Airline Schedule Planning and Operations: Optimization-based Approaches for Delay Mitigation

by

Lavanya Marla

B.Tech., Civil Engineering, Indian Institute of Technology, Madras (2004)
S.M., Operations Research and Transportation, Massachusetts Institute of Technology (2007)

Submitted to the Department of Civil and Environmental Engineering

in partial fulfillment of the requirements for the degree of

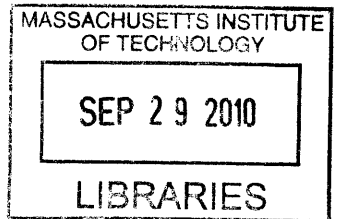Doctor of Philosophy in the field of Transportation Systems

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2010

© Massachusetts Institute of Technology 2010. All rights reserved.

Author . . . .
Department of Civil and Environmental Engineering
August 18, 2010

Certified by . . . .
Cynthia Barnhart
Ford Professor of Civil and Environmental Engineering and Engineering Systems
Associate Dean for Academic Affairs, School of Engineering
Thesis Supervisor

Accepted by . . . .
Daniele Veneziano
Chairman, Departmental Committee for Graduate Students

Airline Schedule Planning and Operations: Optimization-based Approaches for Delay Mitigation

by

Lavanya Marla

Abstract

We study *strategic and operational measures* of improving airline system performance and reducing delays for aircraft, crew and passengers. As a strategic approach, we study *robust optimization* models, which capture possible future operational uncertainties at the planning stage, in order to generate solutions that when implemented, are less likely to be disrupted, or incur lower costs of recovery when disrupted. We complement strategic measures with operational measures of managing delays and disruptions by integrating two areas of airline operations thus far separate - *disruption management and flight planning*.

We study different classes of models to generate *robust airline scheduling* solutions. In particular, we study, two general classes of robust models: (i) extreme-value robust-optimization based and (ii) chance-constrained probability-based; and one tailored model, which uses domain knowledge to guide the solution process. We focus on the aircraft routing problem, a step of the airline scheduling process. We first show how the general models can be applied to the aircraft routing problem by incorporating domain knowledge. To overcome limitations of solution tractability and solution performance, we present budget-based extensions to the general model classes, called the *Delta* model and the *Extended Chance-Constrained programming* model. Our models enhance tractability by reducing the need to iterate and re-solve the models, and generate solutions that are consistently robust (compared to the basic models) according to our performance metrics. In addition, tailored approaches to robustness can be expressed as special cases of these generalizable models. The extended models, and insights gleaned, apply not only to the aircraft routing model but also to the broad class of large-scale, network-based, resource allocation. We show how our results generalize to resource allocation problems in other domains, by applying these models to pharmaceutical supply chain and corporate portfolio applications in collaboration with IBM's Zurich Research Laboratory. Through empirical studies, we show that the effectiveness of a robust approach for an application is dependent on the interaction between (i) the robust approach, (ii) the data instance and (iii) the decision-maker's and stakeholders' metrics. We characterize the effectiveness of the extreme-value models and probabilistic models based on the underlying data distributions and performance metrics. We also show how knowledge of the underlying data distributions can indicate ways of tailoring model parameters to generate more robust solutions according to the specified performance metrics.

As an operational approach towards managing airline delays, we integrate *flight planning with disruption management*. We focus on two aspects of flight planning: (i) flight speed changes; and (ii) intentional flight departure holds, or delays, with the goal of optimizing the trade-off between fuel costs and passenger delay costs. We provide an overview of the state of the practice via dialogue with multiple airlines and show how greater flexibility in disruption management is possible through integration. We present models for aircraft and passenger recovery combined with flight planning, and models for approximate aircraft and passenger recovery combined with flight planning. Our computational experiments on data provided by a European airline show that decrease in passenger disruptions on the order of 47.2%-53.3% can be obtained using our approaches. We also discuss the relative benefits of the two mechanisms studied - that of flight speed changes, and that of intentionally holding flight departures, and show significant synergies in applying these mechanisms. We also show that as more information about delays and disruptions in the system is captured in our models, further cost savings and reductions in passenger delays are obtained.

Thesis Supervisor: Cynthia Barnhart
Title: Ford Professor of Civil and Environmental Engineering and Engineering Systems
Associate Dean for Academic Affairs, School of Engineering

Credits

Chapter 3 and Chapter 4 of this thesis are a result of collaborations with IBM Research's Zurich Research Laboratory and with Jeppesen Commercial and Military Aviation. I sincerely thank them for these collaborations.

I thank Dr. Eleni Pratsini for facilitating my internship at IBM Research Zurich, which led to Chapter 3. I am grateful to Dr. Eleni Pratsini and Dr. Gautier Stauffer for our many valuable discussions during my internship at IBM Research Zurich; and Alexander Rikun for his contribution to the corporate portfolio problem and insights on robust optimization from our discussions.

Chapter 4 resulted through a collaboration with Bo Vaaben, an employee of Jeppesen Commercial and Military Aviation, who is currently pursuing an industrial PhD program at the Technical University of Denmark under the supervision of Prof. Jesper Larsen. I thank Bo for providing me with data from Jeppesen, for facilitating discussion with airlines on the state-of-the-practice of flight planning, and for help with the Jeppesen simulator.

I thank Prof. Cynthia Barnhart, Prof. Patrick Jaillet, Dr. Hamsa Balakrishnan and Dr. Marta Gonzalez for their valuable advice as members of my thesis committee.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Prof. Cynthia Barnhart, for her encouragement and support during my PhD. I have benefited from Cindy's unique research perspective, her depth of knowledge and her approach to solving real-world problems with a blend of theory and practice. Cindy has given me complete freedom and unwavering support to pursue my own research interests, and has encouraged all collaborations. Chapters 3 and 4 of this thesis, which are a result of collaborations with IBM Research Zurich and Jeppesen, would not have been possible without Cindy's strong backing. Her wisdom, patience, and insights have taught me in more ways than one. I sincerely thank her for the wonderful experience she has facilitated for me.

Special thanks to Patrick for his valuable suggestions and guidance through my research, and for valuable advice during my job search. I would like to thank Hamsa and Marta, for their comments and suggestions on my work, and for taking the time to speak with me on a wide range of topics - research, job talks and other career advice. You have guided me through advice as well as example.

I warmly thank Dr. Eleni Pratsini for providing me with a great internship experience at IBM Research Zurich. This translated into Chapter 3 of my thesis. I am also thankful for your valuable perspectives, feedback and mentorship. Thanks to Gautier and Alex for their collaboration and the fun discussions. I also owe thanks to Bo Vaaben and Steve Altus from Jeppesen for our research discussions on three-way conference calls. These discussions helped me define the problem that ultimately led to Chapter 4. Special thanks to Bo, for our long research discussions, for being so quick to respond on email, and providing me with flight plans and data. It was great having you visit MIT as part of our collaboration.

I owe much to the teachers who inspired me in undergrad, and encouraged me on the journey to a PhD. I also thank faculty members Nigel Wilson, Mikel Murga, Georgia Perakis and others in the CEE department and the ORC for being generous with their time. Thanks, Maria, Ginny, Patty, Kris and Jeanette for smoothing out all administrative issues.

Thanks to Vikrant, Ta, Pavithra, and Doug for all our interesting research discussions. Vikrant and Ta, your enthusiasm is infectious, and I had great fun in our chats at the office,

7

research and otherwise. Thanks to the colleagues from the NSF-EFRI, CTL and Transportation@MIT seminar series for your research perspectives. To Niklaus and Nicoleta for collaborating on the airline competition and UPS problems. To friends in the Transportation Students Group and Operations Research Center who have made my stay at MIT a wonderful experience. To my officemates over the years - Andre, Liz, Vikrant, Ta, Miao, and others - for some great memories; and especially Andre for the hike in Switzerland!

I thank all the friends who have inspired, challenged and encouraged me through this journey. To Viji, Madhu, Padma, Naveen, Pavithra, Vikram, and many others for their support. To KP, Prabha, Mythili, Anima, Jayku, AT, Vivek, Varun, Anna for our enjoyable chats. To friends from the MIT India Reading Group for the times spent discussing developmental issues in India, for inspiring me with your passion, commitment and ideas. Also a big thank you to Xiaolu for sharing her tips on thesis writing!

I am grateful to have had this unforgettable experience, and for the many lessons I learned along the way. I thank the numerous people who have contributed in ways small and large, seen and unseen, to the culmination of this work and this period of learning.

Finally, a big thank you to Mum, Dad and Ramya for their constant support, help, and confidence in my ability to succeed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, we study *strategic and operational techniques* to improve airline system performance and reduce delays for aircraft, crew and passengers. As part of strategic measures, we study *robust optimization* techniques, by which possible future operational uncertainties are modeled at the planning stage in order to generate solutions that when implemented, are more likely to be executed or easier to repair when disrupted. We complement strategic measures using operational measures of managing delays and disruptions by integrating two areas of airline operations hitherto separate - *disruption management and flight planning*.

1.1 Motivation

Aviation is an integral part of the international economy, with steady growth in developed countries and explosive growth in developing countries. In 2008, there were 26,245 aircraft departures and 4,282,870 million revenue passenger kilometers worldwide, with revenues of $563,640 million [Air]. In the United States, aviation generates $1.2 trillion in economic output, and is 5.2% of the US Gross Domestic Product [Fed09].

Air travel has increased tremendously in the past decade. 4,282 billion passenger revenue kilometers were traveled across the world in 2008 compared to 2,797 billion passenger revenue kilometers in 1999, a growth of 53% [Air]. While the revenue passenger kilometers traveled in North America rose 22% from 1999 to 2008, an even higher growth of 71% was observed in Europe and the Asia-Pacific regions [Int10b]. Though this growth slowed somewhat towards

the end of 2008 and in 2009 due to the recent economic crisis, it is expected that there will be a recovery in traffic to pre-recession levels [TP09]. With these high growth rates, the percent of US aircraft arriving late (as defined by the 15-minute on-time performance metric) has increased from 17.01% in 2003 to 25.96% in 2008 [Bur09a]. These delays are highly detrimental in an industry where the profit margins are typically less than 2%, with delay costs forming a major component of operating costs [Int08]. As a case in point, the total costs of U.S. domestic air traffic delays crossed the $40 billion mark [JEC08]. Of these delay costs, $19.1 billion represents incremental operating costs for the airlines (including additional fuel costs due to delays of $1.6 billion, releasing 7.1 million metric tons of carbon dioxide into the atmosphere), $12 billion represents the estimated passenger costs due to low productivity and lost business, and $10 billion represents losses to other industries that rely on air traffic [JEC08]. It is evident that if delays can be reduced, society and the airline industry would benefit tremendously.

The airline system operates as a highly interconnected network, with aircraft, crew and passengers interacting closely. Planning an airline's operations involves capturing the complex interactions between airports, gates, airport slots, aircraft fleet types and associated maintenance restrictions, crew restrictions and passenger demands. For large airlines which daily operate thousands of flights, with thousands of aircraft and crew, this this can be a considerable computational challenge. Due to this reason, the process of designing an airline's operations, called airline schedule planning, is usually decomposed into four (usually) sequential sub-problems: (1) schedule design, (2) fleet assignment, (3) aircraft maintenance routing and (4) crew scheduling. We briefly describe these sub-problems here. Given a set of markets desired to be served, *schedule design* is the process of determining the set of flight legs (as described by origin, destination and departure and arrival times) to be operated by an airline so that its profitability is maximized. Following schedule design, the *fleet assignment* problem is to assign to each flight leg a type of aircraft so that passenger travel demands can be met. The operating cost of each flight leg is the cost of flying a particular aircraft type on that leg, plus a spill cost equal to the lost revenue of passengers who could not book the flight leg due to insufficient numbers of available seats. The fleet assignment cannot use more aircraft of each type than available in the airline's fleet, and the assignment must assure flow balance of each fleet type. Once fleet assignment is completed, each *individual* aircraft is assigned a routing, or path in the network,

by solving the *aircraft maintenance routing* problem. The crucial objective of the maintenance routing problem is to ensure that each aircraft receive periodic maintenance checks at a maintenance station, as required by the Federal Aviation Administration (FAA). Given the fleeting and routing decisions made in the previous steps, the *crew scheduling* problem is to generate cost-minimizing activity schedules for the cockpit and cabin crew so that each flight leg is assigned the appropriate crew members. The crew schedule should follow numerous restrictions that are a result of government-mandated work rules, as well as bargaining agreements between the airline and its employees. For a detailed description of the airline schedule planning process, we refer the interested reader to Barnhart [Bar09a].

Because each flight leg can be successfully operated when its resources - the aircraft, crew, airport gates, runways, etc. - are *all* available at the scheduled time, there is a close interaction of all these resources in the system. Each of the complex interactions described can be subjected to uncertainty and delay in the system. Inherent *uncertainty* in operations, manifesting as weather, airport and airspace congestion, crew sickness, aircraft maintenance, security, etc., can be reasons for any one of the resources required to operate a flight to not be available on time, with resulting delays or disruptions. Due to this uncertainty, planned schedules are rarely, if ever, executed. In addition, the close interconnections of these resources results in strong *network effects* - that is, a delay occurring in one part of the network or to one of the resources can propagate downstream to multiple other resources and other parts of the network. These effects are referred to as propagated [LCB06] or reactionary delays [CT09]. For example, a flight arriving late can cause both its aircraft and crew to be delayed in downstream operations. According to the Bureau of Transportation Statistics [Bur10], a distribution of delays by cause, for the period of May 2005 - May 2010 is as shown in Figure 1-1.

When delays or disruptions occur in the network, airlines undertake *disruption management*. Disruption management is usually under the purview of the Airline Operations Control Center (AOCC). The goal of disruption management is *schedule recovery*, that is, to bring the plan back on schedule as quickly and cost-effectively as possible, and minimize the additional operating costs incurred due to the disruptive events. The scope of the recovery problem spans flight schedules, aircraft routes, crew assignments and passenger routes. Due to its complexity, again, the recovery problem is solved in stages, with typically aircraft recovered first, followed by

On Time - 76.46%

Air Carrier Delay - 6.00%

Weather Delay - 0.84%

National Aviation System Delay - 7.53%

Security Delay - 0.06%
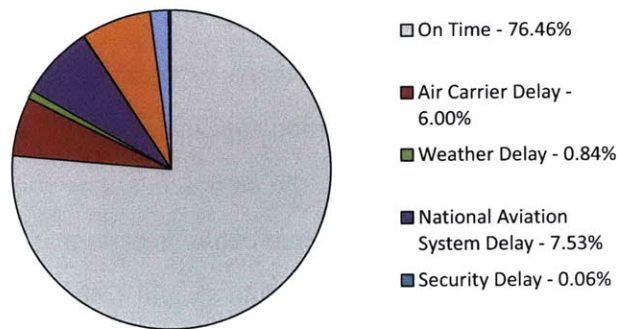
Figure 1-1: US National Airlines' On-Time Arrival Performance (May, 2005 - May, 2010)

crew (which are both resources required for the system to operate); and finally, passengers. For a detailed description of the recovery process, we refer the reader to Barnhart [Bar09b].

To minimize the additional operational costs resulting from these disruptions, different approaches may be used, including:

1. *Strategic approaches*: These approaches focus on *robust schedule design*, by which a schedule that is less sensitive to operational uncertainty is designed at the planning stage, before the day of operation. Understanding that delays and disruptions in the system are inevitable, robust airline scheduling *pro-actively* considers possible delays and disruptions as schedules and plans are developed; with the objective of building plans that are less susceptible to disturbances or, are easier to repair once disrupted. This might be achieved by building plans that require fewer recovery (disruption management) actions, or that decrease the complexity or cost of recovery. Robust airline scheduling, then, is a proactive planning technique aimed at reducing total *realized* costs, including both plan and recovery costs. This is in contrast to early practices, in which responses to disruptive events were *reactive*, that is, actions to manage delays are taken only after an event occurs, resulting in schedule recovery actions which can be costly and complex to implement.

2. *Operational approaches:* Operational approaches are undertaken by the airline on the day of operations, either in anticipation of or after a delay or disruption of the schedule, after information about the disruption or delay is revealed. These measures come under the category of disruption management. Operational approaches are dynamic in their

20

implementation, and their goal is often to bring the plan back on schedule as quickly as possible, while incurring minimal costs, and disrupt as few resources or passengers of the system. Measures such as flight cancelations, flight re-timing, aircraft swaps, use of reserve crews, etc. can be used.

Strategic measures and operational measures complement each other. Ideally, strategic and operational approaches should be synergistic, and work together in order to minimize the realized costs of the system. In this thesis, we focus on both strategic and operational measures of disruption management.

1.2 Thesis Contributions and Structure

In this section, we present the context for our approaches of robust schedule design and enhanced disruption management. We describe the contributions of our work and structure of this thesis.

1.2.1 Chapter 2: Robust aircraft routing

Deterministic models of airline schedule planning that are commonly used in practice do not capture information about potential future uncertainty, and therefore, render the system vulnerable to delays and disruptions. During the past decade, there has been considerable interest in exploring the benefits of robust approaches which pro-actively make the system less vulnerable to uncertainty on the day of operations.

Several types of approaches have been studied in the broader literature for modeling uncertainty and building more robust solutions. These approaches can be categorized as (i) probability distribution-free models, (ii) probability distribution-based models, and (iii) problem-specific models. The first two categories are general approaches that can be applied to any mathematical program. In Chapter 2, we ask the question if more general robust approaches belonging to the probability distribution-free and probability distribution-based categories can be applied successfully to airline schedule planning.

21

We focus on the aircraft routing problem, a step of the airline scheduling process. In particular, we study two general classes of robust models and one tailored approach that uses domain knowledge to guide the solution process. The first class of models is a distribution-free, extreme-value based approach, proposed by Bertsimas and Sim [BS04]. The second class of models is a probability-based chance-constrained approach, proposed by Charnes and Cooper [CC59].

In order to meaningfully apply the general paradigms of capturing uncertainty to aircraft routing, we show that domain knowledge about the problem should be captured. Through experiments conducted on data from a major US hub-and-spoke carrier, we discover that existing models face limitations in modeling the performance metrics and in solution tractability. To overcome these limitations, we present new models in both the extreme-value and the probabilistic paradigms, which we call the *Delta model* and the *Extended Chance-Constrained Programming model (ECCP)* respectively. Our extended models avoid the need to repeatedly re-solve to gain robust solutions, which was an issue for the basic models. Also, the run times of our extended models are comparable to a single iteration of the models. Both these features enhance solution tractability. Our extended models re-define robustness as maximizing a robust parameter within a budget, and consequently also generate consistently more robust solutions than the basic models according to our performance metrics.

We show that the tailored model for robust aircraft routing is a special case of the chance-constrained programming model. The solutions of the ECCP model, reflecting the focus of the model on high-probability delay events, are robust with respect to our metrics of interest. Extreme-value based models, on the other hand, due to a focus on worst-case delays, can generate solutions with good worst-case performance, *and* a high degree of variability in our performance metrics. This is due to the underlying data distributions for the hub-and-spoke US carrier under consideration, where bi-modal delays are seen with delay either at the lower end of the scale (with a probability of 85-90%) or at the higher end of the range (with a probability of 10-15%).

Our work thus underscores the importance of choosing an approach that aligns well with both the data distributions for the aircraft routing problem, as well as the the metrics of interest to the Department of Transportation (DoT), the airline and passengers. The extended models

and insights gleaned in this work apply not only to the aircraft routing model but also to the broad class of large-scale, network-based, resource allocation problems.

1.2.2 Chapter 3: Robust optimization - other applications

The goal of using general methods such as the Bertsimas and Sim and Chance-Constrained Programming method in Chapter 2 is to be able to apply them to various applications, within the airline scheduling context as well to other domains. In this chapter, we study and compare the application of various generally applicable robust approaches to multiple problems, namely, strategic supply chain design for a Top-50 pharmaceutical manufacturer, portfolio optimization for a global corporation, and aircraft routing for a US carrier (studied in Chapter 2). The pharmaceutical supply chain problem and the corporate portfolio problem arose in collaboration with IBM Research's Zurich Research laboratory. We consider various approaches of robustness and develop insights that can help in applying these methods to a broad variety of problems.

From empirical studies conducted on real-world data available for the three applications, we observe that the effectiveness of solutions generated is affected by the robust modeling approach, the underlying data and the performance metrics of interest. Extreme-value-based models are seen to be best applied when the underlying data distribution is known with less certainty - that is, the type of distribution, or the spread of data cannot be well-estimated. As more information about the system is available with some certainty (even in the form of quantiles rather than complete distributions), probabilistic models that can capture partial/full distribution information produce more effective results. Because extreme-value based models focus on the worst-case, they produce conservative solutions geared towards worst-case metrics, whereas probabilistic models produce less conservative solutions geared towards average-case metrics. We also show how knowledge of the underlying data distribution, even if partial, or empirically derived, can indicate ways of modifying input parameters of extreme-value and probabilistic models to produce more robust solutions, according to the specified performance metrics.

1.2.3 Chapter 4: Integrated Disruption Management and Flight Planning

In this chapter, our focus is on operational approaches for airline operations. Disruption management procedures are in place at airlines to bring operations back on track when disruptive events occur, and to reduce recovery costs (which contribute to operating costs).

We include flight planning in an *enhanced disruption management* tool, by providing optimization models that combine flight planning with traditional disruption management models during operations. In particular, we focus on two aspects of flight planning: (i) flight speed changes; and (ii) intentional flight departure holds, or delays, with the goal of optimizing the trade-off of fuel costs and passenger delay costs. Our approach represents an integration of two aspects of airline operations before studied separately, namely, disruption management and flight planning.

Through dialogue with multiple airlines, we provide an update of the current state-of-the-practice with regards to flight planning approaches. We also discuss the current practice in the disruption management area. We identify opportunities for enabling greater flexibility in disruption management using flight planning, possible by integrating these elements, and show the need for optimization-based decision support.

We present models for aircraft and passenger recovery combined with flight planning, and models for approximate aircraft and passenger recovery with flight planning. With these models, we provide a means for optimizing trade-offs between delayed passenger costs and fuel costs, with the goal of minimizing total realized costs.

Our experiments involve the hub operations of an international carrier. In comparison with conventional disruption management, we demonstrate that our enhanced disruption management strategy helps decrease passenger-related operating costs for the airline by reducing passenger misconnections by 47.2% - 57.3%. We demonstrate the dynamic nature of the trade-off frontier between passenger costs and fuel burn costs and discuss in detail the interactions involved in this trade-off under different disruption scenarios. We also discuss the relative benefits of the two types of mechanisms studied - that of flight speed changes, and that of intentionally holding flight departures - and show significant synergies in applying the two mechanisms simultaneously.

We conclude the thesis in Chapter 5 by summarizing the contributions and findings of this research and providing directions for future research.

Chapter 2

Robust Aircraft Routing

2.1 Introduction

Robust airline scheduling is a way of *pro-actively* considering delays and disruptions and creating schedules with the objective of building plans that are less susceptible to disturbances or easier to repair once disrupted. This is in contrast to prior practice, where responses to delays were *reactive*, that is, after an event occurred, schedule recovery actions which can be costly and complex to implement, were taken. Robust airline scheduling, then, is a proactive planning technique aimed at reducing total *realized* costs, including both plan and recovery costs.

To evaluate the robustness of solutions obtained, we use *simulation*, as the objective function values of the planning optimization models do not indicate the *realized costs* or *robustness* of the solution. Through simulation, we measure solution performance with respect to a host of relevant robustness metrics.

2.1.1 Robust Airline Scheduling

Several approaches to build robust airline scheduling solutions, tailored to the airline industry, have been developed.

Ehrgott and Ryan [ER02] provide a bicriteria optimization framework to develop pareto optimal solutions for the crew scheduling problem. The two criteria of interest are cost and robustness. Robustness into crew schedules in built by incorporating sufficient ground times

if crews have to change aircraft, or keeping the crew with the aircraft in case of tight ground times. They define a linear non-robustness penalty function based on the expected delay of each flight in the schedule, and if the crew are required to change aircraft for a tight ground time. However, they do not provide details on the calculation of expected delay. Ehrgott and Ryan's formulation trades off the cost function and the penalty function for non-robustness. Within allowable budgets of deviation from the minimum cost solution, they find the most robust solution. Further, they show that this can be solved effectively by modeling the constraints as 'elastic' constraints, which generate solutions that are part of the pareto optimal frontier. Their results clearly demonstrate the trade-off between cost and robustness. However, the authors do not evaluate the pareto optimal frontiers and their true performance through simulation.

Ageeva [Age00] creates robust airline schedules by focusing on the aircraft routing part of the airline scheduling process. Aircraft routings are considered to be more robust if they contain more opportunities to 'swap'. Two aircraft meet if their routes contain the same location within a specific time-window. The aircraft may be swapped if they meet twice along their routes, allowing for a switch in routes at the first meeting point and swapped back to the original routes at a later meeting point. Such swaps increase flexibility of aircraft availability in disruptive scenarios. Ageeva examines multiple optimal solutions to the aircraft routing problem and ranks them by their robustness, as measured by the number of swaps. The results indicate that robustness of the resulting aircraft routing, as measured by the number of swaps, can be increased as much as 35% compared to the original routing. However, this work stops short of evaluating the resulting robust solutions through simulation.

Rosenberger, Johnson and Nemhauser [RJN04] develop a robust model for fleet assignment and aircraft routing that allows for many 'short cycles'. A cycle is a sequence of flights that begins and ends at the same airport. When flights are canceled in disruptive scenarios, airlines cancel not just one flight, but a cycle containing that flight, in order to reposition the aircraft correctly. By increasing the number of *short* cycles containing fewer flights, the number of flights canceled when one flight is canceled is decreased. In addition, the goal of such short cycles is also to decrease hub connectivity. Hub connectivity indicates the number of aircraft rotations that include more than one hub. Higher hub connectivity means that disruptions can spread from hub to hub in the network. Evaluating the solutions via simulation, the authors

show that the incorporating robustness in the form of short cycles produces fleet assignment solutions that decrease planned operating costs and passenger spill.

Schaefer, Johnson, Kleywegt and Nemhauser [SJKN05] introduce a measure to evaluate crew schedules in practice. Their measure approximates both the planning cost as well as the operational cost of a crew schedule. They propose two methods based on (i) expected operating cost, which is calculated using SimAir, a MonteCarlo simulation of airline operations, and (ii) penalizing pairing properties that may result in poor performance. Using simulation, they illustrate that solutions generated using these approaches perform better under uncertain conditions compared to deterministic approaches that do not take uncertainty into account.

Yen and Birge [YB06] build robust crew schedules by modeling crew scheduling as a two-stage stochastic program. The first stage of the model solves the standard crew scheduling model that minimizes expected costs. The second stage recourse model minimizes the expected costs of crews being swapped between planes. They introduce a novel branching scheme to solve the stochastic program. The results exhibit the trade-off between planned crew costs and recourse costs.

To address the issue of demand uncertainty, Jiang [Jia06] introduces fleet re-timing as a dynamic scheduling mechanism and supplements re-fleeting with re-timing. Re-timing the schedule and re-fleeting of aircraft increase or decrease the number of connecting itineraries available to passengers (compared to the original schedule) and increase or decrease the number of seats available in the affected markets. Jiang shows that this can help to reduce passenger spill by better matching capacity. The dynamic scheduling approach modifies the existing flight schedule and fleet assignments, keeping existing bookings still feasible (though possibly re-timed), so that realized demand can be accommodated as much as possible. Jiang [Jia06] shows that through the dynamic mechanisms of flight re-timing and re-fleeting, even 'optimized' schedules can be improved by re-designing the schedule at regular intervals.

Shebalov and Klabjan [SK06] propose robust approaches tailored to specific instances of crew scheduling problems by exploiting the specialized structure of the problem. The authors introduce the concept of move-up crews and improve costs by swapping crews, and show the resulting benefits.

Lan, Clarke and Barnhart [LCB06] propose an 'intelligent routing model' for aircraft routing

to reduce *delay propagation* along the downstream flight legs. They show that aircraft routings can be made significantly more robust by re-arranging the slack in the schedule to place it where it is needed to a greater extent. Thus robustness can be improved without changing the total slack and adding to planning costs. In addition, they introduce a new approach to minimize the number of passenger mis-connections by re-timing the departure times of flight legs within a small time window. Their approach helps to improve passenger connection times without significantly increasing costs. Their approach is very relevant to this work, and will be discussed in greater detail in later sections. AhmadBeygi, Cohn and Lapp [ACL10] expand on this notion of propagated delay and model propagation of delay using the concept of propagation trees. They measure delay propagation to aircraft, passengers and crew, and measure the extent that each delay propagated down a tree. Using simulation, they show that airline schedules that consider delay propagation can significantly decrease operational delays without increasing planning costs.

2.1.2   Motivation

In this chapter, our focus is on the aircraft routing step of the airline scheduling process. The aircraft routing problem is to find a feasible sequence of flight legs, called aircraft routings or rotations, to be operated by each aircraft so that maintenance restrictions on aircraft are satisfied. Each flight is required to be assigned to (or covered by) exactly one aircraft, using no more than the available number of available aircraft. and meeting all maintenance requirements. Though robust planning is required at every step of the airline scheduling process, we choose aircraft routing because of its high impact on schedule reliability and relatively low impact on crew costs, flight costs and passenger revenues [LCB06].

We demonstrate how aircraft routings differ and what we mean by robust aircraft routings, with an example. In Table 2.1, we report performance for 7 aircraft routings as measured by the percent of flights in the routing that arrive within 15 minutes, 30 minutes, 60 minutes, 120 minutes and 180 minutes of their respective scheduled arrival times. Note that these percentages were calculated over 22 days of operations of a major U.S. airline. For the instances under consideration in this paper, all of which are drawn from actual airline operations, we compare

metrics of interest, as detailed in §2.4.1. The reported variability in flight delays is significant, as even small differences in the range of 1% can improve/deteriorate the airline's ranking with respect to the DoT's 15-minute on-time performance metric [Bur09c]. Because airlines do not typically explicitly consider delays in selecting aircraft routings, the airline effectively might choose at random any of these routings, and thus, can incur high delays. To illustrate, for this instance, the aircraft routing operated by the airline is Routing 5, with DoT on-time performance ranking third from the bottom. Moreover, in addition to aircraft delay disparities, different routings can lead to different levels of *passenger* disruptions and delays. A passenger is considered to be disrupted if one or more flight legs on his itinerary are canceled, or if delays cause insufficient connection time to the next flight leg in his/her itinerary. The percentage of passenger disruptions decreased relative to the airline's routing (% D-pax reduced)are shown in Table 2.1. Routings 1 and 2 can vastly improve upon the airline's routing without any additional resources, while Routing 7 can deteriorate the airline's performance greatly.

Routing	Flight Delays					Pax Disruptions	
	$\leq$15 min	$\leq$30 min	$\leq$60 min	$\leq$120 min	$\leq$180 min	#D-pax	%D-pax reduced
Routing 1	79.1	86.7	93.4	98.0	99.1	988	10.14
Routing 2	78.8	86.8	93.2	98.2	99.2	986	10.30
Routing 3	78.3	86.2	92.9	98.1	99.0	1028	6.50
Routing 4	78.3	86.0	92.9	97.5	98.6	1047	4.80
Routing 5	77.7	85.8	92.8	97.7	98.9	1100	0.00
Routing 6	77.6	85.7	92.4	97.4	98.6	1057	3.90
Routing 7	76.5	84.7	92.0	97.2	98.5	1223	-11.20

Table 2.1: Flight Delay Percentages and Passenger Disruptions of Feasible Routings, N_2

The relationship between aircraft routings and delays and disruptions can be explained by the phenomenon of *propagated delays*. In network structures, flight delays can be divided into two components [LCB06]: *independent delays* that originate at the flight's origin or during the flight, and *propagated delays* resulting from delays in upstream flights that are not absorbed by slack time between flight legs. Delay propagation is illustrated in Fig. 2-1. The solid arrows show the planned schedule for flights f_1 and f_2; and the dotted arrows the operated schedule. PDT, ADT, PAT and AAT are the planned departure time, actual departure time, planned arrival time and actual arrival time respectively, of flight f_2. In Figure 2-1, flight f_1 is delayed, and its delay causes the remaining slack time between flight f_1's arrival and f_2's scheduled

departure to be less than the minimum connection time required for the same aircraft to fly both flight legs. This causes propagated delay PD for flight leg f_2. In addition, independent delay is incurred by f_2, both at its departure (IDD) and its arrival (IAD), resulting in total departure delay (TDD) and total arrival delay (TAD). However, by changing the sequence of flights operated by each aircraft, propagated delay can be reduced (Figure 2-2). This involves changes in the routings of aircraft, but because *no* new aircraft are being added and the flight schedule remains unchanged, the total slack in the system is not altered, instead only the positioning of the slack is changed.
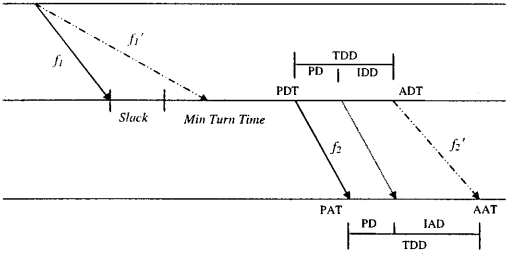
Figure 2-1: Delay Propagation along an Aircraft Route
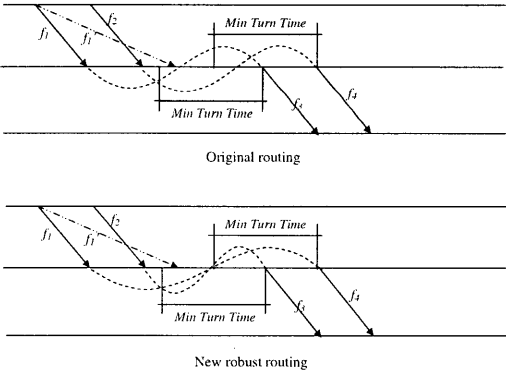
Figure 2-2: Robust Routing with Optimal Slack Allocation

For the airline we study, propagated delay typically represents 20% to 30% of total flight delay [LCB06]. Because total independent delay is a constant for the flight schedule, reducing delay propagation by choosing Routing 1 instead of, for example, Routing 7 has the effect of

32

reducing *total* delay. The differences in propagated delays for Routings 1-7 are shown in Fig 2-3. Different aircraft routings are not very different on 'good' days like Day 10, but differences become apparent on 'bad' days like Day 5.
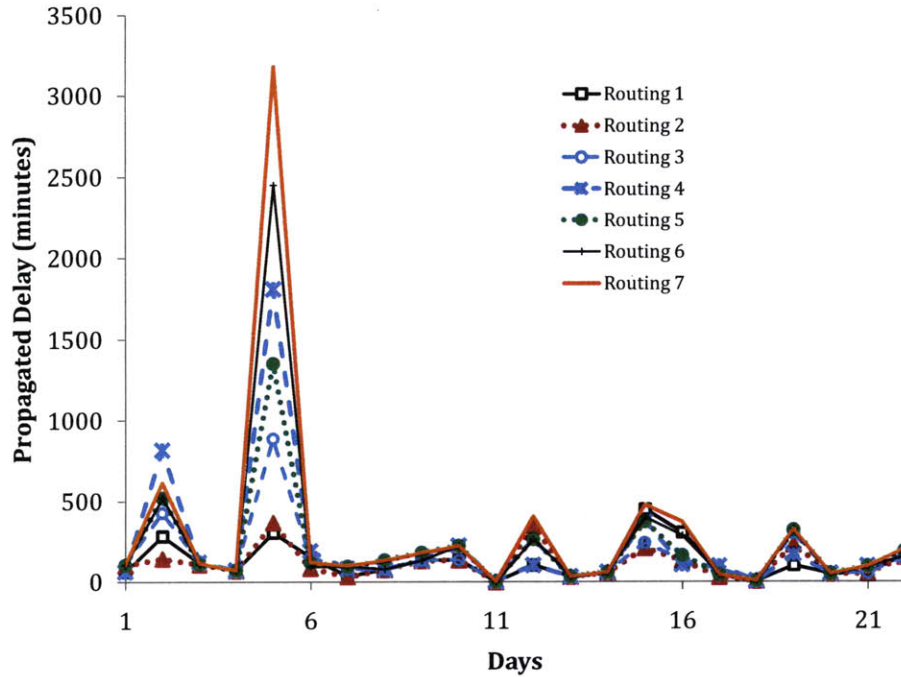
Figure 2-3: Propagated Delays of Feasible Aircraft Routings, N_2

2.1.2.1   Discussion of Metrics

Ideally, good solutions to airline scheduling problems ensure low levels of delay for flights, and good travel experience with low passenger delays and disruptions. Metrics, then, such as total flight delay minutes, total cost of delay, 15-minute on-time performance, 30-min on-time performance, and 60-minute on-time performance, are all examples of measures that reflect airline schedule reliability and robustness. A difficulty, however, is that these metrics are not always aligned with each other. For example, the 15-minute on-time performance metric does not reflect delays greater than 15 minutes, and therefore maximizing 15-minute on-time performance is not the same as minimizing total delay minutes. Similarly, minimizing aircraft delay minutes is different from minimizing passenger delay minutes or passenger disruptions because fewer passengers can be disrupted by holding flights to allow passengers to make their connections, thus increasing total aircraft and passenger delays.

