_{1j}/r_j)(1-r_{1j}/r_j)(r_j-s_j)}{r_j^{-1}} \).

4) Each \( s_{1j} \) is compared to its expectation \( E_{1j} \) in the log-rank test statistic, \( Z = \frac{\sum_{i=1}^{r_j} (s_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^{r_j} V_j}} \). The square of log-rank test statistic follows the chi-squared distribution, \( Z^2 \sim \chi^2 \). The log-rank splitting rule thus leads to a node split that maximizes the log-rank test statistic. This leads to the formation of daughter nodes with the highest difference in capacity behavior.

### 3.3. Model Evaluation

It is important to compare the different capacity models using validation measures that account for censored observations at an airport. One of the biggest challenges to validating capacity prediction models is that airport capacity is not directly observable. In this section, we present some validation measures from the censored model literature (Müller, 2004; Steyerberg et al., 2010) that are applicable for capacity analysis. We also develop validation metric based on observed throughput. In this chapter, all models are trained on 80% of the data and evaluated on 20% of holdout (out-of-sample) data. The validation measures used in this chapter are detailed below:

**Integrated Brier Score:** For models where predictions are in the form of probabilities of events, it is useful to measure the accuracy of these predictions. For example, suppose that for a given hour the predicted capacity CDF for a throughput of 30 is 0.95. If the observation were uncensored by demand it is highly likely that the throughput would be less than 30. Likewise, if the CDF for throughput of 10 is 0.05, it is highly likely that a demand-uncensored observation would have a throughput greater than 10. Thus, a better model will have a smaller score defined by: \( BS(q) = \sum_{h=1}^{N} w(h) (i(h) - \hat{F}(q|X(h)))^2 \), where \( i(h) = 1(q(h) \leq q) = 1 \) if the observation is at capacity, 0 otherwise. \( \hat{F}(q|X(h)) \) is the predicted cumulative probability of capacity at throughput \( q \) for an observation with covariates \( X(h) \). Censored observations with throughput less than \( q \) do not have an estimable capacity status \( i(h) \) and thus cannot contribute to the BS estimation. We have to account for this information loss from censored observations when estimating the BS. Assuming that covariates play no role in the censoring behavior of the data, the contributions of observations to the IBS are reweighted using weights \( w(h) \) based on the Kaplan Meier estimator for the censoring distribution. Graf et al., 1999 discusses in detail how these weights are estimated.
The Brier score is estimated at every unique observed throughput value \( q \) and measures the weighted mean squared error (MSE) between the capacity status of the observation and its predicted CDF at that throughput. We average the Brier score over all throughput values to get the integrated Brier score, \( IBS = \frac{1}{\max(q(h))} \sum_{0}^{\max(q(h))} BS(q) \). A lower Brier score implies a better predictive model. IBS score ranges from 0 for a perfect model and 0.25 for a non-informative model. Thus, when validating a model, a score closer to 0 indicates a better model.

**Concordance Index (C-Index):** A unit-less measure of the discriminative ability of a survival model is the rank correlation between the predicted capacity and the observed capacity of the data. For a randomly selected pair of observations, the C-Index measures the probability that the observation with the lower predicted capacity has a lower observed capacity. It can be thought of as classification accuracy for censored data. An advantage of the C-Index is that its estimation is not tied to a specific throughput like the Brier Score and also accounts for the censoring in the capacity data. For estimating the C-Index, it is necessary to have a predicted outcome for every observation that we can compare with its observed status. The expected predicted capacity aggregates the information contained in the cumulative capacity distribution for an observation. The discrete probability function is estimated from the cumulative distribution and the expected predicted capacity for an observation is estimated as \( E(c(h) | X(h)) = \sum_{i=1}^{n} q_i \cdot p(q_i | X(h)) \), where, \( p(q_i | X(h)) \) is the predicted probability mass function for an observation \( X(h) \) at \( q_i \) and \( n \) is the number of distinct throughput values. We refer to \( E(c(h) | X(h)) \) as the predicted expected capacity for an observation \( h \) when detailing the C-Index calculation. Heagerty and Zheng, 2005 shows that C-Index is related to the area under the Receiver Operating Curve (ROC). A value of \( c = 0.5 \) implies that the model has no predictive ability and value of \( c = 1 \) implies that the model provides perfect predictions. Ishwaran et al., 2008 details the steps for calculating the C-Index using right-censored data. The C-Index captures the discriminative power of the model and we first need to create all possible pairs of the data for estimating it. This entails the following steps.

1) Identify Permissible pairs: We first need to find the pairs of observations which can be ordered based on prediction and outcome to make a valid comparison. For this, we omit the pairs of data where the smaller throughput observation is censored. We also omit any pair of observations \( i, j \), if \( q(i) = q(j) \), unless at least one of the observations is at capacity. These two filters leave us with comparable data, which we call the permissible pairs. The total number of permissible pairs is represented by the ‘Permissible’ variable.

2) Concordant pairs: The prediction model can discriminate well if the observation with the lower predicted expected capacity has a lower observed capacity. Note that in the permissible pairs, the observation with the smaller
throughput is always at-capacity and in the case of pairs with equal throughputs, at least one observation is at-capacity. For each valid pair with unequal throughputs, the model discriminates correctly if the observation with the smaller throughput (which is at capacity) has the lower predicted expected capacity. However, if the predicted expected capacities are equal, the model is non-informative. If the model predicts a lower predicted expected capacity for the observation with the higher throughput, then the model is wrong. Keeping this in mind, count 1 if the smaller throughput has lesser predicted expected capacity and count 0.5 if the observations have equal predicted capacities. If the throughputs are equal and both the observations are at capacity then the model performs well if it predicts the same predicted expected capacity for both the observations, otherwise, the model is non-informative. To reflect this, for at-capacity pairs with tied throughputs, count 1 if the predicted expected capacities are equal and 0.5 otherwise. For the case where throughputs are equal but one of the observations is not at capacity, the model performs well if the predicted expected capacity for the at-capacity observation is lower and is non-informative otherwise. Thus, for the case with tied throughputs, where one is censored and the other at capacity, count 1 if the observation at capacity has a smaller predicted expected capacity and 0.5 otherwise. The sum of the counts over all permissible pairs is defined as the ‘Concordance’ variable.

The Concordance index (C-Index) is given by $C = \frac{\text{Concordance}}{\text{Permissible}}$ and gives a measure of the discriminative accuracy of the survival model. Accurate models will have a C-index close to 1 and non-informative models will have C-index close to 0.5. C-index less than 0.5 means that the model is performing worse than random predictions.

**Predicted Throughput:** The frequently used validation measures for censored data models have been developed and applied broadly in the areas of bioinformatics and infrastructure maintenance. The validation measures discussed so far in this section 1) compare at-capacity status of an observation with its predicted CDF and 2) capture the discriminative ability of the model in predicting which observation has higher capacity. A third approach is to compare predicted and observed throughput. As shown in equation 3.1, the throughput for an observation is the minimum of its capacity and demand. The models discussed in section 3.2 predict the capacity distribution for a given hour. Given the demand, it is thus possible to predict the throughput distribution for that hour using equation 3.1. The expectation of the predicted throughput distribution can then be compared with the observed throughput for all observations to provide validation measures such as the coefficient of determination, $R^2$, and the Root Mean Squared Error (RMSE) as shown in equation 3.6 below:

\[
R^2 = 1 - \frac{\sum_{n=1}^{N}(E(\min(c(h), d(h))) - q(h))^2}{\sum_{h=1}^{N}(q(h) - \bar{q})^2}
\] (3.6)
$$RMSE = \sqrt{\frac{\sum_{h=1}^{N} (E(\min(c(h), d(h))) - q(h))^2}{N}}$$

where, $c(h), d(h), q(h)$ represent the capacity, demand and throughput for observation $h$, $N$ is the number of observations and $\bar{q} = \frac{\sum_{h=1}^{N} q(h)}{N}$ is the average throughput. $E(\min(c(h), d(h)))$ represents the expectation of the predicted throughput distribution. These validation measures are more clearly tied to an observable variable—throughput. We then compare the performance of throughput prediction derived from capacity models against the throughput predictions derived directly from trained throughput models. A good capacity model should provide a more accurate estimate of the throughput at an airport than a model trained to predict throughput directly on the observed throughput and covariates.

### 3.4. Data Preparation

We apply the models discussed in section 3.2 on EWR airport data and then use the validation measures in section 3.3 to evaluate the estimated models. First, we describe each component of the input data and the processing needed in more detail.

**Weather data:** Observed weather data is stored in the METAR database (Aviation Routine Weather Report) and is highly standardized and followed consistently throughout the world. The Performance Data Analysis and Reporting System (PDARS) processes the raw semi-structured METAR data to derive information on different weather variables. Every hour or when there are changes in the weather, the weather data at the airport is provided. In cases when weather changes over the hour, multiple observations are provided for one hour. In such cases, if there was snow (or thunderstorm) in any one of the observations for an hour, we record snow (or thunderstorm) for that hour. For other weather variables, we choose the weather data that is observed for the longest part of the hour as the final weather data for that hour. In our study, we use the ceiling, visibility, wind speed, wind direction, temperature, snow and lightning data thus processed.

**Throughput:** Terminal Radar Approach Control (TRACON) data contains flight track points and flight plans from surveillance radars. The PDARS system synthesizes these TRACON data into flight trajectories. During this process, additional information such as airports and runway threshold crossing are computed based on these trajectories. The runway threshold crossing time is used to compute airport throughput.
Scheduled demand: The runways at an airport are utilized by both arriving and departing flights. Thus, the capacity of an airport is shared between arrivals and departures. Scheduled arrival demand provides information on how much of the airport capacity is planned to be used by arrivals. Thus, scheduled arrival demand is important in modeling arrival capacity. Also, during times of higher scheduled demand, controllers have a higher pressure to utilize the airports efficiently, which may influence the maximum number of flights that can land at an airport. The demand that can be expected at an airport before the the flight plans incorporate impacts of changing weather or traffic management actions is defined as the scheduled demand. A day in aviation need not coincide with the midnight-midnight definition we use everyday, but depends on when the airport activity restarts. Discussions with aviation subject matter experts indicates 4 AM – 4 AM (next day) is an appropriate definition of a day at an airport.

We use the ADL (Aggregate Demand List) dataset to provide information on flight schedules. The ADL dataset contains files generated at very high frequency (5 minutes or less) with each file containing information of flights arriving and departing at an airport between an hour before and 36 hours after the time the file is generated. We assume that the flight schedules are unaffected by Traffic Management Initiatives (TMIs) at 4 AM in the morning, local time. Therefore, the ADL file at 4 AM local time is used to measure the scheduled demand for the next 24 hours (until 4 AM next day). We estimate scheduled arrival and departure demands by aggregating runway wheels-on/wheels-off time into hourly bins.