33

2.1.3 Contributions

In this chapter, we study three different approaches to robustness in aircraft routing, and hence, airline scheduling - two that are generally applicable, the extreme-value based approach and a probabilistic chance-constrained programming approach; and one that is a tailored approach proposed by Lan, Clarke and Barnhart [LCB06]. The extreme-value approach considered is the robust optimization approach of Bertsimas and Sim [BS04], [BS03] and the probabilistic approach is the Chance-Constrained Programming approach of Charnes and Cooper [CC59], [CC63]. We begin by showing how to model the robust aircraft routing problem using these three approaches, and identify their respective limitations; suggesting extensions and enhancements to the models to address these limitations. We then evaluate the similarities and differences in models and solutions generated by these different approaches, using a simulation-based evaluator. The findings and extensions from this work are generally applicable to the broad class of network-based resource allocation problems.

2.1.4 Outline

In §2.2 we discuss Charnes and Cooper's Chance-Constrained Programming and Bertsimas and Sim's extreme-value approach. In §2.3, we present the three classes of robust models for aircraft routing. For this application, we present Charnes and Cooper's Chance-Constrained Programming formulation, the robust optimization formulation of Bertsimas and Sim, and Lan, Clarke and Barnhart's robust aircraft routing formulation. In addition, we propose extensions and enhancements to the general classes of models in §2.3. We present the experimental set-up for our computations, and details of the simulator built to evaluate the performances of the models in §2.4. In §2.5, we compare the models and solutions generated by the different approaches in terms of complexity and run times (§2.5.1), model parameters (§2.5.2) and modeling paradigms (§2.5.3), and show how robust solutions may be generated by all classes of models. In §2.6, we summarize our findings.

2.2 Robust Models of Bertsimas and Sim, and Charnes and Cooper

2.2.1 Robust Formulation of Bertsimas and Sim

Consider a standard linear program, that is:

$$\max \quad \mathbf{c}^T \mathbf{x} \tag{2.1}$$

$$\text{s.t.} \quad \sum_{j \in J} a_{ij} x_j \leq \mathbf{b} \tag{2.2}$$

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u}. \tag{2.3}$$

Soyster [Soy73] considers column-wise uncertainty, where each column A_j of the constraint matrix belongs to a convex set K_j. He shows that the above problem is equivalent to the following robust formulation:

$$\max \quad \mathbf{c}^T \mathbf{x} \tag{2.4}$$

$$\text{s.t.} \quad \sum_{j \in J} \bar{a}_{ij} x_j \leq \mathbf{b} \tag{2.5}$$

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u}. \tag{2.6}$$

where $\bar{a}_{ij} = \sup_{A_j \in K_j}(a_{ij})$. This means that extreme (or worst-case) values of coefficients that effectively maximize the amount of slack for the nominal problem are used in the 'robust' model. The use of worst-case values results in solutions that are far from optimal for many realizations of the constraint matrix coefficients.

Bertsimas and Sim [BS04] argue that worst-case approaches such as that of Soyster, are too conservative, and hence, expensive. Instead, they suggest an approach aimed at avoiding the overly conservative tendencies of Soyster's approach by providing a mechanism to control the 'degree of conservatism'.

In the approach of Bertsimas and Sim, all uncertainty is assumed to be located in the coefficients of the $\mathbf{A}$ matrix. By performing some simple transformations and rewriting $\mathbf{A}$, uncertainty

in **c** and **b** can also be captured. By changing the objective function to maximize z and adding the constraint $z - \mathbf{c}^T\mathbf{x} \leq 0$, the objective function can be moved into the **A** matrix, thus enabling uncertainty in the objective function coefficients to be captured. Similarly, if we have uncertainty in the right-hand-side **b**-vector, the **b**-vector values can be subtracted from the left-hand side and the right-hand side can be replaced by zero. The assumption of uncertainty in the **A**-matrix therefore incurs no loss of generality.

Each entry of of the left-hand side of the constraint matrix, **A**, is assumed to be a random variable with $\tilde{a}_{ij}$ being the symmetric, unbounded variable corresponding to the (i,j)th entry of **A**. No actual probability distribution of the random variable is assumed, only an interval of values that $\tilde{a}_{ij}$ can assume. Specifically, a_{ij} denotes the nominal value of $\tilde{a}_{ij}$, which is used in the deterministic formulation, and $\hat{a}_{ij}$ is the half-interval of $\tilde{a}_{ij}$. Hence, $\tilde{a}_{ij}$ can take on values in the interval $[a_{ij} - \hat{a}_{ij}, a_{ij} + \hat{a}_{ij}]$ and the nominal value a_{ij} is the mean value of the symmetric distribution. The *extreme values* that $\tilde{a}_{ij}$ can take are $a_{ij} - \hat{a}_{ij}$ and $a_{ij} + \hat{a}_{ij}$.

Let J_i be the set of coefficients for constraint i that are subject to parameter uncertainty, that is, $\tilde{a}_{ij}, j \in J_i$ takes values from a symmetric distribution as described above. For each constraint i, there is a parameter Γ_i which can take a (possibly continuous) value in the interval $[0, |\mathbf{J}_i|]$. Because it is unlikely that all $|J_i|$ coefficients will assume their worst-case (or extreme) values, Γ_i is used as a means of adjusting the 'level of protection'. The Bertsimas-Sim formulation protects against the case when up to γ_i of the $|J_i|$ coefficients are allowed to assume their extreme values, for all constraints i.

The corresponding robust non-linear model according to the Bertsimas-Sim model can then be written as:

$$\max \quad \mathbf{c}^T\mathbf{x} \tag{2.7}$$

$$\text{s.t.} \sum_j a_{ij}x_j$$

$$+ \max_{\{S_i \cup \{t_i\}|S_i \subseteq J_i, |S_i| = \lfloor \Gamma_i \rfloor, t_i \in J_i \notin S_i\}} \left\{ \sum_{j \in S_i} \hat{a}_{ij}y_j + (\Gamma_i - \lfloor \Gamma_i \rfloor)\hat{a}_{it_i}y_t \right\} \leq b_i \quad \forall i \tag{2.8}$$

$$- y_j < x_j < y_j \qquad\qquad \forall j \tag{2.9}$$

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u} \tag{2.10}$$

$$y \geq 0 \tag{2.11}$$

Because Γ_i can take on continuous values, up to $\lfloor \gamma_i \rfloor$ of the coefficients $\tilde{a}_{ij}$ in constraint i are allowed to take on their worst-case values, and one coefficient a_{it} changes by $(\Gamma_i - \lfloor \Gamma_i \rfloor)\hat{a}_{it}$. In the above formulation, S_i represents the set of uncertain parameters in constraint i that take on their extreme values, such that $|S_i| = \lfloor \Gamma_i \rfloor$, $S_i \subseteq J_i$. $\{t_i\}$ indicates the coefficient a_{it_i}, for constraint i, that changes by $(\Gamma_i - \lfloor \Gamma_i \rfloor)\hat{a}_{it_i}$.

For the ith constraint, the term $\displaystyle\max_{\{S_i \cup \{t_i\} | S_i \subseteq J_i, |S_i| = \lfloor T_i \rfloor, t_i \in J_i \notin S_i\}} \left\{ \sum_{j \in S_i} \hat{a}_{ij} y_j + (\Gamma_i - \lfloor \Gamma_i \rfloor)\hat{a}_{it_i} y_t \right\}$ is a *protection function* that protects against the worst-case realizations of all $\tilde{a}_{ij}, j \in J_i$. The parameterized protection function thus uses Γ_i to offer various *levels of protection*. $\lfloor \Gamma_i \rfloor$ indicates the minimum number of coefficients in constraint i that can assume their worst case values without destroying feasibility of the solution. $\Gamma_i = 0$ represents the deterministic or nominal case, whereas $\Gamma_i = |J_i|$ reduces this formulation to the Soyster formulation.

Bertsimas and Sim [BS04] prove that the above non-linear formulation (2.7) - (2.11) can be cast as a deterministic linear program, as follows:

$$\max \mathbf{c}^T \mathbf{x} \tag{2.12}$$

$$\text{s.t. } \sum_{j \in J} a_{ij} x_j + z_i \Gamma_i + \sum_{j \in J_i} p_{ij} \leq b_i \qquad \forall i \in I \tag{2.13}$$

$$z_i + p_{ij} \geq \hat{a}_{ij} y_j \qquad \forall i \in I, \forall j \in J_i \tag{2.14}$$

$$-y_j \leq x_j \leq y_j \qquad \forall j \in J \tag{2.15}$$

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u} \tag{2.16}$$

$$p_{ij} \geq 0 \qquad \forall i \in I, \forall j \in J_i \tag{2.17}$$

$$y_j \geq 0 \tag{2.18}$$

$$z_i \geq 0 \tag{2.19}$$

The detailed proof of the equivalence of (2.12)-(2.19) with (2.7)-(2.11) is in [BS04].

Thus, the Bertsimas-Sim robust optimization approach ensures that the form of the math program remains linear, and hence more tractable than formulations with non-linearities. Bertsimas and Sim [BS04] also provide probabilistic guarantees on the feasibility of constraints when more than Γ_i coefficients take on their worst-case values. Moreover, they show how this

formulation can be applied to portfolio optimization, knapsack problems, supply chain management [BT03], and network flows [BS03] in order to obtain robust solutions.

The **advantages** of the Bertsimas-Sim model are:

- It is generally applicable to linear programs and integer programs.

- Linear integer programs remain linear integer programs, but contain more variables, degrading tractability minimally.

- Probability distributions for the uncertain data are not required to be known. Uncertainty can be captured knowing the symmetric bounds of variation alone.

- Adjustments to the 'level of robustness' can be made using the Γ parameter, thereby providing measures of the *price of robustness*, that is, the changes in planned objective function value with changes in protection level. Robustness involves backing off from optimality to gain solutions less vulnerable to uncertainty, implying that there is a price associated with achieving each level of robustness.

- This model, with minor alterations, can capture simple correlations between uncertain data in a constraint [BS04]. However, it cannot capture correlations among uncertain data across constraints.

The approach, however, also has some **limitations**:

- To determine the change in planned costs (or profits) as a function of the level of 'protection', the problem has to be re-solved multiple times, once for each different value of Γ_i, for all i. Because the bounds are also not tight, there are very few guidelines to the choice of Γ_i. This poses computational challenges for large-scale problems.

- It assumes symmetric and bounded distributions of uncertainty of parameters about their nominal values.

- It does not incorporate knowledge of probability distributions, if known. This can result in lack of inclusion of problem knowledge in the model.

38

- Probability bounds of constraint violation are derived for each constraint, and cannot be easily extended to an overall protection level for the system.

- This approach is not particularly well-designed for the solution of very large-scale resource allocation problems [Mar07].

2.2.2 Delta Model

The Delta model is designed to address the basic practical issue encountered in the Bertsimas and Sim approach; that of selecting an appropriate protection parameter Γ_i for each constraint i. This is a potentially cumbersome task for large-scale problems. Given this, it might be necessary to solve the Bertsimas-Sim robust optimization model repeatedly for varying values of the Γ_i parameters before a satisfactory solution is identified. In the case of large-scale network-based resource allocation problems, solving the model even once can be computationally challenging; and therefore the requirement to solve it multiple times is likely to be impractical for large problems. Network-based resource allocation problems are often formulated as binary integer programs, and our Delta model is particularly designed for such formulations.

The standard binary integer program that is required to be made robust is:

$$\max \quad \sum_{j \in J} c_j x_j \tag{2.20}$$

$$\text{s.t.} \quad \sum_{j \in J} a_{ij} x_j \leq b_i \quad \forall i \in I \tag{2.21}$$

$$x_j \in \{0, 1\} \quad \forall j \in J \tag{2.22}$$

In (2.20)-(2.22), we use the following notation. I is the set of constraints, and J the set of variables, c_j is the profit coefficient for variable j and b_i is the right-hand side value for ith constraint, $\forall i \in I$. a_{ij} for all $i \in I, j \in J$ is the coefficient of variable j in constraint i. a_{ij} for all $i \in I, j \in \bar{J}_i$, is subject to uncertainty, with $\tilde{a}_{ij}$ its realized value. a_{ij} is the nominal value of $\tilde{a}_{ij}$, and also the mean value of its symmetric range of variation. $\hat{a}_{ij}$ is the half-interval of the symmetric range of variation of $\tilde{a}_{ij}$, for all $i \in I, j \in J$. $\hat{a}_{ij} = 0$ for $j \in J \backslash \bar{J}_i$. x_j for all $j \in J$, is a binary decision variable that equals 1 if variable is present in the solution and 0 otherwise.

To avoid the need to specify Γ values, we modify the Bertsimas-Sim formulation to include

a constraint requiring the total profit of the robust solution to be within a difference of δ from the nominal optimal value. Additionally we change the objective to one of minimizing the maximum number of variables that *must* assume their nominal, rather than extreme, values to satisfy all constraints. We define *variable* Δ_i equal to the maximum number of variables x in the solution with $x = 1$ whose coefficient values are subject to uncertainty and must assume their nominal values for constraint i to remain feasible. We sort, for each constraint i, its associated columns $j \in \bar{J}_i$, in increasing order of their $\hat{a}_{ij}$ values (ties are broken arbitrarily). After ordering, the rank of the jth column in the ith constraint is denoted by $l(i, j)$. Also, the original index (j) of the variable that takes the lth position in the *sorted* $\hat{a}_{ij}$ values for constraint i is denoted by $j(i, l)$. For example, the variable j in constraint i with the smallest $\hat{a}_{ij}$ value has $l(i, j) = 1$. The variable with the largest $\hat{a}_{ij}$ value has $l(i, j) = N$, with N equal to the number of binary variables in J.

Let y_j^* be the optimal value of variable j for all $j \in J$ for the nominal problem (2.20)-(2.22). Then δ is the user-specified incremental cost that is acceptable for increased robustness, that is, the profit of a robust solution from the Delta formulation is at least $\sum_{j \in J} c_j y_j^* - \delta$. Let variables v_{ij} equal 1 if the uncertain coefficient $\tilde{a}_{ij}$ is not allowed to take on its extreme value, and takes on its nominal value in the solution of the Delta model. Variables w_{il} equal 1 for all $l \geq |J| - |\bar{J}_i| + 1$ in constraint i for which there exists a $k \geq l$ with $v_{ik} = 1$, for $l = |J| - |\bar{J}_i| + 1, ..., N + 1$. w_{il}s for $l = |J| - |\bar{J}_i| + 1, ..., N + 1$ in each constraint i follow a step function.

This leads to the following Delta formulation:

$$\min \sum_{i \in I} \Delta_i \tag{2.23}$$

$$\text{s.t.} \sum_{j \in J} c_j x_j \leq \sum_{j \in J} c_j y_j^* + \delta \tag{2.24}$$

$$\sum_{j \in J} (a_{ij} + \hat{a}_{ij}) x_j - \sum_{j \in J} \hat{a}_{ij} v_{ij} \leq b_i \qquad \forall \, i \in I \tag{2.25}$$

$$\Delta_i \geq \sum_{l=|J|-|J_i|+1}^{|J|} [(l - |J| + |J_i|)(w_{i,l} - w_{i,l+1}) + v_{i,j(i,l)} - w_{i,l}] \qquad \forall \, i \in I \tag{2.26}$$

$$\Delta_i \geq \sum_{j \in J} v_{ij} \qquad \forall \, i \in I \tag{2.27}$$

40

$$v_{ij} \leq x_j \qquad \forall j \in J, \forall i \in I \qquad (2.28)$$

$$v_{i,j(i,l)} \leq w_{i,l} \qquad \forall l \in J \qquad (2.29)$$

$$v_{ij} \geq x_j + w_{i,l(i,j)} - 1 \qquad \forall j \in J, \forall i \in I \qquad (2.30)$$

$$w_{i,l+1} \leq w_{i,l} \qquad \forall l \in |J| - |\bar{J}_i| + 1, ..., |J|, \forall i \in I \qquad (2.31)$$

$$w_{i,|J|-|J_i|+1} \leq 1 \qquad \forall i \in I \qquad (2.32)$$

$$w_{i,|J|+1} = 0 \qquad \forall i \in I \qquad (2.33)$$

$$w_{il} = 0 \qquad \forall l \in 1, ..., |J| - |\bar{J}_i|, \forall i \in I \qquad (2.34)$$

$$x_j \in \{0, 1\} \qquad \forall j \in J \qquad (2.35)$$

$$v_{ij} \in [0, 1] \qquad \forall j \in 1, ..., |J|, \forall i \in I \qquad (2.36)$$

$$w_{il} \in \{0, 1\} \qquad \forall l \in 1, ..., |J| + 1, \forall i \in I \qquad (2.37)$$

The formulation is described as follows: The objective (2.23) is to minimize the sum of the maximum number of coefficients in each constraint that must assume their nominal values to satisfy all constraints. This serves the purpose of trying to maximize the number of coefficients that can take their extreme values and still maintain feasibility. Constraints (2.24) require that the profit from the 'robust' solution not differ from the profit of the deterministic solution by more than a 'robustness budget' of δ. Feasibility is assured by constraints (2.25). We set v_{ij} equal to 1 in constraint i, for all coefficients j that *must* be set to their nominal values. Inequalities (2.28) prevent a coefficient from not taking on its extreme value unless the variable is present in the solution, that is, s_{ij} to zero if x_j is zero in the solution. Constraints (2.26) and (2.27) provide different mechanisms to count the maximum number of variables whose coefficients in constraint $i \in I$ subject to uncertainty and must take on nominal values. The explanation for this constraint lies in the realization that when the columns are sorted in increasing order of $\hat{a}_{ij}$ values for each row i, the *maximum number* of coefficients that must assume nominal values to maintain feasibility is determined by forcing the smallest $\hat{a}_{ij}$ values ($\hat{a}_{ij} \neq 0$) in the solution to have their associated v_{ij} values set to 1, if x_j is in the solution. Though constraints (2.26) and (2.27) are essentially the same, we retain both in order to provide better bounds on the solution. (2.29) requires the w variable to be at least as large as the corresponding v variable for that j and i. Also, (2.30) forces v to be 1 if both the corresponding x and w are 1. Constraints (2.31),

(2.32) and (2.33) require the w_{il} variables corresponding to uncertain coefficients in each i to form a step function. The w_{il} variables corresponding to coefficients that are not subject to uncertainty are zero (2.34). The x and w variables are binary, as required by (2.35) and (2.37) respectively. They thus force the v variables to take on binary values as well. Alternatively, one can think of this model as maximizing the minimum number of coefficients, summed over all constraints, that can take on their worst-case values, under budget limitations.

Alternative objective functions to this model include 1) minimizing a weighted sum of coefficients not allowed to realize their extreme values $\left(\sum_{i \in I} w_i \Delta_i \right)$ with weight i assigned to constraint $i \in I$; or 2) minimizing the maximum number of uncertain coefficients in any constraint that must assume nominal values rather than their extreme values, to satisfy the constraints; that is $\min \nu$, with additional constraints $\nu \geq \Delta_i$, for all $i \in I$.

2.2.3   Chance-Constrained Programming

The standard linear program that is required to be made robust is:

$$\max \quad \mathbf{c}^T \mathbf{x} \tag{2.38}$$

$$\text{s.t.} \quad \sum_{j \in J} a_{ij} x_j \leq b_i \quad \forall \, i \in I \tag{2.39}$$

$$\mathbf{x} \geq 0, \tag{2.40}$$

where the notation used is the same as for (2.20)-(2.22). Its general chance-constrained formulation is as follows:

$$\max \quad f(\mathbf{c}, \mathbf{x}) \tag{2.41}$$

$$\text{s.t.} \quad P(\sum_{j \in J} a_{ij} x_j \leq b_i) \geq \alpha_i \quad \forall \, i \in I \tag{2.42}$$

$$\mathbf{x} \in \mathbf{X}, \tag{2.43}$$

where 'P' means 'probability', and $\mathbf{X}$ is the feasible set for (2.38)-(2.40). α_i ($0 \leq \alpha_i \leq 1$) for all i, is a user-specified *protection level* for constraint i, and $(1 - \alpha_i)$ specifies the maximum degree of violation of constraint i.

Charnes and Cooper translate (2.41)-(2.43) into different models, for varied types of objective functions and correspondingly different constraints. They present models for three types of objective functions, further details of which are available in [CC63], [CK67]. In each of these models, **b** and **c** are assumed to be uncertain. We retain these assumptions in our model.

Ben-Israel [BI62] shows that the chance-constraint (2.42) can be linearized as follows:

$$P(\sum_{j=1}^{N} a_{ij}x_j \leq b_i) \geq \alpha_i \Leftrightarrow \sum_{j=1}^{N} a_{ij}x_j \leq F_{b_i}^{-1}(1 - \alpha_i), \tag{2.44}$$

with $y = F_{b_i}^{-1}(1 - \alpha_i)$ equal to the *quantile* value in the cumulative distribution function (CDF), F_{b_i}, of b_i such that the probability that b_i takes on values less than or equal to y is (1-α). That is, if $f(b_i)$ is the probability distribution function of b_i, $\int_{-\infty}^{y} f(b_i)db_i = 1 - \alpha_i$.

Further, from the distributions of the elements in **c**, we can determine a vector of stipulations β such that $P(\mathbf{c} \leq \lambda_c) \geq \beta$ for some λ. Ben-Israel shows that we can also write the above linear program as

$$\max(F_{\mathbf{c}}^{-1}(\beta))^T \mathbf{x} \tag{2.45}$$

$$\text{s.t.} \mathbf{Ax} \leq F_{\mathbf{b}}^{-1}(1 - \alpha) \tag{2.46}$$

$$\mathbf{x} \geq 0, \tag{2.47}$$

with $F_{\mathbf{c}}^{-1}(\beta)$ is defined similarly to $F_{\mathbf{b}}^{-1}(1 - \alpha)$ above.

We capture uncertainty in the cost function using the expected values of the **c** vector, or by using quantiles of **c** that we want to protect against, and uncertainty in the RHS can be captured by using the relevant quantiles of the **b**-vector.

Thus, given the CDF of the right-hand-side (RHS) vectors, (or even certain *quantile* values of the distribution), we convert the stochastic problem into a deterministic linear programming problem of the same size as measured by the number of variables and constraints. Quantiles of the probability distribution for uncertain parameters can be obtained by analyzing historical data and incorporating additional knowledge of the system behavior.

The CCP model assumes the RHS (= **b**) and **c** alone to be uncertain, and adjusts the values of these uncertain parameters to create a solution with more constraint slack than the nominal problem. Unfortunately, chance-constrained programming encounters serious computational issues as we try to capture uncertainty in multiple coefficients per constraint. For most of their models, Charnes and Cooper limited uncertainty to one random variable per constraint (the **b**-matrix value or RHS value). To incorporate uncertainty in the **A**-matrix (left-hand-side), we must calculate a joint probability distribution for all uncertain coefficients in the constraint, making the deterministic program cumbersome to solve. Miller and Wagner [MW65] discuss chance-constrained methods for the case of multiple random variables (per constraint) generated by a multinomial distribution. However, most chance-constrained programming has been limited to uncertainty in the constraints only in the right-hand-side due to the difficulties associated with multiple random variables.

Modifying the right-hand side **b** vector in (2.44) is sufficient, however, to provide the entire constraint a protection of α_i. Therefore, though capturing uncertainty explicitly in the **A** matrix is cumbersome, constraint (2.44) is implicitly protecting, to some extent, against changes in the left-hand-side matrix.

The following are some of the **advantages** of this model:

- The structure of the CCP model is generalizable to all linear/integer programs.

- The model of capturing uncertainty has intuitive appeal. The deterministic formulation is also easy to understand and interpret.

- The CCP model does not require complete knowledge of the distribution that the uncertain data follows. In fact, knowledge of the quantile value of the distribution, corresponding to the required protection level for the constraint, is sufficient. In general, knowledge of a few discrete quantiles of the uncertain data for each constraint allows the user to approximate the distribution without requiring too much data about the distribution. Such information is also usually available through statistical analysis of the historical data of the system.

- Finer knowledge of the behavior of the system, as compared to simply the bounds of variation, can be captured through this model. Distributions other than the uniform distribution can be easily incorporated without an increase in complexity.

However, this model also has some **limitations**, as mentioned below:

- Uncertainty in the left-hand side **A** matrix, including correlations among uncertain data, is difficult to model explicitly.

- Approximate probability distributions or some quantiles of the distribution of the RHS have to be known. If unknown, the extreme-value bounds, as used in the Bertsimas-Sim model, can be considered as the bounds of a uniform distribution.

2.2.4 Extended Chance-Constrained Programming

The Chance-Constrained Programming or CCP approach faces the problem of specifying a probability of satisfaction for each constraint. In order to find the appropriate constraint protection parameters that result in overall solution protection (or robustness), we might need to solve the problem repeatedly. This is potentially a limitation of the approach when applying to large problems. The Extended Chance-Constrained Programming ($ECCP$) approach avoids the need to specify the protection level for each constraint explicitly.

In our extended model, to avoid the need to specify the protection level for each constraint explicitly, we include a constraint on the overall expected profit of the robust solution and change the objective to one of maximizing the sum total of protection level provided for all constraints. The ECCP formulation of (2.38) - (2.40) is as follows:

$$\max \sum_{i \in I} \alpha_i \tag{2.48}$$

$$\text{s.t.} P(\mathbf{Ax} \leq \mathbf{b}) \geq \alpha \tag{2.49}$$

$$E(\mathbf{c}^T \mathbf{x}) \geq \mathbf{c}^T \mathbf{y}^* - \delta \tag{2.50}$$

$$\mathbf{x} \geq 0 \tag{2.51}$$

$$\alpha \geq 0, \tag{2.52}$$

where $\mathbf{c}^T \mathbf{y}^*$ is the expected profit of the nominal optimal solution $\mathbf{y}^*$ to (2.38)-(2.40). Alternative objective functions may include 1) maximizing the minimum protection level α_{min} with an added constraint $\alpha_{min} \leq \alpha_i$ for all $i \in I$; or 2) maximizing a weighted sum of the constraint protection levels ($\sum_{i \in I} w_i \gamma_i$), w_i being the non-negative weight assigned to constraint $i \in I$.

To linearize (2.48) - (2.52), we require the knowledge of some quantiles and their associated values of the probability distribution of the right-hand-side b_i, for each constraint i, and the expected values of the profit function. Let b_i be the nominal value for the right-hand-side of the ith constraint, and c_j the expected profit coefficient corresponding to the jth variable, for all $j \in J;$. Let K_i represent the set of discretized protection levels known for b_i. We now set the protection levels α_i for each constraint i as variables, representing the quantile that is chosen from among the K_i available quantiles. Let b_i^k be the kth quantile value of the RHS parameter of the ith constraint, for all $k \in K_i, i \in I$; and z_j^* be the optimal solution to (2.38) - (2.40) found using nominal values of the b and c parameters, for all $j \in J$. p_i^k is the protection level probability associated with quantile $k \in K$ for constraint $i \in I$. The objective function value, denoted α_{min}, equals the minimum protection level achieved over all constraints $i \in I$. To capture the trade-off of robustness with profits, we assume that the planner is willing to forego a (user-specified) profit of δ to instead gain a robust plan.

Decision variables y_i^k are binary variables that equal 1 if the protection level (expressed as a probability p_i^k, with $0 \leq p_i^k \leq 1$) represented by the kth quantile ($k \in K_i$) is attained in constraint $i \in I$; and 0 otherwise. This means that if the kth quantile value is protected against, the $(k + 1)$st quantile is also automatically protected against. This follows from the fact that constraints (2.55) are 'less than' inequalities. The y_i^k values for any constraint i, follow a step function. Variables α_i represent the protection level attained for the ith constraint, for all $i \in I$.

The extended chance-constrained model (ECCP) is as follows:

$$\max \sum_{i \in I} \gamma_i \qquad (2.53)$$

$$\text{s.t.} \sum_{j \in J} c_j x_j \geq \sum_{j \in J} c_j z_j^* - \Delta \qquad (2.54)$$

$$\sum_{j \in J} a_{ij} x_j \leq \sum_{k=1}^{K} b_i^k (y_i^k - y_i^{k-1}) \qquad \forall\, i \in I \qquad (2.55)$$

$$y_i^k \geq y_i^{k-1} \qquad\qquad\qquad \forall\, k = 1, ..., K_i, i \in I, \qquad (2.56)$$

$$y_i^0 = 0 \qquad\qquad\qquad\qquad\qquad \forall\, i \in I \qquad (2.57)$$

$$y_i^K = 1 \qquad\qquad\qquad\qquad\qquad \forall\, i \in I \qquad (2.58)$$

$$\gamma_i \leq \sum_{k=1}^{K_i} p_i^k (y_i^k - y_i^{k-1}) \qquad\qquad \forall\, i \in I \qquad (2.59)$$

$$x_j \geq 0 \qquad\qquad\qquad\qquad\qquad \forall\, j \in J \qquad (2.60)$$

$$y_i^k \in \{0, 1\} \qquad\qquad\qquad \forall\, k \in 1, ..., K_i, i \in I \qquad (2.61)$$

$$0 \leq \gamma_i \leq 1 \qquad\qquad\qquad\qquad \forall\, i \in I \qquad (2.62)$$

The objective function (2.53) maximizes the total probability that each constraint $i \in I$ is feasible. It can also be re-written to maximize the minimum value of γ_i over all constraints. (2.54) ensures that the solution's expected profit is within Δ units of the expected profit associated with the nominal optimal solution (found by solving the problem using nominal values of the **b** vector). For all constraints $i \in I$, (2.55) forces the left-hand-side (LHS) to be less than or equal to b_i^k if y_i^k equals 1, thereby ensuring constraint satisfaction with at least the probability associated with quantile k. For the smallest quantile k^* that can be satisfied, the $y_i^{k^*}$ value is 1, and quantiles $k < k^*$ have $y_i^k = 0$. Thus, the RHS value of this constraint is selected as the smallest one that can be satisfied by the solution. (2.56) ensures that the y_i^ks are monotonically increasing and follow a step function, such that if a smaller quantile (higher protection) is achieved, the larger quantile (lower protection) is automatically achieved. (2.57) and (2.58) set the boundary values of the y_i^k step functions. Constraints (2.59) set γ_i to be no greater than the highest protection level provided to constraint i by the solution. x_j is non-negative for each $j \in J$; y_i^k is binary for each $j \in J$ and each $k \in K_i$ for all $i \in I$; and γ_i is non-negative for all $i \in I$ as required by (2.60), (2.61) and (2.62) respectively.

2.3 Robust Models Applied to Aircraft Routing

2.3.1 The Standard Deterministic Aircraft Routing Model

Following is the standard deterministic aircraft maintenance routing formulation, denoted AR, which we attempt to make robust in later sections.

AR :

$$\min 0 \tag{2.63}$$

$$\text{s.t.} \sum_{s \in S} a_{is} x_s = 1 \qquad \forall\, i \in F \tag{2.64}$$

$$\sum_{s \in S_i^+} x_s - y_{i,d}^- + y_{i,d}^+ = 0 \qquad \forall\, i \in F^+ \tag{2.65}$$

$$- \sum_{s \in S_i^-} x_s - y_{i,a}^- + y_{i,a}^+ = 0 \qquad \forall\, i \in F^- \tag{2.66}$$

$$\sum_{s \in S} r_s x_s + \sum_{g \in G} p_g y_g \leq N \tag{2.67}$$

$$y_g \geq 0 \qquad \forall\, g \in G \tag{2.68}$$

$$x_s \in \{0, 1\} \qquad \forall\, s \in S \tag{2.69}$$

The decision variables in this formulation correspond to strings (sequences of flight legs) with each string starting and ending at a maintenance station and obeying FAA and other regulatory rules regarding maximum time between maintenance. Strings capture multiple decisions simultaneously, and thus are *composite variables* [ABW02]. These are typically used when they yield strong formulations, and/or when they remove the need to include complex, difficult to model, constraints, as in the case of aircraft routing.