Fleet mix: Different weight classes of flights arrive at the airport in a given time interval. Aircraft weight classes are classified into three categories: heavy, large and small. A heavier aircraft creates more wake turbulence behind it, which can destabilize aircrafts too close to it, especially smaller aircrafts. The weight classes of the aircrafts determine the minimum safe separation necessary between them during different stages of flight. This is especially important in landing and take-offs when flights can be in close proximity of each other. If a small aircraft is scheduled to land at an airport behind a heavy aircraft, then a larger separation is needed near the runway thresholds than if a heavy aircraft is scheduled behind a small aircraft. The number of flights that can arrive at an airport in an hour are thus affected by the mix of the weight classes of flights arriving in that hour. The ADL dataset contains the aircraft weight class for each flight and are aggregated hourly to generate estimates of the proportion of heavy aircrafts and proportion of small aircrafts in that hour. To capture possible interactions between the number of heavy and small aircrafts, we also consider the product of the proportion of heavy and small aircrafts in that hour.

Convective weather: A binary SVM classifier is trained on the National Convective Weather Forecast (NCWF) data. The input convective weather data is transformed into a vector of indicator variables denoted by \( x \) signifying the presence of convective weather in different parts on the airspace. We use GDP occurrence
indicators, which indicates if a GDP is in effect in an airport for a given hour as the GDP label for that hour. The GDP label data is used as the output variable to train the weights of different parts of airspace. Liu et al., 2017 estimates the vector of weights $\mathbf{w}$ for the airspace around EWR airport and an offset $b$ specific to the GDP label data at EWR. A higher weight $w$ for a given part of the airspace suggests that convective weather in that part has a higher impact on the incidence of a GDP. The bias term, $b$ is necessary to allow for a separating hyperplane with maximum margin that does not pass through the origin. The convective weather score $\mathbf{w}_x + b$ reflects the likelihood of GDP at EWR airport and is used to represent convective weather in capacity prediction.

**Event status:** The event status for every observation (hour) is an indicator variable to capture if the observation is at capacity or demand censored. We infer the event status using demand and throughput data and equation 3.3. The efficiency demand variable and the throughput variable from ASPM (Aviation System Performance Metrics) data are used in the event status estimation. The event status of an observation is 1 when its ASPM throughput is less than ASPM demand, indicating that the observation is at capacity and zero when throughput is equal to demand, indicating that the observation is demand censored. ASPM demand and throughput data are constructed so that throughput can never exceed demand. Note that the ASPM throughput and demand data are used only to determine event status. The throughput data and demand data used for model training come from different sources, as previously described.

There are two demand estimates in the ASPM data, one corresponding to System Airport Efficiency Rate (SAER) estimation and the other corresponding to the Terminal Arrival Efficiency Rate (TAER) estimation. The SAER is estimated to assess overall system efficiency in allowing the flights intending to arrive, to operate at the airport. The TAER is designed to measure TRACON efficiency. In the SAER estimation, a flight starts to contribute towards the airport arrival demand from the time of estimated wheels-on calculated at takeoff. The wheels-on time is estimated as the sum of actual wheels-off time and the filed en route time. The flight is included in the arrival demand until its actual wheels-on time. This demand will be referred to as the SAER demand and the corresponding estimated status will be referred to as SAER status. If, however, the flight is assigned a ground delay as the result of a TMI, the demand time begins at its estimated wheels-on time at take-off, minus the assigned ground delay—i.e. at the time it would have been expected to arrive without the TMI. Traditionally in the aviation literature, SAER demand is used to represent the number of flights intending to land at the airport. In TAER estimation, a flight starts contributing towards the airport demand from the wheels-on time estimated at the 100-mile crossing time. The start of demand time is the sum of three components: the 100-mile crossing time, estimated time from 100-mile boundary crossing point to the 40-mile point, and the unimpeded mean time from 40-mile point to wheels on. This flight contributes towards the demand until its actual wheels on time. This demand will be referred to as TAER demand and the...
corresponding estimated status will be referred to as TAER status. In this study, we estimate capacity models using both SAER and TAER status values and compare their results.

Before proceeding to applying the models discussed in section 3.2 to airport data, we present some summary statistics of the variables used to predict airport capacity in table 3.3. In this study, we use busy hours data 7-22 for EWR from 2012-2014 (3 years of data) for modeling airport capacity and generate some summary statistics below.

Table 3.3: Summary statistics of EWR hourly airport data, 2012-2014 (3 years)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceiling (100 feet)</td>
<td>1</td>
<td>300</td>
<td>76.67</td>
<td>50</td>
<td>76.98</td>
</tr>
<tr>
<td>Wind Direction (degrees)</td>
<td>0</td>
<td>360</td>
<td>194.15</td>
<td>220</td>
<td>109.84</td>
</tr>
<tr>
<td>Wind Speed (knots)</td>
<td>0</td>
<td>39</td>
<td>8.66</td>
<td>8</td>
<td>4.75</td>
</tr>
<tr>
<td>Visibility (statute miles)</td>
<td>0</td>
<td>10</td>
<td>9.21</td>
<td>10</td>
<td>2.15</td>
</tr>
<tr>
<td>Temperature (°C)</td>
<td>-16</td>
<td>39</td>
<td>14.61</td>
<td>16</td>
<td>10.47</td>
</tr>
<tr>
<td>Thunderstorm Indicator</td>
<td>0</td>
<td>1</td>
<td>0.01</td>
<td>0</td>
<td>0.11</td>
</tr>
<tr>
<td>Snow Indicator</td>
<td>0</td>
<td>1</td>
<td>0.02</td>
<td>0</td>
<td>0.14</td>
</tr>
<tr>
<td>Throughput</td>
<td>0</td>
<td>51</td>
<td>30.85</td>
<td>33</td>
<td>8.11</td>
</tr>
<tr>
<td>Scheduled Arrival Demand</td>
<td>0</td>
<td>58</td>
<td>32.24</td>
<td>33</td>
<td>9.15</td>
</tr>
<tr>
<td>VFR Indicator</td>
<td>0</td>
<td>1</td>
<td>0.93</td>
<td>1</td>
<td>0.26</td>
</tr>
<tr>
<td>Proportion of Heavy Aircrafts</td>
<td>0</td>
<td>1</td>
<td>0.08</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>Proportion of Small Aircrafts</td>
<td>0</td>
<td>1</td>
<td>0.007</td>
<td>0</td>
<td>0.018</td>
</tr>
<tr>
<td>Proportion of Heavy Aircrafts*</td>
<td>0</td>
<td>0.02</td>
<td>0.0005</td>
<td>0.00</td>
<td>0.002</td>
</tr>
<tr>
<td>Proportion of Small Aircrafts*</td>
<td>0</td>
<td>0.02</td>
<td>0.0005</td>
<td>0.00</td>
<td>0.002</td>
</tr>
<tr>
<td>Convective weather</td>
<td>-1.03</td>
<td>1.41</td>
<td>-0.53</td>
<td>-0.56</td>
<td>0.15</td>
</tr>
<tr>
<td>SAER Status</td>
<td>0</td>
<td>1</td>
<td>0.66</td>
<td>1</td>
<td>0.47</td>
</tr>
<tr>
<td>TAER Status</td>
<td>0</td>
<td>1</td>
<td>0.74</td>
<td>1</td>
<td>0.44</td>
</tr>
</tbody>
</table>
3.5. Results – SAER event status

We now apply the three models discussed in section 3.2 to EWR airport data using the SAER event status. In section 3.5, the word demand is used synonymously with SAER demand and event status is used synonymously with SAER event status. In the remaining paper, the word capacity is used synonymously with arrival capacity of the airport.

3.5.1. Kaplan–Meier estimator

In section 3.2, we detailed the estimation of the Kaplan-Meier curve. A single CDF is estimated for the entire data using throughput and event status data and does not incorporate covariates. The major advantage of using the KM estimator is that it is an empirical methodology that reflects the aggregate behavior of data well. However, the major disadvantage of the KM estimator is that it doesn’t control for covariates. For further analysis, we only consider data between in busy hours starting at 7 to 22 hours at EWR airport (Gorripaty and Hansen, 2017). Using the survival library (Therneau and Lumley, 2016) in R, we estimate the cumulative probability distribution for arrival capacity using data at EWR airport and plot it in Fig. 3.1. In Fig. 3.1, we observe that for throughput below 10, the probability of being at capacity is close to 0 and for throughput above 45, the probability is close to 1.

![Figure 3.1: CDF of arrival capacity, EWR airport](image)
The capacity distribution in Fig. 3.1 doesn’t capture the effects of covariates such as visibility and cloud ceiling on the arrival capacity distribution. To explore such effects with the KM estimator, we must disaggregate the data. As an illustration, we divide the observations into those for which the ceiling is greater than and less than or equal to 3000 feet and apply the KM estimator separately to each group. The results are shown in Fig. 3.2. The results clearly show that arrival capacities tend to be lower on the hours with low cloud ceiling. However, it is not possible to capture the combined effects of a large set of covariates using this approach.

![Figure 3.2: Comparing capacity CDF between ceiling-filtered data, EWR airport](image)

We next look at the performance of the KM estimator on test data using the validation techniques discussed in section 3.3. The Integrated Brier score (IBS) for the KM estimator is 0.076. The IBS score ranges between 0 for a perfect model and 0.25 for an uninformative model. The C-index is not applicable for the KM estimator, since it does not consider covariates and thus assumes that the capacity distribution is the same for every hour. Comparing the observed throughput with that predicted from the demand and capacity CDF obtained from the KM estimator, we find the RMSE to be 4.9 flights and $R^2 = 0.64$. Fig. 3.3 shows the scatter plot of the predicted and observed throughput values. Note that all of the variation in the predicted throughput values of Fig. 3.3 is derived from variation in demand, since the capacity distribution is assumed to be the same for all observations.
3.5.2. Cox Proportional Hazards Model

The Cox model has the advantage of controlling for covariates. We use the implementation of the Cox model in the survival package in R to estimate the coefficients for covariates. The model is specified in equation 3.7, where hazard \( z_a(q, X_a(h)) \) is a measure of the risk of being at capacity for observation \( X_a(h) \). The coefficients \( \beta_a \) estimated using EWR airport data are detailed in Table 3.4.

\[
z_a(q, X_a(h)) = z_{b,a}(q) \exp(X_a(h)\beta_a)
\]  

(3.7)

Table 3.4: Cox proportional hazards (PH) model estimation results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate, ( \beta )</th>
<th>Z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local Hour</td>
<td>-0.0606</td>
<td>-17.85</td>
</tr>
<tr>
<td>Ceiling (100 ft)</td>
<td>-0.0004</td>
<td>-2.62</td>
</tr>
<tr>
<td>Visibility (nm)</td>
<td>-0.119</td>
<td>-18.17</td>
</tr>
<tr>
<td>VFR Indicator</td>
<td>-0.371</td>
<td>-7.57</td>
</tr>
<tr>
<td>Temperature (°C)</td>
<td>-0.0109</td>
<td>-8.88</td>
</tr>
<tr>
<td>Wind Speed (knots)</td>
<td>0.036</td>
<td>12.81</td>
</tr>
<tr>
<td>Wind Direction</td>
<td>-0.006</td>
<td>-4.51</td>
</tr>
<tr>
<td>Scheduled Arrival Demand</td>
<td>-0.0495</td>
<td>-33.46</td>
</tr>
<tr>
<td>Thunderstorm Indicator</td>
<td>1.300</td>
<td>14.12</td>
</tr>
<tr>
<td>Snow Indicator</td>
<td>0.676</td>
<td>8.32</td>
</tr>
</tbody>
</table>
The coefficients in the Cox PH model have two important pieces of information, one being the magnitude of the coefficients and the other being the sign. It is important to note that the Cox PH model captures the influence of the covariates on the hazard/risk of an observation being at capacity. From the estimation results in Table 3.4, we observe that the local hour coefficient is negative, which indicates that the risk of being at capacity at EWR is lower at later hours of the day. This may arise from the local hour variable capturing information about the local weather patterns at the airport. The negative coefficients for ceiling and visibility indicate that lower ceiling and visibility (indicates bad weather conditions) increase an observation's risk of being at capacity. The VFR indicator coefficient being negative indicates that when the observation is at VFR, it has a lower risk of being at capacity, as is expected. The negative coefficient of temperature variable suggests that higher temperatures have a lesser hazard of being at capacity, which is expected for EWR airport owing to snow related capacity constraints. We also observe that more the wind speed, higher is the risk of being at capacity, which may reflect the destabilizing effect of high crosswind speeds. The negative coefficient for scheduled arrival demand indicates that for larger scheduled arrival demand, the airport has a lower hazard. This may reflect how higher arrival demand puts pressure to prioritize arrivals over departures. The snow and thunderstorm coefficients are positive which indicates that when an hour experiences snow or thunderstorm, the risk of being at capacity increases, as is expected.

The coefficients for proportion of heavy aircrafts and the product of proportion of small and heavy aircrafts have negative coefficients. This suggests that when these proportions are larger, there is a lower risk of being at capacity, which is a surprising result. The coefficient for proportion of small aircrafts is positive, which indicates that for a large proportion of small aircrafts, there is a higher risk of being at capacity, which is as expected. Though the sign of the coefficient for product of proportion of small and heavy aircrafts is surprising, the positive coefficient for the proportion of small aircraft may capture the larger separation needed for trailing small aircrafts, thus reducing the capacity of the airport. The coefficient for convective weather is positive, and indicate that higher convective weather scores increase the risk of being at capacity, which is as expected.

We explore the performance of the Cox model using the validation measures discussed in section 3.3. The model is evaluated on test data. The integrated Brier

<table>
<thead>
<tr>
<th>Proportion of Heavy Arriving Aircrafts</th>
<th>-2.97</th>
<th>-12.93</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of Small Arriving Aircrafts</td>
<td>3.21</td>
<td>3.14</td>
</tr>
<tr>
<td>Product of Proportion of Small &amp; Heavy Aircrafts</td>
<td>-38.6</td>
<td>-3.31</td>
</tr>
<tr>
<td>Convective Weather Score</td>
<td>0.772</td>
<td>10.46</td>
</tr>
</tbody>
</table>
score (IBS) for Cox model is 0.058. The estimated score of 0.058 for Cox model is closer to 0 as compared to Kaplan Meier estimator’s score of 0.076. This suggests that Cox model performs better than KM estimator based on IBS. The C-index that ranges between 1 for perfect model and 0.5 for non-informative model is estimated to be 0.72 for the Cox model. This indicates that the model has good discrimination power and compares about 72% of the pairs of observations accurately based on their capacity behavior.

We also estimate the throughput distribution for every observation in test data using equation 3.1. The Cox model predicted throughput is plotted against the observed throughput for all observations in the test data in Fig. 3.4. The coefficient of determination, $R^2$ between the observed and predicted throughputs is 0.74 and the square root of RMSE is 4.1. These values indicate an improved performance of the Cox model in predicting the throughput as compared to the KM estimator.

![Figure 3.4: Cox-predicted vs observed throughput](image)

3.5.3. Random Survival Forest

The Random Survival Forest (RSF) model is non-linear, non-parametric and controls for covariates, thus proving to be more applicable for capacity modeling than KM estimator and Cox model. Using randomForestSRC (Ishwaran and Kogalur, 2015) implementation in R, we estimate the RSF model for EWR airport. One of the limitations of non-parametric models such as RSF is that we do not obtain easily interpretable “coefficients” of the covariates when the model is trained. We can however quantify the importance of covariates in predicting the capacity for an observation. To understand the influence of different variables in predicting airport
capacity, we estimate the minimal depth of every variable from the root node, averaged over all the trees in the forest (Ishwaran et al., 2011). The minimal depth measures the importance of different variables in predicting airport capacity. The variable importance estimated using the minimal depth can also be used to guide variable selection in models with a large number of covariates.

Using the three model evaluation frameworks described in section 3.3, we find the optimal depth for the RSF model over a range of tree depths from 1 to 25. The IBS plotted against the tree depth is shown in Fig. 3.5 and indicates an optimal depth of 11 with the minimum IBS of 0.055. A lower IBS score closer to 0 indicates a better model with lesser deviation between the predicted capacity probabilities and the observed capacity status. The IBS for the best RSF model is lesser than IBS of 0.058 for Cox model and 0.076 for Kaplan Meier estimator. Based on IBS, the RSF model performs better than both Cox and KM estimator.

![Figure 3.5: IBS vs tree depth for RSF model](image)

As discussed in section 3.3, we use the predicted capacity CDFs on test data to estimate the C-Index. The CDFs are estimated using observations in the terminal nodes that the test observation falls into. When we train deeper trees, owing to more node splits, the number of observations in the terminal nodes decreases. In some cases, we may not have sufficient information in one terminal node to estimate a complete CDF and is one of the limitations of deeper trees. Thus, the CDFs estimated for some test observations using deeper trees may be incomplete and cannot be used to estimate their expected capacity. This problem also arises in
estimating the throughput CDFs used in calculating the $R^2$ measure. We explore the truncation in the predicted capacity CDFs and design a framework to address it.

First, we investigate the degree of incompleteness of the capacity CDFs predicted on the test data. We observe the last predicted cumulative probability beyond which the model cannot predict for a given observation, which is where the CDF got truncated. We then filter all observations where the last predicted cumulative probability is less than a given truncation filter. In Fig. 3.6 we plot the proportion of observations in the filtered subsample for different truncation filters varying from 0.5 to 0.95 as shown in the legend. As an example, we observe that approximately 90% of observations have predicted CDFs (using a trained RSF model of depth 25) where the last predicted cumulative probability is >0.5.

![Figure 3.6: Proportion of data above truncation filter](image)

Fig. 3.6 visualizes the trade-off made in choosing a truncation filter: choosing a small truncation filter, allows us to retain more data, but forces us to make assumptions on the behavior of a larger part of the predicted CDF. For a subsample with a higher truncation filter, we loose more data, but have more complete CDF information for the observations that remain. Thus, the optimal truncation filter should balance between loosing fewer observations and minimizing the extrapolation on the CDFs. The maximum capacity of the EWR airport is the highest observed throughput in the EWR airport data. When a CDF is truncated, we only have information on the capacity distribution until a certain value. Thus, for incomplete CDFs, we can extrapolate the CDF linearly between the last known value.
to a cumulative probability of 1 expected for the maximum capacity of the airport. As an example, Fig. 3.7 below shows an example where an observation's CDF is truncated at a value of 0.93 for a capacity of 38. The CDF is thus extrapolated linearly between 39 and a maximum capacity of 52.

![Figure 3.7: Example extrapolation of incomplete predicted CDF](image)

We filter the data such that our subsample only contains observations that are truncated above the value of the chosen truncation filter. We then extrapolate the CDFs to 1 as described in the above paragraph for the observations in the subsample. We then estimate the C Index using the extrapolated subsample of CDF data. The C-index, which evaluates the discriminatory power of the model, is estimated for different tree depths ranging from 1 to 25. C Index ranges between 1 for perfect model and 0.5 for non-informative model and the optimal RSF model has the highest C Index. Fig. 3.8 shows the C Index estimated on subsamples with varying truncation filters. We can observe that for smaller truncation filters, where we loose only some observations, the highest C Index is obtained for larger tree depths as compared to subsamples with higher truncation filters. Also, the highest C Index is also lower for subsamples with higher truncation filters.

From Fig. 3.6 we observe that for a truncation filter of 0.65, we retain more than 80% of the predicted CDFs across models with tree depths 1 to 25. We thus use a truncation filter of 0.65 and extrapolate the CDFs to 1 accordingly for estimating evaluation measures in the study. The plot in Fig. 3.9 shows the C Index plotted
against the tree depths and suggests that 11 is the optimal depth for the RSF model with a C Index of 0.77, which is greater than C-Index of 0.74 observed for Cox model.

Figure 3.8: C-Index computed for truncation-filtered data

Figure 3.9: C Index vs tree depth for RSF model
We now consider the third validation measure discussed in section 3.3 where the throughput distribution is estimated on test data using equation 3.1. The predicted throughput is compared against the observed throughput for the test data and the coefficient of determination $R^2$ is plotted for different tree depths of the RSF model. Fig. 3.10 plots $R^2$ vs tree depth and we observe that the optimal depth of 10 gives $R^2$ of 0.8 and RMSE of 3.6. However, we can observe that $R^2$ corresponding to a depth of 11 is very close to the highest value of 0.8. The RSF model predicted throughput is plotted against observed throughput for all observations in the test data in Fig. 3.11. The validation measures indicate an improved performance of the RSF model in predicting throughput as compared to Cox model and KM estimator.

![Throughput prediction R^2 vs tree depth](image)

Figure 3.10: Throughput prediction $R^2$ vs tree depth

We also predict the throughput at an airport by training a random forest (RF) model directly on the airport data described in Table 3.3. The highest $R^2$ obtained by the RF regression model is 0.64. When contrasted with highest $R^2$ of 0.8 as seen in Fig. 3.10, we can better appreciate the impact of explicitly representing airport capacity in predicting airport throughput. We can see that the quality of throughput predictions increases from $R^2$ of 0.64 to 0.8 by adding the capacity layer in throughput prediction.

The variable importance for the optimal RSF model (depth 11) with 200 trees, for EWR airport data is shown in Fig. 3.12. We observe that local hour is the most important variable in predicting the capacity at EWR airport, followed by scheduled arrival demand, ceiling, visibility and temperature. The high importance of weather and fleet mix variables in as expected. However, the highest importance
of local hour and scheduled arrival demand is surprising. While the physical mechanism is unclear, the results imply that these two factors should be considered in predicting airport capacity. As discussed in the estimation results of the Cox model in section 3.5.2, the influence of the local hour may arise from capturing local weather patterns at the airports. The airport infrastructure serves both arrivals and departures and thus the capacity is shared between them. Larger scheduled arrival demand may put pressure on the controllers to allow more arrivals, thus showing a higher variable importance in predicting the arrival capacity.

Figure 3.11: RSF-predicted vs observed throughput

Figure 3.12: RSF variable importance using minimal depth
3.6. Capacity Prediction Case Study

In this section we discuss a case study where we focus on a day of interest at the EWR airport and make predictions on its hourly capacity behavior using the optimal RSF model. For this case study, we consider August 13th, 2014 and we make an OOB prediction of the capacity CDF for all the busy hours (7-22) at EWR airport. The predicted CDFs for these hours are plotted in Fig. 3.13 and indicate: low capacity for hour 7 with improvements in capacity each hour until late afternoon (hour 16) followed by a decrease in the capacity until the end of the day (22 hour). The capacity at 22 is predicted to be higher than the capacity at 7.