Let F be the set of daily flights, F^+ be the set of flight legs originating at a maintenance station, F^- be the set of flight legs ending at a maintenance station, and S be the resulting set of possible strings. The set of ground arcs (including wraparound arcs, beginning in one day and ending in another day at the same location) is denoted by G, the set of flight legs ending with flight leg i is S_i^-, and the set of flight legs beginning with flight leg i is S_i^+. Ground variable $y_{i,d}^-$ represents the number of aircraft on the ground before flight leg i departs and $y_{i,d}^+$ is the

48

number of aircraft on the ground after flight leg i departs, for all flight legs i. Similarly $y_{i,a}^-$ is the number of aircraft on the ground before flight leg i arrives and $y_{i,a}^+$ is the number of aircraft on the ground after flight leg i arrives, for all flight legs i. a_{is} is 1 if flight leg $i \in F$ is contained in string $s \in S$ and 0 otherwise. r_s is the number of times string $s \in S$ crosses the count line, p_g is the number of times ground arc $g \in G$ crosses the count line, and N is the number of aircraft available.

Constraints (2.64) are the cover constraints that require each flight leg to be covered exactly once. Constraints (2.65) and (2.66) balance the number of aircraft at each location and constraints (2.67) count the number of aircraft. x_s takes on value 1 if string $s \in S$ is selected to be operated by an aircraft, and 0 otherwise. y_g is the number of aircraft on ground arc $g \in G$.

2.3.2   Modeling

The first challenges associated with achieving robust solutions are deciding what constitutes robustness and how to capture these in the optimization model. In the following sections, we illustrate how different modeling paradigms will force different modeling approaches, resulting in different solutions and different computational challenges.

Using schedule data and details of operated aircraft routings (typically available from the ASQP database [Bur09b]), Lan, Clarke and Barnhart compute the independent and propagated delay for each flight leg. Because independent delays are, by definition, independent of aircraft routings, independent delays can be applied to any sequence of flights forming a string to estimate propagated delay (PD_i), total departure delay ($TDD(i)$) and total arrival delay (TAD_i) for each flight leg i along that string, or along *any* possible string, even those not operated by the airline. With this, it is also possible to compute the probability of a flight leg being delayed to a certain extent, or the probability of a string experiencing a specified level of propagated delay, or the range of delays experienced by a flight leg or a string. These characterizations of uncertainty are employed in the following sections.

2.3.3 Tailored Approach

Lan, Clarke and Barnhart [LCB06] attempt to make an aircraft routing robust by driving AR using the metric of total expected propagated delay. Recall from §2.1.2, that total propagated delay differs from total aircraft delay by a constant value. Thus, minimizing expected propagated delay is equivalent to minimizing expected total aircraft delay.

Let pd_{ij}^s be the propagated delay from flight leg i to flight leg j when i immediately precedes j in string s. Using the notation introduced in 2.3.1, and assuming that the different strings in the network are independent of each other, the expected propagated delay in the network is:

$$E\left[\sum_{s\in S}\left(\sum_{(i,j)\in s} pd_{ij}^s\right) x_s\right] = \sum_{s\in S}\left(x_s \times \sum_{(i,j)\in s} E[pd_{ij}^s]\right) = \sum_{s\in S} d_s x_s \qquad (2.70)$$

where $d_s = \sum_{(i,j)\in s} E[pd_{ij}^s]$.

In terms of problem structure and complexity, the Lan, Clarke and Barnhart tailored model (LCB) is the same as AR, except that the feasibility objective is replaced with the objective of minimizing total expected propagated delay, specifically:

$$\text{LCB}:$$

$$\min \quad \sum_{s\in S} d_s x_s \qquad (2.71)$$

$$\text{s.t. Cover, Balance, Count, and Integrality } (2.64) - (2.69) \qquad (2.72)$$

2.3.4 Probabilistic Chance-Constrained Programming Approach

In Chance-Constrained Programming, the chance or probability that a constraint of the model is satisfied is required to exceed some specified threshold level. The chance-constrained formulation of aircraft routing is as follows:

$$\max 0 \tag{2.73}$$

$$\text{s.t. } \mathbf{P}(\sum_{s \in S} a_{is} x_s = 1) \geq \alpha_i \qquad \forall\, i \in F \tag{2.74}$$

$$\text{Balance, Count, and Integrality } (2.65) - (2.69), \tag{2.75}$$

where $\mathbf{P}$ denotes probability, and α_i is a user-defined 'protection' parameter indicating the minimum probability that the aircraft routing solution will satisfy constraint i. Charnes and Cooper [CC59] describe how to model the non-linear constraints of (2.74) as linear constraints.

We model uncertainty in the flight cover constraints (2.64) by defining p_{is} as the probability obtained from historical data, that flight leg i in string s is covered, that is, is delayed by fewer than t minutes when string s is operated. We let $t = 90$ minutes to indicate the threshold beyond which flight cancelations, and hence flight non-coverage, can occur. Because each flight leg is present only in a single string in the solution, the probability of flight i being delayed less than t minutes is $p_i = \sum_{s \in S} p_{is} x_s$. Using this, we re-write (2.73) - (2.75) as:

$$CCP :$$

$$\max 0 \tag{2.76}$$

$$\text{s.t. } \sum_{s \in S} a_{is} x_s = 1 \qquad \forall\, i \in F \tag{2.77}$$

$$\sum_{s \in S} p_{is} x_s \geq \alpha_i \qquad \forall\, i \in F \tag{2.78}$$

$$\text{Balance, Count, and Integrality } (2.65) - (2.69)$$

Constraints (2.78) are the 'robustness constraints' that limit to α_i the probability that flight leg i is delayed more than t minutes in the operation of string s.

In structure and complexity, CCP is similar to AR, though it adds constraints (2.78), one for each flight leg, and requires specification of the value of α_i for each constraint $i \in F$. As discussed in §2.2.4, this model faces challenges associated with the specification of α values. Too high values can lead to infeasibilities and too low can lead to lead to inadequate levels of

51

robustness or protection. As a result, CCP might have to be solved repeatedly to find appropriate α-values. Repeated solution of CCP, however, might be both impractical and ineffective in identifying the best α-values. The $\alpha - CCP$ model (2.79) - (2.82), which is a special case of the general ECCP model (2.53) - (2.62), overcomes these limitations. In the $\alpha - CCP$ model, the protection levels α for constraints (2.78) need not be specified apriori and instead are decision variables in the model. The objective of $\alpha - CCP$ (2.79) is to maximize the sum of protection levels of all the constraints.

$\alpha - \textbf{CCP} :$

$$\max \sum_{i \in F} \alpha_i \tag{2.79}$$

$$\text{s.t.} \sum_{s \in S} a_{is} x_s = 1 \qquad \forall\, i \in F \tag{2.80}$$

$$\alpha_i \leq \sum_{s \in S} p_{is} x_s \qquad \forall\, i \in F \tag{2.81}$$

$$\text{Balance, Count, and Integrality} \ (2.65) - (2.69) \tag{2.82}$$

Alternative objective functions include: 1) maximizing a weighted sum of flight probabilities $\left(\sum_{i \in F} \alpha_i w_i \right)$ with weight w_i assigned to flight i; or 2) maximizing the minimum probability α_i ($\max \alpha_{min}$) with additional constraints $\alpha_{min} \leq \sum_{s \in S} p_{is} x_s$, for all $i \in F$.

2.3.5 Extreme-Value Robust Optimization Approach

We adapt the extreme-value robust optimization approach of Bertsimas and Sim to the aircraft routing problem by letting $\hat{a}_{is} = -1$ if flight $i \in F$ in string $s \in S$ has extreme value of delay exceeding t minutes, based on historical data. Because the Bertsimas and Sim approach considers realizations of the uncertain parameters at their extreme (or worst-case) values ([BS04], [BS03]), if a flight i in string s has extreme delay exceeding t minutes using historical data, the extreme value $\hat{a}_{is} = -1$ results in $a_{is} + \hat{a}_{is} = 1 - 1 = 0$, reflecting the extreme occurrence that flight i is canceled, and hence, uncovered.

In the cover constraint for flight $i \in F$, let S_i be the set of strings $s \in S$ whose coefficients a_{is} are subject to uncertainty. For each flight i, the Bertsimas and Sim robust optimization approach defines a 'robustness' or 'protection' parameter Γ_i taking on (possibly continuous) values in $[0, |S_i|]$, representing the number of coefficients in the solution in constraint i that can assume worst-case or extreme values and still satisfy feasibility of the constraint. The model ensures constraint feasibility in the case when up to $\lfloor \Gamma_i \rfloor$ coefficients take on extreme values, and one coefficient a_{it} changes by $(\Gamma_i - \lfloor \Gamma_i \rfloor)\hat{a}_{it}$. For our purposes, integer values of Γ_i are the most meaningful because of the uncertainty definition. Γ_i represents the number of strings, for each flight $i \in F$, in which flight i cannot experience delays greater than or equal to t minutes in the extreme case. The resulting extreme value formulation from the Bertsimas and Sim model is:

$$\min 0 \tag{2.83}$$

$$\text{s.t.} \sum_{s \in S} a_{is}x_s$$

$$+ \max_{\{S_i' \cup t_i | S_i' \subseteq S_i, |S_i'| = \lfloor \Gamma_i \rfloor, t_i \in S_i \setminus S_i'\}} \left\{ \sum_{s \in S_i} \hat{a}_{is}w_s + (\Gamma_i - \lfloor \Gamma_i \rfloor)\hat{a}_{it_i}w_t \right\} = 1 \quad \forall\, i \in F \tag{2.84}$$

$$x_s \leq w_s \quad \forall\, s \in S \tag{2.85}$$

$$w_s \geq 0 \quad \forall\, s \in S \tag{2.86}$$

$$\text{Balance, Count, and Integrality } (2.65) - (2.69). \tag{2.87}$$

If $\Gamma_i = 0$ then the cover constraint for flight i reduces to $\sum_{s \in S} a_{is}x_s = 1$, with robustness concerns effectively ignored. If $\Gamma_i \geq 1$, the constraint ensures that each flight $i \in F$ is covered by at least one string s that both contains flight i and has extreme value of delay less than t for i. EV ensures that each flight is covered, in the worst-case, by placing it in more than one string if needed, thus *over-covering* flights. Because the values of $\hat{a}_{is}$ are either 0 or -1, (2.83)-(2.87) simplifies to:

$$EV:$$

$$\min 0 \tag{2.88}$$

$$\text{s.t.} \sum_{s \in S} a_{is}x_s + \max\left\{ \sum_{s \in S} \hat{a}_{is}x_s, -\Gamma_i \right\} = 1 \tag{2.89}$$

$$\text{Balance, Count, and Integrality } (2.65) - (2.69), \tag{2.90}$$

with the second term in (2.89) representing the protection level, and can be linearized easily.

Note that the level of robustness can be varied by selecting different values of the extreme delay parameter t. Reducing the value of t will have the effect of generating more conservative solutions, those that permit little delay, while increasing the value of t will have the opposite effect. This change in t produces a similar effect in the CCP and $\alpha - CCP$ as well.

As in the CCP model, the EV model requires the specification of a parameter value (in this case, Γ_i) for each cover constraint (2.89). To avoid the need to repeatedly solve EV to determine the best Γ-values, likely an impractical exercise, we propose an alternate model, denoted $\Delta - EV$, in which Γ parameters are modeled as decision variables. This model is a special case of our general *Delta* model presented in §2.2.2, tailored to aircraft routing. The goal of this formulation is to minimize, for each constraint i, the number of variables in the solution whose coefficients are *not* allowed to realize their extreme values (denoted by Δ_i), in order to ensure feasibility of the constraint. (Note that Δ_i is the converse of Γ_i.) We minimize sum of Δ_i for all flight legs i, thus maximizing the protection level. Due to the special structure, the $\Delta - EV$ model can be formulated without any Δ variables, as follows:

$\Delta - \mathbf{EV}$:

$$\max \sum_{i \in F} \sum_{s \in S} \hat{a}_{is} x_s \qquad\qquad (2.91)$$

$$\text{s.t.} \sum_{s \in S} a_{is} x_s \geq 1 \qquad\qquad \forall\, i \in F \qquad (2.92)$$

$$\text{Balance, Count, and Integrality (2.65) -- (2.69)} \qquad\qquad (2.93)$$

The objective (2.91) effectively maximizes $\sum_{i \in F} \Gamma_i$ or minimizes $\sum_{i \in F} \Delta_i$ by maximizing the total value of the protection function in (2.89) summed for all flights $i \in F$.

2.3.5.1 Capturing Uncertainty in the Objective

Because the extreme value robust optimization framework also allows uncertainty to be modeled in the objective function, an alternative to EV and $\Delta - EV$ is to capture uncertainty in a manner similar to that of the Lan, Clarke and Barnhart tailored LCB model, but within the

extreme value framework. This can be translated into minimizing the total propagated delay when a user-defined number of strings Γ in the formulation realize their *worst-case values* of propagated delay. The detailed formulation is presented below.

The protection parameter Γ is defined in the interval $[0, |S|]$. Let $\hat{d}_s$ be the worst-case (maximum or 100th percentile) value of propagated delay of string s observed in the historical data. (Note that this is not the worst possible realization of propagated delay, but only that in the selected period of historical data.)

The formulation allocates slack such that it minimizes the effect of the maximal delay caused by *any* set of Γ strings in the solution realizing their worst-case values, and the other strings attaining their nominal values of propagated delay (equal to zero). That is, it minimizes the propagated delay caused by the maximal subset of Γ strings in the solution realizing their extreme propagated delay values, as shown in (2.94). Variable w_s for string $s \in S$ takes on value 1 if string s is present in the solution, and the model plans for its extreme value being realized. The extreme-value formulation, according to Bertsimas and Sim [BS03], is as follows.

Obj − EV :

$$\min \left\{ \sum_{s \in S} 0 x_s + \max_{\{S' \cup \{t\} | S' \subseteq S, |S'| = \lfloor \Gamma \rfloor, t \in S \setminus S'\}} \left\{ \sum_{s \in S'} \hat{d}_s w_s + (\Gamma - \lfloor \Gamma \rfloor) \hat{d}_t w_t \right\} \right\} \qquad (2.94)$$

s.t. Cover, Balance, Count, and Integrality (2.64) − (2.69) $\qquad$ (2.95)

$$x_s \leq w_s \qquad \forall\, s \in S \qquad (2.96)$$

$$w_s \geq 0 \qquad \forall\, s \in S \qquad (2.97)$$

Obj − EV can be linearized, and cast as a mixed integer program, as follows:

Obj − EV :

$$\min z\Gamma + \sum_{s \in S} p_s \qquad (2.98)$$

$$\text{s.t. } z + p_s \geq \hat{d}_s w_s \qquad \forall\, s \in S \qquad (2.99)$$

Cover, Balance, Count, and Integrality (2.64) − (2.69) $\qquad$ (2.100)

$$x_s \leq w_s \qquad \forall\, s \in S \qquad\qquad (2.101)$$

$$w_s \geq 0 \qquad \forall\, s \in S \qquad\qquad (2.102)$$

$$z \geq 0. \qquad\qquad\qquad\qquad (2.103)$$

These mixed integer programs have a very different structure from AR [Mar07], and face computational challenges in solving (§2.5.1).

The difficulty with this model, again, is specifying the 'best' value of Γ. Because a value of Γ does not indicate the overall robustness of the solution, it is difficult to specify apriori. We therefore present the $\Delta - Obj - EV$ model for which we find a solution such that the maximum propagated delay D exceeds the sum of the extreme delay values of any subset of Γ strings, with Γ maximal. The $\Delta - Obj - EV$ solution, therefore, is an aircraft routing that allows the largest number of strings to realize their worst-case propagated delay values without exceeding the maximum allowable system propagated delay D. This model, like the $\Delta - EV$, is a special case of our general Delta model, presented in §2.2.2.

D is a threshold on total propagated delay, obtained by analyzing the historical occurrences of propagated delays. A possible value of D could be 50% of the average total propagated delay on a bad day. We set to zero the nominal propagated delay value for any string s, and let $\hat{d}_s$ represent the expected extreme or worst-case value of propagated delay for string s, as computed using historical data. Let $\bar{S}$ be the set of strings $s \in S$ with realizations of non-zero propagated delays in the historical data, that is, with non-zero $\hat{d}_s$. We set $v_s = 1$ if string $s \in S$ has to take on its nominal propagated delay value of 0 and not its worst-case (historical) value for the solution to be feasible. To maximize the size of the minimal subset of strings that can realize their worst-case values and ensure feasibility, we sort the strings in increasing order of their $\hat{d}_s$ values such that $\hat{d}_1 \leq \hat{d}_2 \leq ... \leq \hat{d}_{|S|}$. Δ is the maximum number of strings with propagated delays subject to uncertainty that must assume nominal, not worst-case, values to be feasible.

$$\Delta - \mathbf{Obj} - \mathbf{EV}:$$

$$\min \Delta \tag{2.104}$$

$$\text{s.t.} \sum_{s \in S} \hat{d}_s x_s - \sum_{s \in S} \hat{d}_s v_s \leq D \tag{2.105}$$

$$\Delta \geq \sum_{s \in S} v_s \qquad \forall\, s \in \bar{S} \tag{2.106}$$

$$v_s \leq x_s \qquad \forall\, s \in \bar{S} \tag{2.107}$$

$$v_s \leq w_s \qquad \forall\, s \in \bar{S} \tag{2.108}$$

$$v_s \geq x_s + w_s - 1 \qquad \forall\, s \in \bar{S} \tag{2.109}$$

$$w_s \geq w_{s+1} \qquad \forall\, s \in |S| - |\bar{S}| + 1, ..., |S| - 1 \tag{2.110}$$

$$w_{|S|-|\bar{S}|+1} \leq 1 \qquad \forall\, i \in I \tag{2.111}$$

$$w_{|S|} \geq 0 \qquad \forall\, i \in I \tag{2.112}$$

Cover, Balance, Count, Integrality

$$(2.64) - (2.69) \tag{2.113}$$

$$v_s \in [0, 1] \qquad \forall\, s \in S \tag{2.114}$$

$$w_s \in \{0, 1\} \qquad \forall\, s \in S \tag{2.115}$$

Constraints (2.105) require that the total worst-case propagated delay be limited by D when any $|\bar{S}| - \Delta$ strings take on worst-case propagated delay values. $v_s = 1$ for all strings $s \in \bar{S}$ *in the solution* whose delay value *must* be set to 0 to achieve feasibility. $w_s = 1$ for all strings $s \in \bar{S}$ for which there exists a $k \geq s$ such that $v_k = 1$. Inequalities (2.107) force $v_s = 0$ unless string s is present in the solution. Inequalities (2.108) allow v_s to be 1 only if w_s is 1. (2.109) allow v_s to be 1 only if both $w_s = 1$ and $x_s = 1$. Constraints (2.106), in combination with the objective (2.104) count the maximum number of strings $s \in \bar{S}$ whose coefficients must take on the nominal propagated delay value of 0. The explanation for this constraint lies in the realization that when the strings in S are sorted in increasing order of $\hat{d}_s$ values, the *maximum number* of coefficients that must assume nominal values to maintain feasibility is determined

by forcing the strings with the smallest $\hat{d}_s$ values, for $s \in \bar{S}$, to have their associated v_s values set to 1 if x_s is in the solution. Constraints (2.110), (2.111) and (2.112) require w_s for $s \in \bar{S}$ to form a step function, so that the maximal set of w_s can be set to 1. Constraints (2.114) and (2.115) restrict w and v variables to take on values of 0 or 1 only.

2.4 Evaluation

2.4.1 Experimental Set-up

We conduct our experiments using data representing the flight network of a major US airline that operates a hub-and-spoke network. We identify the sub-networks for two different aircraft types, with each representing a different aircraft routing problem, denoted N_1 and N_2. The schedules of both networks are daily schedules, that is, the same set of flight legs is operated every day by each aircraft type. The characteristics of N_1 and N_2 are shown in Table 2.2. Historical flight leg delay and cancelation data are obtained from the Airline Service Quality Performance (ASQP) database [Bur09b] for two of the busiest months of the year. Both the months have similar schedules, load factors, and levels of delay. Delay data consisting of 19 days in the first month (referred to as historical data) is used to derive delay information (distributions and expected values) of flights. These data are used as inputs to the aircraft routing models. The solutions are then evaluated using delay data for 22 days of the following month (referred to as future data). Our models are implemented in C++ and OPL Studio v6.0 on a Dell PC with 1 GB RAM.

In evaluating robust routings, we assume that flight delays are allowed to propagate along the string, without any recovery interventions such as cancelations or swaps. This allows us to estimate the levels of delay propagation and robustness of the strings that may occur *without* intervention. Because cancelation and swapping strategies are different for different airlines, assuming particular swapping and cancelation strategies could lead to bias.

2.4.2 Metrics and Simulator

We assess the robustness of an aircraft routing solution by the following metrics:

Fleet Type	Daily Flights	Locations
N_1	38	10
N_2	50	16

Table 2.2: Fleet Network Characteristics

1. Expected on-time performance for all legs in the flight schedule for 15 minutes, 60 minutes, 120 minutes, and 180 minutes;

2. Total expected number of passenger disruptions; and

3. Total expected daily flight delay.

These metrics reflect DoT performance and passenger-centric metrics. Because total flight delay is equal to propagated delay plus independent delay (which is constant), comparison results related to propagated delay also apply to total delay.

For our simulator, we use Lan, Clarke and Barnhart's algorithm to compute from the airline's solution, for each day in the 'future' month of data, the propagated delay ($PD(i)$) and independent delay ($ID(i)$) from the total departure and arrival delays $TDD(i)$ and $TAD(i)$ for each flight i. Following that, we also compute metrics such as total flight delay for each day, 15 minute on-time performance, 30 minute on-time performance, and the total number of passengers disrupted for the airline's routing. To compute passenger disruptions, we enumerate all pairs of flights f_1, f_2 between which passengers connect. Let $C(f_1, f_2)$ be the scheduled time available for the passenger to make the connection and let $m_{f_1 f_2}$ represent the minimum time needed for a passenger to connect between flights f_1 and f_2. Then, if $C(f_1, f_2) - TAD(f_1) + TDD(f_2) \leq m_{f_1 f_2}$, the actual connection time between f_1 and f_2 is too short and passengers scheduled to make this connection are disrupted. Then, given independent delay for each day in the 'future' data for each flight leg i, we re-employ Lan, Clarke and Barnhart's algorithm (in the reverse order of steps) to compute the same delay and disruption metrics for each of the solutions we generate using our models. Note that the models use historical data to generate the solution, and then the 'future data' is used to evaluate them.

2.5 Results

In this section, we present the results of our experiments, studying similarities and differences in the solutions obtained in terms of complexity, run time and robustness as measured by our metrics.

2.5.1 Typical Computation Times

Table 2.3 reports average computation times for the airline instances solved in this work.

Model	Parameters	Iterations	Run time per iteration
AR	None	1	5 sec
LCB	None	1	10 sec
CCP	$\alpha_i \forall i$	53	7 sec
$\alpha - CCP$	None	1	7 sec
EV	$\Gamma_i \forall i$	50	35 sec
$Obj - EV$	Γ	15	45 sec-10 hrs (sometimes out of memory)
$\Delta - EV$	None	1	28 sec
$\Delta - Obj - EV$	D	5	3 hrs

Table 2.3: Complexity and Run Times

For the CCP and EV models, multiple iterations are required to determine the appropriate α and Γ values, respectively. Because there is no prior indication if a particular protection α_i or Γ_i (for flight i) renders the model infeasible, experimentation with different values is necessary. We found, for example, that for N_2, CCP was infeasible with α values of 99% and 95%. Also, some flights can be protected to a greater extent than others. To obtain the 'right' protection levels, the model had to be solved multiple times. We overcame this limitation using our $\alpha - CCP$, $\Delta - EV$ and $\Delta - Obj - EV$ models. Although they each had to be solved only once, the computation time of $\alpha - CCP$, $\Delta - EV$ and $\Delta - Obj - EV$ models are comparable to the time required to solve a single iteration of CCP, EV and $Obj - EV$ respectively.

2.5.2 Correlations between protection levels and robustness metrics

The protection parameters α and Γ in the CCP, EV and $Obj - EV$ models are designed to represent the extent of robustness desired, with larger values of α and Γ representing higher

	% Flight Delays (min)					Pax Disruptions	
	≤ 15	≤ 60	≤ 90	≤ 120	≤ 180	Num. D-pax	%D-pax reduced
$\alpha_i = 90 \, \forall \, i$	78.54	93.10	95.63	97.82	98.91	1025	6.77
$\alpha_i = 92 \, \forall \, i$	77.54	92.54	95.00	97.36	98.54	1209	-9.90
$\alpha_i = 94 \, \forall \, i$	79.54	93.73	96.00	98.18	99.18	987	10.20
Airline's Routing	77.72	92.82	95.30	97.73	98.91	1100	0.00

Table 2.4: Robustness metrics for N_2 do not improve with increasing protection parameters in the CCP model ($t = 90$)

levels of solution robustness and improved robustness metrics values. Notice, from Table 2.4, that for individual flight protection levels α_i for flight i for a delay threshold $t = 90$, we get a network on-time performance for $t = 90$ minutes, of at least $\alpha = \alpha_i$ for all i. We observe, however, that the values of the robustness metrics do not necessarily increase for solutions to the EV, $Obj - EV$ and CCP models using increased values of protection parameters.

Table 2.4 and Figure 2-4 show that with increases in α, the solutions to CCP can worsen with respect to flight on-time performance, passenger disruptions and total aircraft delay minutes. There are several explanations for this. First, optimal solutions to the CCP with $\alpha_i = 90$ for all $i \in F$ can include solutions that satisfy $\alpha_i = 90$ or $\alpha_i = 94$ or $\alpha_i = 96$ for all $i \in F$. All of these solutions are considered optimal although, intuitively, the solution with the highest α value has the most slack and should therefore be the most 'robust'. As a result, non-monotonicity of robustness metrics occurs, as shown in Table 2.4. In the case of $\alpha_i = 90$ for all $i \in F$, 95.63% of the flights are delayed less than t=90 minutes and for $\alpha_i = 92$ for all $i \in F$, 95% of the flights are delayed less than t=90 minutes.

Another reason that the robustness metric values do not always increase with increasing values of α is that the CCP model focuses on selecting routings that limit the likelihood of occurrence of 'long' flight delays as a proxy for robustness, and is not (and cannot) be formulated such that its solution will optimize simultaneously the three different (and sometimes opposing) robustness metrics that we evaluate.

Similarly, as illustrated in Figure 2-5 and Table 2.5, higher values of Γ in EV do not always produce better solutions. The explanations for this occurrence are similar to that for CCP. Multiple optimal solutions to EV and $Obj - EV$ for given Γ-values satisfy protection values Γ, even though the solutions might have very different levels of slack and hence, exhibit very
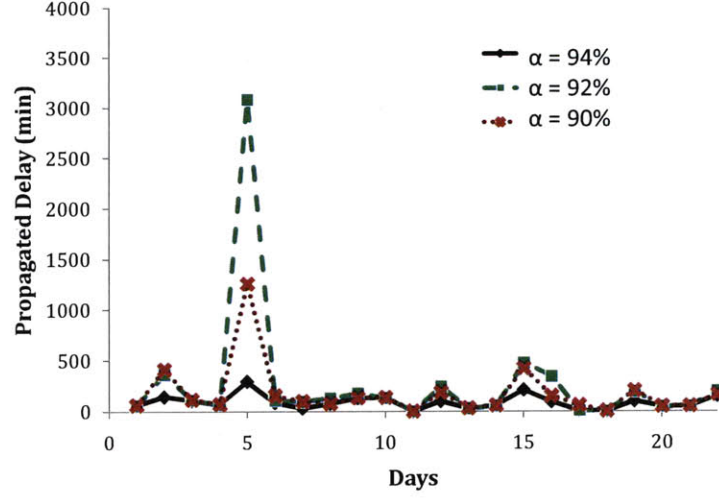
Figure 2-4: CCP model solutions for network N_2 do not show improved total delay minutes with increased protection ($t = 90$)

	% Flight Delays (min)				Pax Disruptions	
	≤ 15	≤ 60	≤ 120	≤ 180	#D-pax	%D-pax reduced
$\Gamma_i = 1 \forall i$	78.63	93.00	97.36	98.54	1222	-11.0
$\Gamma_i = 2 \forall i$	79.54	93.54	98.09	99.18	987	10.3
$\Gamma_i = 3 \forall i$	74.18	91.00	96.54	97.90	1353	-23.0
Airline's Routing	77.72	92.82	97.73	98.91	1100	0.0

Table 2.5: Non-monotonicity in robustness metrics for N_2 with increase in Γ in EV ($t = 90$)

different performance with respect to our robustness metrics. In fact, some of the optimal EV solutions are less robust with respect to our robustness metrics than the airline's routing, as is the case with $\Gamma = 3$ in Figure 2-5 and Table 2.5. The second explanation for why higher levels of Γ in EV and $Obj - EV$ models can lead to less robust solutions is the same as that for CCP. EV, like CCP, does not capture all the robustness metrics precisely, but rather builds robustness into the solution by over-covering flight legs with multiple aircraft if some have associated delays that can be long in the extreme-case. Similar is the case with $Obj - EV$. We, however, see slightly less variability in the alternate optima because the $Obj - EV$ model is based on extreme propagated delay minutes for each string and accounts for every minute of delay, instead of thresholds of t minutes.

The issue of choosing solutions with the maximum α and Γ values among multiple optimal
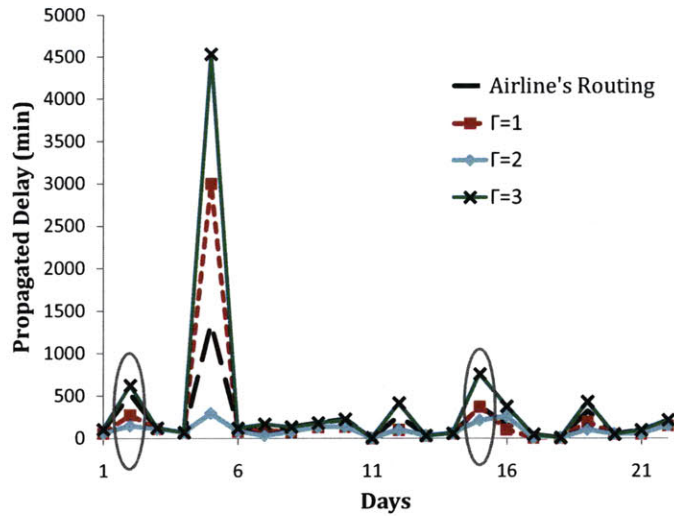
Figure 2-5: Bertsimas-Sim model solutions for N_2 show non-monotonic relationship of propagated delay with Γ ($t = 90$)

solutions is addressed by the $\alpha - CCP$, $\Delta - EV$ and $\Delta - Obj - EV$ respectively, with their objective functions to maximize the protection levels. Table 2.6 in comparison with Tables 2.5 and 2.4 show that these models can select, among different optimal solutions to the CCP and EV models, those with the highest levels of protection parameters, and it turns out, the greatest values of our robustness metrics.

2.5.3   Solution Differences due to Modeling Paradigms

In this section, we compare the three modeling paradigms: extreme-value based, probabilistic chance-constrained-based and the tailored approach studied in this paper. To avoid issues with specification of robustness parameters and to focus on the modeling paradigms, we compare the LCB, $\alpha - CCP$, $\Delta - EV$ and $\Delta - Obj - EV$ models.

Table 2.6 compares the airline's routing with solutions obtained from the $\Delta - EV$, $\Delta - Obj - EV$, $\alpha - CCP$ and LCB models. First, both the $\Delta - EV$ and $\alpha - CCP$ model solutions improve upon the airline's routing, overcoming the drawback with some of the solutions generated by the EV and CCP models. In fact, $\Delta - EV$, $\Delta - Obj - EV$, $\alpha - CCP$ and LCB perform similarly with respect to our metrics. The improvements in the 15 minute on-time performance results, for this data, result in the airline's ranking improving to place second among US carriers. The delay

63

	%Flight Delays (min)				Pax Disruptions	
	≤ 15	≤ 60	≤ 120	≤ 180	#D-pax	%D-pax reduced
$\Delta - EV$	79.54	93.54	98.09	99.18	987	10.3
$\Delta - EV$ (alt)	78.20	93.10	97.73	98.82	1056	4.03
$\Delta - Obj - EV$	79.27	93.54	98.10	99.10	988	10.14
$\Delta - Obj - EV$ (alt)	78.36	93.10	97.90	99.00	1012	7.93
$\alpha - CCP$	79.54	93.73	98.18	99.18	987	10.2
LCB	79.54	93.73	98.20	99.18	986	10.3
Airline's Routing	77.72	92.82	97.73	98.91	1100	0

Table 2.6: $\Delta - EV$, $\Delta - Obj - EV$ and $\alpha - CCP$ identify robustness parameters to improve upon the airline's routing for N_2 ($t = 90$)

minutes saved for this fleet result in a savings of \$120,000 for the 22 days, when the per-minute costs are according to the Air Transport Association [Air08]. Second, the $\Delta - EV$ model, like EV, can still have multiple optimal solutions with significant differences in performance. Although alternative optimal solutions have the same Δ-value, they differ significantly in the associated values of the robustness metrics.

Consider $\Delta - EV$ and $\Delta - Obj - EV$ solutions in Table 2.6. We see that the $\Delta - EV$ and $\Delta - Obj - EV$ solutions perform similar to the LCB and $\alpha - CCP$ model solutions; however their alternate optima $\Delta - EV$ (alternate) and $\Delta - Obj - EV$(alternate) perform significantly differently. In fact, $\Delta - EV$ and $\Delta - EV$(alternate) both ensure that of 44 of 50 flights in N_2 are covered (have delays less than t) in the extreme case, but have very different *probabilities* of flight delay less than $t = 90$, of 91% and 87% respectively. This results in a total delay difference of 2091 minutes over 22 days between the two solutions. A similar difference is observed between $\Delta - Obj - EV$ and $\Delta - Obj - EV$(alternate). These differences are much lower, however, in the case of worst-case delay metrics (such as delays more than 180 minutes), because these models are driven by the extreme value delays. Because the $\Delta - EV$ and $\Delta - Obj - EV$ models are formulated to avoid low probability values of worst-case delay or non-coverage, optimal solutions can have very differing values of average performance as measured by on-time performance and number of disrupted passengers averaged over several days. To illustrate, consider Figure 2-6 in which realization probabilities of two flight-string combinations $i_1 s_1$ and $i_2 s_2$ are shown. $\Delta - EV$ and EV do not distinguish between both these realizations, because both $i_1 s_1$ and $i_2 s_2$ have non-zero historical probability of realizing a value

64
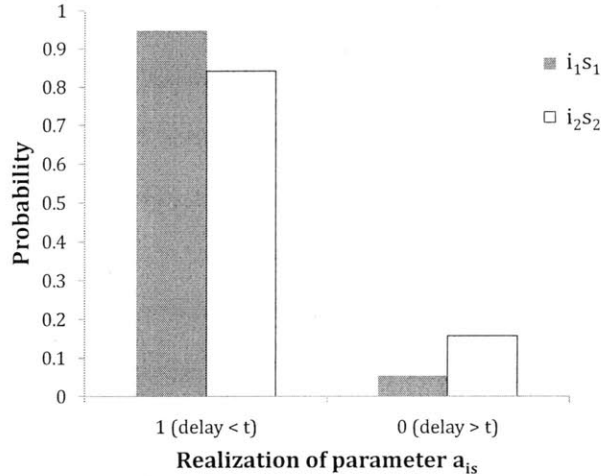
Figure 2-6: a_{is} realization probabilities, N_2

of 0. Also consider Figure 2-7(b) in which the distributions of propagated delays for strings in N_1 are shown. The propagated delay distribution is seen to be approximately bi-modal, with a large ($\approx$88%-95%) probability of the propagated string delay being on the lower end of the scale, and a small ($\approx$5%-12%) probability of the propagated delay being close to its worst-case value. The probability of interim values occurring is very small. Strings in which the worst-case propagated delay value is very large are correlated with the aircraft flying through multiple (usually) congested airports and the string containing a large number of flights (which can mean tightness in turn times). Because the probability of occurrences of such extreme propagated delays is small, the emphasis on extreme values of delays does not necessarily drive the extreme-value models towards good values of the average-case robustness metrics.