![Arrival Capacity CDF Graph](image)

Figure 3.13: OOB arrival capacity CDF for 7-22 hours, 08/13/2014 at EWR airport

It would be useful to contrast these OOB predictions with the observed capacity of these hours at EWR airport. Since we do not have a ground truth for capacity at the airport, we visualize the weather around the airport and the arriving and departing flights at EWR on 08/13/2014 in Fig. 3.14. Each panel in Fig. 3.14 is a snapshot of the airport at certain hours starting from 7 to 22. For the sake of brevity, the snapshots of some hours are only shown. The yellow dot represents the EWR airport, the green arrows indicate arriving flights, red arrows indicate departing flights from the airport and the green area represents bad weather. The consecutive panels in Fig. 3.14 represent hours 7, 8, 10, 12, 14, 17, 18, 20, 22 hours respectively at EWR airport. At hours 7, 8 and 10, the airport is surrounded by large sections of bad weather and highly reduced flows. The next row shows three panels for hours 12, 14 and 17 during which the weather clears off and the flow near the airport increases.
increases. The hour 17 panel and the next three panels for hours 18, 20 and 22 show a new weather front approaching the airport from the left and shows how the flows decrease again. In the last panel at 22, we can see that the weather is directly at the airport further reducing the capacity, but depicts better conditions than at the beginning of the day. These visualizations in Fig. 3.13 and Fig. 3.14 illustrate how the RSF airport capacity model predictions for the day of interest reflect the evolution of weather and capacity for that day.

Figure 3.14: Weather over EWR airport on 08/13/2014 from 10-22 hours
3.7. Results – TAER event status

In section 3.5, we discussed the capacity models and model validation results estimated using SAER event status data. In this section, we discuss the capacity model estimation and validation results using TAER event status data. In this section, capacity models estimated using the SAER and TAER event status data will be referred to as SAER based models and TAER based models respectively. Again, the models in this section are estimated using EWR airport data from 2012-2014 (3 years) for arrivals. We will specifically focus on comparing the estimation results and model performance between the SAER and TAER based models.

3.7.1. Kaplan–Meier estimator

We estimate the cumulative probability distribution for arrival capacity using TAER event status and compare it to the capacity CDF estimated using SAER event status in Fig. 3.15. The TAER and SAER based CDFs are very similar, with the TAER based CDF indicating a slightly lower airport capacity. In TAER demand estimation, flights contribute to demand only after crossing the 100-mile radius of the airport. This provides a more accurate estimate of the actual number of flights physically near the airport that are intending to land. On the other hand, in the SAER demand estimation, the flight contributing to the demand may sometimes be physically away from the airport owing to factors such as en route airspace constraints or ground delay programs. Thus, the SAER demand data may over-estimate flights that are actually available to land in a given time period, which could lead to a lower capacity estimate. On the other hand the SAER adjusts demand counts so that any flight that is less than 15 minutes late only counts toward demand in the period it actually landed, while TAER sets this limit to be 6 minutes. It appears that the latter difference offsets the former one.

The integrated Brier score (IBS) for the TAER based KM estimator is 0.071 as compared to 0.076 for the SAER based KM estimator. Comparing the observed throughput with that predicted from the TAER demand and predicted capacity CDF obtained from the TAER based KM estimator, we find the RMSE to be 4.3 flights and $R^2 = 0.71$. This can be compared against RMSE of 4.92 and $R^2 = 0.64$ obtained for the SAER based KM estimator. The validation measures indicate the TAER based KM estimator is a better model than the SAER based KM estimator. Fig. 3.16 shows the scatter plot of the predicted and observed throughput values. Compared to Fig. 3.3 showing the observed vs predicted throughput derived from SAER data, we see less prediction error for observations with low observed throughput in Fig. 3.16.
3.7.2. Cox Proportional Hazards Model

As discussed in section 3.5.2, Cox model is specified as $z_a(q, X_a(h)) = z_{b,a}(q) \exp(X_a(h)\beta_a)$, where hazard $(z_a(q, X_a(h)))$ is a measure of the risk of being at capacity for observation $X_a(h)$. The coefficients $\beta_a$ of the TAER
based Cox model are detailed in Table 3.5 and are compared to the coefficients of the SAER based Cox model. The TAER based coefficients are comparable to the SAER based coefficients and can be interpreted similarly as described in section 3.5.2. The main difference in the estimation results is that the ceiling coefficient is insignificant and the local hour coefficient is two magnitudes larger in the TAER based model. One hypothesis to explain this result is that the local hour in the TAER model captures the local patterns of ceiling changes characteristic to the EWR airport, thus leading to an insignificant ceiling coefficient.

Table 3.5: Cox proportional hazards model estimation results (using TAER data)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate, $\beta$ (SAER)</th>
<th>Estimate, $\beta$ (TAER)</th>
<th>Z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local Hour</td>
<td>-0.0606</td>
<td>-9.23</td>
<td>-28.86</td>
</tr>
<tr>
<td>Ceiling (100 ft)</td>
<td>-0.0004</td>
<td>0.00007</td>
<td>0.45</td>
</tr>
<tr>
<td>Visibility (nm)</td>
<td>-0.119</td>
<td>-0.108</td>
<td>-16.73</td>
</tr>
<tr>
<td>VFR Indicator</td>
<td>-0.371</td>
<td>-0.309</td>
<td>-6.32</td>
</tr>
<tr>
<td>Temperature (°C)</td>
<td>-0.0109</td>
<td>-0.0162</td>
<td>-14.26</td>
</tr>
<tr>
<td>Wind Speed (knots)</td>
<td>0.036</td>
<td>0.028</td>
<td>10.59</td>
</tr>
<tr>
<td>Wind Direction</td>
<td>-0.0006</td>
<td>-0.0006</td>
<td>-5.63</td>
</tr>
<tr>
<td>Scheduled Arrival Demand</td>
<td>-0.0495</td>
<td>-0.0501</td>
<td>-36.62</td>
</tr>
<tr>
<td>Thunderstorm</td>
<td>1.300</td>
<td>1.29</td>
<td>13.76</td>
</tr>
<tr>
<td>Snow</td>
<td>0.676</td>
<td>0.623</td>
<td>7.95</td>
</tr>
<tr>
<td>Proportion of Heavy Arriving Aircrafts</td>
<td>-2.97</td>
<td>-3.16</td>
<td>-14.54</td>
</tr>
<tr>
<td>Proportion of Small Arriving Aircrafts</td>
<td>3.21</td>
<td>2.91</td>
<td>3.05</td>
</tr>
<tr>
<td>Product of Proportion of Small &amp; Heavy Aircrafts</td>
<td>-38.6</td>
<td>-37.3</td>
<td>-3.42</td>
</tr>
<tr>
<td>Convective Weather Score</td>
<td>0.772</td>
<td>0.582</td>
<td>7.69</td>
</tr>
</tbody>
</table>

The IBS for TAER based Cox model is 0.053 as compared to IBS of 0.058 for SAER based Cox model. The C-index is estimated to be 0.72 for the TAER based Cox model, which is consistent with the C Index obtained from the SAER based model. The TAER based Cox model predicted throughput is plotted against the observed throughput for all observations in the test data in Fig. 3.17. The coefficient of determination, $R^2$ between the observed and TAER based predicted throughputs is 0.78 and the square root of RMSE is 3.7 as compared to the $R^2 = 0.74$ and RMSE of 4.1 for the SAER based model. These values indicate an improved performance of
the TAER based Cox model as compared to the SAER based model. Again, compared to Fig. 3.4, we see lesser prediction error for observations with low throughput. This is consistent with the greater accuracy of the TAER-based demand estimate discussed earlier.

![Figure 3.17: Cox-predicted vs observed throughput (using TAER data)](image)

3.7.3. Random Survival Forest

The validation measures discussed in section 3.3 are estimated for the TAER based RSF model for tree depths ranging from 1 to 25. Fig. 3.18 shows that from the IBS criterion, 12 is the optimal depth with an IBS of 0.052 as compared to an optimal depth of 11 with minimum IBS of 0.055 for the SAER based RSF model. However, we observe that the IBS curve is relatively flat between tree depths of 12 and 14 and the corresponding RSF models have similar IBS estimates. Again, we use a truncation filter of 0.65 for subsampling the data and extrapolate the predicted CDFs to 1 as discussed in section 3.5.3. We use these extrapolated CDFs to estimate evaluation measures for the RSF models with different depths. The plot in Fig. 3.19 shows that 14 is the optimal depth for the RSF model with a C Index of 0.77 as compared to an optimal depth of 11 and C Index of 0.77 using the SAER demand data.
The average throughput is estimated using the extrapolated capacity CDF and is compared to the observed throughput on test data using the coefficient of determination $R^2$. Fig. 3.20 plots TAER based $R^2$ vs tree depth and suggests an optimal depth of 11, with $R^2 = 0.82$ and RMSE of 3.5. This is compared to optimal depth of 10, $R^2 = 0.8$ and RMSE of 3.6 for the SAER based RSF model. In Fig. 3.20, we observe that TAER based $R^2$ corresponding to a depth of 14 is very close to the highest value of 0.82. The RSF model predicted throughput is plotted against the
observed throughput for all observations in the test data in Fig. 3.21. As noted above in Section 3.5.3, predicting the throughput by training a random forest (RF) model directly on the airport data gives a highest $R^2$ of 0.64. When contrasted with highest $R^2$ of 0.82 from Fig. 3.20, we see that a model that explicitly takes capacity into account can more accurately predict airport throughput. The variable importance for the optimal TAER based RSF model (depth 14), for EWR data is shown in Figure 3.22. The variable importance is very similar to the results obtained for the SAER based RSF model and thus can be interpreted similarly.

![Figure 3.20: R²-predicted throughput vs tree depth (using TAER data)](image1)

![Figure 3.21: RSF-predicted vs observed throughput (using TAER data)](image2)
3.8. Conclusions

In this chapter we employ Kaplan Meier estimator, Cox Proportional Hazards model and Random Survival Forest model to predict airport capacity using weather, fleet mix, demand and throughput data at EWR airport. The two main contributions of the chapter are: (1) we build airport capacity prediction models based on weather, fleetmix and scheduled demand data, independent of AAR data which may be unreliable. These prediction models, commonly used in the fields of bioinformatics and reliability analysis, have never before been used to model airport capacity. (2) We also applied three validation measures evaluating aspects of the model such as accuracy of the probabilistic predictions, its discriminative power and its ability to accurately predict throughput at the airport. On evaluating the three capacity models, we found that Random Survival Forest model is superior across the three validation frameworks. Both quantitatively and in terms of the model properties, RSF model is well-suited for airport capacity modeling owing to its non-linear and non-parametric nature. Also, the models estimated using the TAER event status data perform better than those estimated from the SAER event status data. The important insights derived from each model and the comparison of the models in terms of three evaluation frameworks are highlighted in this section.

Kaplan Meier Estimator: The KM estimator is a non parametric empirical model and estimates one capacity distribution curve for all observations as shown in Fig. 3.1. In Fig. 3.2, we observed that this aggregation over all observations is not
accurate since days with a lower ceiling have a lower capacity. This limitation is reflected in higher IBS score and low $R^2$ scores. The KM estimator is useful if we are interested in capturing the capacity distribution of a specific subset of data. However, it is impractical on a large scale to create “relevant” filtered subsets of data for modeling.

**Cox Proportional Hazards Model:** The Cox model is a semi-parametric, linear model that captures the influence of airport weather and scheduled demand variables on the observation’s risk of being at capacity for a given throughput. The signs of most of the coefficients are consistent with the expected influence of these variables on airport capacity, with the exception of local hour and, the product of proportion of small and heavy aircrafts variables. Some interesting conclusions from the model are that snow indicator has a relatively large positive significant coefficient, indicating that when the airport experiences snow, its risk of being at capacity for a given throughput increases. The coefficients for VFR indicator, ceiling, visibility, temperature and scheduled arrival demand are negative. The evaluation metrics show an improvement of the Cox model over the KM estimator.

**Random Survival Forest Model:** The RSF model is a non-parametric, non-linear model which controls for covariates. An optimal tree depth of 11 and 14 are suggested by all the model evaluation methods for the SAER and TAER based RSF models respectively. The optimal RSF model shows an improvement in all the three validation measures over the KM estimator and the Cox model and has advantages of being non-parametric and non-linear. The relative importance estimated for the RSF model shows that local hour, scheduled arrival demand, ceiling, visibility and temperature are the five most predictive variables when modeling airport capacity. A case study where we make OOB predictions on the busy hours (7-22) of 08/13/2014 at EWR airport shows that the optimal SAER-based RSF model predicts the daily evolution of hourly airport capacity in a manner consistent with the observed weather patterns.

Table 3.6 summarizes the validation measures estimated for the three models discussed above using both SAER and TAER event status data. As explained earlier, IBS measures the accuracy of the probabilistic predictions with a range between 0 for perfect models and 0.25 and non-informative model. A lower IBS score indicates a better model. The C Index evaluates the discriminatory power of the model with a range between 1 for perfect model and 0.5 for non-informative model. The $R^2$ and RMSE for the prediction of throughput based on the predicted capacity distribution using equation 3.1 is also tabulated. $R^2$ closer to 1 and a lower RMSE indicates a better model. We observe that RSF model is consistently better across all validation measures for both the SAER and TAER based models.
The capacity prediction models discussed in this study can be used to develop reliable capacity forecasts, without depending on AARs. The input variables consist of scheduled demand, fleet mix and weather information, which can be forecasted, thus making it easy to obtain capacity forecast for a new scenario or day of operation. An important application of these capacity models is in estimating similarity between two days, which can then be used as an input to decision support tools in air traffic management. The models can also be used directly to predict when demand is likely to exceed capacity at an airport, suggesting that a ground delay program may be needed. In either context, the models developed in this study can be used for decision support in air traffic management.

### Table 3.6: Validation measures summary

<table>
<thead>
<tr>
<th>Demand Data</th>
<th>Model</th>
<th>IBS</th>
<th>C Index</th>
<th>$R^2$-Throughput Prediction</th>
<th>RMSE – Throughput Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAER</td>
<td>KM Estimator</td>
<td>0.076</td>
<td>-</td>
<td>0.64</td>
<td>4.9</td>
</tr>
<tr>
<td></td>
<td>Cox Model</td>
<td>0.058</td>
<td>0.72</td>
<td>0.74</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>RSF Model</td>
<td>0.055</td>
<td>0.77</td>
<td>0.8</td>
<td>3.6</td>
</tr>
<tr>
<td>TAER</td>
<td>KM Estimator</td>
<td>0.071</td>
<td>-</td>
<td>0.71</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>Cox Model</td>
<td>0.053</td>
<td>0.72</td>
<td>0.78</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td>RSF Model</td>
<td>0.052</td>
<td>0.77</td>
<td>0.82</td>
<td>3.5</td>
</tr>
</tbody>
</table>
4. Validation of Similar Days

4.1. Introduction

Air traffic managers and flight operators are faced with decision-making scenarios everyday due to the highly dynamic, stochastic nature of weather and air traffic. The decisions made to manage air traffic play a significant role in increasing the efficiency of an airport and the National Airspace System as whole. Decisions such as the duration, start times of a GDP (Ground Delay Program) and other Traffic Management Initiatives (TMI) can make the difference between an efficient airport, wasted airport capacity or expensive airborne delay (Cook and Wood, 2010). With increasing demand, airports around the world are getting saturated (Eno, 2013) and airport expansions are becoming increasingly expensive and time taking. This calls for improved decision support that can help maximize the utilization of airport.

Demand, weather and capacity information is crucial in deciding the appropriate TMI. It is important to note that in aviation, demand is managed to avoid expensive and unsafe queues or ‘traffic jams’ in the air at the airports. This demand feedback makes it difficult to observe and thus predict the capacity at airport. Furthermore, the emphasis on safety and human involvement in aviation needs to be considered when incorporating new tools to improve air traffic management. (Gorripaty et al., 2017; Gorripaty et al., 2016) identify similar days based on weather, demand and capacity information that can intelligently augment TMI decision-making experience.

(Gorripaty et al., 2017; Gorripaty et al., 2016) generate similarity measures based on capacity and demand data. Gorripaty et al., 2017 uses Aviation System Performance Metrics (ASPM) estimates and Gorripaty et al., 2016 explores using capacity distributions generated from Random Survival Forest (RSF) model. Chapter three details the background and estimation of the RSF model and other censored regression models that predict the capacity distribution for a given hour based on weather, demand and fleet mix data. Chapter three also develops accurate estimates for scheduled demand using Aggregate Demand List (ADL) data, which contains fine-grained flight schedule data of all the flights arriving and departing at an airport. The capacity and demand similarities that are generated from capacity and demand data need to be combined in a meaningful manner to capture the similarity between days well. (Gorripaty et al., 2017; Gorripaty et al., 2016) have explored ways of combining the capacity and demand similarities and conclude that the maximum of the two distances forms the most appropriate overall similarity matrix.

However this combination approach is limited in the number of alternative combinations that can be explored to combine the capacity and demand similarities.
These studies make restrictive assumptions on the interaction of capacity and demand similarity in generating the final similarity. Capacity and demand similarities may differ in their contribution towards the combined similarity measure. This work develops a framework to combine the capacity and demand similarities in a data-driven manner. This framework accounts for the differences in capacity and demand similarity generation and does not make assumptions on their interaction to generate the final similarity metric.

The goal of the validation framework is to determine how well the similarity metrics match reality. Owing to the lack of ground truth similarity measures, we devise a validation framework that compares operational outcomes similarity with the combined capacity, demand and TMI similarity. On average, days with similar demand, capacity and TMI behavior should have comparable operational outcomes. In this chapter, we generate a combined similarity matrix where each cell corresponds to overall capacity, demand and TMI similarity between two days and an operational outcomes similarity matrix. The correlation between these matrices provides a measure of the validity of the estimated capacity, demand and TMI based similarity measure. Fig. 4.1 summarizes the relationship between the capacity, demand, TMI and operational outcomes similarity measures. In this chapter, we discuss the results of the similarity validation framework for EWR airport data.

Figure 4.1: Capacity, demand, TMI and operational outcome similarities relationship

For a given reference day, the capacity and demand-based distance can be used to identify similar historical days. The traffic management initiatives taken on
past similar days and their resulting outcomes can augment controller experience to guide decision-making on the reference day at an airport. In this work, the words similarity and distance are used in the same vein to quantify resemblance between two observations. This chapter is organized as follows: Section 4.2 describes the data used in this study to estimate capacity, demand, TMI and operational outcomes similarities. Section 4.3 discusses the frameworks to generate capacity based similarity, section 4.4 discusses demand similarity estimation and section 4.5 discusses operational outcome similarity calculation. Section 4.6 discusses the validation of capacity similarity by comparing the correlation between capacity-demand-TMI combined similarity and the operational outcome similarity. Section 4.7 discusses some conclusions of the study.

4.2. Data

To estimate accurate capacity, demand, TMI and operational outcome based similarities, we need reliable estimates for these variables. We discuss the sources of these variables in this section:

**Capacity:** Airport capacity depends on numerous factors such as weather, fleet mix, scheduled demand and human factors. Chapter three details censored data models such as Kaplan Meier estimator, Cox Proportional Hazards and Random Survival Forest models to model airport capacity. These models predict the distribution of capacity for a given hour, instead of a single capacity estimate. The predicted capacity distributions can be used to estimate similarity between two hours. By using distributions instead of single estimates for capacity, we are able to incorporate more information into the similarity measure. The OOB (out-of-box) distributions predicted using the RSF model described in chapter three will be used in this work for capacity based similarity calculations. The variables used to predict capacity in the RSF model include local hour, scheduled arrival demand, ceiling, visibility, temperature, proportion of heavy arrivals, convective weather score, wind speed, thunderstorm, proportion of small arrivals, wind direction, snow, VFR indicator and the product of the proportion of small and heavy arrivals in the decreasing order of importance.

**Demand:** For two days to be similar, it is important that they have similar demand patterns across different hours of the day. Also, the interaction of the demand scheduled at the airport with the available capacity is important to understand similarity between days. The scheduled demand data is processed from Aggregate Demand List (ADL) dataset, which contains fine-grained information of the flight schedules of all flights arriving and departing from an airport. In chapter three we processed the ADL data to generate estimates for scheduled arrival demand. This data will be used to estimate demand based similarity measure between two days.
**TMI:** The distance between two days based on the TMI implemented is important in validating the capacity similarity. The features of the TMI data used in the distance calculation include Ground Stop (GS) duration, average called rate, file time, start time, duration, scope and relative rate. Estes and Lovell, provides more information on generation of these features from National Traffic Management Log (NTML) data. The pairwise distances are calculated separately for each feature using the L-1 norm distance measure. The pairwise distances are normalized across all features: the normalized distance in a feature for a pair of days is the proportion of pairs in the data that have a distance smaller than that pair. In other words, the normalized distance for a pair of days is its quantile compared to all the pairs of points in the data. The final TMI distance between two days is estimated by taking the maximum of all the feature-wise distances between the two days. The final TMI distance estimated in Estes and Lovell, for EWR airport is used in this study.

**Operational Outcomes:** The operational outcomes at an airport are a result of the interaction of weather, capacity, demand and TMI. This dependence is used as the basis for designing the validation framework for similar days. Different performance indicators capture different aspects of the operation of an airport. The most widely used operational outcomes include number of cancellations, diversions, holdings and the total delay. The hourly cancellations data is derived from Bureau of Transportation Statistics (BTS) on-time performance data. The hourly holding and diversions data is derived from Performance Data Analysis and Reporting System (PDARS) data. We use indicators in individual flight data to identify if they were cancelled or diverted or put into holding. Subsequently, we aggregate the information hourly using the times when these initiatives were implemented on the flight. The hourly average arrival delay data is aggregated from ASPM individual flight data.

We use the TMI data from 2011 to 2014 at EWR, where every observation is a TMI (GDP/GS). We use the hourly cancellation, diversion, holding and delay data from 2011-2014 for busy hours (7-22). We tabulate summary statistics of these TMI and operational outcome variables for EWR airport in Table 4.1 before discussing their similarity estimation framework.