Because the $\alpha - CCP$ model captures the probabilistic nature of events, its emphasis on high-probability events results in optimal solutions which correlate on average to improved values of the metrics under consideration (Table 2.6). Alternative optima to $\alpha - CCP$ exist, but there is little difference in their performance with respect to the values of robustness metrics.

The LCB model uses average values of string propagated delay, shown in Fig. 2-7(b). Because the average is at about the 85th or 90th percentile of string propagated delay, LCB ignores the extreme value occurrences forming the remaining 5-10% of the distribution, and its objective of minimizing total average propagated delay seems to correlate well with our robustness metrics, with the LCB solutions performing well for all metrics (Table 2.6). Although

65

(a) String Propagated Delays

(b) Average and Extreme Values of Propagated Delays of Strings
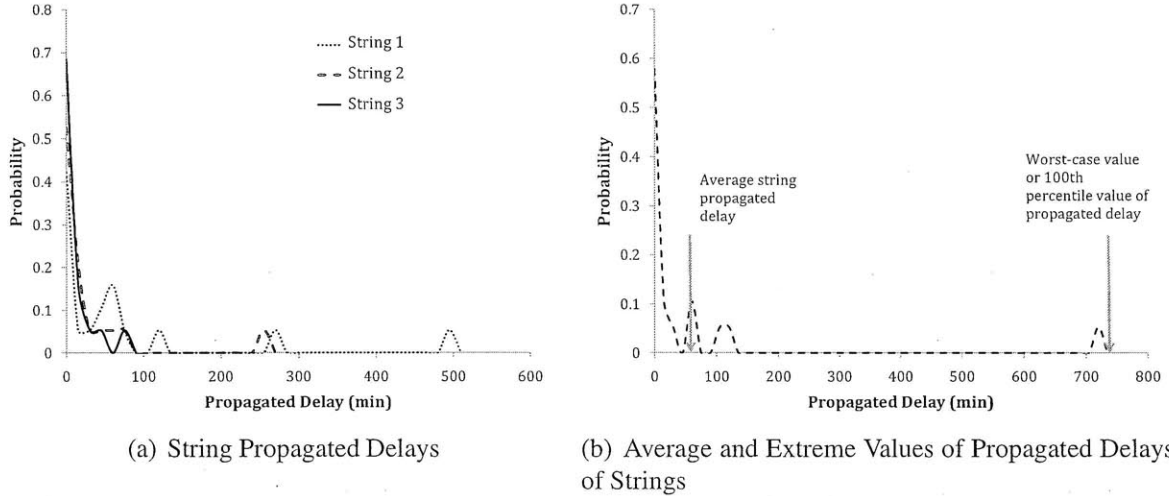
Figure 2-7: Propagated Delay Distributions of Strings

the LCB model does not capture the probability distribution of delays, by using average delay values, it is similar to a string-based $\alpha - CCP$ model when its protection level is set, for these problems, to the 90th percentile level. The advantage of LCB over $\alpha - CCP$ is that it is able to capture uncertainty in a simpler way, resulting in a highly tractable model.

We can apply the insights gained in solving the aircraft routing problem with LCB and $\alpha - CCP$ models to improve the extreme value-based robust optimization models. Specifically, by adjusting the 'extreme' values in the $Obj-EV$ and $\Delta-Obj-EV$ models to, say, the average values of string propagated delay, we can generate solutions using these models that perform more like those of the LCB and $\alpha - CCP$ models. This, of course, underscores the difficulty of setting these extreme values a priori, and the sensitivity of the solutions to these model inputs.

2.6 Conclusions

In this chapter, we study the application of three types of models - extreme-value based, probabilistic and tailored approaches, to the problem of aircraft routing. These three robustness mechanisms lead to different models with different solutions which have different robustness performances with respect to various metrics of interest.

The extreme-value models (EV and $Obj - EV$) were based on the Bertsimas and Sim robust optimization approach, the probabilistic model (CCP) on Charnes and Cooper's chance-

66

constrained programming approach and the tailored model (LCB) was Lan, Clarke and Barnhart's robust aircraft routing approach. Increased complexity and solution times are associated with the extreme value and probabilistic models, when compared to a deterministic model, because the models have to be solved several times for different values of the robustness parameters Γ and α. To avoid iterative re-solving, we developed extensions to these models: $\Delta - EV$, $\Delta - Obj - EV$ for the extreme-value approach, and $\alpha - CCP$ for the probabilistic approaches respectively. Our extended models can be solved in a single iteration, with runtimes equivalent or lower than a single iteration of the basic models. We evaluated solutions to the different models through simulation, and measure performance via total aircraft delay, on-time performance metrics, and passenger disruption metrics. The extended extreme value and probabilistic models can consistently lead to the generation of more robust solutions (compared to the basic models, and the solution currently operated by the airline), as defined by the metrics of interest. These models are also generally applicable, as described in §2.2.2 and §2.2.4.

Our findings show that extreme-value based models EV, $Obj - EV$, and $\Delta - EV$ and $\Delta - Obj - EV$ have optimal alternative solutions that exhibit very different performances according to our robustness metrics; varying from good improvement compared to the airline's routing, to no significant improvement or even deterioration. This behavior is because the robustness mechanism (Γ or Δ) is driven by *extreme* values of delay. This dependence on extreme delay values, ignoring probabilistic information, leads in some cases to a large disparity in the performance of the alternative optimal solutions. In such cases, extra care should be taken in evaluating alternative optimal solutions to these approaches. From this, we conclude that it is not effective to drive the solution process with extreme values that are rare.

The tailored approach LCB and the probabilistic CCP robustness approaches are very similar, in that expected values of string propagated delay, used to drive the LCB, are at about the 85th to 90th percentile of the string propagated delay distribution; and cause the formulation to focus on higher-probability events. Similarly, probabilistic approaches also focus on higher-probability delay events, and produce improved routings according to our metrics. These approaches thus capture more information about the system and focus on more likely delay events, and thus are more in line with our metrics of interest. Though the tailored approach in itself does not explicitly capture knowledge of probability distributions, by simplistically incorpo-

rating the 'right' delay quantile in its objective, it can achieve improved results through a less complex model. The probabilistic approaches (CCP, $\alpha - CCP$) allow more fine-tuning of robustness using the α and t parameters, but at a cost of a larger, more complex formulation (albeit very tractable for this application) and the implementation cost that additional distribution knowledge is needed for the $\alpha - CCP$ compared to LCB.

In conclusion, the efficacy of any given robust approach is determined not by the approach or model alone, but by the interaction between the model, data and evaluation metrics. Our work underscores the importance of choosing an approach that aligns itself well with the data distributions for the aircraft routing problem and metrics of interest to the DoT, industry and passengers. When applying general robust approaches to more specific problems, care should be taken to understand the nature of uncertainty and in choosing robustness parameters, in relation to the metrics, especially when metrics involving multiple stakeholders are involved.

Chapter 3

Robust Optimization Insights from Three Applications

A variety of robust approaches exist in the literature, but there has been limited work into (i) articulating useful guidelines to their application; and (ii) developing validation criteria for effectiveness across a variety of application domains. This work takes a step in that direction, by studying the application of three types of robust approaches to resource allocation problems drawn from the application areas of corporate portfolio optimization, pharmaceutical supply chain design and aircraft routing. Using a computational approach, we model, solve, and evaluate solutions from the robust optimization approaches. While the aircraft routing problem has been discussed in detail in Chapter 2, the corporate portfolio optimization and pharmaceutical supply chain problems arose in collaboration with IBM Research's Zurich Research Laboratory [MPRS09]. We present modeling perspectives for the three problems, comment on the importance of data and provide general insights into the application of the models considered. We expect that these insights can provide value in the case of other robust approaches as well.

We do not present a comprehensive analysis of each problem here. They can be obtained in Masters theses of Gallay [Gal05] and Epiney [Epi07]. Neither is the goal to present a 'recipe' for robust optimization. Instead, by combining observations from the three domains, we try to provide insights into the nature and application of different robust optimization approaches.

3.1 Robust Optimization

3.1.1 Defining Robust Optimization

Robustness in the airline context has been discussed in Chapter 2. In this section, we briefly discuss the notion of robustness in a more general context. Robust optimization may be conceptualized in two ways - (i) decreasing the susceptibility of a system to volatility; and (ii) creating flexibility when the decision making process is sequential, that is, in the recourse or repair actions that become available when uncertainty is encountered and results in disruptions to the plan.

Robust optimization problems are multi-criteria optimization problems. Fundamentally, the concept of robustness refers to the ability of a system to be less vulnerable to uncertainty and thus, more reliable. Because the different metrics of interest - cost, robustness, etc. - might not necessarily be positively correlated with each other, we expect that there will exist a trade-off between the different metrics of interest. In different systems, the concept of 'less vulnerable' is defined by different, sometimes conflicting, metrics of robustness. Ideally, we would like to obtain a pareto-optimal trade-off curve among the different robustness and cost metrics of interest to the stakeholders and decision-maker. Therefore, the goal in building robust solutions is to identify a trade-off frontier or a pareto-optimal frontier between these metrics. However, because robust approaches are at best approximations of the actual robustness metrics, the trade-off curves that are generated by the robust approaches are different from the trade-off curves we are trying to generate. We would like to be able to use approaches that most closely approximate the 'true' trade-off between the metrics, however, again, that is difficult to determine apriori.

3.1.2 Challenges in Building Robust Solutions

In building robust models and solutions, we encounter the following questions and challenges:

1. What does robustness mean? How can we model it in different contexts?

2. How can a problem be formulated using different robust optimization techniques?

3. What is the knowledge of uncertainty (data requirements) that allows a particular model to be applied?

4. What are the major factors that distinguish solutions generated from different modeling techniques?

5. Does more information/data about a system help generate better models and find more robust solutions?

6. In general, it is difficult to estimate realized costs or realized reliability during the optimization stage. How do we evaluate 'robustness'?

Building robust solutions first requires identifying the sources and nature of uncertainty - noise, ambiguity, measurement or statistical error, or unknown future events. In addition, as Greenberg [Gre07] points out, it is important to identify the time horizon of the decision, the price of constraint violation and undesirable outcomes, and if recourse action is possible. Approaches to model uncertainty are chosen based on a number of criteria, such as tractability and data availability, as we will discuss in the following sections. To evaluate the solutions obtained from different approaches, we should identify the metrics of robustness and costs of interest to the stakeholders and the decision-maker. These metrics may be ones that reflect volatility (e.g.: variance) or other metrics of importance to the different stakeholders (e.g.: passenger delays, passenger disruptions, aircraft delay).

Simulation is used to evaluate how well the trade-off made by the robust approach approximates the true trade-off between the robustness metrics for the problem. This is because we cannot always capture the true robustness metrics in the formulation due to modeling constraints, or the need to use proxies to capture features of robustness. Moreover, even if the true metrics are captured in the formulation, we are interested in capturing the realized values of the different robustness metrics instead of the deterministic value of the robustness metric from the optimization model.

3.1.3 Literature on Robust Optimization Approaches

Notable in collecting and classifying approaches under the umbrella of robust optimization is Greenberg and Morrison's summary of robust optimization approaches [GM08]. Stochastic optimization, chance-constrained programming, the Bertsimas and Sim approach, Ben-Tal and Nemirovski's approach and recourse-based optimization are brought under the umbrella of robust optimization. Robust approaches aim to build in resistance to uncertainty by defining

different measures of robustness. *Robust measures* are mathematically tractable representations of risk, (or conversely, reliability), which aim to approximate risk or reliability metrics of interest in real-world scenarios. A robust measure is typically modeled using parameters called protection levels, where increased values of the protection level parameters indicates higher robustness.

Deterministic models capture uncertainty by assuming that all uncertain parameters realize their average, or nominal values. This does not necessarily result in optimizing the objective function returns because uncertain parameters can be correlated, and the objective functions may not be linear. Several studies, including Mulvey et al. [MVZ81] and Ben-Tal and Nemirovski [BTN99] show that in the presence of uncertainty, in practice, such solutions can exhibit sub-optimal performance, and often are even infeasible.

Markowitz [Mar52] proposed the first mathematical programming model of uncertainty: the mean-risk model. For the portfolio selection problem, Markowitz suggested that investors not only wish to maximize expected returns but also minimize risk. Because these objectives conflict with each other, he suggested a model that trades off mean expected return and risk, using a bi-objective optimization approach. Risk is described as the variance of the portfolio, or using other metrics of variability. This allows the decision-maker to choose the most acceptable point on the expected return - risk curve based on the risk tolerance.

Stochastic optimization with recourse was proposed by Dantzig [Dan55]. This method provides a way of modeling not only the uncertainty, but also recourse actions in future scenarios of uncertainty through a two-stage optimization model. The decision variables of this model are of two kinds - one set is specified before the realizations of uncertain parameters are known, and the others after the uncertain parameters are realized. Typically this model is solved by enumerating the number of possible realizations (scenarios) and creating a decision vector for each scenario. This modeling approach can also extend similarly to the n-stage recourse model. This approach has been used in investment portfolio optimization to model investments over time. However, its application has been limited in the case of large-scale problems because it can rapidly enlarge as the number of scenarios and time stages increase. This is made easier due to advances in computation. In addition, these models do not capture ideas of tolerance to risk (risk-aversion) [MVZ81].

Charnes, Cooper and Symonds [CCT64] propose the chance-constrained programming approach, which regulates the probability of violation for each constraint under uncertain realizations, as described earlier in Chapter 2. From the above three classical models of capturing uncertainty (mean-risk, stochastic programming, chance-constrained programming), several other modeling paradigms emerged [GM08].

Mulvey et al. [MVZ81] present a paradigm of robust optimization that 'integrates goal programming with scenario descriptions of uncertain data' and generates solutions that protect against realizations of uncertain data from an uncertainty set. This framework integrates ideas from stochastic programming and multi-objective programming. Using ideas from robust statistics, Mulvey et al. propose a method by which higher moments of uncertainty distribution can be captured, and risk-aversion of users can be captured. This model thus results in solutions that are qualitatively different and have different operating strategies from those from stochastic programming.

Savage [FJ48] describes the decision-making process as one that is actuated by regret. He describes the utility analysis of choices involving risks. He postulates that among the choices of action available, the decision-maker picks the one that will minimize his regret of not having chosen another course of action under different possible scenarios of realization. Savage uses the criterion of *minimax regret*, that is, he hypothesizes that the decision-maker chooses the course of action that will minimize the maximum regret possible under different scenarios of real-world realizations [Sav51] [Sav54].

Gupta and Rosenhead [GR68] describe robustness as allowing flexibility in the kinds of recourse actions that are subsequently available. These models do not require complete information regarding the probability distributions of uncertain data or the exact cost structure of the model. Using qualitative information, the course of action is selected based on the flexibility it allows after the uncertain parameters are realized.

Several of these approaches assume that uncertainty is well-known, and can be expressed in the form of a probability distribution (by means of its mean and variance). There are other approaches to robust optimization that do not assume probability distributions for the uncertainty data. Instead, uncertainty is defined by uncertainty sets, ranges of values that each uncertain parameter can take. Notable among these are Sosyter [Soy73], Ben-Tal and Nemirovski [BTN99],

Bertsimas and Sim [BS04] and other work. Because the methods of Charnes and Cooper, Bertsimas and Sim and CVaR are of special interest, we discuss them in the following section.

Bertsimas and Thiele [BT06] describe the approach of Bertsimas and Sim in detail, while providing details on its application in practical applications, and in defining uncertainty in real data. Thiele, Terry and Epelman [TTE09] provide a model by which the robust optimization method of Bertsimas and Sim can be used to capture recourse decisions. They provide results for applications where uncertainty is focused on the right-hand-side of the uncertain constraints.

Fischetti and Monaci [FM09] propose an approach of capturing uncertainty that they refer to as *light robustness*, which combines the ease of modeling of the robust optimization approaches of Ben-Tal and Nemirovski [BTN99] and Bertsimas and Sim [BS04] with the ability of stochastic programming to model second-stage (recourse) actions using slack variables. They first formulate the nominal problem using the robust optimization approach, and add a budget constraint on the objective function of the nominal problem to create an optimization model with no objective function. Because this might be subject to infeasibility, the authors introduce slack (second-stage) variables that can account for local violations of the robustness requirements and add an auxiliary objective function corresponding to minimizing the slacks.

3.1.4   Approaches of Particular Interest

Of particular interest for the applications in this chapter are the approaches of using uncertainty sets as proposed by Bertsimas and Sim [BS04] [BS03], probabilistic constraints in Chance-Constrained Programming (CCP) proposed by Charnes and Cooper [CC59] [CC63] [CCS58], their extensions in §2.2.2 and §2.2.4, and Conditional Value-at-Risk (CVaR) proposed by Uryasev et. al. [RU00], [KPU02]. These approaches are chosen primarily due to modeling and tractability considerations. They satisfy the requirements of capturing uncertainty, and creating formulation structures that aid tractability.

We briefly revisit the approach of Bertsimas and Sim and Charnes and Cooper described in Chapter 2, and describe the CVaR measure of robustness in some detail.

The robust optimization approach of Bertsimas and Sim defines a robustness measure based on uncertainty sets [BS03], [BS04]. Within a defined uncertainty set, this model protects a

74

selected Γ number of uncertain parameters from realizing their extreme values at the bound of the range of values specified by the uncertainty sets. Specifically, this model assumes each uncertain parameter in the model to assume values within a symmetric, bounded interval. It does not assume any specific probability distribution for the uncertain parameters. Among the parameters thus subject to uncertainty, the model protects against the case when any number of coefficients, defined by the protection parameter Γ, assume worst-case values in their defined uncertainty sets. Bertsimas and Sim also relate their measure of robustness, which is defined in terms of uncertainty sets, to the measure of 'probability of constraint violation' by deriving bounds of constraint violation in terms of their protection level parameters Γ.

The Chance-Constrained Programming (CCP) approach of Charnes and Cooper defines a robust measure based on the probability of constraint violation. Charnes and Cooper allow random data variations and permit constraints to be violated up to specified probability limits. The level of protection afforded to a constraint subject to uncertainty is defined by a user-specified probability of constraint violation α. Unlike the approach proposed by Bertsimas and Sim, there is no single way of converting the chance-constraint into deterministic form. Multiple methods have been proposed in the literature to create a deterministic formulation, under several scenarios of uncertainty occurrences and probability distributions [CC59] [MW65] [NS06a] [Kuc09]. Perhaps the easiest case is when uncertainty occurs only in the right-hand-side of each chance-constraint, and this uncertainty can be captured by using the appropriate quantile of the right-hand-side parameter [CC59], [Mar07]. Chance-Constrained Programming encounters computational issues as we try to capture multiple uncertain coefficients per constraint, primarily because joint probability distributions need to be known [MW65]. Scenario generation under partial knowledge of joint probability distributions, or assumptions on joint probability distributions of the uncertain parameters are used to create deterministic equivalents under specific assumptions [NS06b]. Another, more complex model of CCP, is joint chance-constraints [MW65] [Baw73] [JR73] [Jag74]. Here an allowable probability of violation is specified for a joint set of multiple constraints. This case is the most complex to model and often can prove intractable.

Conditional Value at Risk or CVaR is a measure of risk that focuses on the more costly (extreme) outcomes of the probability distribution of the function of interest. (If the objective is

to maximize return, it focuses on the cases with low return, and if the objective is to minimize losses, it focuses on cases with high loss. Here we present the discussion for the loss function.) CvaR is based upon VaR, a measure of uncertainty that is defined as follows:

$$VaR(\alpha, \boldsymbol{x}) = \min\{\beta : P\left(f(\boldsymbol{x}, \zeta) \geq \beta\right) \leq \alpha\}, \tag{3.1}$$

where $f(\boldsymbol{x}, \zeta)$ is the probability distribution of the return function with input parameters ζ, and α a chosen probability. While VaR measures worst loss with a certain probability, it does not address how large the losses can be when the 'bad' events we want to protect against occur. VaR though easier to model, lacks some properties such as coherence and convexity. CVaR therefore measures the expected loss given the loss exceeds the VaR. CVaR is defined as:

$$CVaR(\alpha, \boldsymbol{x}) = E\left(f(\boldsymbol{x}, \zeta) | f(\boldsymbol{x}, \zeta) \geq Var(\alpha, \boldsymbol{x})\right) \tag{3.2}$$

Therefore, CVaR measures the expected return in the worst α percent of cases of the probability distribution of return, for a chosen α. Uryasev et al. [RU00] [KPU02] show that it offers stronger mathematical properties, compared to measures such as probability of violation used by the CCP. Uryasev et. al. also provide methods to approximate the CVaR function when it cannot be directly written in closed form, using discretization by scenario generation. This involves the ability to sample from the uncertainty distribution of the uncertain parameters, or the full knowledge of the joint probability distribution.

By applying these models to problems involving corporate portfolio optimization, pharmaceutical supply chain design, and aircraft routing (studied in Chapter 2), which encompass different domains, problem sizes and uncertainty types, we attempt to evaluate the efficacy of robust models and gain insights into performance of different robust approaches.

3.2 Problems of Interest

This section presents the three resource allocation problems, each from a different domain, we use as test-beds for developing our understanding of robustness and robust models. The problems under consideration belong to the areas of corporate portfolio optimization, pharmaceutical supply chain design and aircraft routing. The aircraft routing problem has been described in detail in Chapter 2. In this section, we will describe the other two problems.

These application domains span a broad range of resource allocation problems in operations research. These problems are of different scales - the corporate portfolio problem is a small-scale problem, the pharmaceutical supply chain problem a medium-sized one and the aircraft routing problem is larger-sized. Uncertainty in these problems also arises from different factors - the underlying model, data sensitivity and network structure, respectively. The metrics associated with measuring uncertainty in these problems are also different, as we will describe in the following subsections.

3.2.1 Corporate Portfolio Optimization

The objective of this problem is to allocate optimally a sales and marketing budget among twenty-six business units of a major corporation. For each unit, we are given sixteen quarters of past historical investments and corresponding quarterly revenues. In addition, for each business unit we have a minimum and a maximum feasible investment amount based on historical investments and business constraints.

To model an investment-return relationship, we make the assumption that the more money we invest in a business unit, the more we expect to make in revenue. The causal relationship between sales and marketing investment and the corresponding revenue has been studied extensively in business investment theory, and is typically assumed to follow an S-curve [LS82] [HPS03] [LR02]. In the data available to us, the manager's observations, existence of limited data and modeling considerations all dictate that we assume the investment-return relationship to lie in the linear portion of the S-curve [Epi07]. That is, the revenue r_i is described by $r_i = a_i + b_i x_i$ where a_i and b_i are parameters that describe the return for an investment of x_i. We estimate the parameters a_i and b_i using simple linear (least squares) regression, and the

corresponding estimates are $\bar{a}_i$ and $\bar{b}_i$ with covariance matrix C_i. We also estimate the standard deviation of the residuals in the regression. Under these assumptions, the problem can be formulated as a linear program.

We define the following notation:

- I: set of business units among which budget is allocated

- B: budget available for sales and marketing

- a_i, b_i: parameters describing the rate of return on investment for unit $i \in I$

- $\bar{a}_i, \bar{b}_i$: parameters describing the average rate of return on investment for unit $i \in I$

- $\bar{r}_i$: average rate of return on investment per unit for $i \in I$, equal to $\bar{a}_i + \bar{b}_i x_i$

- C: covariance matrix for a and b estimated using least squares regression

- x_i: budget allotted to business unit $i \in I$

- l_i: lower bound on investment in unit $i \in I$

- u_i: upper bound on investment in unit $i \in I$

The nominal problem, which does not explicitly model uncertainty, is as follows:

$$\max \sum_{i \in I} \bar{r}_i = \sum_{i \in I} \left(\bar{a}_i + \bar{b}_i x_i \right) \tag{3.3}$$

$$\text{s.t.} \sum_{i \in I} x_i \le B \tag{3.4}$$

$$l_i \le x_i \le u_i \qquad \forall\, i \in I \tag{3.5}$$

$$x_i \ge 0 \qquad \forall\, i \in I \tag{3.6}$$

The objective (3.3) of the nominal formulation (3.3) - (3.6) is to maximize the return on investment subject to a budget B on the total investment (3.4). (3.5) ensures that the investment x_i does not change 'too much' from previous investments, where the allowable change described by the limits l_i and u_i. (3.6) ensures non-zero investments in each business unit.

Uncertainty in the return on investment in this situation, which we plan to capture using robust models, arises from two sources: (i) possible mis-estimation of the linear model parameters due to limited historical data availability, and (ii) noise around the assumed linear trend. We capture these sources of uncertainty by modeling a_i and b_i as random variables via robust

models. For this problem, we define a good solution as one with high average return and low variance. We evaluate the cost-robustness trade-off through the mean-variance-trade-offs of the solutions generated, and generate these trade-off curves using Monte-Carlo simulation.

3.2.1.1 Robust Models

Bertsimas and Sim Model

The Bertsimas and Sim model assumes that the uncertain parameters a and b realize values in an interval that is symmetrically distributed around the nominal values $\bar{a}$ and $\bar{b}$. In this formulation, we define the interval of uncertainty to be one standard deviation away from the nominal values $\bar{a}$ and $\bar{b}$, as described by (3.8). The formulation protects against the scenario when a_i and b_i realize values $\tilde{a}_i$ and $\tilde{b}_i$. y_1^i and y_2^i are variables that determine if the uncertain parameters realize values at the extremes of their ranges of uncertainty. The number of parameters thus realizing extreme values is summed by the variables z^i and is limited by a user-defined parameter Γ. Γ is the parameter that measures the robustness measure of this approach, which controls the degree of conservatism of the solutions.

The robust formulation of (3.3) - (3.6) using the Bertsimas and Sim methodology is:

$$\max_x \left(\min_{(\tilde{a},\tilde{b})} \sum_{i \in I} \left(\tilde{a}_i + \tilde{b}_i x_i \right) \right) \tag{3.7}$$

$$\text{s.t.} \quad \begin{pmatrix} \tilde{a}_i \\ \tilde{b}_i \end{pmatrix} = \begin{pmatrix} \bar{a}_i \\ \bar{b}_i \end{pmatrix} + G \begin{pmatrix} y_1^i \\ y_2^i \end{pmatrix} \tag{3.8}$$

$$\sum_{i \in I} x_i \le B \tag{3.9}$$

$$l_i \le x_i \le u_i \qquad \forall\, i \in I \tag{3.10}$$

$$0 \le |y_j^i| \le z^i \le 1 \qquad \forall\, j \in \{1,2\}, \forall\, i \in I \tag{3.11}$$

$$\sum_i z^i \le \Gamma \tag{3.12}$$

$$x_i \ge 0 \qquad \forall\, i \in I \tag{3.13}$$

$$y_j^i \ge 0 \qquad \forall\, j \in \{1,2\}, \forall\, i \in I, \tag{3.14}$$

where G is the standard deviation matrix and $GG' = C$ (the covariance matrix), and Γ is the

number of uncertain parameters that are protected against the realizations of their worst-case, in the defined uncertainty range. This can be linearized by dualizing as described in [BS04]. We do not describe it here for the sake of brevity. This model, thus, protects against a specified uncertainty set, described by the range of uncertainty - here, the standard deviation G, and the protection parameter Γ.

Chance-Constrained Programming (CCP) Model

The CCP formulation enhances the nominal formulation of this problem by adding a constraint that requires the return to be greater than a specified critical value ($c.v._{CCP}$) with probability α ($\alpha \geq 0.5$). That is,

$$P\left[\sum_{i \in I}(a_i + b_i x_i) \geq c.v._{CCP}\right] \geq \alpha \tag{3.15}$$

As described in [Lin03], the above constraint can be expressed in closed form under the assumption of a bivariate normal distribution for the a_i and b_i. Under the normality assumption, (3.15) is converted into (3.17), which is a second order conic constraint. The chance-constrained formulation for (3.3) - (3.6) is as follows:

$$\max_x \sum_{i \in I}\left(\bar{a}_i + \bar{b}_i x_i\right) \tag{3.16}$$

$$\text{s.t.} \sum_{i \in I}\left(\bar{a}_i + \bar{b}_i x_i\right) + \Phi^{-1}(1 - \alpha)\sqrt{\sum_{i \in I}\langle 1, x\rangle^T C_i \langle 1, x\rangle} \geq c.v._{CCP} \tag{3.17}$$

$$\sum_{i \in I} x_i \leq B \tag{3.18}$$

$$l_i \leq x_i \leq u_i \qquad \forall\, i \in I \tag{3.19}$$

$$x_i \geq 0 \qquad \forall\, i \in I, \tag{3.20}$$

where $c.v._{CCP}$ is the critical value (obtained from statistical data) that the return from the portfolio should exceed, with probability α. Φ^{-1} in (3.17) indicates the inverse cumulative

standard normal function. Note that $\alpha = 0.5$ is equivalent to the nominal problem, and for an $\alpha \geq 0.5$, we obtain a second-order convex formulation. If the size of the problem is large, this formulation might face issues of tractability. In this case, because the size of the problem is small, the second-order program proves tractable.

Conditional Value-at-Risk (CVaR) Model

CVaR maximizes the expected return of the investment subject to the CVaR constraint, which states that under those scenarios when the return is less than the VaR, we get an expected return at least as much as $c.v._{CVaR}$. That is, the average value of return among those α percent of cases with the worst return, is at least $c.v._{CVaR}$. $c.v._{CVaR}$ is a specified critical value, determined by statistical methods. The CVaR constraint is as follows:

$$E\left[\text{Return}|\text{Return} \leq VaR_\alpha\right] \geq c.v._{CVaR}, \qquad (3.21)$$

where α is a user-specified tail probability of the return function.

Krokhmal, Palmquist and Uryasev [KPU02] show that (3.21) can be approximated linearly through the generation of a large number (M) of scenarios of uncertain parameter realizations. Each scenario is introduced into the formulation as a constraint. In order to generate such scenarios, we require the ability to sample from the joint probability distribution of the uncertain parameters. Note that we do not actually require to know the true distributions, an ability to sample is sufficient.

(3.22) - (3.28) is the CVaR formulation of the nominal formulation (3.3) - (3.6). Here, $\tilde{a}_{ij}$ and $\tilde{b}_{ij}$ represent the realizations of the uncertain parameters as observed in scenario j of the M sampled scenarios. (3.26) together with (3.25) approximate the true CVaR equation (3.21). z_j is a dummy variable that helps in the approximation [KPU02]. β is the approximation of the VaR value when CVaR is constrained as shown in this formulation. The remaining constraints are from the nominal formulation.

81

$$\max_{x,z_j,\beta} \sum_{i \in I} \left(\bar{a}_i + \bar{b}_i x_i \right) \tag{3.22}$$

$$\text{s.t.} \sum_{i \in I} x_i \leq B \tag{3.23}$$

$$l_i \leq x_i \leq u_i \qquad \forall\, i \in I \tag{3.24}$$

$$z_j \geq \beta - \sum_{i \in I} \left(\tilde{a}_{ij} + \tilde{b}_{ij} x_i \right) \qquad \forall\, j \in 1,...,M \tag{3.25}$$

$$\beta + \frac{1}{(1-\alpha)M} \sum_{j=1}^{M} z_j \geq c.v._{CVaR} \tag{3.26}$$

$$z_j \geq 0 \qquad \forall\, j \in 1,...,M \tag{3.27}$$

$$x_i \geq 0 \qquad \forall\, i \in I. \tag{3.28}$$

Clearly, to be able to approximate well the uncertainty and the true value of CVaR, we would like to generate a large number of scenarios. However, this formulation may run into tractability issues as more and more scenarios are sampled. Unlike CCP formulations [NS06a], there are no guidelines on the number of scenarios to generate for accurate representation of the constraints. This may be an issue as the computation time grows exponentially with the number of scenarios, as shown in Figure 3-1. Thus, we face a trade-off between accuracy and tractability for the CVaR formulation.

This trade-off is seen to be more pronounced in larger problems such as the pharmaceutical supply chain design problem and the aircraft routing problem. In the case of portfolio optimization, because the number of uncertain parameters is small, we do not experience tractability issues before obtaining a convergence in the value of the expected return and CVaR.