Before we proceed to describing the similarity estimation we briefly discuss an important visualization technique for the estimated similarity/distance matrix between all pairs of days in the dataset. Metric Multidimensional Scaling (MDS) is a technique used to visualize data in lower dimensions (2 or 3 dimensions generally) based on the pairwise similarities/distances between observations. The points are arranged mutually to minimize the error between the actual distances and the distances represented on the MDS plot. Similar days appear nearby in the plot, therefore allowing quick visual and interactive analysis of the obtained results. The color and size/shape of the markers of the MDS plot can be used to visualize additional dimensions of the data. The axes of the MDS plots do not have any physical meaning and should only be interpreted for the relative positions of the
points (days). The MDS plots are an approximate visualization of the relative positions and the distances between the days. The MDS algorithm, given the constraint of a two-dimensional plot, tries to preserve the distances between two days as estimated using the similarity metric.

Table 4.1: Summary statistics for TMI and operational outcome data (2011-2014)

<table>
<thead>
<tr>
<th>Data</th>
<th>Variable</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMI</td>
<td>Average Called Rate during GDP (flights/hour)</td>
<td>17.3</td>
<td>43.85</td>
<td>35.37</td>
<td>36</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td>GDP Duration (minutes)</td>
<td>205</td>
<td>1079</td>
<td>614.4</td>
<td>629</td>
<td>153.7</td>
</tr>
<tr>
<td></td>
<td>Number of Airports in Scope of GDP</td>
<td>16</td>
<td>29</td>
<td>25.11</td>
<td>29</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>GDP Start Time (minutes after 4:00 AM local time)</td>
<td>180</td>
<td>924</td>
<td>516.93</td>
<td>483.5</td>
<td>103.66</td>
</tr>
<tr>
<td></td>
<td>GDP Entry Time (minutes after 4:00 AM local time)</td>
<td>179</td>
<td>929</td>
<td>435.55</td>
<td>411</td>
<td>130.46</td>
</tr>
<tr>
<td></td>
<td>Ground Stop Duration (minutes)</td>
<td>0</td>
<td>115</td>
<td>8.48</td>
<td>0</td>
<td>20.26</td>
</tr>
<tr>
<td></td>
<td>Relative Rate</td>
<td>0.18</td>
<td>1</td>
<td>0.96</td>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td>Operational Outcome</td>
<td>Cancellations (flights/hour)</td>
<td>0</td>
<td>29</td>
<td>0.69</td>
<td>0</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>Diversions (flights/hour)</td>
<td>0</td>
<td>203</td>
<td>0.25</td>
<td>0</td>
<td>2.35</td>
</tr>
<tr>
<td></td>
<td>Holdings (flights/hour)</td>
<td>0</td>
<td>58</td>
<td>1.3</td>
<td>0</td>
<td>3.91</td>
</tr>
<tr>
<td></td>
<td>Average Delay (minutes/flight)</td>
<td>0</td>
<td>313.5</td>
<td>21.04</td>
<td>13.1</td>
<td>25.36</td>
</tr>
</tbody>
</table>
4.3. Capacity Similarity

The OOB capacity cumulative distribution functions (CDFs) are predicted for every busy hour (7-22) at EWR airport for 3 years of data from 2011 to 2014 in chapter three. The best model corresponds to the RSF model with maximum depth of 14 using terminal demand data. For incorporating capacity data into finding similar days, we need to develop a similarity measure between every pair of days in a dataset. The capacity information of a day consists of the set of CDFs over its busy hours. Thus, to estimate the capacity-based similarity between a pair of days, we need to device a framework that can estimate the similarity between the corresponding hours of the two days. The similarity between two days can then be found by aggregating the similarity across its busy hours. In this section, we present several alternative methods to capture the similarity between two CDFs.

**Euclidean Distance:** The CDF of an hour \( i \) is given by \((F_{i1}, F_{i2}, \ldots, F_{ik})\) where \( F_{iq} \) is the probability that the capacity for hour \( i \) is less than or equal to some specified value \( q \) and \( k \) is the maximum capacity observe in the dataset. The distance between two hours \( i \) and \( j \), \( d_{ij} \) estimated using the Euclidean formula is shown below.

\[
d_{ij} = \sqrt{\sum_{q=0}^{k} (F_{iq} - F_{jq})^2}
\]  

(4.1)

For two days \( a \) and \( b \), the distance between the CDFs of their corresponding hours \( i(a) \) and \( i(b) \) is estimated for all busy hours. These distances are then aggregated into the capacity distance between two days \( a \) and \( b \), \( D_{ab} \) as the square root of the sum of squared distances for every hour as shown below. \( h_1 \) and \( h_2 \) are the start and end of busy hours and are 7 and 22 respectively for EWR airport.

\[
D_{ab} = \sqrt{\sum_{i=h_1}^{h_2} (d_{i(a)i(b)})^2}
\]  

(4.2)

We estimate a Euclidean based capacity distance matrix where the cell \((a, b)\) represents the distance \( D_{ab} \). MDS plot for the Euclidean capacity distance matrix based on predicted CDFs from chapter three is plotted in Fig. 4.2. Each point in the plot is a day and the color of the plot represents the date of that day, with blue points indicating the beginning of 2012 and red points indicating the end of 2014.
We can see that days are temporally mixed in the MDS plot with some days closer to 2012 being similar to days closer to 2014. This implies that we can learn from days in the “distant history” that are similar to a given day based on capacity information. Again, we don’t see the presence of any clusters in the data based on the arrangement of the points in the MDS plots. This also suggests that we use continuous distance measures to quantify the similarity between two days in accordance with the findings of Gorripaty et al., 2017.

![MDS plot](image)

**Figure 4.2:** MDS plot of Euclidean Distance of OOB Predicted Capacity CDFs

**Hellinger Distance:** We estimate the probability mass function (PMFs) from the CDFs of all hours in the dataset. Let \( f_{i1}, f_{i2}, \ldots, f_{ik} \) denote the PMF of hour \( i \) where \( f_{iq} \) is the probability that the capacity for hour \( i \) is equal to some specified value \( q \) and \( k \) is the maximum capacity observed in the dataset. The distance between two hours \( i \) and \( j \), \( d_{ij} \) estimated using the Hellinger formula is shown below.
\[ d_{ij} = \frac{1}{\sqrt{2}} \sqrt{\sum_{q=0}^{k} (\sqrt{f_{iq}} - \sqrt{f_{jq}})^2} \] 

(4.3)

The maximum Hellinger distance is 1 and is observed when the two hours \( i, j \) have complementing PMFs, i.e, \( f_{iq} * f_{jq} = 0 \) and \( f_{iq} + f_{jq} = 1 \) for all \( q \). For two days, the Hellinger distance between the PMFs of their corresponding hours is estimated for all busy hours. The final distance between two days is the square root of sum of squares of the distances between their corresponding busy hours. We estimate a capacity distance matrix based on Hellinger formula where the cell \((a, b)\) represents the final distance between days \(a\) and \(b\). An MDS plot for this distance matrix estimated using predicted CDFs from chapter 3 is plotted in Fig. 4.3. We observe that the distribution of days using the Hellinger distance is different from the Euclidean distance based spread found in Fig. 4.2, but again show mixing of days from different years.

![MDS plot of Hellinger Distance of OOB Predicted Capacity CDFs](image)

Figure 4.3: MDS plot of Hellinger Distance of OOB Predicted Capacity CDFs
**Bhattacharyya Distance:** We estimate the probability mass function (PMFs), $(f_{i1}, f_{i2}, \ldots, f_{ik})$ from the CDFs of all hours. Here, $f_{iq}$ is the probability that the capacity for hour $i$ is equal to some specified value $q$ and $k$ is the maximum capacity observed in the dataset. The Bhattacharyya distance between two hours $i, j$ is related to the overlap between their PMFs quantified by the Bhattacharyya coefficient $BC(i, j)$. The Bhattacharyya distance $d_{ij}$ between two hours $i, j$, is shown in equation 4.4. The Bhattacharyya distance is unbounded and is related to Hellinger distance as $Hellinger - d_{ij} = \sqrt{1 - BC(i, j)}$. We estimate a capacity distance matrix based on Bhattacharyya formula where the cell $(a, b)$ represents the final distance between days $a$ and $b$. An MDS plot for this distance matrix for EWR airport is plotted in Fig. 4.4. We observe some outliers in the MDS plot for Bhattacharyya distance, which may be explained by its unbounded nature.

\[
BC(i, j) = \sum_{q=0}^{k} \sqrt{f_{iq} * f_{jq}}
\]

\[
d_{ij} = -\ln (BC(i, j))
\]

Figure 4.4: MDS plot of Bhattacharyya Distance of OOB Predicted Capacity CDFs
**Proximity Distance:** The distance measures discussed so far are based on the predicted capacity distribution from the RSF model in chapter three. The OOB prediction for an observation is generated by running it down each tree in the forest of the trained RSF model. For each tree in the forest, the data in the terminal node where the observation falls is used to estimate the CDF. This estimate is then aggregated across all the $T$ trees in the forest. The samples that fall in the same node are similar whereas the samples that fall in different nodes are relatively dissimilar. This approach provides a $0$-$1$ discrete distance measure, $d_{ijt}$ for each tree $t$.

$$d_{ijt} = \begin{cases} 0, & i,j \text{ fall in the same terminal node in tree } t \\ 1, & i,j \text{ fall in different terminal nodes in tree } t \end{cases}$$

The proportion of trees where two observations fall in different nodes provides a measure of overall proximity distance, $d_{ij}$ across the trained forest. The proximity distance between two days $a, b$ can be aggregated as the mean of proximity distance over busy hours ($h_1 = 7$ & $h_2 = 22$ for EWR airport) of two days.

$$d_{ij} = \frac{\sum_{t=1}^{T} d_{ijt}}{T}$$

$$D_{ab} = \frac{\sum_{t=h_1}^{h_2} d_{(a)(b)}}{(h_2 - h_1)}$$

We estimate a proximity based capacity distance matrix where the cell $(a, b)$ represents the distance $D_{ab}$. The MDS plot for this distance matrix is plotted in Fig. 4.5. The distribution of days based on the proximity distance is very distinct from other measures of capacity distance discussed above. Owing to the fact that every hour has some neighbours that fall into same terminal nodes, we do not observe any outliers in the MDS plot. Again, there is no evidence for clusters based on proximity based distance measure, and substantial mixing of days from different time periods.
Figure 4.5: MDS plot of Proximity Distance using RSF model

4.4. Demand Similarity

To capture the influence of demand patterns on the similarity between two days, we explore the interaction of scheduled arrival demand with the worst-case arrival capacity of the airport (under IFR conditions). We quantify this interaction using a deterministic queuing model (Daganzo, 1997) described in this section. In Table 4.2, we show an example scheduled demand and worst-case capacity profile for 14th January 2013 at EWR airport. Here, I refers to IFR conditions. To interpret the table, for a start hour of 0, the scheduled arrival demand on the 14th January 2013 was 4 flights. The worst-case capacity of 34 flights/hour is used across all hours. Please note that the demand profile is different for different days in the year.
The cumulative counts are estimated by cumulatively adding the components of a vector, such that the nth cumulative count is the sum of the first n components of the vector. Given the 24-dimensional demand profile, we estimate its cumulative demand curve. Due to capacity limitations, an airport is able to either satisfy demand entirely or operate at a limited throughput below the observed demand. Thus throughput is the minimum of the total demand seen at that hour (the residual demand from previous hours that was not accommodated added to the scheduled demand at that hour) and the capacity of the airport at that hour. The throughput is estimated for each hour using the effective demand (scheduled demand + residual demand from previous hour) and the capacity estimates and the corresponding 24 dimensional throughput cumulative curve is constructed. Note that the cumulative demand curve is always equal to or more than the cumulative throughput curve, since at the airport, it only possible to operate as many flights or less as there is total demand. The sum of the differences between the demand and throughput curves for each hour gives the total queuing delay across the 24 hours.

In Fig. 4.6 below, we plot the cumulative demand \(D(t)\) and throughput curves \(T(t)\) observed on 14th January 2013 at Newark (EWR) airport. The x-axis shows the time of day which ranges from 0-24 hours and the y-axis shows the cumulative demand/throughput at the airport at that time of the day. The total delay estimated as the area between the curves \(D(t)\) and \(T(t)\) will be used to capture the effect of the scheduled arrival demand juxtaposed with the worst-case IFR capacity. The demand distance between two days is defined as the absolute difference between their total delays estimated using deterministic queuing model. For EWR airport, we use the worst-case scenario arrival capacity of 34, which is the IFR capacity of a single runway, and the demand is derived from the ADL data.
4.5. Operational Outcomes Similarity

The cancellation, diversions, holdings and average delay data are aggregated from different sources of individual flight data as discussed in 4.2. The operational outcomes in the busy hours only are used in the similarity estimation. Different operational outcomes have different scales and in some cases different units. It is thus important to standardize features before applying the distance formula on them. The similarity between two hours \((i,j)\), \(d_{ij}\) is estimated as the Euclidean distance between the standardized vectors containing the operational outcome data, \(O_i, O_j\). The distance between two days \((a,b)\), \(D_{ab}\) is obtained by aggregating the distances between their corresponding busy hours \(d_{i(a)}, d_{j(b)}\), as the square root of sum of squared hourly outcomes distances.

\[
d_{ij} = \sqrt{\sum_{v \in \{\text{cancellation, diversion, holding, delay}\}} (O_{iv} - O_{jv})^2}
\]

(4.6)
\[ D_{ab} = \sqrt{\sum_{i=h_1}^{h_2} (d_{i(a)}d_{i(b)})^2} \]

We estimate an operational outcome distance matrix where the cell \((a, b)\) represents the distance \(D_{ab}\). MDS plot for the operational outcome distance matrix is plotted in Fig. 4.7. The spread of data points indicates that days across different years can be similar to one another and do not provide any evidence for the presence of clusters.

Figure 4.7: MDS plot of Operational Outcomes based Distance

4.6. Validation

Sections 4.3, 4.4 and 4.5 discussed the frameworks used to develop capacity, demand and operational outcome based similarity measures. Estes and Lovell
develop a TMI based similarity measure. One aspect that stands out in these discussions is the diversity in the origin of data and methodology used to generate the different similarities. Deciding which capacity similarity best represents the EWR airport data is an important question to be answered. To guide these decisions, we need a validation framework that can quantitatively evaluate different capacity similarity alternatives.

Owing to the lack of a ground truth capacity similarity measure, we investigate the relationship between different variables for possible alternative validation frameworks. As discussed in section 4.1, days that are similar in terms of their capacity and demand, and are similar in terms of their TMIs, should have similar operational outcomes. On the other hand, days that are dissimilar in terms of their capacity and demand or in terms of their TMIs, should have dissimilar operational outcomes. Also, the extent of similarity of days based on capacity, demand and TMI should influence the extent of similarity based on their operational outcomes. We use this assumption, which has been discussed with numerous subject matter experts (SMEs) on Air Traffic Management (ATM) to design our validation framework. Translating this to similarity measures, this means that days with high (low) similarities based on capacity, demand and TMI should have high (low) operational outcome similarity. Mathematically, the correlation between the combined capacity, demand and TMI similarity matrix and, the operational outcome similarity matrix can be used to evaluate the estimated capacity similarity metrics. This can be represented using the Fig. 4.8 below.

![Figure 4.8: Evaluation of capacity similarity](image)

The two main parts of the validation framework are: the estimation of an operational outcome based similarity matrix which has been discussed in section 4.5 and the estimation of the combined capacity, demand and TMI based similarity matrix, which will be discussed in this section. It is important that the combination of capacity, demand and TMI similarity matrices is supervised by the ability of the combined matrix to correlate well with the operational outcomes matrix. Mathematically, the capacity, demand and TMI similarities should be regressed on
the operational outcome similarity. The combination framework can be represented visually as shown in the Fig. 4.9 below.

![Diagram of similarity matrices](image)

Figure 4.9: Regression model of the distance metrics

As noted earlier, the similarity metrics used in the regression framework have different scales owing to the fact that they are derived from different estimation methods and variables. Also, the specification of the regression model and the nature of interaction of variables are not previously known. This motivates using a Random Forest (RF) regression model to supervise the combination of the similarity matrices. Every pair of days is treated as an observation and using a training subset of the upper triangular parts of the similarity matrices, we train a RF regression model with capacity, demand and TMI similarities as the covariates and the operational outcome similarity as the dependent variable. Since distances generally add as the sum of squares, we use the squared distances in the RF regression framework.

For this study, the convective weather data is available for EWR airport only from 2012 to 2014 and is the constraining factor in the number of observations for which capacity CDFs can be predicted. To cover a larger subset of days in training and testing the RF model, we use the OOB CDF predictions from a capacity model trained without convective weather data. We observe that the performance of the RSF capacity model is not affected significantly by removing the convective weather variable. We train a RF regression model using capacity, demand, TMI and operational outcomes similarity matrices estimated for EWR airport for 4 years from 2011 to 2014 with the scikit-learn package in python (Pedregosa et al., 2011). An optimal maximum depth of a decision tree strikes a balance between learning the trends in the data without over-fitting. To find the optimal maximum depth of the trees in the RF model, we estimate the resulting correlation between the combined similarity matrix and the operational outcomes matrix on test data for a range of tree depths form 1 to 25. The highest correlation between the combined similarity matrix and the operational outcome based similarity matrix on test data is used to choose the optimal tree depth.
As discussed in section 4.3, we estimate four capacity-based similarity measures using the OOB predicted CDFs. Below, we evaluate the four estimated similarity matrices using the correlation measure described in this section. Fig. 4.10, 4.11, 4.12 and 4.13 show the correlation score plotted against different tree depths of the RF regression model. Each figure shows two plots, one for observations with TMIs and the other for observations without TMIs. The RF model and the capacity distance with the highest correlation are the optimal combination model and capacity distance respectively for EWR airport. We observe that all the plots show similar trends that the correlation increases with depth until a certain point, beyond which the correlation starts decreasing. Overall, we see that the combined similarity metric using Hellinger capacity distance has the highest correlation with the operational outcomes similarity matrix for with and without TMI scenarios. The decreasing order of correlation of the other three capacity distances are the Euclidean capacity distance, Bhattacharya capacity distance and proximity based capacity distance for both the with and without TMI scenarios. Thus, the Hellinger distance measure between the predicted capacity distributions is used to estimate the best capacity distance between days.

![Figure 4.10: Correlation vs Tree Depth for Euclidean Capacity Distance for days with (a) TMIs and (b) without TMIs](image-url)
Figure 4.11: Correlation vs Tree Depth for Hellinger Capacity Distance for days with (a) TMIs and (b) without TMIs

Figure 4.12: Correlation vs Tree Depth for Bhattacharya Capacity Distance for days with (a) TMIs and (b) without TMIs
Figure 4.13: Correlation vs Tree Depth for Proximity based Capacity Distance for days with (a) TMIs and (b) without TMIs

The relative importance of the capacity, demand and TMI similarity measures in predicting the combined similarity measure is useful to understand the interaction between different components of the similarity at an airport. The higher the relative importance of a similarity measure, the more important it is in predicting the combined similarity measure. The relative importance measures vary from 0 to 1 and sum up to 1 across all the covariates of the RF regression model (Breiman et al., 1984.). We tabulate the relative importance of the capacity, demand and TMI similarities in predicting the combined similarity for days with and without TMIs in Table 4.3 and Table 4.4 respectively. For the RF model on days with TMIs, the relative importance of the Hellinger capacity similarity is 0.7, of demand similarity is 0.08 and of TMI similarity is 0.22. We see that the capacity similarity is the most important component of the model in predicting the combined similarity measure, followed by TMI similarity and further followed by demand similarity, which is less important by an order of magnitude. For the model on days without TMIs, the relative importance of the capacity similarity is 0.8 and of demand similarity is 0.2. This again shows the dominant role of capacity similarity in predicting KPI similarity between days.

In Table 4.3 and 4.4, we can see that the Euclidean, Hellinger and Bhattacharya distance measures yield comparable results in terms of the variable importance and the highest observed correlation across different tree depth of the RF combination model. These distance measures are all based on the capacity CDFs predicted from the RSF model. The other distance measure based on the proximity of observations in the terminal node of the trees of the RSF model performs worse in terms of correlation for both days with and without TMIs. For days with TMIs, the importance of the capacity similarity is significantly lower than the other three
distance measures, indicating that it does not capture important aspects of similarity between two days. This is expected since the proximity criterion is based on two observations occupying the same terminal node, and does not account for cases where the two observations occupy “nearby” terminal nodes. The proximity-based distance doesn’t take into account similarities in the paths followed by the observations in reaching the terminal node, which is important information.

Table 4.3: Relative importance (days with TMIs)

<table>
<thead>
<tr>
<th>Capacity distance</th>
<th>Capacity importance</th>
<th>Demand Importance</th>
<th>TMI Importance</th>
<th>Highest Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>0.71</td>
<td>0.07</td>
<td>0.22</td>
<td>0.64</td>
</tr>
<tr>
<td>Hellinger</td>
<td>0.7</td>
<td>0.08</td>
<td>0.22</td>
<td>0.66</td>
</tr>
<tr>
<td>Bhattacharya</td>
<td>0.62</td>
<td>0.09</td>
<td>0.28</td>
<td>0.63</td>
</tr>
<tr>
<td>Proximity</td>
<td>0.15</td>
<td>0.18</td>
<td>0.67</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table 4.4: Relative importance (days without TMIs)

<table>
<thead>
<tr>
<th>Capacity distance</th>
<th>Capacity importance</th>
<th>Demand Importance</th>
<th>Highest Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>0.78</td>
<td>0.22</td>
<td>0.59</td>
</tr>
<tr>
<td>Hellinger</td>
<td>0.8</td>
<td>0.2</td>
<td>0.61</td>
</tr>
<tr>
<td>Bhattacharya</td>
<td>0.86</td>
<td>0.14</td>
<td>0.57</td>
</tr>
<tr>
<td>Proximity</td>
<td>0.78</td>
<td>0.22</td>
<td>0.49</td>
</tr>
</tbody>
</table>
4.7. Conclusions

In this study, we estimate four different capacity similarities using the OOB predicted CDFs from a RSF model trained on weather, fleet mix and scheduled demand data. In the Euclidean capacity similarity framework, we compare two hours using the Euclidean distance between their predicted CDFs. In the Hellinger and Bhattacharya capacity similarity estimation, we compare two hours using their predicted PMFs. In the proximity based capacity similarity, we compare two hours based on their relative positions in the terminal nodes of the trees in the forest. We estimate demand similarity between two days by constructing deterministic queuing models for each day based on the airport capacity at EWR airport under IFR conditions. We then compare two days using the difference in their total delays, estimated as the area between the cumulative demand and throughput curves in the queuing diagram. For TMI similarity, the results of Estes and Lovell are used for this study. The operational outcomes similarity is estimated using the Euclidean measure between the standardized operational outcome data, which consists of cancellations, holdings, diversions and delay data.