For this problem, the Delta model and the Extended Chance-Constrained Programming model (ECCP) [MB09] are inapplicable, as these models were developed for cases when the Bertsimas and Sim model and CCP model are binary integer programs and linear programs respectively.
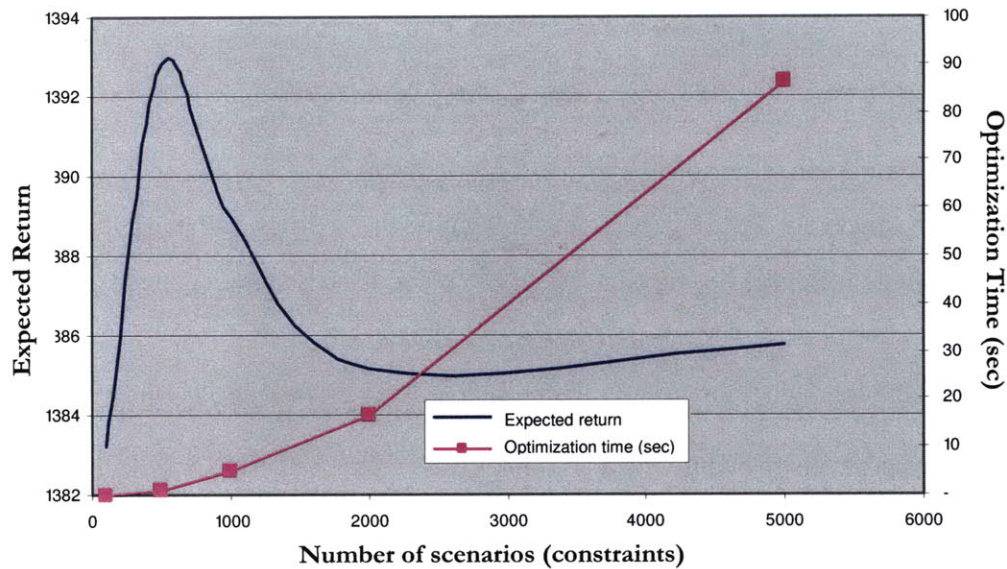
Figure 3-1: Accuracy - Tractability Trade-off for CVaR

3.2.2 Pharmaceutical Supply Chain Design

This is a strategic supply chain planning problem for a pharmaceutical company. The client manufactures 17 broad classes of products using different technologies at different manufacturing plants. The goal is to arrive at an optimal configuration for the supply chain for the following 5-10 years. Allowed changes to the existing network include closing or opening a plant, improving the technology used at the plant, moving a product from one plant to another, or in some cases, adding or discontinuing a product. Because the FDA introduced strict regulations in 2002, there is a hazard rate of failed inspections associated with a particular action of production, that is, with producing a particular product at a site using a technology. The formulation maximizes profit, while limiting the risk level to be borne due to the hazard of inspection. Constraining the design of the supply chain are the following: (i) limit on the amount of hazard (risk) to which the network is subject, (ii) the production technologies that can be adopted or introduced at a particular site, (iii) exit of a technology from a site/re-introduction of a technology at a site, and (iv) stopping production of a product at a particular site. We do not present the complete formulation with details of these constraints for the sake of brevity. Instead, we focus on the constraints that contain uncertain parameters to explain the robustness requirements.

83

Because the values of the hazard rates are determined by statistical methods, they are themselves subject to uncertainty. The goal of robust models, therefore, is to obtain solutions that are less sensitive to changes in the hazard rates. The solutions obtained are evaluated using Monte-Carlo simulation, a reasonable approximation in this case study. Metrics of interest in measuring robustness are the mean, variance, 5th and 95th quantiles of the profit. In particular, we study the trade-off curve between the mean profit and its variance as generated by different solutions.

We define the following notation for the problem, based on [Gal05]:

- P: set of products to be produced

- S: set of sites (locations) available for production

- T: set of technologies to be used for producing $p \in P$

- E: set of discrete time periods into which time horizon is divided

- X: set of feasible supply chain configurations, derived from the constraints (i) - (iv) described above

- $x(p, t, s, e)$: decision variables that take on value 1 if product p is produced using technology t at site s during period e

- $H(t, s, e)$: hazard rate, which is subject to uncertainty, and equal to the probability that an inspection of technology t at site s during period e results in a failure

- $\bar{H}(t, s, e)$: is the expected value of $H(t, s, e)$, the mean hazard rate of failed inspections

- $Rev(p, e)$: revenue generated by producing product p during period e

- $c.v.$: critical value of the revenue at risk, estimated by statistical methods

The nominal problem for the production supply chain, without considering uncertainty in the hazard rates, is as follows. The formulation in this case is a mixed binary integer program.

$$\max_{x \in X} \sum_{p,t,s,e} Rev(p,e)x(p,t,s,e) - \sum Costs(\mathbf{x}) \tag{3.29}$$

$$\text{s.t.} \sum_{p,t,s,e} \bar{H}(t,s,e)Rev(p,e)x(p,t,s,e) \leq c.v. \tag{3.30}$$

$$\sum_{t,s} x(p,t,s,e) \leq 1 \qquad \forall\, p,e \tag{3.31}$$

$$x(p,t,s,e) \in \{0,1\} \qquad \forall\, p,t,s,e \tag{3.32}$$

In the above formulation, (3.29) maximizes the expected profit, subject to the expected revenue that is at risk being limited by a critical threshold $c.v.$ (3.30); all products being produced (3.31); and all variables being binary (3.32). Let us denote the profit obtained from the solving the nominal problem as $Profit^*$. We will use this notation in the Delta and ECCP models described in the following section.

The motivation for building robust models for this problem is as follows. $H(t, s, e)$ values are estimated from historical data using statistical methods and the true realizations are inherently subject to *uncertainty*. Especially, in this data, it is found that with small changes in the values $H(t, s, e)$, the configuration of the supply chain that is optimal as obtained by solving the nominal problem (3.29) - (3.32) changes drastically. To illustrate, the supply chain configuration obtained by solving the nominal problem generated an expected profit of \$61000 (numbers have been scaled), but is extremely sensitive to these values of hazard rates. If the values of the hazard rates are perturbed a little from their statistical averages $\bar{H}(t, s, e)$, we find that the configuration of the supply chain changes drastically. In fact, if the hazard rates increase by 2% (from their original values), the profits drop by 40% and the number of product types produced drops from 17 to 14. To avoid the volatility exhibited by the nominal problem solution, we build the following robust models.

3.2.2.1 Robust Models

We formulate this problem according to the Bertsimas and Sim model, the Delta model, the Chance-Constrained Programming model and the Extended Chance-Constrained Programming model. The CVaR formulation of this problem runs into tractability issues because of its size, and the number of scenarios required to reasonably capture the uncertainty.

Bertsimas and Sim Model Formulation

According to this model, the parameters $H(t, s, e)$ are assumed to realize values in a range of uncertainty $\hat{H}(t, s, e)$ around the mean hazard rates $\bar{H}(t, s, e)$. That is, the hazard rates take values in the interval $[\bar{H}(t, s, e)\text{-}\hat{H}(t, s, e), \bar{H}(t, s, e)+\hat{H}(t, s, e)]$. The uncertainty set in this model is defined as the case when Γ of the hazard rates realize values at the worst-case bounds of their uncertainty ranges. The formulation, according to [BS04] and [BS03] is as follows:

$$\max_{x \in X} \sum_{p,t,s,e} Rev(p,e)x(p,t,s,e) - \sum Costs(\mathbf{x}) \tag{3.33}$$

$$\text{s.t.} \sum_{p,t,s,e} \bar{H}(t,s,e)Rev(p,e)x(s,e) \leq c.v. \tag{3.34}$$

$$\sum_{p,t,s,e} \bar{H}(t,s,e)Rev(p,e)x(s,e) + z\Gamma + \sum_{t,s,e} v(t,s,e) \leq c.v. \tag{3.35}$$

$$z + v(t,s,e) \geq y(t,s,e) \qquad\qquad \forall\, t,s,e \tag{3.36}$$

$$- y(t,s,e) \leq \left(\sum_{p} x(p,t,s,e)Rev(p,e) \right) \hat{H}(t,s,e) \qquad\qquad \forall\, t,s,e \tag{3.37}$$

$$\left(\sum_{p} x(p,t,s,e)Rev(p,e) \right) \hat{H}(t,s,e) \leq y(t,s,e) \qquad\qquad \forall\, t,s,e \tag{3.38}$$

$$\sum_{t,s} x(p,t,s,e) \leq 1 \qquad\qquad \forall\, p,e \tag{3.39}$$

$$x(p,t,s,e) \in \{0,1\} \qquad\qquad \forall\, p,t,s,e \tag{3.40}$$

$$v(t,s,e) \geq 0 \qquad\qquad \forall\, t,s,e \tag{3.41}$$

$$y(t,s,e) \geq 0 \qquad\qquad \forall\, t,s,e \tag{3.42}$$

Delta Model

This model is based on similar assumptions of uncertainty and uncertainty sets as the Bertsimas and Sim model. However, it drives the trade-off between optimality and robustness based on a budget (δ) for the profit. Additionally we change the objective to one of minimizing the maximum number of coefficients that must assume their nominal, rather than extreme, values to satisfy all constraints.

For this model, we order the ranges of the uncertain coefficients $\hat{H}(t,s,e)Rev(p,e)$ in increasing order. After ordering, the rank of the (p,t,s,e)th coefficient is denoted by $k; p,t,s,e$. Also, the original index (p,t,s,e) of the variable that takes the kth position in the sorted $\hat{H}(t,s,e)Rev(p,e)$ values is denoted by $p,t,s,e; k$. Thus, the value K of the index in the last position in the sorted list is described by $K = |P| + |T| + |S| + |E|$. We define variable Δ equal to the maximum number of variables $x(p,t,s,e)$ in the solution with $x = 1$ whose

coefficient values must assume their nominal values for the solution to remain feasible.

Let $Profit^*$ be the optimal profit of the nominal problem (3.29) - (3.32). Then δ is the user-specified incremental cost that is acceptable for increased robustness, that is, the profit of a robust solution from the Delta formulation is at least $Profit^* - \delta$. Let variables $v(p, t, s, e)$ equal 1 if the uncertain coefficient $H(t, s, e)Rev(p, e)$ is not allowed to take on its extreme value, and takes on its nominal value in the solution of the Delta model. Variables $w(k)$ equal 1 for all k for which there exists a $l \geq k$ with $v(p, t, s, e; l) = 1$. $w(k)$ variables in the sorted order of $\hat{H}(t, s, e)Rev(p, e)$ values follow a step function. This leads to the Delta formulation as follows:

$$\min_{x \in X} \Delta \tag{3.43}$$

$$\text{s.t.} \sum_{p,t,s,e} Rev(p, e)x(p, t, s, e) - \sum Costs(\mathbf{x}) \geq Profit^* - \delta \tag{3.44}$$

$$\sum_{p,t,s,e} (\bar{H}(t, s, e) + \hat{H}(t, s, e))Rev(p, e)x(p, t, s, e)$$

$$- \sum_{p,t,s,e} \hat{H}(t, s, e)Rev(p, e)v(p, t, s, e) \leq c.v. \tag{3.45}$$

$$\Delta \geq \sum_{p,t,s,e} v(p, t, s, e) \tag{3.46}$$

$$v(p, t, s, e) \leq x(p, t, s, e) \qquad\qquad \forall\, p, t, s, e \tag{3.47}$$

$$v(p, t, s, e; k) \leq w(k) \qquad\qquad \forall\, k = 1, ..., K \tag{3.48}$$

$$v(p, t, s, e) \geq x(p, t, s, e) + w(k; p, t, s, e) - 1 \qquad\qquad \forall\, p, t, s, e \tag{3.49}$$

$$w(k + 1) \leq w(k) \qquad\qquad \forall\, k = 1, .., K \tag{3.50}$$

$$w(0) = 1 \tag{3.51}$$

$$w(K + 1) = 0 \tag{3.52}$$

$$\sum_{t,s} x(p, t, s, e) \leq 1 \qquad\qquad \forall\, p, e \tag{3.53}$$

$$x(p, t, s, e) \in \{0, 1\} \qquad\qquad \forall\, p, t, s, e \tag{3.54}$$

$$v(p, t, s, e) \in [0, 1] \qquad\qquad \forall\, p, t, s, e \tag{3.55}$$

$$w(k) \in \{0, 1\} \qquad\qquad \forall\, k = 0, .., K \tag{3.56}$$

The constraints in this formulation are of the exact form as (2.23) - (2.37), and so we do not describe them in detail.

Chance-Constrained Programming (CCP) Model

We would like to capture the chance-constraint:

$$P\left(H(t,s,e)Rev(p,e)x(s,e) \leq c.v.\right) \geq \alpha. \tag{3.57}$$

CCP encounters issues when trying to capture multiple uncertain parameters per constraint. Therefore, we adopt a different way to constrain the uncertainty. As in the case of the corporate portfolio problem, we add an additional constraint (3.60). Here the risk is described by the left-hand-side of the constraint(s) (3.60), which assume that some βth quantile realizations of the uncertain parameters $H(t,s,e)$, denoted by $F^{-1}_{H(t,s,e)}(\beta)$, occur. Thus, using the left-hand-side, the type of realizations (scenarios) we want to protect against can be captured. This is appropriate for this problem, as sensitivity was found for such scenarios. The right-hand-side represents the quantile α of the critical value $c.v.$ against which we want to protect.

$$\max_{x \in X} \sum_{p,t,s,e} Rev(p,e)x(p,t,s,e) - \sum_{\text{All Costs}} \text{Costs}(\mathbf{x}) \tag{3.58}$$

$$\text{s.t.} \sum_{p,t,s,e} \bar{H}(t,s,e)Rev(p,e)x(s,e) \leq c.v. \tag{3.59}$$

$$\sum_{p,t,s,e} F^{-1}_{H(t,s,e)}(\beta)Rev(p,e)x(s,e) \leq F^{-1}_{c.v._{CCP}}(\alpha) \tag{3.60}$$

$$\sum_{t,s} x(p,t,s,e) \leq 1 \qquad\qquad \forall\, p,e \tag{3.61}$$

$$x(p,t,s,e) \in \{0,1\} \qquad\qquad \forall\, p,t,s,e \tag{3.62}$$

Extended Chance-Constrained Programming (ECCP) Model

The extended chance-constrained model builds on the chance-constrained formulation (3.58) - (3.62), as described in (2.53) - (2.62). We assume that some quantiles $\alpha_k, k = 1,..,K$ of the critical value $c.v._{CCP}$ are known, from analysis of historical data. Instead of choosing one par-

88

ticular value of α_k, we try to attain the highest protection level possible, within a budget δ on the profit. y_k are binary variables that equal 1 if the protection level α_k is attained by the solution. The objective (3.63) maximizes the protection level realized by the solution, which is described by (3.71) and (3.66). The protection level variables y_k take the form of a step function.

$$\max \ \alpha \tag{3.63}$$

$$\text{s.t.} \sum_{p,t,s,e} Rev(p,e)x(p,t,s,e) - \sum Costs(\mathbf{x}) \geq Profit^* + \delta \tag{3.64}$$

$$\sum_{p,t,s,e} \bar{H}(t,s,e)Rev(p,e)x(s,e) \leq c.v. \tag{3.65}$$

$$\sum_{p,t,s,e} F^{-1}_{H(t,s,e)}(\beta)Rev(p,e)x(s,e)$$

$$\leq \sum_{k=1}^{K} F^{-1}_{c.v.CCP}(\alpha_k)(y_k - y_{k-1}) \tag{3.66}$$

$$\sum_{t,s} x(p,t,s,e) \leq 1 \qquad\qquad \forall\, p, e \tag{3.67}$$

$$y_k \geq y_{k-1} \qquad\qquad \forall\, k = 1,..,K \tag{3.68}$$

$$y_0 = 0 \tag{3.69}$$

$$y_K = 1 \tag{3.70}$$

$$\alpha \leq \sum_{k=1}^{K} \alpha_k(y_k - y_{k-1}) \tag{3.71}$$

$$x(p,t,s,e) \in \{0,1\} \qquad\qquad \forall\, p,t,s,e \tag{3.72}$$

$$y_k \in \{0,1\} \qquad\qquad \forall\, k = 1,..,K \tag{3.73}$$

The constraints in this formulation are of the exact form as (2.53) - (2.62) and so we do not describe them in detail.

Conditional Value-at-Risk (CVaR) Model The CVaR constraint for the nominal problem is given by:

$$CVaR(\alpha, \boldsymbol{x}) =$$

$$E\left(\sum_{p,t,s,e} \bar{H}_j(t,s,e)Rev(p,e)x(p,t,s,e) \middle| \sum_{p,t,s,e} \bar{H}_j(t,s,e)Rev(p,e)x(p,t,s,e) \geq VaR(\alpha, \boldsymbol{x}) \right) \tag{3.74}$$

89

The CVaR model minimizes the expected revenue at risk, when the realized revenue at risk is greater than the VaR corresponding to the αth level of protection. The constraints for the pharmaceutical supply chain problem are along the lines of (3.22)-(3.28), as below. We do not describe the constraints in detail because they are almost the same as those for the corporate portfolio problem ((3.22) - (3.28)).

$$\max_{x \in X} \sum_{p,t,s,e} Rev(p,e)x(p,t,s,e) - \sum Costs(\mathbf{x}) \tag{3.75}$$

$$\text{s.t. } z_j \geq -\beta + \sum_{p,t,s,e} \bar{H}_j(t,s,e)Rev(p,e)x(p,t,s,e) \qquad \forall\, j \in J \tag{3.76}$$

$$\beta + \frac{1}{(1-\alpha)M} \sum_{j=1}^{M} z_j \leq c.v._{CVaR} \tag{3.77}$$

$$\sum_{t,s} x(p,t,s,e) \leq 1 \qquad \forall\, p,e \tag{3.78}$$

$$x(p,t,s,e) \in \{0,1\} \qquad \forall\, p,t,s,e \tag{3.79}$$

$$z_j \geq 0 \qquad \forall\, j = 1,..,M \tag{3.80}$$

This formulation proves to be intractable when the number of scenarios is more than 400. In order to approximate the uncertainty accurately, a number of scenarios, at least in the tens of thousands, is necessary. We do not report the solutions for the smaller number of scenarios because they might be flawed and misrepresentative of the true realizations.

Aircraft Routing

Robust models for the aircraft routing problem are presented in detail in Chapter 2. We solve this problem using the Bertsimas and Sim model, the Delta model, the CCP model and the ECCP model. CVaR is difficult to model for the aircraft routing problem because we need to generate millions of scenarios to comprehensively capture the possible scenarios. Because of the size of the problem as well as the network structure, generating real scenarios proves difficult. In addition, even if we could generate the scenarios, the CVaR model for aircraft routing is expected to be intractable (similar to the pharmaceutical supply chain problem), due

90

to its size.

Table 3.1 summarizes the methods applied and solved for the three applications.

Models	Corporate Portfolio	Pharma Supply Chain	Aircraft Routing
Bertsimas and Sim	×	×	×
Delta	Inapplicable	×	×
CCP	×	×	×
ECCP	Inapplicable	×	×
CVaR	×	Intractable	Intractable

Table 3.1: Summary of Robust Approaches and Applications

3.3   Results

As discussed in the previous sections, robust approaches generate solutions that trade off between different metrics of interest, such as cost and robustness. In doing so, they aim to approximate the 'true' trade-off frontier between different metrics of interest. In this section, we present some observations and insights from the application of the various robust approaches to the three applications. We show computationally that the generation of truly robust solutions for a problem instance is dependent on the interaction between the robust approach, the data distributions and the stakeholders' metrics for robustness.

3.3.1   Role of Robust Approach

3.3.1.1   Comparison of the Trade-off Curves from Different Approaches

Because the robustness measures used by the robust approaches are different, and moreover, the relationships of these measures to the decision-maker's robustness metrics are different, the trade-off curves generated by the approaches vary. Figure 3-2 compares solutions to the pharmaceutical supply chain problem obtained using the measures of 'number of uncertain coefficients at worst-case' (Bertsimas and Sim and Delta approaches) and 'probability of constraint violation' (CCP and ECCP approaches), and measures their trade-off curves according to two criteria - mean profit (cost metric) and standard deviation of profit (robustness metric).
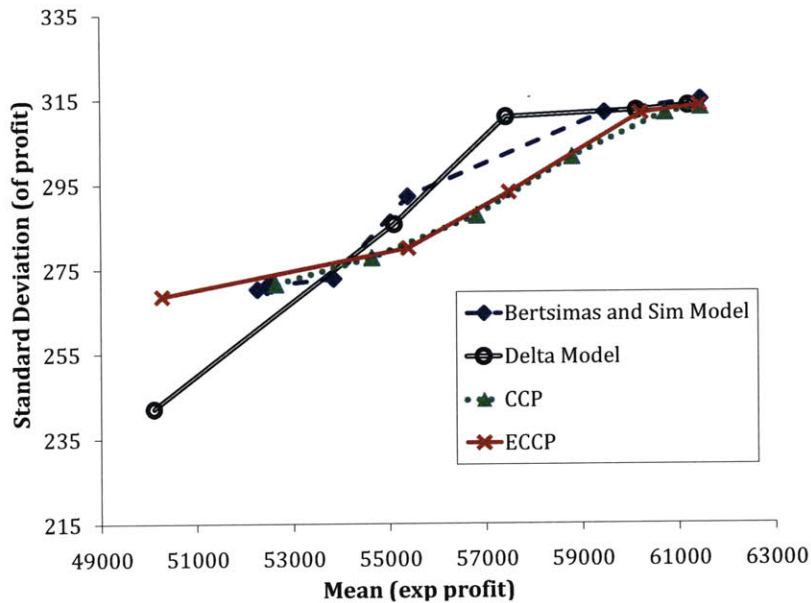
Figure 3-2: Mean-variance trade-off curves of extreme-value and chance-constrained models

The measure of 'number of uncertain coefficients at worst-case' used by the Bertsimas and Sim and Delta approaches is consistently seen to have a less smooth trade-off than the 'probability of violation' metric used by CCP and ECCP, because the latter methods measure the actual constraint violation probability, whereas the former methods exhibit step-like behavior in the protection levels (described by Bertsimas and Sim as phase transitions [BS04]) because they are based on the *number* of uncertain coefficients protected. Due to this reason, the trade-off curves generated by these approaches intersect, resulting in different methods being more effective over different ranges. Similar behavior is observed when risk adjusted profit is plotted against the standard deviation.

For different requirements of robustness (in this case, defined by the standard deviation of the profit), solutions generated by different approaches might become more valuable in terms of other criteria such as profit. For the data instance solved in Figure 3-2, the Delta approach is more valuable if the decision maker requires the standard deviation of the profit to be lower than $275, and the CCP and ECCP are more profitable on average if a higher variance is acceptable. It is difficult, however, to apriori predict where the trade-off curves produced by these models can intersect, and which method is better in a particular range.

Similarly, in the aircraft routing problem, as discussed in Section 2.5.3, we see that the robustness metrics - flight delays and passenger delays - are different for the different approaches. Remember, that the aircraft routing problem is a feasibility problem. Thus the trade-off is such that the cost (feasibility) does not change, but we are able to find solutions with different values of robustness metrics from the different modeling paradigms.

Thus, a wise choice of models and solutions requires simulation of the different methods to choose the more robust models according to the metrics of interest.

3.3.1.2 Conservatism of the Bertsimas and Sim and Delta Approaches

The measure of 'number of uncertain parameters at worst-case boundaries' (Γ and Δ), has been observed to exhibit a high degree of conservatism.

In [BS04], Bertsimas and Sim derive a relationship between the parameter Γ and the probability of constraint violation, which is a helpful starting point to choose a good value of Γ. However, it is important to note that this relationship is sensitive to the type of uncertainty distribution, because it assumes a symmetric probability distribution. Most importantly, this relationship is an upper bound and is not guaranteed to be tight. It was found not to be tight for the pharmaceutical supply chain problem [Gal05], as well as in the UAV routing problem of Sakamoto [Sak06]. This was true even in the case of discrete distributions, where the bound is expected to be tight. The probability of actual violation from experiments was much lower than that predicted by the bound in [BS04], rendering this approach conservative. Gallay [Gal05] shows that for integer/mixed integer programs in particular, the bound can be tightened to some extent by guiding the choice of protection level Γ using the possible number of non-zero variables in the solution N', instead of N (the total number of variables) in the formula given in [BS04]. N' can be reasonably estimated using the rank of the matrix and the expected number of basic variables in the solution. Repeated re-solving with several different values for Γ is possible for problems that are small and tractable in terms of solution and simulation times. This was possible for the corporate portfolio problem and the pharmaceutical supply chain problem. For problems that are of larger size, such as the aircraft routing problem, repeated re-solving is cumbersome, if not intractable. When limiting solution conservatism is a priority, the Delta model that drives the trade-off between cost and robustness through the budget constraint can

Solutions from the Bertsimas and Sim approach: Mean-variance tradeoff of profit
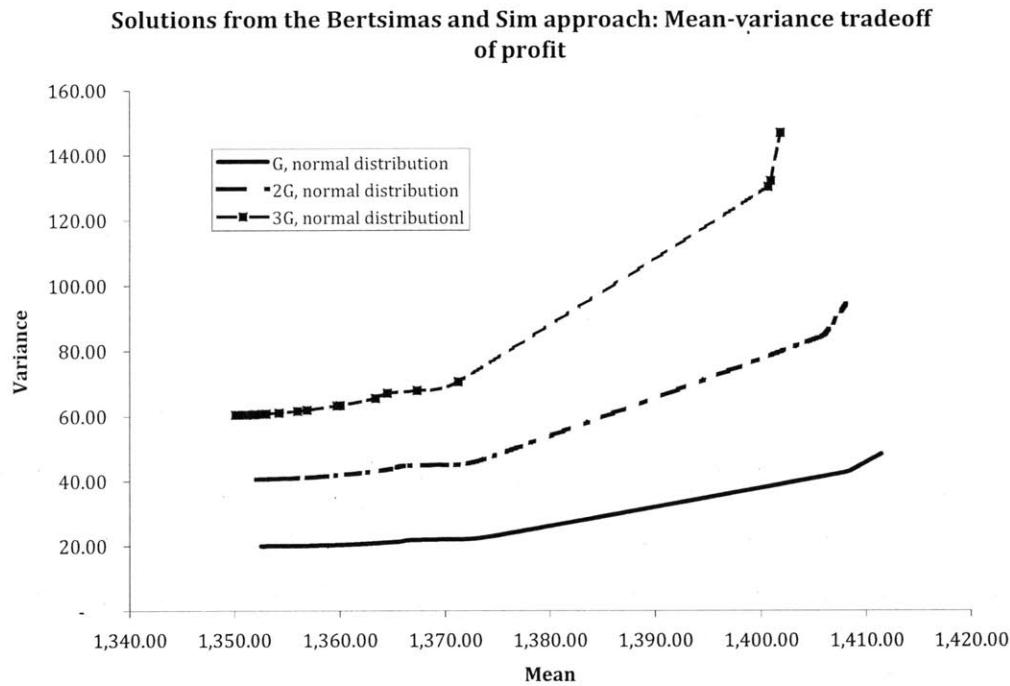
Figure 3-3: Sensitivity to uncertainty range

prove useful [MB09].

The degree of conservatism observed in the Bertsimas and Sim approach leads it to be most effective and to produce solutions with good trade-offs when we have very little confidence in the distribution of the uncertain parameters, and the uncertainty arises from a distribution with more variance than expected. This is illustrated by Figure 3-3. The solutions shown all arise from assuming a deviation of one standard deviation (G) from the nominal value. However, these solutions are less conservatively robust when the realized uncertainty is much larger, that is, 2 standard deviations away from the normal ($2G$) or 3 standard deviations away from the normal ($3G$) as shown.

From our experiments we see that the solutions to the Bertsimas and Sim and Delta models are sensitive to the choice of the ranges of uncertainty. For the pharmaceutical supply chain problem, Table 3.2 shows the performance of solutions to Bertsimas and Sim's model when the underlying uncertainty distribution is normal and input ranges of uncertainty in the hazard rates are 0.04 and 0.02, for the same values of Γ. Solutions are evaluated for the true uncertainty range of 0.04. We observe that the degree of conservatism with respect to the mean values

94

	Bertsimas and Sim, range = 0.04		Bertsimas and Sim, range = 0.02	
	Mean	Standard Deviation	Mean	Standard Deviation
$\Gamma = 0$	61447.40	314.6154	61447.40	312.7749
$\Gamma = 8.18$	59467.25	311.6346	61447.40	312.7749
$\Gamma = 20$	55374.25	292.0533	59465.35	310.7554
$\Gamma = 31.78$	53832.47	272.7504	58547.33	310.1444
$\Gamma = 51$	52544.58	271.2527	57662.30	300.3998
$\Gamma = 80$	52449.74	270.8838	57068.12	300.1223
$\Gamma = 120$	52253.36	270.2679	57059.65	300.1865

Table 3.2: Solutions to the Bertsimas and Sim model and sensitivity to uncertainty range

can be changed by changing the uncertainty range used in the optimization, though the general trend in the trade-off curve remains of the same shape (Figure 3-4). Fig 3-5 shows the same for solutions to the Delta model, for the same values of δ. Similar behavior is observed when simulated with uniform and discrete distributions.

We see that the solutions trace different trade-off curves between the mean and variance of profit, for the same values of Γ and δ, when different ranges of uncertainty are input to the models. Thus changing the range of uncertainty for the model can prove to be a way of controlling the conservatism or trade-offs made by the approach.

3.3.1.3 Degree of Conservatism in Chance-Constrained Programming

Solutions to the three problems produced by the CCP approach, unlike the extreme-value based approaches (Bertsimas and Sim and Delta approaches), typically encounter degrees of constraint violation when simulated that are similar to those specified in the optimization model, as discussed in Section 2.5.2 for aircraft routing. However, this comes at the cost of the need for more distribution information, or the ability to sample from the true realizations. When limited distribution/quantile information is available, this method can still be applied, but may result in conservatism if too little information is available. In addition, capturing multiple uncertain coefficients also can require assumptions on distributions and inaccurate capture of the uncertainty. CCP, however, allows the protection of uncertain coefficients to some quantile of uncertainty, as in the case of (3.58) - (3.62), thus controlling the degree of conservatism to a greater extent than that of the Bertsimas and Sim approach.
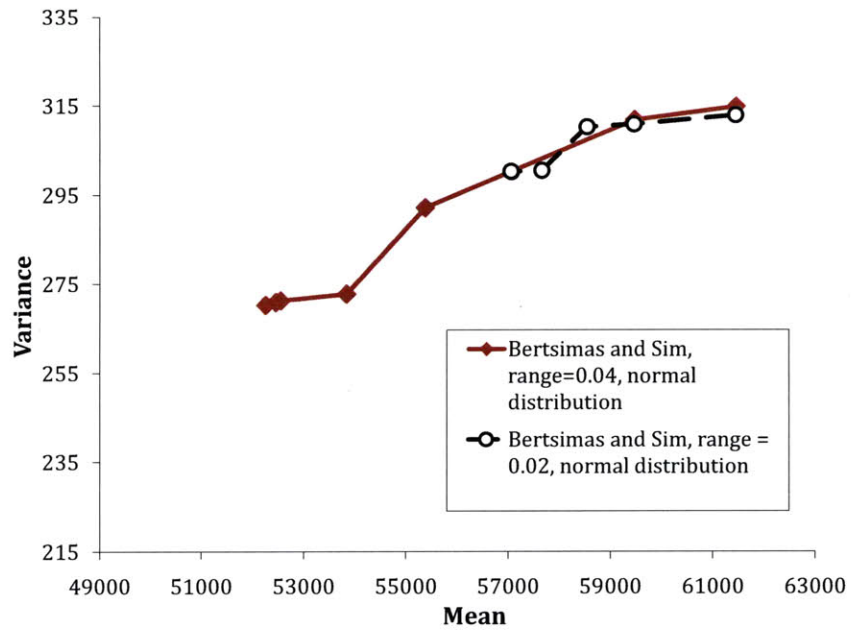
Figure 3-4: The Bertsimas and Sim model's sensitivity to uncertainty range
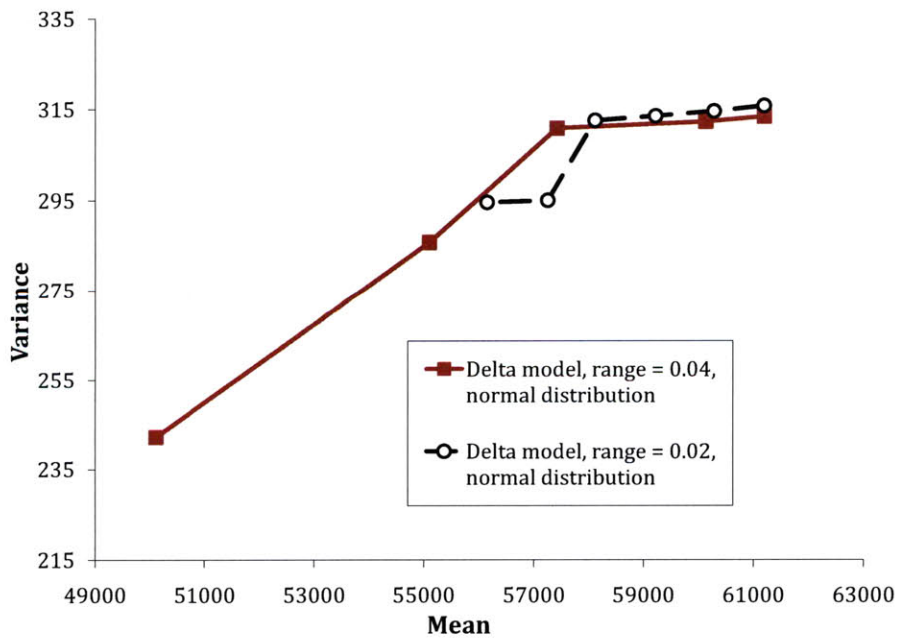
Figure 3-5: The Delta model's sensitivity to uncertainty range

3.3.2   Relationship between Robust Approach and Metrics

3.3.2.1   Optimizing model's robustness metric differs from optimizing decision maker's robustness metric

Key to understanding the behavior of the solutions generated using various robust approaches is understanding the relationship between robustness as defined by the approach, versus robustness as defined by the stakeholders' metrics. When we optimize with respect to a robust measure that approximates (sometimes not very closely) the metrics of the decision-maker, we may encounter unexpected behavior.

For example, when solving the problem of robust aircraft routing using the Bertsimas and Sim and Chance-Constrained Programming approaches, we find (as described in Chapter 2) that as the respective robust measures increase, the solutions do not necessarily increase in robustness with respect to flight delays and passenger disruptions (the decision-maker's metrics). This appears as a non-monotonicity in solution performance, with respect to our metrics of interest. Sakamoto [Sak06] and Bryant [Bry06] observe similar phenomena in problems of UAV scheduling and routing under uncertainty. Underlying this behavior are two reasons. First, the measure of robustness that is being optimized by the robust model is not the same as the decision maker's metrics. In fact, capturing explicitly the decision maker's metrics in the robust formulation proves difficult. Because the Bertsimas and Sim and CCP classes of models each have their own robustness measures, different degrees of non-monotonicity are observed. Second, for an 'optimized value' of the robust measure (protection level) set by the approach, there might be multiple (optimal) solutions with different values of the decision-maker's robustness metric, as was seen in the case of the aircraft routing problem in §2.5.3. Existence of multiple optimal solutions is very commonly observed in the case of problems involving network design or flows, because these problems tend to have a very large number of multiple paths or routes available, resulting in the availability of multiple optimal solutions.

The existence of such non-monotonic behavior does not mean that solutions generated using these methods are not robust. In fact, several of the solutions produced in the aircraft routing problem are more 'robust' than the airline's routing. Rather, this indicates that care should be taken in ranking the solutions as 'more' or 'less' robust simply based on the model's robustness

parameters or measures, such as Γ or α. This observation again points to the importance of simulation, and evaluation based on *multiple* criteria, in choosing robust solutions.