Based on the interaction of the different components of an airport, the combination of capacity, demand, TMI similarities should yield high correlation with the operational outcomes similarities. The capacity, demand and TMI similarities are combined using a RF regression model trained separately for days with and without TMIs. We find that the combined similarity measure predicted from the RF model has the highest correlation with operational outcome data when Hellinger capacity similarity is used. The highest correlation between the combined capacity-demand-TMI similarity matrix estimated for days with TMIs with the operational outcomes similarity matrix is 0.66. For days without TMIs, the highest correlation between the capacity-demand similarity matrix and the operational outcomes similarity matrix is 0.61 again using the Hellinger similarity measure. In the final RF models trained using the Euclidean, Hellinger and Bhattacharya measures, capacity similarity is the most important variable in predicting the combined similarity matrix for both days with and without TMIs.
5. Conclusions

Air traffic delays around the world cost billions of dollars and adversely affect people and businesses. Delays are in large part caused by imbalance in airport capacity and demand, more specifically, when there is insufficient capacity to accommodate flights intending to operate at the airport. One solution to the delay problem, airport expansion, is difficult to pursue owing to constraints in aviation funding, physical space near the airports and environmental objections. Another solution, limiting airport demand, is considered a last resort. This brings our focus to improving air traffic management (ATM) to reduce air traffic delays. Decision support tools can augment controller experience and lead to efficient traffic management initiatives (TMIs). From similar historical days controllers can better understand the impact of different TMIs and opportunities to reduce capacity underutilization. In a day-of-operations scenario, information on the performance of different TMIs on similar historical days can allow controllers to make better tradeoffs on the best strategy to manage a given day. In a post-operations scenario, finding similar historical days can help evaluate the strategies used in managing air traffic. Accurate information on capacity and demand at an airport would also help controllers understand the risks of different TMIs and thus make less conservative decisions when feasible. This will reduce capacity underutilization and thus improve airport capacity and reduce delays.

5.1. Summary

In this dissertation, we explore capacity and demand estimates at an airport and use them to calculate similarity between days. In chapter 2, we first explore similarity as a discrete measure based on clustering using existing capacity and demand data. AAR capacity data and demand data from ASPM are used in the similarity estimation. To ensure that we are only capturing important information at airports, we use the capacity and demand data for busy hours of the day. We find that the AAR data is correlated in neighboring hours of the day. Since weather strongly influences airport capacity and weather in contiguous hours is correlated, it is expected that hourly airport capacity is correlated. On the other hand, we find that airport demand is uncorrelated across the hours of the day. Owing to the high capacity correlation, we find an efficient reduced dimensional representation of capacity data that explains 90% of the variance of the data. Due to the lack of significant demand correlation, we are unable to find an efficient reduced dimensional representation of demand data. We use K means clustering algorithm to find clusters in demand and reduced dimensional capacity data. Using within sum of squares (WSS) and the average silhouette width (ASW) plots, we conclude that capacity and demand data
do not have clusters for EWR, JFK, ORD and SFO airports. We then explore a continuous measure of similarity between two days by combining the capacity and demand distances for these airports.

Visualizing the distance matrices using Metric Multidimensional Scaling (MDS), we find that days across different years, with varying TMI activity are similar to each other. This finding provides evidence that we can identify historical days similar to a given day and gain insights to guide TMI decision-making on the reference day. One of the limitations of the work in chapter 2 is that we depend on AARs to capture airport capacity. Discussions with Air Traffic Controllers (ATCs) reveal that the recorded AARs may not be reliable estimates of airport capacity. In chapter 2, we also do not evaluate the estimated similarity measures owing to the lack of ground truth similarity estimates. The lack of an evaluation measure hinders in learning the right combination of the estimated capacity and demand similarities. In the next two chapters in the dissertation, we address these limitations.

In chapter 3, we use techniques from bioinformatics to develop predictive models of airport capacity using throughput and demand data. We estimate the scheduled arrival demand using the fine-grained Aggregate Demand List (ADL) data. To predict airport capacity, we use the Kaplan Meier (KM) estimator, Cox model and Random Survival Forest model using weather, fleet mix, demand and throughput data at EWR airport. To evaluate the trained capacity models, we employ three validation measures evaluating the accuracy of the probabilistic predictions using IBS, their discriminative power using C Index and their ability to accurately predict throughput at the airport using $R^2$. We find that Random Survival Forest model outperforms the other capacity models across all the validation measures. Owing to its non-linear and non-parametric nature and based on the validation results, we use the trained RSF model for predicting capacity for an hour using demand, weather and fleet mix data. Also, we find that the models trained using the event status derived from terminal demand data are better. The capacity prediction models developed in this study can be used to predict capacity similarity. These capacity models can also be used in numerous applications such as validating AARs and other existing capacity models.

In chapter 4, we use the predicted capacity CDFs from the trained RSF model discussed in chapter 3 and estimate four different capacity similarities. The Euclidean capacity similarity estimates the Euclidean distance between the predicted CDFs of two hours. The Hellinger and Bhattacharya capacity similarity estimate the distance between the predicted PMFs of two hours. The proximity based capacity similarity compares the relative positions of two hours in the terminal nodes of the trees in the forest. A demand similarity between two days is estimated by comparing their total delay resulting from their demand interacting with the worst-case capacity scenario (IFR capacity). The total delay is calculated using deterministic queuing models for each day. This chapter estimates an operational outcome based similarity as the Euclidean distance between the hourly
operational data vectors consisting of cancellations, holdings, diversions and delay data. Days with similar capacity, demand TMI data should be similar in terms of their operational outcomes data. Based on this, we design a validation framework for similarity estimates that measures the correlation between the combined capacity-demand-TMI similarity and the operational outcome similarity. We combine the capacity, demand and TMI similarities, supervised by the operational outcome similarities using a Random Forest (RF) regression model. We find that the Hellinger capacity similarity, when combined with the demand and TMI similarity has the highest correlation with the operational outcome similarity. Capacity similarity measured using the Euclidean, Hellinger or Bhattacharyya methods has the highest variable importance in the trained RF combination model.

5.2. Contributions

This dissertation explores ways to estimate similarity between two days at an airport. We also develop capacity and demand estimates necessary to calculate the similarity between two days, using EWR airport as a case study. Finally, we design validation frameworks to evaluate estimated similarities and train a model to combine the similarity measures. We restate the three main contributions of this dissertation: 1) This dissertation applies dimensionality reduction techniques on capacity data that capture about 90% of the variance. We observe little correlation in demand data and the lack of effective dimensionality reduction for demand data. Using quantitative and qualitative validation measures we find the lack of inherent clusters in both capacity and demand data. In this work, we thus develop continuous measures of similarity between days. 2) This dissertation develops a capacity prediction model independent of AARs. The model incorporates observations censored by insufficient demand. The trained models can be used to generate hourly capacity distributions based on weather, demand and fleet mix data for varied applications. The study of the capacity models, complete with detailed validation frameworks tailored towards censored capacity data is a key contribution of this dissertation. We also generate scheduled demand estimates from fine-grained ADL data for the first time. 3) This dissertation also contributes to the validation literature for similarity measures. We design a validation framework that relies on the interaction between different components of the airspace: capacity, demand, TMI and operational outcomes. We also develop a model to combine the different similarity measures in a data-driven manner. Overall, this dissertation harnesses advances in machine learning to model airport capacity from weather, demand and fleetmix data and to find historical similar days using capacity, demand, TMI and operational outcome data.
5.3. Future Work

In this dissertation, we explored capacity models and similarity estimation for EWR airport. It would be useful to extend this work to other airports in the US and in different countries of the world. The capacity models across different airports can provide insights on the varied factors influencing airport capacity. The capacity models are currently trained on hourly weather, demand and fleet mix data. Each airport is treated as an independent entity in the modeling framework described in this dissertation. However, metroplex airports such as EWR, LGA and JFK have significant interactions with each other. The capacity models described in this dissertation can be extended to incorporate interactions between the airports. Numerous questions need to be answered to facilitate this extension: which variables can capture the interactions between the airports, which airports need to be included for modeling a given airport and should the metroplex capacity be estimated instead of the airport capacity for airports with strong interactions. It is also necessary to understand the influence of the interactions between airports in the similarity estimation framework.

This dissertation uses observed weather, demand, fleet mix, TMI and operational outcomes data. The various models and frameworks are also evaluated using observed data. However, important applications of the trained capacity models and the similarity measures are in the day-of-operations scenario. It is thus important to evaluate the performance of these models when using forecasted data instead. Also, the capacity models discussed in this dissertation are valid for a given infrastructure of the airport. It would be useful to study how infrastructure changes affect the capacity of the airport. It would be valuable to develop more case studies to compare the predicted capacity CDFs with the observed weather and operations data at numerous airports for different days. Also, systematic feedback from air traffic controllers, managers and airline operators through surveys would help understand the limitations of the capacity model and the similarity estimation framework from a practical point of view. This feedback is especially important in ensuring that this work can be implemented into ATM and can be used to develop decision support tools.

5.4. Concluding Remarks

Efficient airports are central to ensuring the smooth functioning of the aviation systems across the world. The aviation industry is opening its doorways to advances in technology and machine learning with initiatives such as FAA’s NextGen. Machine learning methods when applied to the vast archives of data collected on flight operations, trajectories, weather, demand and passenger movement across all the airports in the world can revolutionize our understanding of the air traffic management. With increasing pressure on the existing aviation infrastructure and limited physical space that inhibits airport expansions, it is even more important to
improve our existing operations. As discussed in this dissertation, many studies in the past decade have developed numerous models to better understand capacity, demand and air traffic management. However, owing to the emphasis placed on human judgment and experience, and safety, many research contributions fail to get implemented into the aviation technology ecosystem that is used by air traffic specialists and traffic operators for decision-making.

This dissertation works specifically on ideas that allow air traffic specialists to harness their own experience in the past, which may increase their confidence on the results and recommendations of this work. In this dissertation we develop algorithms to identify similar days at an airport using capacity, demand, TMI and operational outcomes data. Identifying historical similar days can augment the experience of air traffic specialists and help them make data-driven and systematic decisions to better manage an airport. The dissertation presents a first attempt at developing capacity prediction models that generate capacity distributions used for estimating similarity measures. In section 5.3, we discussed the numerous extensions desirable to the capacity prediction model and the similarity framework discussed in this dissertation. When these extensions are incorporated, the improved capacity prediction models and similarity estimation frameworks can better capture the interaction of different components of an airport and its role in the aviation system. The similar days from such comprehensive models will provide efficient decision-support to air traffic specialists and assist in their efforts to manage uncertainty in the aviation systems in a safe and efficient manner.
Bibliography


Gorripaty et al., 2017. Identifying similar days for air traffic management. Journal of Air Transport Management (accepted)