3.3.3 Importance of data

3.3.3.1 Using the nature of data distributions to guide the approach

For the application of any robust approach, it is important to know the nature of the data distributions. Implementing the robust model itself may not require exact knowledge of the data distributions, however, in order that the approach be effective, input parameters to the model must be modified to reflect the nature of the distribution.

For instance, the Bertsimas and Sim and Delta models use an extreme-value based approach that assume symmetric and bounded distribution of uncertainty around the mean. Bertsimas and Sim's robust optimization model is suitable when the decision-maker is risk-averse [KT09] and is found to be most effective for heavy-tailed distributions where the worst-case values have a significant probability of occurrence. For example, in the pharmaceutical supply chain problem, we see that as the true distribution of uncertainty becomes more heavy-tailed in the model's range of values, the solutions might be in better accordance (though still not tight) with the probability of constraint violation predicted by Bertsimas and Sim [Gal05].

So, if the *true* realization of uncertainty is described best by thin-tailed distributions (the worst-case values have low probabilities of occurrence), we re-solve the model after 'shrinking' the range of uncertainty, to obtain less conservative solutions with respect to the extreme-values. For instance, in the corporate portfolio problem, we see that the Bertsimas and Sim approach produces solutions with better trade-offs between expected profit (cost) and standard deviation of profit (robustness metric) when uncertainty (thin-tailed distribution) input into the model is for a smaller range than actually realized (Figure 3-3).

This underscores the difficulty of setting apriori a set of 'extreme values', 'uncertainty ranges' or probability distributions as inputs into a robust approach. It must be kept in mind, however, that this can be done when the extreme values are both low probability as well as low cost. If the cost associated with the realization of the worst-case is very high, or the decision-maker (or stakeholders) are highly risk-averse, then it might be more valuable to incorporate

them into the model and err on the side of conservatism.

3.3.3.2 Knowledge of empirical data distributions

Knowledge of empirical data distributions can add value. In the aircraft routing problem, we use historical data to generate empirical data distributions that are used as inputs into approaches for generating robust solutions. In particular, for the aircraft routing problem, this was useful in understanding the bi-modal nature of the probability distribution. In this case, delays were observed to be small or moderate for 85-90% of the time, and very large about 10-15% of the time. This information helps the decision-maker to understand which kind of delays to target, set model parameters accordingly, and interpret performance of solutions generated from various robust approaches [MB09].

3.4 Conclusions

From the results discussed above, we see that some approaches such as the approach of Bertsimas and Sim and the Delta model are particularly applicable if we know very little about the underlying uncertainty distribution, and we are likely to encounter more uncertainty than expected. In such cases, it may be be worth taking the risk of conservatism using this approach. Other approaches such as CCP and CVaR require the ability to sample the uncertainty distribution, via historical data. If sufficient historical data (and estimates of future uncertainty to validate these data distributions) are available, these approaches might be valuable, especially when the decision maker's metrics may not be completely risk-averse.

Finally, in choosing a solution among the multiple trade-off frontiers generated, the **multi-criteria nature of robustness** comes into play. The decision-maker compares the simulated performances of the nominal solution, and the solutions obtained from multiple robust approaches, and examines the trade-offs made by the different solutions between cost and the various robustness metrics. Using the nominal solution as a reference point, the decision-maker chooses among the different solutions based on acceptable levels of conservatism of the objective function. Sometimes the solutions from a particular robust approach, for some data in-

99

a) Robust approach produces solutions with slightly smaller mean, but far lower variance compared to nominal

b) Robust approach produces solutions with slightly smaller mean, but also only slightly lower variance compared to nominal
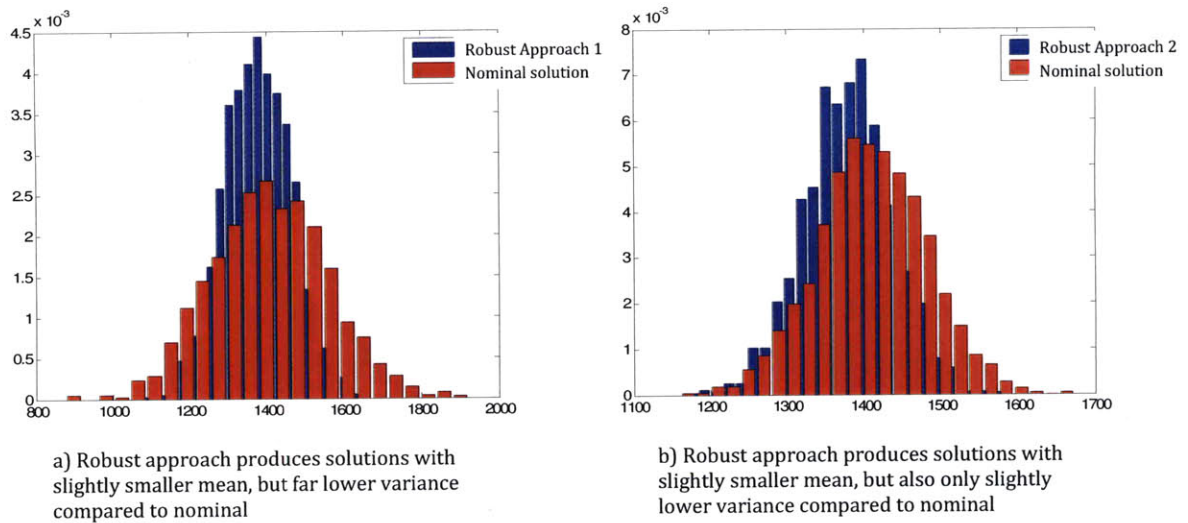
Figure 3-6: Multi-criteria nature of robustness

stances, may offer a less favorable trade-off than the solution to the nominal problem. Figure 3-6 illustrates how in case (a), the robust solution produces solutions with a slightly smaller mean profit, but far lower variance in profit than the nominal problem solution. Thus the decision-maker would lose little profit by choosing the solution from the robust approach. In contrast, the robust solution in Figure 3-6 (b) also reduces the mean profit as well as the variance of the profit compared to the nominal problem; however, the decrease in the mean profit is significantly higher than the decrease in variance, causing the solution to not gain much reliability. If case (b) is observed, the decision-maker may decide that the solution from the robust approach is too expensive and decide against implementing it in practice.

However, this does not mean that the robust models using the various approaches were of no value in case (b). Indeed, the value provided is in evaluating and validating the efficacy of the solutions available, identifying a trade-off frontier between the multiple objectives confronting the decision-maker and stakeholders, and choosing a solution that provides the 'best' balance among the multiple criteria.

Chapter 4

Integrated Disruption Management and Flight Planning

4.1 Introduction

In this chapter, we discuss the integration of *disruption management* and *flight planning*. Our goal is to reduce flight delays and disruptions using mechanisms facilitated by flight planning. To our knowledge, this is the first work that integrates these aspects of airline operations.

Inherent uncertainty in airline operations makes delays and disruptions inevitable. Because the airline system operates as a closely interconnected network, it is subject to 'network effects', that is, a disruption in one place can quickly propagate to multiple other parts of the network, as discussed in Chapter 1. Therefore, managing these delays as they arise is crucial. *Disruption management* is the process by which, on the day of operation, when a disruption occurs, airlines try to bring the plan back on schedule as quickly as possible, while incurring minimal costs and disrupting as few resources and passengers in the system as possible. Measures such as flight cancelations, flight re-timing, aircraft swaps, crew duty swaps and use of reserve crew are used as part of the disruption management process.

Flight planning is the process of determining, at the pre-departure stage of each flight, its three-dimensional travel route, speed and fuel burn as the aircraft flies from its origin to destination. Because the ability to make speed changes directly impacts the actual block time of a flight, and thus, its actual arrival time, which in turn can impact network connectivity of aircraft,

crew and passengers from the flight, flight planning can be used as a mechanism to change the lengths of the block and ground times of a flight. Flight plans, thus trade-off the costs associated with the flight arrival time (network connectivity costs associated with delays and disruptions), and the fuel burn (fuel costs). For this reason, flight plans are identified by a measure called the Cost Index (CI), described in detail in §4.3. CI is an indicator of the trade-off between fuel and network connectivity costs made by changing aircraft speed. The use of CI is now standard practice in the industry, and is used as a rule-of-thumb.

In this chapter, we will see how mechanisms of (i) flight speed changes and (ii) intentional flight departure holds, enabled by flight planning, can enhance the process of disruption management.

We study flight speed changes as a way to trade off fuel cost of a flight and its network connectivity costs; by adjusting the flying time of the flight and its passenger connectivity during a delayed or disrupted scenario. Our objective is to discuss the potential for using operational flight speed changes (flight planning) as a tool for dynamic scheduling and disruption management and thus present an *enhanced disruption management approach using flight planning*. For example, a flight experiencing departure delay at its origin can dynamically exploit the additional flexibility of operating at increased speeds (and increased fuel burn), in addition to conventional techniques such as aircraft swaps and cancelations, in order to absorb delays at the flight destination, and decrease costs associated with passenger delays and mis-connections. Further, if connectivity is unaffected, flights may be slowed down to decrease fuel burn, the associated fuel costs and emissions. Our goal is thus, to identify the operational trade-offs between (i) aircraft and passenger delay costs and (ii) fuel burn costs, via flight planning; in order to further decrease costs incurred during airline operations.

4.1.1 The Problem

We briefly describe the problem setting in this section.

We consider scenarios where a flight is delayed at its origin due to a disruption in the network. Our decision time is about a half-hour prior to flight departure, when we are in a position to make operational decisions regarding the choice of flight plans. We consider disruption man-

agement techniques that combine aircraft swaps, flight cancelations and passenger recovery, with flight planning. Through this process, we trade-off network connectivity costs and delay costs associated with flight arrival times, with the fuel costs associated with flight speed changes. In particular, we re-allocate slack in block and ground times in the network, using the following mechanisms:

- by changing aircraft speeds to preserve connections. A disrupted flight may be sped up (or not) to preserve passenger connections if needed, thus decreasing passenger-related costs to the airline but consuming more fuel; or, a flight may be slowed down further to save on fuel if fuel dominates passenger-related costs. Thus slack is generated or absorbed; and

- by changing aircraft speeds, combined with delaying downstream passenger connections. The ability to hold flights adds slack to 'tight' connections if needed, and may not require flights to necessarily be sped up to generate slack.

We discuss this in further detail in §4.1.2.

4.1.2 Opportunities for Integrating Flight Planning and Disruption Management

In this section, we illustrate through an example, the trade-off frontier between fuel burn costs and time-related passenger costs. Then, we present information about the state-of-the-practice in airline operations, obtained from discussions with multiple airlines. We show how the state-of-the-practice falls short in an operational context, and incurs higher costs than the 'optimal' course of action.

4.1.2.1 Flight speed changes

Consider the example from §4.3 of flight a from airport A into hub H. We consider a disrupted schedule in which a is delayed by Δ when departing from A. If the aircraft flies at the same speed that was planned for in the schedule, the flight will also reach the hub Δ units of time later than scheduled. Due to this, passengers are delayed in reaching H and if the connecting time available is less than the minimum connecting time $MinCT$, they can misconnect to subsequent

103

flights in their itineraries. By changing the speed at which the flight is operated, the block time can be decreased and ground time increased or, the block time can be increased and the ground time decreased. Figure 4-1 shows how using alternate flight plans which operate at different speeds can create different amounts of slack in the schedule, which can be exploited to make passenger connections from a to flights b, c, d, and e; which would not have been possible if alternate flight plans were not used.

Original flight plan for flight a

Alternative flight plans for flight a
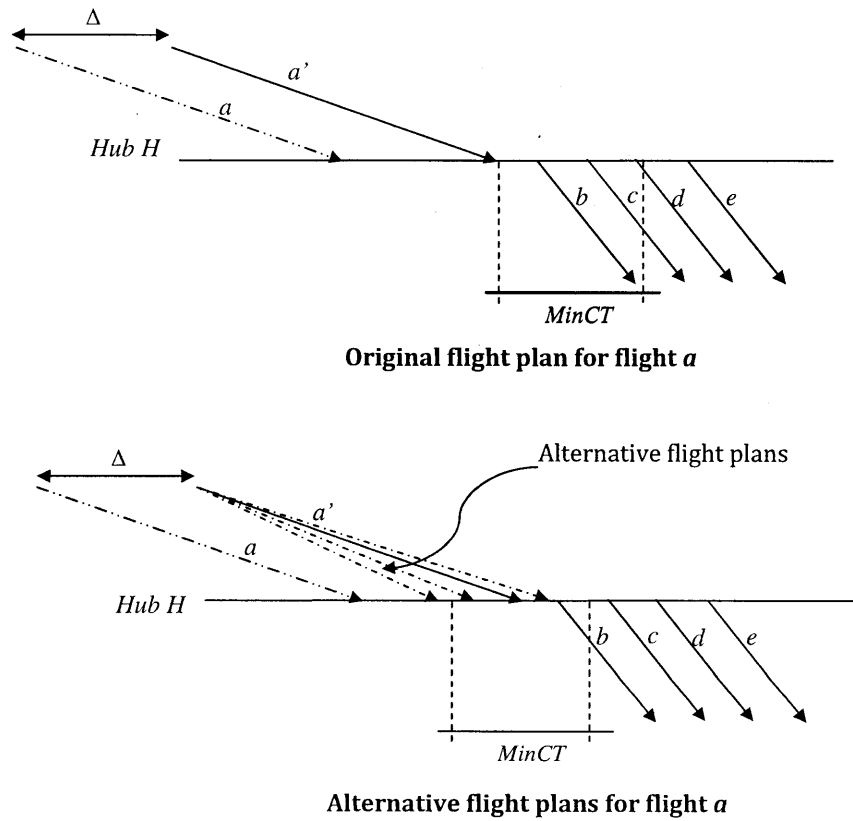
Figure 4-1: Flexibility provided in disruption management by choosing alternate flight plans

Table 4.1 shows the changes in fuel costs of flight a and its corresponding realized passenger costs by operating flight a at different speeds, when Δ is a departure delay of one hour. The flight speeds here are expressed in terms of their *cost index (CI)* values, which can be briefly defined as the amount of additional fuel worth burning (relative to the minimum fuel burn to operate the flight) to save one unit of time. Cost index will be discussed in detail in

104

§4.3. The fuel consumption of the flight and corresponding fuel costs are obtained from the flight plan information, as described in §4.3. Passenger costs are obtained using an airline disruption management simulator [DCT] [Vaa], part of the Integrated Operational Control System developed by Jeppesen[Jepa], and capture the realized costs of passenger delays due to delayed flight arrivals as well as delay costs due to passenger mis-connections and re-accommodation on alternative itineraries. Thus for each value of cost-index, the flying time associated, the corresponding fuel burn and the simulated passenger delay cost are shown. These are summarized in Figure 4-2, which shows the trade-off curves between the flying time and total cost.

As the speed is increased by generating flight plans from cost-index 20 to 1500, the fuel burn increases non-linearly. However, observe that in comparison with the passenger costs, however, the fuel burn curve is quite flat, due to the large number of passengers affected by the disruption.

Cost Index (CI)	Flight Time	Fuel burn ($)	Passenger delay cost ($)	Total Cost ($)
20	455	53772.8	103396.5	157179.3
40	454	53776.1	103337.1	157113.2
60	454	53777.0	103337.1	157114.1
80	454	53838.8	103337.1	157175.9
100	455	53799.8	103396.5	157196.3
300	442	55962.1	102010.0	157972.1
500	431	58401.0	42715.3	101116.4
700	427	60013.0	41308.6	101321.6
900	426	60551.6	41249.2	101800.8
1100	424	61651.3	38361.6	100012.9
1500	423	62942.7	38302.2	101244.9

Table 4.1: Flight time - cost trade-offs associated with different flight plans

Note that if the aircraft operates with the originally planned flight plan, which is at CI 40, many more passengers will be disrupted and need to be re-accommodated, compared to the case of flying at CI 500. At this point, there is a sharp dip in the passenger cost function as several passengers are prevented from misconnecting. Far more fuel is also burned, and this gives rise to a trade-off between these elements of airline operations. Similarly, from CI 900 to CI 1100 again, there is a (much smaller) dip in the passenger delay cost function as some more passenger misconnections are prevented. It turns out that the sum of passenger delay and fuel burn costs
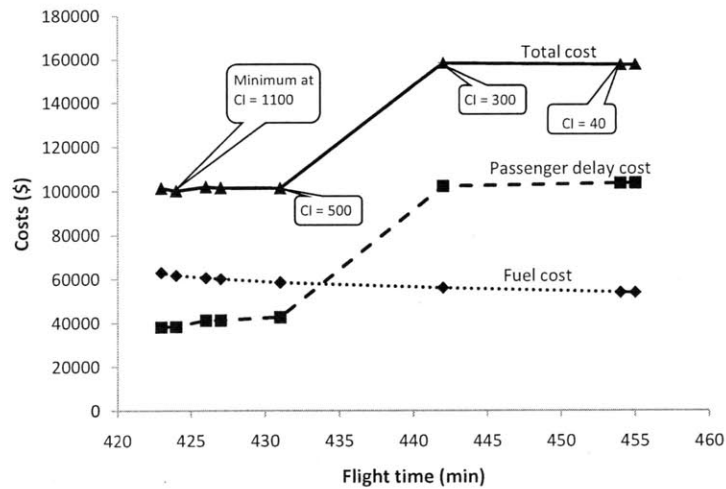
Figure 4-2: Trade-off between flight time and associated costs

is minimized at CI 1100.

This example indicates two important points. First, we observe that it is possible to decrease total costs by flight speed increases, relative to the 'static' flight plan (CI 40) where the airline operates with its planned cost-index. Not only it is possible to speed up flights in case passenger connections become the driver of total costs, but also it is possible to slow down to save on fuel when fuel burn drives the total cost function. This is indicative of the potential of using flight planning as a tool for absorbing delays and complementing traditional disruption management approaches. Second, the actual benefit of flight speed changes as a tool for disruption management depends on the network connectivity of the passenger itineraries from flight a. For a different value of the initial disruption Δ, the trade off between fuel costs and time-related passenger costs can be very different. Both these observations indicate that the choice of CI for each flight is to be made pre-departure, once departure delay is known, taking into account the current state of the system and its network effects relative to the fuel costs.

Traditional disruption management practice does not capture elements of speed changes as a means to add slack during operations. Flight planners also do not capture the network impacts of the schedule during operations [Alta]. Combining these elements, as we have seen, leads to improvement in total cost. In §4.3.2, we discuss the state-of-the-practice of using flight planning during operations at various airlines.

106

4.1.2.2 Flight departure re-timing

Additional flexibility in the schedule can be exploited by re-timing flights. Figure 4-3 illustrates that instead of speeding up flight a to CI 1100 as shown in the example, it may be more cost efficient to speed up a to a lesser extent, while holding flights b, c, d (to which there are possible misconnections from a). The trade-off we make here is between the fuel cost saved in not speeding up a to CI 1100, preventing passenger misconnections between $a - b, a - c, a - d$, and the arrival delay cost of passengers on flights b, c, and d.

Original flight plan for flight a

Alternative flight plans for flights a, b, c, d
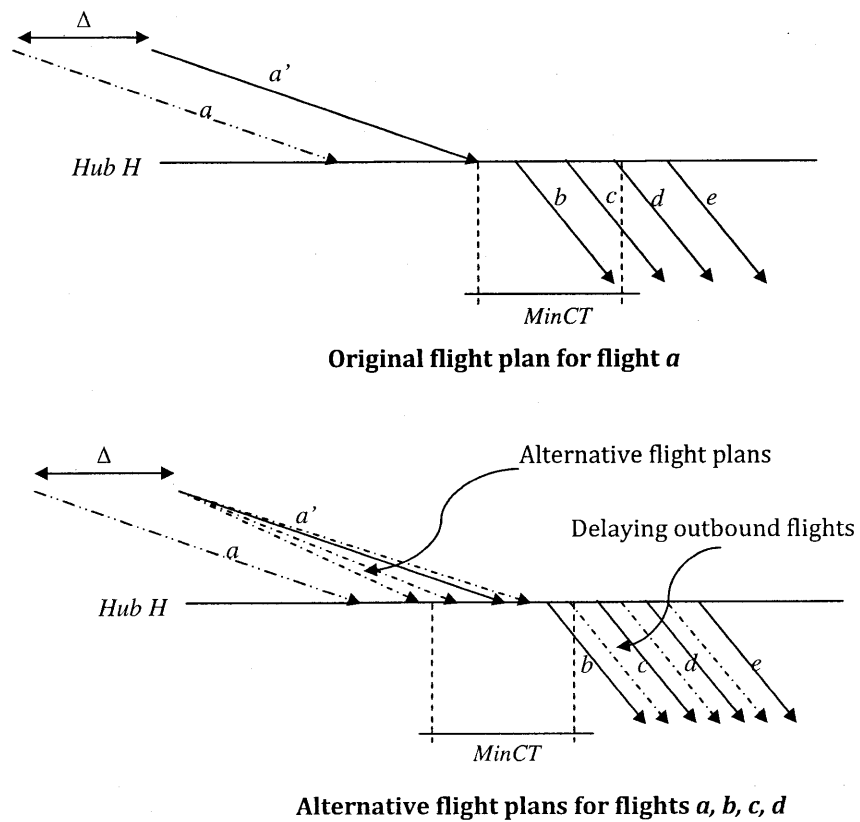
Figure 4-3: Flexibility provided in disruption management by choosing alternate flight plans

4.1.3 Contributions

The contributions of our research are as follows:

- We introduce flight planning as a *enhanced disruption management* tool, and provide optimization models that combine flight planning with traditional disruption management models during operations. In particular, we focus on two aspects of flight planning (i) speed changes; and (ii) flight departure delay, in order to trade off fuel costs and passenger delay costs. Our approach represents an integration of two aspects of airline operations hitherto studied separately, namely, disruption management and flight planning.

- Through dialogue with multiple airlines, we provide an update of the current state-of-the-practice with regards to flight planning approaches. We also discuss the current practice of dynamic scheduling and disruption management. We identify opportunities for improving disruption management and dynamic scheduling using flight planning by integrating both these aspects, and show the need for optimization-based decision support.

- We present models for aircraft and passenger recovery combined with flight planning. Our experiments focus on hub operations and opportunities for improved trade-offs between passenger costs and fuel costs, with the goal of minimizing total realized costs.

- In comparison with conventional disruption management, we demonstrate that our enhanced disruption management strategy helps decrease passenger misconnections by 47.2%-53.3%, and the passenger-related operating costs associated with misconnections, for the airline under consideration. We demonstrate the dynamic nature of the trade-off frontier between passenger costs and fuel burn costs and discuss in detail the interactions involved in this trade-off in different disruption scenarios. We also discuss the relative benefits of the two types of mechanisms studied - that of flight speed changes, and that of delaying flight departures - and show significant synergies in applying the two mechanisms simultaneously.

4.1.4   Organization of the chapter

This chapter is organized as follows. In §4.1.2 we illustrated using an example, opportunities for integrating flight planning and disruption management to minimize costs. We show how system performance can be improved using alternate flight plans with different operating speeds

and different departure times. In §4.2, we present an overview of disruption management practice and mathematical models commonly used. In §4.3, we provide an overview of the flight planning process, provide some background and introduce some terms relevant to this work. We also provide information about the current state-of-the practice of operational flight speed changes in §4.3.2, obtained through discussions with airlines and from published material, and show that they vary considerably from optimal cost practices, indicating a knowledge gap that can be filled using decision-support. In §4.4, we present our modeling architecture to integrate flight planning with dynamic scheduling and disruption management, which we refer to as our *enhanced disruption management* approach. Our models provide a way to trade off passenger delay costs and fuel burn costs. We provide models that capture passenger connectivity exactly, and approximations that are faster to solve and improve passenger connectivity. We describe our experimental setup in §4.5. In §4.6, we present our results and compare them with the current state-of the practice to estimate cost savings to airlines.

4.2 Disruption Management

Airline Operations Control Centers (AOCC) manage airline operations on a day-to-day basis, including (i) management of aircraft, passenger and crew operations, (ii) maintaining operational safety considerations and (iii) interfacing with the FAA and other airlines to exchange critical information including during air traffic flow management initiatives [Bra03].

During operations, operational procedures of dynamic scheduling, routing and disruption management vary among carriers. The first priority in a disrupted situation is to bring operations back on track as soon as possible. For this, the operations controllers re-assign the resources of the airline in order to minimize the costs associated with the disruption. Three types of decisions are made: (i) whether or not a flight is canceled, (ii) departure times of flights that are to be operated, and (iii) which aircraft and crew is assigned to each operated flight. Following aircraft and crew recovery, passenger recovery and re-accommodation is performed typically.

Though the aircraft and crew recovery actions significantly affect passenger costs, these may not be explicitly considered apriori because (i) higher importance is given to on-time performance of flights, (ii) ensuring quick operational recovery can prevent additional passengers

from being disrupted, (iii) passenger delays in themselves do not contribute to loss of operations, as passengers are not system resources, and (iv) an assumption that quick operational recovery may also mean quick passenger recovery.

However, Bratu and Barnhart [BB05] show that flight delays and passenger delays in the system can vary significantly - primarily due to delays incurred by passengers whose flights are canceled and whose itineraries are disrupted. Bratu and Barnhart [BB06] also propose models that apriori consider passenger delays in addition to system operating costs, and can provide solutions that not only have low operating costs but also reduce passenger delays significantly. Due to the difficulty of implementing passenger-centric recovery optimization models, and inherent stochasticity in passenger re-accommodation, airlines typically rely on conventional aircraft and crew recovery models. At AOCCs with more sophisticated systems, large groups of passengers or valuable passengers may be considered apriori in recovery models to facilitate their connectivity. However, passenger-centric recovery models are still a rarity rather than the norm in most resource recovery decisions.

4.2.1 Mathematical Formulation

In this section, we present the recovery models typically adopted at AOCCs.

Given a disrupted schedule, an airline defines a recovery time-window of duration T starting from the current time, beyond which normal operations should be resumed. In current disruption management practice, aircraft swaps, flight delays and flight cancelations are used to recover from the disruption and restore the original schedule.

The recovery model is based on a *time-space network* representation of the airline's schedule. The nodes in a time-space network are associated with both time and location, and an arc between two nodes indicates a possible movement between the two locations (or same location) and times. Given the state of the system at time t, we create time-space networks within the time-window T whose arcs are based on (i) expected departure and arrival times of disrupted (or enroute) flights in the system, and (ii) scheduled departure and arrival times of non-disrupted flights. For the two stages of disruption management, we create two different time-space networks: (i) aircraft flow networks for aircraft recovery; and (ii) a passenger flow network for

passenger recovery, as described below.

The *aircraft flow network* N_a for aircraft a tracks its movement over the flight schedule. Each node in the aircraft flow network represents either a possible departure time of the flight f or a possible arrival time of the flight *plus* the minimum turn time of the aircraft type that is assigned to flight f. Given the airline system state, the nodes and arcs are created to represent the expected departure and arrival times of disrupted and enroute flights, and the scheduled departure and arrival times of non-disrupted flights. The aircraft flow network need only contain those flights in the schedule to which an aircraft of the same fleet type as a is assigned, because aircraft a can only operate those flights to which the same aircraft type is assigned. We refer to the arcs that represent scheduled flights and their copies as flight arcs. For each aircraft in the recovery time window, we create copies of each flight f every r minutes (with the same block time as f) until a maximum departure delay of R minutes that represent possible flight departure. Ground arcs connecting successive nodes (in time) at the same location are added to this network to represent the aircraft remaining at the same place over time.

Let F be the set of flight legs f operated by the airline and A the set of aircraft a available. C_f is the set of copies of flight f, where the copies are generated by alternative possible departure times of flight f. In the aircraft flow network, each flight leg copy $k \in C_f$ connects a possible departure time of flight f to a possible arrival time (corresponding to a specific flight plan) plus the minimum turn time of aircraft a. G_a is defined as the set of ground arcs in the aircraft flow network for individual aircraft a and N_a' the set of nodes in the aircraft flow network for aircraft a. s_a^n is the supply of aircraft a at node n in the aircraft flow network (a demand is specified as a negative supply). For each individual aircraft, a supply $s_a^n = 1$ is associated with the node where the aircraft is known to start at the beginning of the time window T, and a demand of $s_a^n = -1$ where it finishes the last flight of the time window. Ground arcs at each location connect the successive nodes, which allows feasible aircraft paths to be defined.

Let x_f^k be a binary variable that takes on value 1 if copy k of flight leg f is present in the solution and 0 otherwise, y_g be a binary variable that is 1 if ground arc g is present in the solution and 0 otherwise and z_f be a binary variable that is is 1 if flight f is canceled in the solution and 0 otherwise.

The two elements of cost typically captured in aircraft recovery are the delay costs and the

cancelation costs. With each flight copy $k \in C_f$ for each flight f is associated a cost c_f^k. c_f^k is obtained by multiplying the passenger-related cost incurred by the airline per passenger per minute, the number of passengers booked on flight f, and the number of minutes that copy $k \in C_f$ is delayed beyond the scheduled departure time of f. A cost c_f is associated with the cancelation of flight f.

Aircraft flow balance constraints:

In the aircraft flow network N_a for aircraft a, a supply $s_a^n = 1$ is associated with the node where the aircraft is known to start at the beginning of the time window T, and a demand of $s_a^n = -1$ where it finishes the last flight of the time window. s_a^n is derived from our knowledge of the airline system state. n^- is the set of incoming arcs to node $n \in N_a'$ and n^+ is the set of outgoing arcs to node $n \in N_a'$. The aircraft flow balance constraints are then:

$$\sum_{g \in n^-} y_g + \sum_{(f,k) \in n^-} x_f^k + s_a^n = \sum_{g \in n^+} y_g + \sum_{(f,k) \in n^+} x_f^k \qquad \forall\, n \in N_a', \forall\, a \in A \qquad (4.1)$$

Flight coverage constraints:

A flight is either operated using one of the copies created or canceled. The flight coverage constraints are thus:

$$\sum_{k \in C_f} x_f^k + z_f = 1 \qquad\qquad \forall\, f \in F \qquad (4.2)$$

We briefly mention the ability to model scheduled aircraft maintenance. Aircraft maintenance is modeled by creating an artificial 'flight leg' beginning at the start of maintenance at the maintenance station and ending at the end of the scheduled maintenance at the same station. If maintenance can be delayed, we capture it by creating copies of this arc. By modeling 'flight cover' of the maintenance leg using the constraint shown above, with the corresponding z variable set to 0 to disallow cancelation of maintenance, we ensure that compulsory maintenance is carried out. If maintenance is not compulsory within the recovery time window, these constraints are omitted. The aircraft recovery model is thus the following.

$$\min \sum_{f \in F} \sum_{k \in C_f} c_f^k x_f^k + \sum_{f \in F} c_f z_f \qquad (4.3)$$

$$\text{s.t.} \sum_{k \in C_f} x_f^k + z_f = 1 \qquad \qquad \forall f \in F \qquad (4.4)$$

$$\sum_{g \in n^-} y_g + \sum_{(f,k) \in n^-} x_f^k + s_a^n = \sum_{g \in n^+} y_g + \sum_{(f,k) \in n^+} x_f^k \qquad \forall n \in N_a', \forall a \in A \qquad (4.5)$$

$$x_f^k \in \{0,1\} \qquad \qquad \forall k \in C_f, \forall f \in F \qquad (4.6)$$

$$y_g \geq 0 \qquad \qquad \forall g \in G_a \qquad (4.7)$$

After solving the aircraft recovery model (4.3) - (4.7), we construct *passenger flow networks* N_p for each itinerary p to facilitate passenger recovery. The flight legs in the passenger flow network are those which are present in the optimal solution to (4.3) - (4.7), that is, with $x_f^k = 1$. For each flight f, the flight leg in N_p represents the scheduled departure of flight f and the arrival of flight f plus the minimum connecting time for a passenger on itinerary p. Connection arcs at each location connect successive flight legs in a passenger itinerary. Note that some of these arcs can actually be *reverse* arcs, with the time of arrival earlier than that of departure. This arc, if present in the solution, indicates that the passenger is disrupted and cannot make the connection. For simplicity, we will assume that all connecting passenger itineraries consist of two legs. It is straightforward to extend the model to the case with multiple flight legs.

Let P be the set of passenger itineraries p that are operated on the network and n_p the number of passengers on itinerary p, N_p' the set of nodes in passenger flow network for itinerary p, and G_p the set of ground arcs in the passenger flow network for itinerary p. Passenger recovery constraints may be modeled as in Bratu and Barnhart's PDM1 model [BB06]. First, we generate candidate itineraries $R(p)$ for each passenger type p. If passenger itinerary p is not disrupted at time t, $R(p)$ consists only of the originally scheduled itinerary. If passenger itinerary p is disrupted at time t, $R(p)$ is a list of candidate itineraries or paths on the passenger flow network from the itinerary origin to its destination, with each starting after the original itinerary for p. $R(p)$ also includes a virtual itinerary to indicate re-accommodation to another airline's network, or, cancelation of the passenger trip at its origin. Decision variables ρ_p^r indicate the number of

passengers originally on itinerary p who are re-accommodated on itinerary r and d_p^r represents the arrival delay of these passengers. ρ_p^p is the number of passengers on the non-disrupted itinerary p. Parameters Cap_f is the number of seats on flight f and parameter δ_f^r is 1 if flight f is on itinerary r and zero otherwise. Passenger-related costs c_p^r denote the cost of using itinerary r to accommodate passenger p. This is based on the actual arrival time of itinerary $r \in R(p)$ at the destination, and includes delay costs, goodwill costs, and hotel and meal costs if relevant.

The following constraint ensures that all passengers are flown to their destinations.

$$\sum_{r \in R(p)} \rho_p^r = n_p \qquad\qquad \forall\, p \in P \qquad\qquad (4.8)$$

To ensure that no passengers are assigned to a canceled flight leg, and to restrict the number of passengers assigned to a flight leg if the flight is not canceled, we add the following constraint:

$$\sum_{p \in P} \sum_{r \in R(p)} \delta_f^r \rho_p^r \leq Cap_f(1 - z_f) \qquad\qquad \forall\, f \in F \qquad\qquad (4.9)$$

The passenger recovery model is as follows:

$$\min \sum_{p \in P} c_p^r \rho_p^r \qquad\qquad\qquad (4.10)$$

$$\text{s.t.} \sum_{r \in R(p)} \rho_p^r = n_p \qquad\qquad \forall\, p \in P \qquad (4.11)$$

$$\sum_{p \in P} \sum_{r \in R(p)} \delta_f^r \rho_p^r \leq Cap_f(1 - z_f) \qquad\qquad \forall\, f \in F \qquad (4.12)$$

$$\rho_p^r \in \mathbb{Z}^+ \qquad\qquad \forall\, r \in R(p), \forall\, p \in P \qquad (4.13)$$

114

4.3   Flight Planning

A flight plan is a document prepared by the operator (usually an airline) that indicates the movement of the concerned aircraft in time and space, from its origin to its destination. The flight plan specifies the route (ground track) of the aircraft, its profile (altitudes along the route), its speed (which varies along the route) and the fuel burned in operating the flight plan. An example of a flight plan is shown in Figure 4-4 (Source: Altus [Altb]).
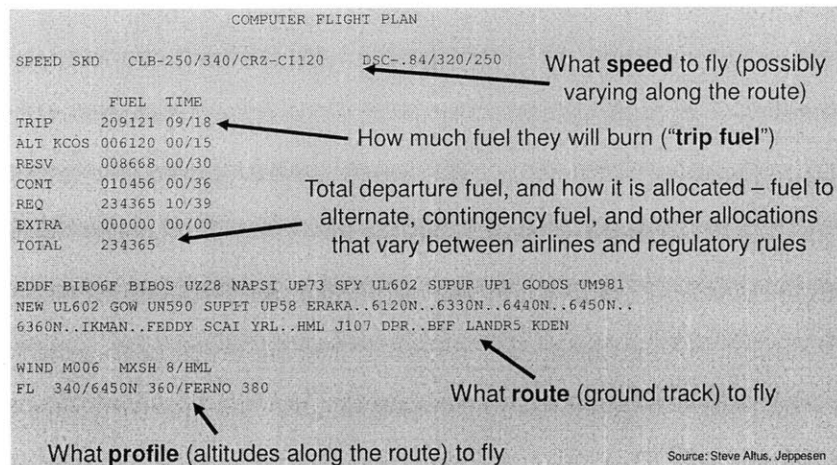
Figure 4-4: Sample flight plan

The goal of a flight plan is to minimize a weighted sum of fuel cost, time-based costs, overflight costs (in countries other than the US) and passenger spill costs subject to constraints of aircraft performance, weather, allowed route and altitude structure, schedule and operations; by varying the route, profile, speed and departure fuel. The generation of flight plans is a complex non-linear optimization problem, due to the following considerations: (i) fuel burn and cruise fuel flow rate are nonlinear functions of the aircraft weight, which varies dramatically during the course of a flight, (ii) weather uncertainty, (iii) 'allowed' ground tracks and altitudes should be used (static and dynamic airway structures exist, and Route Availability Document [EUR06] has to be followed), (iv) optimal paths vary significantly due to current wind conditions, (v) Computational tractability issues lead to decomposition and other heuristics solve the flight plan optimization problem [Altb], (vi) Computational performance trade-offs, such as that between plan robustness and solution time.

Flight planning algorithms also include some fuel and time buffer for expected enroute con-

gestion, using weather/wind forecasts and historical sector-based air traffic delay information. Based on this calculation, a flight plan also includes buffer for emergency fuel (specified by the FAA or by EUROCONTROL), contingency fuel specified by the airline that accounts for enroute congestion along the flight path, and discretionary fuel.

Flight plans are created typically on a flight-by-flight basis by the airline, and must be filed with Air Traffic Control (ATC) before departure. Typically this may be up to 30 minutes before the flight departs from its origin airport. As we discuss in §4.3.2, flight planning models today do not specifically compute flight plans that account for network connectivity of aircraft and passengers on the day-of-operations.

The relationship between fuel burn and flight time (and consequently, block time) is highly non-linear. Figure 4-5 illustrates the relationship between flying time and fuel burn for flight a from airport A to the hub H. For this flight, the duration of the flying time can be varied between 455 min to 420 min by varying the speed. (Further slowing down beyond 455 minutes begins to again increase fuel burn, and therefore airlines typically vary speeds within the range shown in Figure 4-5.) Flight a is a long-haul flight, with duration greater than 6 hours. Long-haul flights are flights that involve long distances, typically more than six hours in length, and are usually operated by wide-body aircraft [Tho09] [Air10]. For such flights, the flexibility in time provided by changing flying speeds is high in comparison with flights that are short-haul. Short-haul flights are typically those that are less than 3 hours in length [Tho09]. For short-haul flights, changing flight speed results in a smaller range of variation in the flying time. Consequently, the difference in fuel burn produced due to operating different flight plans at different speeds is far less, though the shape of the fuel burn curve remains similar. For short haul flights, the change in block times resulting from speed changes is seen to be almost negligible.

Each point in the plot represents a flight plan, with an associated fuel burn and flight time. Fuel burn gives rise to fuel-related costs, while flight time relates to block time of the flight and time-related costs. Typically, a flight plan is selected from this curve by specifying one of the following two metrics: (i) static aircraft cruise speed, or (ii) a more sophisticated measure called the *Cost Index*. Compared to the simplistic measure of specifying the aircraft speed, the cost-index measure takes into consideration the time-value of speeding up or slowing down the aircraft, as described in §4.3.1.

116
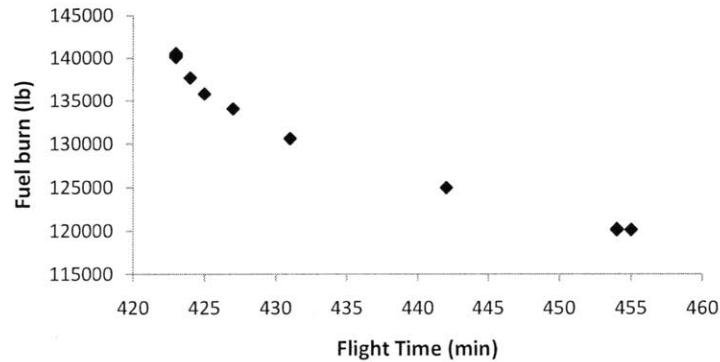
Figure 4-5: Relationship between flight time and fuel burn

4.3.1 Cost Index-based flight planning

Cost Index (CI) is an assumed ratio of the time-related (or delay-related) costs of a flight divided by the fuel cost; that is, it is the ratio of cost per unit time divided by the cost per mass unit fuel. Time-related costs are defined as those that are related to (i) the duration of the flight, such as aircraft maintenance costs and crew duty costs; and (ii) the arrival time of the flight, such as aircraft connectivity, crew connectivity and elapsed time, and passenger connection and delay costs. CI is expressed in units of 100lb/hr (Boeing) or kg/min(Airbus). This measure has two physical meanings: (i) CI is the amount of additional fuel worth burning (relative to the minimum fuel burn to operate the flight) to save one unit of time, or (ii) CI is a way to express time units using the same metric as fuel flow units, allowing us to optimize for a sum of fuel and time-based costs.

Typically, the 'right' CI value to operate at is determined by the airline from its historical data. The airline computes the delay costs to passengers, crew and aircraft, as well as fuel costs incurred, from historical data. These delay costs and fuel burn costs are aggregated over the network or over each fleet type or each market type, and divided, to generate the corresponding cost index values. A CI value of zero means that relative to fuel costs, time-related costs are zero; or the additional fuel worth burning to save one unit of time relative to the minimum fuel burn speed is zero. Then the aircraft is operated at its most fuel-efficient cruise speed, called the *maximum range cruise speed*. When operating at a high CI, time is more costly compared to fuel, and the aircraft is sped up even though more fuel is burnt, in order to incur lower time-

117

related delay costs. This is with an objective of minimizing the sum total of fuel and time-related costs.

Airlines typically associate their flights with the historically derived cost index and schedule their flights under the assumption of operating under that CI and speed. This CI is called the 'normal CI'. To the estimated flying time, additional buffer for taxi times, transit times, delay buffer, etc. are added to find the schedule for the flight. Compared to flying by simply determining a speed, this is a better and more systematic measure that accounts for time-related costs.

A key observation with regard to the usage of cost-index is the following. When operating at a specified cost-index, the underlying assumption is that of linear increase of time-related costs. That is, if the flight were to be delayed by one minute, this measure assumes that a cost of *CI*fuel cost* per minute would be incurred. However, it has been well-studied that time-related costs in airline networks are highly non-linear. In reality, time-related costs are related to aircraft, passenger and crew connectivity, and thus cannot be simplistically expressed as a linear function of the CI values. In the following sections, we present and solve models that capture the relationships between fuel costs and time-related costs in a more accurate manner.

4.3.2   Flight planning: state-of-the-practice

In this section, we discuss the current state-of-the-practice involving operational flight planning at various airlines. This information was obtained from six international carriers. To obfuscate the specific information of each carrier, we simply refer to them as carrier 1, 2, 3, 4, 5 and 6.

It is well-understood in the airline industry that associated with flight speed changes are both fuel impacts as well as network connectivity impacts. For this reason, the cost index (CI) measure, described in §4.3, which is an indicator of this trade-off, has become standard practice in the industry, and is used as a rule-of-thumb.

Each carrier determines its operational CI value by analyzing its historical operations, and aggregating the total realized cost of fuel and the realized cost of time-related effects (delays, connectivity, etc.). This may be done at a network, fleet or market level, resulting in network CI, or fleet CI, or market CI. Calculations of the CI value from historical data are typically done

at great cost, using dedicated software [Alta]. Using this *static* CI value, a rule-of-thumb for the cost of delays is used by dispatchers when calculating the flight plans and by pilots when flying the aircraft to determine the speed at which to fly in order to optimize the time and fuel costs trade-off.

In addition to the static CI value, which indicates normal operation of the flight according to the schedule, airlines also specify a range of CI values that an aircraft can be operated at when speeding up or slowing down may be required. This range of CI serves as a guideline to dispatchers and pilots. The pilot is allowed to speed up or slow down within this range at his or her discretion. (The max CI in the range does not mean that further speed up is not possible, only that this range specifies an allowable limit that can be operated at the pilot's discretion.) The lower limit of the CI range is 0. The higher limit of the CI range is typically determined based on a percentage cap on excess fuel burnt beyond the 'normal' CI, which can differ from carrier to carrier. One reason for the specification of the higher limit is due to the fuel tankering policies of the airline, which would allow speed up to a certain extent with the discretionary fuel carried. (Flying even faster may result in burning emergency fuel, which should be done only in emergency situations.) The increasing marginal cost of fuel burn per minute of time saved in flying is another. Yet another reason for this rule-of-thumb may also be that airlines would like to prevent pilots from 'flying too fast' to reach their destination earlier, and result in high fuel consumption.

Airlines also provide limited guidelines on operating at different CI values than the 'normal' CI value. They indicate that if tailwinds are encountered or if the aircraft has an early start, the pilot may slow down to a lower CI value, and in case of headwind or late start, he or she may speed up to the higher CI value. These guidelines also caution pilots that speeding up the flight will consume excess fuel, and the pilot should assess whether the excess fuel burn is outweighed by the benefits of making up the corresponding time.

To understand further the practice of flight planning, we held discussions with multiple international carriers. The information gleaned from these discussions, as well as from published material by airlines, is summarized as follows.

Airline 1 is an international airline that operates across three continents. This airline operates a hub-and-spoke network centered around a single large hub. It operates a significant

119

number of long-haul flights that represent 10% of departures, and also several short-haul flights that are operated as short cycles around the hub. The normal CI for its schedule is CI 30. In case of delay, or headwind, the pilot may speed up to CI 300 to make up for lost time. However, the 'correct' CI to operate at is not chosen considering system state at the time of filing the flight plan. Airline A recognizes that the CI guidelines may not be sufficient, and is incorporating network connectivity concerns of passengers and crews at a preliminary level (without decision-support models).

Airline 2 is also an international airline that operates a large number of long-haul flights across continents. This airline has been exploring aspects of flight speed changes into operations, primarily slowing down of aircraft, to save on fuel. At this point, Airline B is interested in conducting studies to explore the integration of flight planning and disruption management in its network. These measures are not however, being tested out on the network.

Airline 3 operates a large hub-and-spoke network with multiple hubs. It operates a number of intercontinental and intra-continental long-haul flights. The CI range for this airline is 0-500. The airline issues directives to pilots to slow down to save fuel if extra buffer is present in the schedule and has cautioned against speeding up unless significant benefits of the time are observed.

Airlines 4 and 5 also are hub-and-spoke carriers, each with a single hub. Almost all of the flights operated by these carriers are short-haul, with an extremely small percentage of long-haul flights. Airline 4 operates its flight based on fleet-type-based CI values. The normal CI values for different fleet types used are 60, 80 and 90. Due to the low flexibility in time provided by short-haul operations, these airlines do not consider CI changes in their operations.

Airline 6 has recognized that a static CI for the entire network does not capture the time-related costs effectively, and hence, has been studying the use of market-based CI [SMW]. A market-based CI specification better captures the network structure related to the market and hence, better approximates the associated time-value compared to a network-wide or fleet-wide measure. However, this value is again pre-determined for the market and not during operations.

Another recent trend has been observed in the industry during the recent fuel price spike in 2007. Operational speed changes were highlighted during this fuel crisis, as a number of airlines began to exploit the slack in their schedule to save on fuel. Associated Press articles

[Ass08a] and [Ass08b] report that airlines slowed down flights, resulting in longer flying times but lower fuel burn. Airlines reported about $20 million savings in a year by practicing this policy.

Our work has implications for fuel tankering policies. When flight plans with higher CI are indicated by the optimization model (as in the cas of the example discussed in §4.1.2), the fuel to be carried by the aircraft can be significantly higher than that specified by the standard fuel tankering policy of the airline.

4.3.3 Flight Planning Engine

Flight plans used in this work are generated using JetPlan [Jepb], a flight planning tool developed by Jeppesen Commercial and Military Aviation [Jepa]. Jeppesen's flight planning engine uses information about the flight, weather patterns, allowable fleet type(s) for the flight, payload during the day of interest and generates flight plans at different CI and departure times for each fleet type. The flight plan generator takes into account the fuel burn due to the payload consisting of cargo, passengers, luggage hold, and fuel weight. Included in fuel are contingency and discrepancy fuel.

The actual fuel cost incurred depends on the airport at which the plane is re-fueled. In our computations, we assume standard costs throughout the network based on information from the airline. In practice, an airline will compute the trade-offs between fuel burn costs and delay costs using their actual fuel prices. Fuel hedging is another aspect that could change the fuel costs incorporated in the model if data were available.

4.3.4 Concerns related to state-of-the-practice

While flight planning using static CI values and existing industry guidelines captures the understanding that there is a trade-off between fuel burn and time-related costs, our analysis in §4.1.2 indicates that capturing dynamically changing network connectivity effects is crucial to minimize true costs of fuel burn and (time-related) delay costs. In the example discussed, Figure 4-2 shows that the choice of the minimum cost is at CI 1100. However, Airline 1 (from whose network this is extracted) typically operates at CI 30 and allows its dispatchers and pilots

121

to speed up to a maximum of CI 300. It is clear from the figure that neither of these CI values truly minimizes costs. While speeding up the flight from CI 30 to CI 300 may be viewed by the pilot as 'making up time', in fact it simply burns excess fuel and increases total cost. Most airlines do not have systematic guidelines or decision-support systems for choosing the right CI dynamically for a given scenario. There has been growing understanding of the shortcomings of current practice, as discussed in Burrows [BBT+01] and Altus [Alta], but models to overcome these limitations have not been built.

We present in the following section a framework that allows us to optimize operating fuel costs and performance costs as measured by passenger service reliability, and minimize total costs incurred.

4.4   Integrated Disruption Management and Flight Planning

In our disruption management approach that incorporates flight planning, schedule and flight plan optimization is performed prior to each flight, at the time that the flight plan is filed for the flight. This provides the ability to produce different flight planning solutions during operations, that are designed to the capture the features of aircraft and passenger connectivity for that flight given current schedules, and further network effects that propagate down the network. In our model, we will focus on aircraft and passenger disruption management. With suitable modifications, this can be extended to crew disruption management as described in Bratu [BB06].

Given a disrupted schedule, an airline defines a recovery time-window of duration T starting from the current time, beyond which normal operations should be resumed. In current disruption management practice, aircraft swaps, flight delays and flight cancelations are used to recover from the disruption and restore the original schedule. In addition, in this recovery time window, at the time of filing the flight plan for each flight, we incorporate the ability to include flight planning as a recovery mechanism to improve passenger service reliability and aircraft recoverability.

At a given time t of the day of operations, we assume that we have a snapshot of the airline's schedule and resource allocations at that point in time. That is, we assume knowledge of (i) where aircraft are located, (ii) passenger itineraries (and therefore disrupted itineraries). We

122

refer to this information as the *airline system state* at time t.

The flights within the recovery time-window can be divided, in a temporal fashion, into 3 types:

1. flights that are already in the air/arrived at destination at this time,

2. flight(s) departing in the next 30 minutes to half-hour, whose flight plans have to be filed, and

3. flights departing more than 30 minutes from the current time .

Knowing the airline system state, we first determine aircraft ready times for all aircraft, and based on this, we create the time-space network representations for aircraft and passenger movements. Second, we determine the set of flights to which the disruption may propagate via aircraft or passengers. This we call the *propagation boundary*. Third, we generate flight copies for flights in T. Flight copies represent re-scheduling of existing flights as well as creation of speed change opportunities for the flight. Fourth, we solve the combined aircraft and passenger recovery model, which provides a schedule that minimizes the sum of passenger delay and fuel burn costs.

The above sequence of steps presents a brief description of our enhanced disruption management process. Our model includes simplifications including not capturing aircraft maintenance and crew connectivity. Bratu and Barnhart [BB06] present a comprehensive model that incorporates these aspects but does not include flight planning opportunities. Our model can be expanded to capture these additional aspects as described in Bratu and Barnhart; however, this is not within the scope of this thesis.

4.4.1   Network representation

Our model is based on a *time-space network* representation of the airline's schedule. The nodes in a time-space network are associated with both time and location, and an arc between two nodes indicates a possible movement between the two locations (or same location) and times. Given the state of the system at time t, we create time-space networks within the time-window T whose arcs are based on (i) expected departure and arrival times of disrupted (or enroute)

flights in the system, and (ii) scheduled departure and arrival times of non-disrupted flights. In fact, we create two different types of time-space networks: (i) an aircraft flow network for each aircraft and (ii) a passenger flow network for the passengers, as described below.

The *aircraft flow network* N_a for aircraft a tracks its movement over the flight schedule. Each node in the aircraft flow network represents either a possible departure time of the flight f, or a possible arrival time of the flight *plus* the minimum turn time of the aircraft type that is assigned to flight f. Given the airline system state, the nodes and arcs are created to represent the possible departure and arrival times of disrupted and enroute flights, and the scheduled departure and arrival times of non-disrupted flights. Because we allow for aircraft swaps within a fleet type, the aircraft flow network N_a need only contain those flights in the schedule to which an aircraft of the same fleet type as a is assigned. We refer to the arcs that represent scheduled flights and their copies as flight arcs. The *passenger flow network* is similarly constructed. Each node in the passenger flow network represents either a (scheduled or possible) departure of flight f or an arrival of flight f plus the minimum connecting time for a passenger on that itinerary. Connection arcs at each location connect successive flight legs in a passenger itinerary. Note that some of these arcs can actually be *reverse* arcs, with the time of arrival earlier than that of departure. This arc, if present in the solution, indicates that the passenger is disrupted and cannot make the connection. For simplicity, we will assume that all connecting passenger itineraries consist of two legs. It is straightforward to extend the model to the case with multiple flight legs.

4.4.2 Flight copy creation

Flight copy creation in our model is of three kinds: (i) to represent alternative departure and arrival times of the flight without speed changes compared to the original, (ii) to represent flight plans that involve speed changes, and therefore, block time changes compared to the original flight without changing its departure time, and (iii) a combination of (i) and (ii). These copies are created in both the aircraft flow networks and the passenger flow network. In order to create flight copies, we first define the *propagation boundary*.

4.4.2.1 Copies in the propagation boundary

Because determining optimal flight plans can be done only for those (yet to depart) flights whose departure times are known with a high degree of certainty, we build multiple flight plans for a subset of flights called the *propagation boundary*, within the recovery time window. The propagation boundary includes the flights which might experience time-related delay costs as a result of the initial disruption. For a flight f experiencing a disruption that causes it to be delayed at departure by t minutes, and for which we want to determine an optimal flight plan, we define its propagation boundary. The propagation boundary of f is the subset of flights to which the disruption propagates to aircraft and passengers; when allowed to propagate downstream *without* any recovery measures.

We illustrate this concept using an example. Consider Figure 4-6, where long-haul flights f_1 and f_2 are operated by one aircraft and short-haul flights f_3 and f_4 by another. Passengers connect from flight f_1 to flight f_3. In the scenario shown, f_1 is delayed to f_1'. This disruption, if unchecked, can propagate via insufficient passenger connection time to flight f_3. Because the slack between f_1' and f_2 is greater than the minimum aircraft turn time, the disruption does not propagate to f_2. However, because the disruption propagates to f_3, and due to short turn time between f_3 and f_4, any action taken towards re-scheduling f_3 will propagate to f_4. Thus, the propagation boundary consists of f_1', f_3 and f_4.
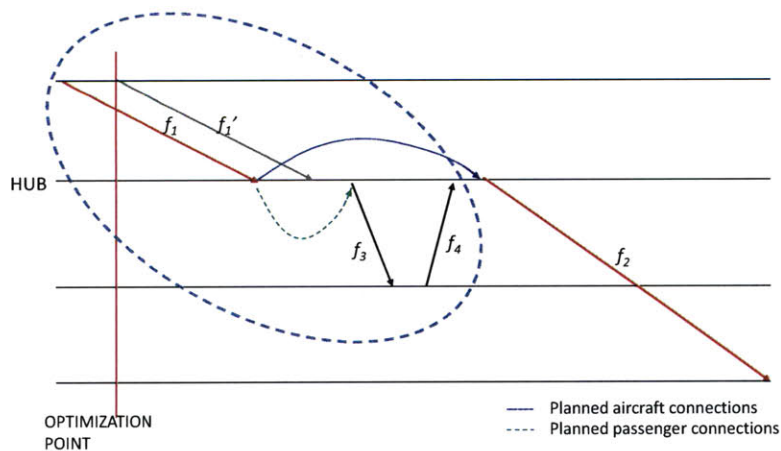
Figure 4-6: Propagation Boundary

125

Once the propagation boundary is defined, we create alternate flight plans for each flight f within the boundary using the following steps:

1. If disruption propagates to f via aircraft connections:

   - Create copies of flight f every r minutes until a maximum departure delay of R minutes. These are copies of type (i) described above.

   - If f is long-haul, as defined in §4.3, create copies also representing change in block times. These are copies of types (ii) and (iii).

2. If disruption propagates to f via passenger connections, that is, some passengers cannot connect to f due to disruptions upstream, we do the following. First, we define Θ (in minutes) as the maximum extent to which downstream flight departures are allowed to be held or delayed in order to facilitate passenger connections. That is, downstream connections f are allowed to arrive at most Θ minutes late at their destination. If $\Theta = 0$, it means that we do not allow for downstream flights to incur departure delays due to upstream passengers, and the copy generation algorithm is complete. If $\Theta > 0$, we follow these steps. The goal of the following is to create flight copies that allow for a flight to depart later, but to arrive no later than Θ minutes at its destination.

   - If f is a long-haul flight, we will capture possible speed changes as well as delaying departure times for this flight. Let the maximum decrease in block time possible due to speeding up f be δ. Create possible departure nodes of flight f every 5 minutes until a maximum departure delay of $\Theta + \delta$. Now create corresponding flight copies of types (ii) and (iii) associated with these departure times, such that the arrival time of the copy at its destination always entails an arrival delay of no more then Θ minutes.

   - If f is a short-haul flight, as defined in §4.3, we do not capture the possibility of speed changes because of the limited impact of changing the speed of the flight. We simply make copies of the flight with its scheduled block time and departure arcs at at 5 minute intervals until a maximum departure delay (and a corresponding arrival delay) of Θ.

In our experiments, we choose values of Θ to be 0, 10 and 15 minutes. We impose a limit of 15 minutes on Θ so that arrival delay of a downstream flight f due to delay propagated to it from an upstream flight via passenger connections is limited to 15 minutes. This is so that the on-time performance of the system is not deteriorated. As discussed in §4.2, it is not a common practice for airlines to hold their downstream flights and risk poor on-time performance.

4.4.2.2 Copies beyond the propagation boundary

In addition, for each flight f in time-window T that is operated by an aircraft of the same fleet type as one of the disrupted flights, we generate copies of the flight every m minutes until a specified number of minutes after its departure time. This allows for aircraft swaps within the same fleet type in order to recover the schedule. These copies belong to type (i) specified above, and represent later departure times of flights. Note that we do not create copies that allow a flight to depart before its scheduled departure time.

Airline scheduling recovery models employ extensively the approach of generating flight copies of type (i) to find good recovered schedules. Thengvall et. al [TYB00], Andersson and Varbrand [AV04], and Yan and Young [YY96] generate flight copies every m minutes for each flight f. We adopt this method in our model. We observe that some of these copies might not be useful, for example, in the case of a flight whose aircraft is ready to depart and passengers have not been disrupted. For this reason, Bratu and Barnhart[BB06] provide a flight copy generation algorithm that limits the number of flight copies by considering resource availability apriori. Though we do not employ this algorithm in the model described in this work for reasons of simplicity, we can adopt an alternative method of generating flight copies without changing the structure of our model.

4.4.3 Definitions

In this section we describe additional details of aircraft and passenger network construction and define additional notation for our models.

Let F be the set of flight legs f operated by the airline and A the set of aircraft a available. C_f is the set of copies of flight f, where the copies are generated from alternate flight plans with

the same departure time, or by alternative departure times or both. This set of flight plans can also contain alternate flight plans using new equipment types if such swaps are allowed. Let c_f^k denote the fuel cost of operating copy k of flight f. G_a is defined as the set of ground arcs in the aircraft flow network for individual aircraft a and N_a' the set of nodes in the aircraft flow network for aircraft a. s_a^n is the supply of aircraft a at node n in the aircraft flow network (a demand is specified as a negative supply).

For each flight leg f, we denote the set of flight copies as C_f. In the aircraft flow network, each flight leg copy $k \in C_f$ connects a possible departure time of flight f to a possible arrival time (corresponding to a specific flight plan) plus the minimum turn time of aircraft a. For each individual aircraft, a supply $s_a^n = 1$ is associated with the node where the aircraft is known to start at the beginning of the time window T, and a demand of $s_a^n = -1$ where it finishes the last flight of the time window. Ground arcs at each location connect the successive nodes, which allows feasible aircraft paths to be defined.

Let P be the set of passenger itineraries p that are operated on the network and n_p the number of passengers on itinerary p, N_p' the set of nodes in passenger flow network for itinerary p, and G_p the set of ground arcs in the passenger flow network for itinerary p.

4.4.4 Assumptions

The following assumptions are considered when building our models:

1. A flight cannot be cleared for departure prior to its scheduled departure time;

2. The decrease in payload (and hence the decrease in fuel burn) due to passengers missing the flight is negligible;

3. If a flight plan with a significantly different arrival time at the destination airport is used, there is a landing slot available at that time.

4.4.5 Aircraft and Passenger Recovery Model

We propose models to minimize the sum of operating costs and disrupted passenger costs. Let x_f^k be a binary variable that takes on value 1 if copy k of flight leg f is present in the solution

and 0 otherwise, y_g be a binary variable that is 1 if ground arc g is present in the solution and 0 otherwise and z_f be a binary variable that is is 1 if flight f is canceled in the solution and 0 otherwise.

4.4.5.1 Resource Allocation Constraints

Aircraft flow balance constraints:

In the aircraft flow network N_a for aircraft a, a supply $s_a^n = 1$ is associated with the node where the aircraft is known to start at the beginning of the time window T, and a demand of $s_a^n = -1$ where it finishes the last flight of the time window. s_a^n is derived from our knowledge of the airline system state. n^- is the set of incoming arcs to node $n \in N_a'$ and n^+ is the set of outgoing arcs to node $n \in N_a'$. The aircraft flow balance constraints are then:

$$\sum_{g \in n^-} y_g + \sum_{(f,k) \in n^-} x_f^k + s_a^n = \sum_{g \in n^+} y_g + \sum_{(f,k) \in n^+} x_f^k \qquad \forall\, n \in N_a', \forall\, a \in A \qquad (4.14)$$

Flight coverage constraints:

A flight is either operated using one of the copies created or canceled. The flight coverage constraints are thus:

$$\sum_{k \in C_f} x_f^k + z_f = 1 \qquad\qquad \forall\, f \in F \qquad (4.15)$$

Scheduled aircraft maintenance may also be modeled as described in §4.2.

4.4.5.2 Passenger Constraints

These constraints are modeled similar to Bratu and Barnhart's PDM1 model [BB06]. First, we generate candidate itineraries $R(p)$ for each passenger type p. If passenger itinerary p is not disrupted at time t, $R(p)$ consists only of the originally scheduled itinerary. If passenger itinerary p is disrupted at time t, $R(p)$ is a list of candidate itineraries or paths on the passenger flow network from the itinerary origin to its destination, with each starting after the original itinerary for p. $R(p)$ also includes a virtual itinerary to indicate re-accommodation to another

airline's network, or perhaps, cancelation of the passenger trip at its origin. Decision variables ρ_p^r indicate the number of passengers originally on itinerary p who are re-accommodated on itinerary r and d_p^r represents the arrival delay of these passengers. ρ_p^p is the number of passengers on the non-disrupted itinerary p. Parameters Cap_f are the number of seats on flight f and parameter δ_f^r is 1 if flight f is on itinerary r and zero otherwise.

The following constraint ensures that all passengers are flown to their destinations.

$$\sum_{r \in R(p)} \rho_p^r = n_p \qquad\qquad \forall\, p \in P \qquad\qquad (4.16)$$

To ensure that no passengers are assigned to a canceled flight leg, and to restrict the number of passengers assigned to a flight leg if the flight is not canceled, we add the following constraint:

$$\sum_{p \in P} \sum_{r \in R(p)} \delta_f^r \rho_p^r \leq Cap_f(1 - z_f) \qquad\qquad \forall\, f \in F \qquad\qquad (4.17)$$

4.4.5.3   Cost coefficients

The two elements of cost we capture using this model are the fuel burn costs and the passenger delay costs. With each flight copy $k \in C_f$ for each flight f is associated a cost c_f^k that is a sum of (i) the fuel burn costs obtained from the flight planning engine, and (ii) incremental costs of delayed departure. Small incremental costs of \$10 per minute are associated with each minute a flight is delayed beyond its scheduled departure time to prevent the model from unnecessarily delaying flights. Passenger-related costs c_p^r denote the cost of using itinerary r to accommodate passenger p. This is based on the actual arrival time of itinerary $r \in R(p)$ at the destination, and includes delay costs, goodwill costs, and hotel and meal costs if relevant.

4.4.5.4   Aircraft Recovery and Passenger Re-accommodation Model

Given the defined notation and constraints in §4.4, the following is our formulation for combined aircraft recovery and passenger re-accommodation including flight planning opportunities.

$$\min \sum_{f \in F} \sum_{k \in C_f} c_f^k x_f^k + \sum_{p \in P} c_p^r \rho_p^r \qquad (4.18)$$

$$\text{s.t.} \sum_{k \in C_f} x_f^k + z_f = 1 \qquad \forall f \in F \qquad (4.19)$$

$$\sum_{g \in n^-} y_g + \sum_{(f,k) \in n^-} x_f^k + s_a^n = \sum_{g \in n^+} y_g + \sum_{(f,k) \in n^+} x_f^k \qquad \forall n \in N_a', \forall a \in A \qquad (4.20)$$

$$\sum_{r \in R(p)} \rho_p^r = n_p \qquad \forall p \in P \qquad (4.21)$$

$$\sum_{p \in P} \sum_{r \in R(p)} \delta_f^r \rho_p^r \leq Cap_f(1 - z_f) \qquad \forall f \in F \qquad (4.22)$$

$$x_f^k \in \{0, 1\} \qquad \forall k \in C_f, \forall f \in F \qquad (4.23)$$

$$\rho_p^r \in \mathbb{Z}^+ \qquad \forall r \in R(p), \forall p \in P \qquad (4.24)$$

$$y_g \geq 0 \qquad \forall g \in G_a \qquad (4.25)$$

4.4.6 Approximate Aircraft and Passenger Recovery Model to Trade-off Fuel Burn and Passenger Cost

Solving the aircraft and passenger recovery model with passenger re-accommodation described in (4.18) - (4.25) can be excessive for real-time decision making. For a similar model, Bratu and Barnhart [BB06] report solution times of 84 minutes for instances drawn from a US airline. Moreover, they report that feasible solutions obtained when the model is stopped after 5 minutes result in high operating costs. Similar behavior can be expected from (4.18) - (4.25) due to the large sizes of the problem (due to flight copies from alternate flight plans and departure times), and due to the capacity constraints (4.22) that can often result in fractional solutions; which are also observed in Barnhart, Kniker and Lohatepanont [BKM02] and Bratu and Barnhart [BB06]. Thus solving (4.18) - (4.25) typically requires excessive solution time and therefore is impractical to solve in real-time [BB06], it is not suitable for application when flight plan selection decisions of choosing the flight plans have to be determined in a few minutes, before filing a flight plan for an aircraft. Therefore, we introduce an alternative model that captures the trade-off between fuel burn and passenger delays approximately.

In addition to the notation in (4.18) - (4.25), let P denote the set of passenger itineraries, C the set of connecting itineraries, $IT(p)$ the set of flight legs in itinerary p, $IT(p, n)$ the nth flight leg in itinerary p, and n_p the number of passengers originally booked on itinerary p. Let n_f be the number of booked passengers whose itineraries terminate with flight leg f; δ_f^p equal 1 if itinerary p terminates with flight leg f, and 0 otherwise; and d_f^k be an incremental delay cost associated with operating copy k of flight f. In the formulation presented in this section, we assume that connecting passenger itineraries have two flight legs, however, this can be extended to itineraries with more than two flight legs in a straightforward manner. We let $MC(p, f, k)$ denote the set of flight leg copies in the passenger flow network N_p for itinerary p to which there is insufficient time to connect from copy k of flight leg f, and which, if chosen, will result in itinerary p passenger misconnects. Let λ_p be a binary variable that is 1 if itinerary p is disrupted and 0 otherwise, and let c_p be the cost of disruption per passenger on itinerary p.

c_p approximates the costs of re-accommodation for each disrupted itinerary p. In our model, we assume that if passenger itinerary p is disrupted, the passengers on itinerary p are re-accommodated on similar flight(s) in the next bank. Based on this assumption, we compute the per passenger average arrival delay cost to the airline for passengers on itinerary p. These costs represent the airline's estimate of the costs it incurs due to passenger delays, including recovery, hotel and meal costs, and goodwill costs corresponding to the arrival delay. Note that setting a cost per itinerary p allows the capture of non-linearity in costs, where higher delays incur higher costs compared to smaller delays.

Our modified aircraft recovery model with passenger disruptions is as follows:

$$\min \sum_{f \in F} \sum_{k \in C_f} c_f^k x_f^k + \sum_{p \in P} c_p n_p \lambda_p + \sum_{f \in F} \sum_{k \in C_f} d_f^k x_f^k \tag{4.26}$$

$$\text{s.t. } \sum_{k \in C_f} x_f^k + z_f = 1 \qquad \forall f \in F \tag{4.27}$$

$$\sum_{g \in n^-} y_g + \sum_{(f,k) \in n^-} x_f^k + s_a^n = \sum_{g \in n^+} y_g + \sum_{(f,k) \in n^+} x_f^k \qquad \forall n \in N_a, \forall a \in A \tag{4.28}$$

$$x_{IT(p,1)}^k + \sum_{m \in MC(p,IT(p,1),k)} x_{IT(p,2)}^m - \lambda_p \leq 1 \qquad \forall k \in C_{IT(p,1)}, \forall p \in P \tag{4.29}$$

132

$$\lambda_p \geq z_f \qquad\qquad \forall\, f \in IT(p), \forall\, p \in P \qquad (4.30)$$

$$x_f^k \in \{0,1\} \qquad\qquad \forall\, k \in C_f, \forall\, f \in F \qquad (4.31)$$

$$z_f \in \{0,1\} \qquad\qquad \forall\, f \in F \qquad (4.32)$$

$$\lambda_p \in \{0,1\} \qquad\qquad \forall\, p \in P \qquad (4.33)$$

$$y_g \geq 0 \qquad\qquad \forall\, g \in G_a, \forall a \in A \qquad (4.34)$$

The objective function (4.26) consists of three terms - the fuel costs of flights, the costs of passenger itinerary disruptions and the incremental costs of flight delays. Constraints (4.29) ensure that itineraries with insufficient connection time are classified as disrupted. Because the value of c_p is greater than zero and by definition of $MC(p, IT(p,1), k)$, this constraint ensures that λ_p is 1 only if both $x_{IT(p,1)}^k$ and $\displaystyle\sum_{m \in MC(p, IT(p,1),k)} x_{IT(p,2)}^m$ are 1, that is, if passengers on itinerary p cannot connect from the first leg on their itinerary to the second leg on their itinerary. Constraints (4.30) similarly ensure that if a flight leg is canceled, all itineraries containing the flight are classified as disrupted. The constraints (4.33) that variables λ_p should be binary can be relaxed to $0 \leq \lambda_p \leq 1$ because x and z variables are binary and λ variables are 1 only if the first two terms in (4.29) are 1 or a corresponding z variable is 1. In all other cases, λ variables will be zero because of the positive cost associated with them in the objective. Thus their binary nature can be enforced even by specifying λ as continuous between 0 and 1. Constraints (4.27), (4.28), (4.31), (4.32) and (4.34) ensure flight cover, aircraft balance, and binary nature of variables x, z and y respectively, as discussed for (4.18) - (4.25).

For all the instances that we solve in our experiments, described in §4.6, (4.26) - (4.34) is solved within a minute of run time, making it suitable for real-time application.

4.5 Experimental Setup

4.5.1 Network Structure and Experiment Design

In this section, we demonstrate the potential impact of disruption management enhanced with flight planning, using data obtained from a major European airline, Airline 1 specified in §4.3.2, that serves multiple continents. The airline operates a hub-and-spoke network with about 250

flights per day serving about 60 cities daily. (This does not include feeder airline flights.) The airline operates a banked schedule at its hub. About 243 flights, or 93% of the flights operated by the airline are into or out of the hub. 10% of the flights (approximately 30 arrivals and departures per day) operated are long-haul, and present significant opportunities for speed changes. The remaining 90% of flights are short-haul. Aircraft rotations on this network are typically designed as cycles originating from and ending at the hub, with each cycle consisting of 2 to 4 flights. This is particularly true of short-haul flights that operate within Europe, which are operated as short cycles around the hub. Long-haul flight operations comprise more than 30% of the flying hours of the airline per day. About 40% of the passengers have at least one long-haul flight on their itinerary. Because these itineraries bring in more revenue than itineraries with only short-haul flights, we estimate that about 50% of the revenue is associated with long-haul flights on the passenger itinerary.

In our experiments, we focus on disruptions of long-haul flights that are inbound to the hub. This is because of two reasons. First, a significant percentage of passengers connect at the hub from international locations to Europe and vice versa, and therefore the hub presents an opportunity to study passenger connectivity. Second, flight planning opportunities, in particular, speed changes, are significant for long-haul flights. These two reasons lead us to concentrate on disruptions of such long-haul flights arriving into the hub.

We compare three different types of disruption management options: (i) disruption management that does not incorporate operational speed changes, (ii) operational speed changes according to the airline's rule of thumb, and (iii) our enhanced disruption management models with operational speed changes and intentional holding of downstream flights to wait for delayed connecting passengers from upstream flights.

Here we add a remark on applicability to other airlines. Our computations in this section are geared towards Airline 1 that has provided us data for this research. As discussed in §4.3.2, this airline operates at a 'normal' CI of 30, and dispatchers and dispatchers and pilots typically speed up flights up to CI 300 based on their discretion. Other airlines operate at different values of CI and the network trade-offs of their operations will be determined by the operational CI. For example, an airline that operates at a higher CI than Airline 1 will observe higher fuel costs for speeding up, as it is already operating at a steeper portion of the flight time-fuel burn curve

(Figure 4-5); however, there is also greater possibility for slowing down of flights and saving fuel if network connectivity is not impaired.

Our models are implemented in Java and C using ILOG CPLEX 9.0. Computational experiments are conducted on a workstation using an Intel Pentium 4 2.8 GHz processor and 1 GB RAM.

4.5.2 Historical Delay Analysis and Scenario Generation

First, we conduct an analysis of delays of long-haul flights that are inbound to the hub and generate distributions of these delays to understand the frequency of such delays. Our historical delay analysis is conducted for data available for the months of August, September and October 2008. In our experiments, we focus on those scenarios that are frequent in a statistically significant manner. For the airline under consideration, we observe that in a given bank, there is typically at most one inbound long-haul flight delayed into the hub. Simultaneous delays of long-haul flights into the hub though existing, are not found to be statistically significant.

Based on this observation, we divide the inbound delay for each long-haul flight into 'buckets' of delay. Beginning from a bucket of 0 minutes of delay, we create buckets in 10 minute increments to the highest level of delay observed in the data. Corresponding to each long-haul flight, we create a histogram of frequency of delay occurrences in each delay bucket. From the histogram for each flight, we construct instances where long-haul flight f is delayed into the hub H by Δ minutes, where Δ takes on values of each bucket of delay. Our instances encompass a representative set of long-haul flights f, for a representative set of delays Δ. We represent each instance of this type by $S(f, \Delta)$.

Though historical flight delay information is available for August-October 2008, passenger information for this period is not available. We have passenger data only for a period of two weeks in November 2008 (for which we do not have flight delay data). We replicate each instance $S(f, \Delta)$ for each day for which passenger data is available. Thus, each delay scenario (from Aug-Oct 2008) is solved multiple times, once for each day (of two weeks in Nov 2008) with its particular passenger connections.

4.5.3 Parameter assumptions

We assume the following values for the parameters in the model:

- Passenger-related delay costs = $1.161/passenger per minute, for 2008. This number is the airline's estimate of *their own cost* incurred for passenger delays, including recovery, re-accommodation and goodwill cost.

- Fuel cost = $6.08/gal. This estimate is a result of the airline AOCC's reported costs of €700 - €800 per metric ton of fuel in February 2010. Jeppesen [Jepa] converts this to a cost range in September 2008. The airline's cost in Feb 2010 is converted to a cost of €2.2-€2.52/gal (with density 0.82 kg/litre and a litre equal to 0.26 US gal). Further, guided by the IATA fuel price development charts [Int10a], a ratio of 1.78 for costs in September 2008 to February 2010 is applied, to convert the price range to €3.91-€4.48/gal in September 2008, or $5.27 - $6.27/gal (using a conversion rate of $1.35 - $1.4 per €[Eur10]).

- T = approximately 1.5 days, encompassing two successive arrival banks at the hub to allow for aircraft swaps.

- normal CI = 30; rule-of-thumb maximum CI, specified by the airline = 300

- c_p, Cost per disrupted passenger in model (4.26) - (4.34) = $384.3. This cost is calculated assuming that disrupted passengers are re-accommodated in the next bank, with an average time of 5 hours, by calculating the average time to the next connecting flight for different passenger itineraries.

4.5.4 Baseline for comparison

We use as a basis for comparison the case where airlines allow for fleet-based aircraft recovery, followed by passenger re-accommodation and recovery. Models for combined aircraft and passenger recovery are not standard practice at most airlines, one reason being that this would involve delaying outbound flights and risking loss of on-time performance. Another reason is that airlines may not have the type of decision support that simultaneously enables both aircraft

and passenger recovery. Some information systems can be used to allow larger groups of passengers to connect, but this is also done in an ad-hoc and manual way rather than using a form of automated decision-support.

The conventional recovery model used as a basis for comparison in this work follows the model (4.3) - (4.7) described in §4.2. We use an aircraft recovery model, with flight departure copies for the fleet types(s) undergoing recovery in order to allow for swaps, departure delays and cancelations. This model does not allow speed changes and passenger-related delay costs are not the objective, rather the objective is to minimize flight departure delay costs and cancelation costs.

4.5.5 Simulation

The solutions to our baseline and our enhanced disruption management models are evaluated using an airline disruption management simulator [DCT] [Vaa], which is part of the Integrated Operational Control System developed by Jeppesen[Jepa]. The purpose of this simulator is to compute the estimated true realized passenger delay costs of our models' solutions on the day of operations. This simulator performs passenger re-accommodation by solving the passenger recovery problem (4.10) - (4.13) with the actual cost values experienced by the airline and provides an estimate of the true passenger-related delay cost to the airline. These are computed using the delay cost specified in §4.5.3, which include passenger-delay related costs to the airline, hotel, meal reimbursements and goodwill costs. Note that the simulated costs obtained using this approach are different from the objective functions to our model (4.26) - (4.34). This is because (4.26) - (4.34) approximates passenger disruption costs, whereas the simulator evaluates the actual cost realizations of the solutions to our models. In §4.6, we present the estimated *true costs* of our model solutions for the simulator, and not simply the objective function values to our models.

4.6 Results

In this section, we present the results of our experiments. In our computations, we compare the following disruption management strategies: (i) Baseline disruption management using the

model (4.3) - (4.7) described in §4.2, (ii) Naive approach of speeding up to the maximum allowable CI when a disruption is encountered, and (iii) Enhanced disruption management with flight planning ((4.26) - (4.34)) with $\Theta = 0$, 10 and 15 minutes.

4.6.1 Case Study 1

In this case study, we describe results for instances of type $S(f, \Delta)$ (described earlier in §4.5.2), which contain a single long-haul inbound delay into the hub. We then scale these up to measure the impact of our approach at a network-wide scale. To discuss the performance of solutions to our model (4.26) - (4.34), we report the *simulated* solution costs.

To illustrate the trends in the solutions, we first present the results for a specific flight f, for different levels of delay Δ on 12 days of operation of f, in November 2008. Flight f is representative of the other flights in the network in that the trends and trade-offs observed with this flight are also seen in the case of other flights. In this case, we vary Δ from 10 minutes to 60 minutes in intervals of 10 minutes.

Figure 4-7 shows the change in fuel burn and passenger cost curves for different levels of Δ, for selected days. The horizontal axis represents the arrival delay of flight f and the vertical axis represents costs incurred. First, observe that for each value of Δ, the fuel cost curve can be plotted to reflect speed changes in f, resulting in different arrival delays and corresponding fuel burn. Fuel cost curves are marked by Δ values from 10 to 60 in the upper portion of the figure. As the value of Δ increases, the curve itself does not change shape, but shifts to the right to reflect increased arrival delay.

In the case when downstream flights are not held on the ground ($\Theta = 0$), the passenger costs incurred for different levels of flight arrival delay are shown, for instances across five days of data. (These are indicated in the lower part of Figure 4-7.) As arrival delay increases, passenger delay increases and more passenger misconnects occur. The delay cost curve incurs a 'jump' when a set of passengers misconnect and require recovery and re-accommodation. The total cost curve that is a sum of fuel costs and passenger costs changes dynamically with changes in the delay Δ of flight f. This we illustrate using Figure 4-8, which demonstrates the change in the total cost curve for different Δ, for one day of operations (represented by Day 2 in Figure
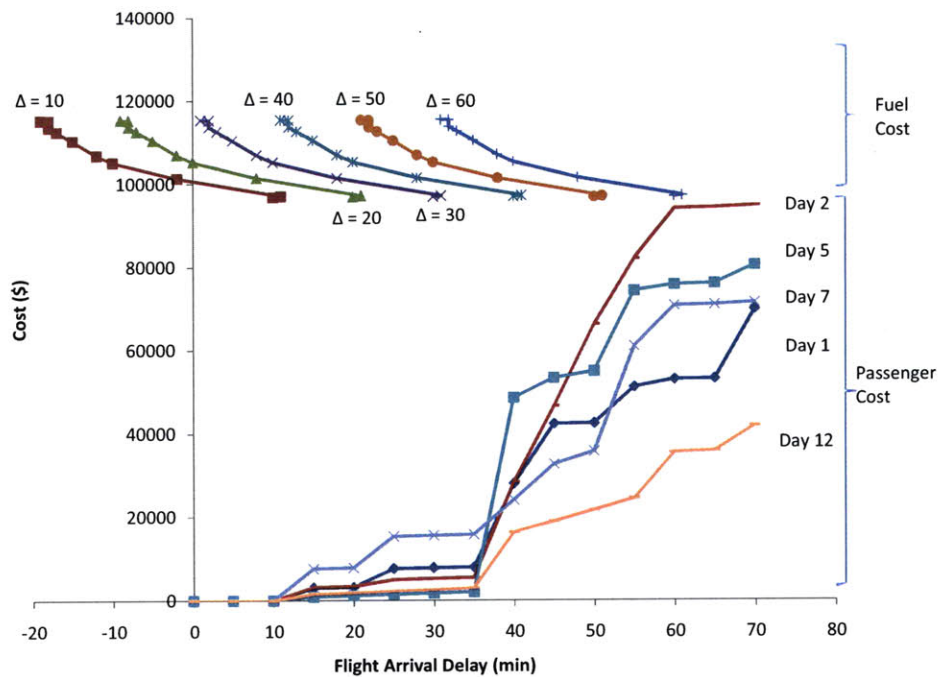
Figure 4-7: Trade-offs between fuel burn and passenger delay costs over multiple days

4-7).

Figure 4-8 serves to illustrate that the choice of the optimal arrival time, and hence the optimal flight plan, depends on the network connectivity at all the possible arrival times of the flight. This dynamic changes when the departure times of downstream flights are allowed to be altered (or are altered in the course of the day, due to plans not operating exactly as planned) so that passengers can make connections. Holding downstream passenger connections opens up the possibility of the upstream flight speeding up to a smaller extent and burning less fuel, but incurring fewer misconnections. The network interactions now become more interesting, as we have the flexibility of changing speeds and departure times of inbound delayed flights as well as the outbound flight departure times.

We now describe the phenomena that occur when flight speeds and departure times are simultaneously modified to mitigate the effects of disruptions. We do so by solving the model (4.26) - (4.34), with different values of Θ. $\Theta = 0$ results in the phenomenon so far discussed and described in Figures 4-7 and 4-8. Now we present the trends in fuel burn and passenger-related

139
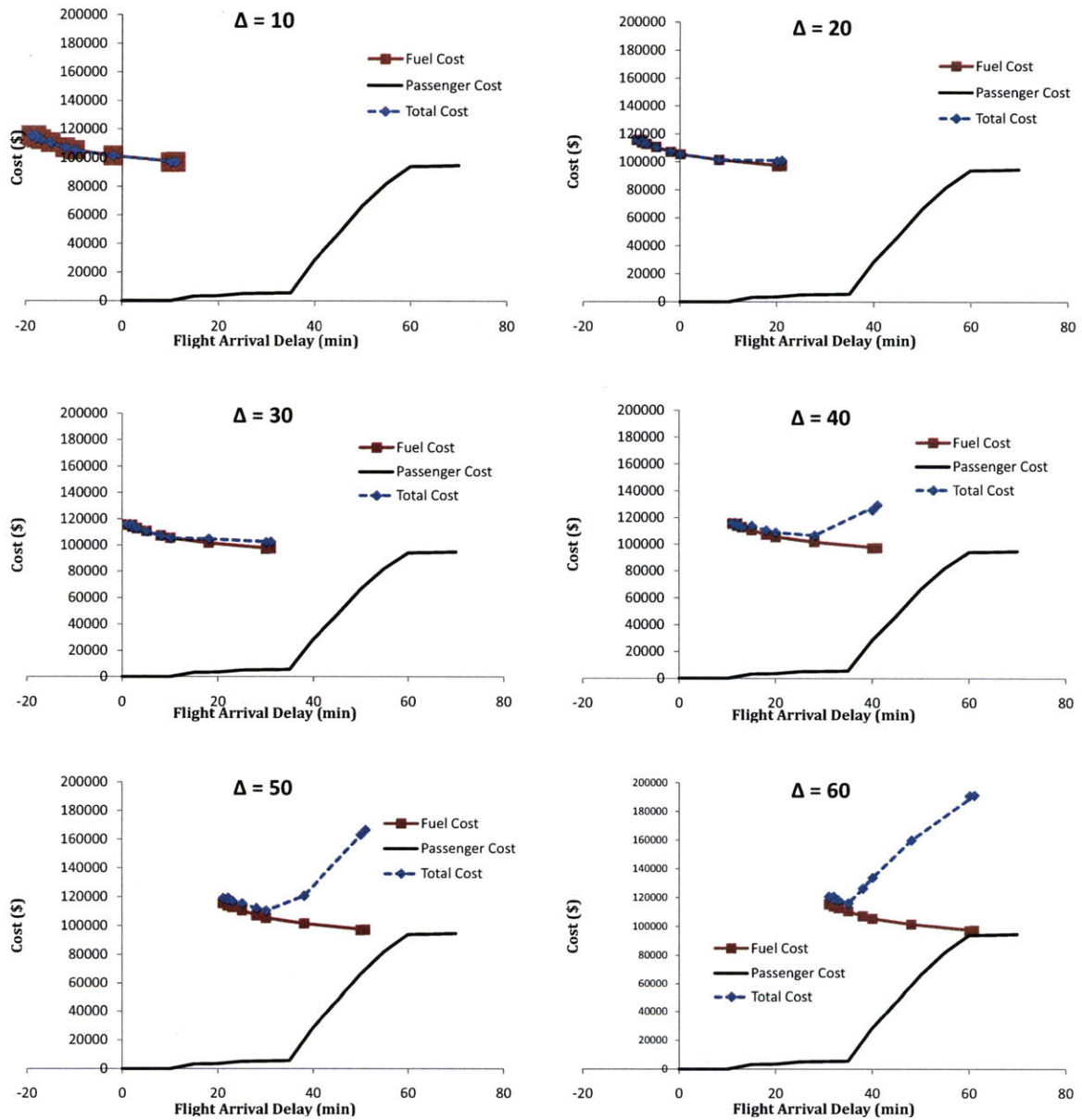
Figure 4-8: Changing optimal trade-off point between fuel and passenger cost with departure delay Δ

airline costs (true costs estimated via simulation) when (4.26) - (4.34) is solved for $\Theta = 0$, 10 and 15, for the representative flight f and each value of Δ; and compare them to our baseline results. The costs presented in this table are over a period of 12 days for which data is available for this flight. Note that the fuel burn costs presented are those related to the disruption, not related to the schedule - that is, we report not the total fuel burned, but the additional fuel

burned compared to the schedule due to the disruption. In Table 4.2, we present the fuel burn, passenger (pax) misconnections and associated costs experienced by the airline, for results from five different strategies of disruption management:

- Column 1: The baseline disruption management strategy described in Section 4.5.4, which does not allow for speed changes

- Column 2: A disruption management strategy that combines the baseline disruption management strategy with a naive speed-up strategy. In the case of delay, the dispatcher always speeds up to the maximum rule-of-thumb CI. In this case, the speed up is to CI 300.

- Column 3: The enhanced disruption management strategy that combines the baseline disruption management strategy with 'optimal' speed changes using (4.26) - (4.34). Θ is set to 0 to prevent downstream flights from being delayed at departure.

- Column 4: The enhanced disruption management model strategy that combines the baseline disruption management strategy with flight speed changes and downstream flight departure delays up to 10 minutes. This is obtained by solving (4.26) - (4.34) with Θ set to 10 minutes.

- Column 5: The enhanced disruption management model strategy that combines the baseline disruption management strategy with flight speed changes and downstream flight departure delays up to 15 minutes. This is obtained by solving (4.26) - (4.34) with Θ set to 15 minutes.

From our analysis, we present the following findings:

1. For all levels of disruption, recovery models enhanced using flight planning will hold constant or decrease total passenger delay costs compared to the baseline approach. In the instances we tested for a proof-of-concept, the cost improvements ranged from 0 to 15%.

2. The naive model that is sometimes adopted by dispatchers, of speeding up to the allowable cost-index (Column 2 solutions), results in improved passenger costs compared to

141

	Recovery without speed changes	Naive speed up (airline's rule-of-thumb to CI 300)	Enhanced recovery: don't hold connecting flights	Enhanced recovery: hold connecting flights up to 10 min	Enhanced recovery: hold connecting flights up to 15 min
$\Delta = 10$ min					
Fuel burn (lb)	0	60036	-624	-624	-624
Fuel cost ($)	0.00	54480.43	-566.26	-566.26	-566.26
Pax misconnects	0	0	0	0	0
Num. flights held	-	-	-	0	0
Delayed pax cost ($)	0.00	0.00	0.00	0.00	0.00
Total cost ($)	0	54480.43	-566.26	-566.26	-566.26
$\Delta = 20$ min					
Fuel burn (lb)	0	60036	34813	34813	-624
Fuel cost ($)	0.00	54480.43	31591.50	31591.50	-566.26
Pax misconnects	351	1	23	12	1
Num. flights held	-	-	-	4	16
Delayed pax cost ($)	55329.75	3505.95	10899.90	9514.80	3505.95
Total cost($)	55329.75	57986.38	42491.40	41106.30	2939.69
$\Delta = 30$ min					
Fuel burn (lb)	0	60036	43489	35021	35021
Fuel cost ($)	0.00	54480.43	39464.64	31780.25	31780.25
Pax misconnects	351	203	101	22	12
Num. flights held	-	-	-	14	20
Delayed pax cost ($)	59695.65	36224.55	25108.65	18719.10	17545.95
Total cost ($)	59695.65	90704.98	64573.29	50499.35	49326.20
$\Delta = 40$ min					
Fuel burn (lb)	0	60036	76233	55458	55458
Fuel cost ($)	0.00	54480.43	69178.60	50326.07	50326.07
Pax misconnects	806	351	239	46	45
Num. flights held	-	-	-	40	42
Delayed pax cost ($)	216265.95	57948.75	59332.50	30870.45	31387.50
Total cost ($)	216265.95	112429.18	128511.10	81196.52	81713.57
$\Delta = 60$ min					
Fuel burn (lb)	0	60036	79954	110916	79954
Fuel cost ($)	0.00	54480.43	72555.27	100652.13	72555.27
Pax misconnects	1239	1069	363	355	342
Num. flights held	-	-	-	71	143
Delayed pax cost ($)	533182.50	413020.35	82301.40	77556.15	189668.25
Total cost ($)	533182.5	467500.78	154856.67	178208.28	262223.52

Table 4.2: Flight $A - H$ disruption costs from simulations for different recovery strategies, summed over 12 days of operation

the baseline recovery model, as it improves the on-time performance of the flight. However, for medium levels of disruption, such as 20-30 minutes, it results in increased fuel consumption even in cases where it may not be required. This rule-of-thumb-based policy may be able to recover passengers for $\Delta = 10$ and $\Delta = 20$, but may fall short for larger disruptions. In comparison with our optimization-based models, the rule-of-thumb to speed up to CI 300 almost always results in higher costs.

3. Because the objective function of the model (4.26) - (4.34) captures different costs than the simulation (it captures approximate costs of disruption and not exact recovery costs), discrepancies may be (rarely) observed in cases where true passenger delay costs from the simulation, are not well approximated by the passenger disruption costs in the objective

142

	Recovery without speed changes	Naive speed up (airline's rule-of-thumb to CI 300)	Enhanced recovery: don't hold connecting flights	Enhanced recovery: hold connecting flights up to 10 min	Enhanced recovery: hold connecting flights up to 15 min
$\Delta = 10$ min					
Fuel burn savings (lb)	-	-5003	52	52	52
Fuel cost savings ($)	-	-4540.04	47.19	47.19	47.19
Pax misconnects saved	-	0	0	0	0
Delayed pax cost savings ($)	-	0.00	0.00	0.00	0.00
Total cost savings ($)	-	-4540.04	47.19	47.19	47.19
Std.dev. total savings ($)	-	0.00	0.00	0.00	0.00
Total cost savings (%)	-	N/A	N/A	N/A	N/A
$\Delta = 20$ min					
Fuel burn savings (lb)	-	-5003	-2901.083333	-2901.083333	52
Fuel cost savings ($)	-	-5448.04	-2632.62	-2632.62	47.19
Pax misconnects saved	-	35.00	27.33	28.25	29.17
Delayed pax cost savings ($)	-	5182.38	3702.49	3817.91	4318.65
Total cost savings ($)	-	-265.66	1069.86	1185.29	4365.84
Std.dev. total savings ($)	-	3380.51	4286.77	4263.51	12503.64
Total cost savings (%)	-	-4.80	23.20	25.71	94.69
$\Delta = 30$ min					
Fuel burn savings (lb)	-	-5003	-3624.1	-2918.4	-2918.4
Fuel cost savings ($)	-	-4540.04	-3288.72	-2648.35	-2648.35
Pax misconnects saved	-	12.33	20.83	27.42	28.25
Delayed pax cost savings ($)	-	1955.93	2882.25	3414.71	3512.48
Total cost savings ($)	-	-2584.11	-406.47	766.36	864.12
Std.dev. total savings ($)	-	1569.33	1218.01	2323.02	2146.41
Total cost savings (%)	-	-51.95	-8.17	15.41	17.37
$\Delta = 40$ min					
Fuel burn savings (lb)	-	-5003	-6352.75	-4621.5	-4621.5
Fuel cost savings ($)	-	-4540.04	-5764.88	-4193.84	-4193.84
Pax misconnects saved	-	37.92	47.25	63.33	63.42
Delayed pax cost savings ($)	-	13193.10	13077.79	15449.63	15406.54
Total cost savings ($)	-	8653.06	7312.90	11255.79	11212.70
Std.dev. total savings ($)	-	7756.53	7936.95	6146.52	6181.36
Total cost savings (%)	-	48.01	40.58	62.46	62.22
$\Delta = 60$ min					
Fuel burn savings (lb)	-	-5003	-6662.8	-9243	-6662.8
Fuel cost savings ($)	-	-4540.04	-6046.27	-8387.68	-6046.27
Pax misconnects saved	-	14.17	73.00	73.67	74.75
Delayed pax cost savings ($)	-	10013.51	37573.43	37968.86	28626.19
Total cost savings ($)	-	5473.48	31527.15	29581.18	22579.91
Std.dev. total savings ($)	-	6300.55	11011.94	9284.81	11884.86
Total cost savings (%)	-	12.32	70.96	66.58	50.82

Table 4.3: Flight $A - H$ simulated average cost *savings* per day for different recovery strategies, averaged over 12 days of operation

function of (4.26) - (4.34). For example, note the negative value (indicating expenditure) of realized total cost value for $\Delta = 30$ in column 3 of Table 4.3. The simulated passenger delay cost savings are not as high as is indicated by the objective function when (4.26) - (4.34) is solved, resulting in costs instead of savings.

4. Depending on the itineraries of passengers, both flight speed changes as well as holding of downstream flights can reduce the number of misconnected itineraries. For example, in the case of a 20 minute initial disruption, simply allowing speed changes without hold-

ing downstream flights is sufficient to recapture 93% of disrupted passengers back onto their original itineraries compared to the baseline case. On the other hand, in the case of 30 and 40 minute delays, more than 33% of passenger misconnects can be prevented by allowing speed changes without holding flights; however, by allowing downstream flight departures to be delayed by 10 minutes, the misconnects are decreased by 95% in this case. In practice, as Δ increases above 30 minutes, speed ups are considered in an ad-hoc manner by the aircraft controllers as means of mitigating disruption, though not integrated with the disruption management system. So, in the case of $\Delta = 30$ and $\Delta = 40$, the true savings in misconnections lies somewhere between the 'recovery without flight planning' and the 'naive speed-up' cases. This translates into a decrease in passenger misconnections compared to airline practice of between 95% and 87%.

5. For lower levels of disruption, fuel burn costs dominate and drive the trade-off between fuel burn, as seen in the cases of 10 - 20 minutes of delay. In these cases, because fewer passengers are impacted, the balance in the optimization model tilts in the favor of decreasing fuel costs. The decision in such cases is to slow down the flight because passengers are not disrupted by the slow down. Occurrences of these levels of delays provide an opportunity to save fuel in comparison to the baseline recovery approach.

6. For very low levels of disruption (for example, 10 minutes), enough slack is present in the system to absorb the disruption, and flight planning mechanisms such as speed increases and holding downstream flights, are not required. Instead, we might be able to slow down the flight without incurring disruptions, as seen in Table 4.2. However, for even fairly low levels of disruption such as 20 minutes, the interaction between speed changes and passenger delays can come into play. In the case of the flight demonstrated in Table 4.2, some passenger connections have a small amount of slack for which delays of 20 minutes cannot be absorbed, resulting in misconnections if the flight is not sped up. In the case of 20 minutes of delay, we see that almost all disruptions can be absorbed by speeding up the flight, and/or delaying downstream flights to the appropriate extent.

7. For higher levels of initial disruption (more than 30 minutes in the case shown above), passenger delay costs dominate the trade-off between fuel burn and delay costs. This

is because many more downstream flights are impacted by a large initial disruption. To reduce the number of passenger disruptions, the optimal least total cost decision is to speed up the long-haul flight. If allowed, downstream flights are also held in order to facilitate passenger connections.

8. For intermediate levels of delay, such as 20 - 40 minutes, holding downstream flights to wait for connecting passengers can have significant benefits. In Table 4.2, the number of passenger misconnections decreases significantly from Column 3 to Columns 4 and 5. With the decrease in the number of misconnections, we see a corresponding decrease in passenger-related costs, for $\Delta = 20$ and $\Delta = 30$. For $\Delta = 40$ and higher, however, we see that passenger delay costs increase in the case of $\Theta = 15$, even though the number of misconnects decreases. The increase in passenger-related costs connected with holding downstream flights begins to exceed the decrease in passenger-related costs associated with re-accommodation and recovery of disrupted passengers. However, the optimization model (4.26) - (4.34) chooses these solutions which decrease misconnects but increase delayed passenger costs because it is geared towards passenger misconnects and not passenger recovery and re-accommodations. Thus the benefits of holding downstream flights are seen to decrease as the level of the initial disruption Δ increases, because many more flights must be held, and causing increased waiting time for passengers (who are not necessarily connecting from A-H) on the downstream flights. Thus the cost savings of preventing misconnections from the long-haul flight is more than offset by the cost due to holding the downstream flights.

9. We also observe from Table 4.2 that for departure delay levels less than 40 minutes, the number of passenger misconnections and the corresponding passenger costs significantly decrease when downstream flights are held compared to the case when only flight speeds are changed without holding the downstream flights. For higher levels of delay, as shown in the 60 minute case, the decrease in the number of misconnections and the passenger cost is less significant when flights are held compared to when only speed changes are allowed. (In fact, passenger delay costs increase for $\Theta = 15$.) This is because of the model (4.26) - (4.34) being geared towards passenger misconnections; and costs of intentionally

holding downstream flights exceed the cost savings from the decrease is passenger mis-connections. Thus it may be recommended to hold downstream flights when the delay is less than 40 minutes; whereas this is not recommended when the delay is 60 minutes or more. Similar behavior is observed for other flights f for which experiments were conducted. Therefore, for a flight, depending on the connectivity, we can identify such a threshold where passenger flights should be held below the threshold and not held above it.

Table 4.3 shows the average savings in dollars per day and the standard deviation (std.dev.) in cost savings for flight $A - H$, for different values of initial disruption, as compared to the baseline recovery model. Positive values in this table represent cost savings and negative values represent additional expenditure compared to the baseline disruption management model described in Column 1.

While the discussion in this section focused on a single inbound flight delay into the hub, such types of delays occur in more than 85% of cases for the airline under consideration. Summarizing the savings over a representative set of long-haul flights and disruption levels, that is, $S(f, \Delta)$ for different Δ and f; and further weighting these savings with their historical occurrence frequencies, we conclude that the benefit of integrating speed changes and flight departure scheduling with disruption management can be significant, as shown in Table 4.4. Compared to the baseline disruption management approach, for the time period that our experiments were conducted, the three enhanced recovery approaches shown in Table 4.4 result in annual savings of \$ 15,879,801, \$ 15,846,767, and \$ 14,555,865 respectively. These savings are possible with a relatively small increase in fuel burn of 0.2% per long-haul flight. Passenger misconnects from long-haul flights are reduced by 47.2% - 53.3% and delayed passenger costs are reduced by 44.9% - 46.61% compared to the baseline case.

We add a note on the sensitivity of our model solutions and the corresponding savings to fuel costs. Note, from Figure 4-8 that for more than 20 minutes of initial delay, passenger delay costs dominate fuel burn costs, and drive the total cost function. This indicates that the dominant component of the cost savings are due to the reduction in passenger misconnects, especially at higher levels of delay. Therefore, even if fuel costs fluctuate significantly, it is likely that in (4.26) - (4.34), the passenger cost component will drive the objective, and similar savings in

146

	Enhanced recovery: don't hold connecting flights	Enhanced recovery: hold connecting flights up to 10 min	Enhanced recovery: hold connecting flights up to 15 min
Passenger misconnects decreased compared to baseline	47.2%	52.0%	53.3%
Increase in fuel cost per long-haul flight	0.2%	0.22%	0.14%
Delayed passenger delay costs decreased compared to baseline	44.90%	47.60%	46.61%

Table 4.4: Improvements observed using enhanced models, compared to conventional disruption management case

passenger misconnects can be observed compared to the baseline.

4.6.2 Case Study 2

In the previous section, we analyzed cases of a single inbound flight delay into the hub at a given time, and optimize for one inbound delay at a time. This was motivated by the historically observed delay distributions for the carrier from which we have available data. In contrast, cases of multiple inbound delays into a hub have been frequently observed for several US airlines. In this section, we provide some insight into optimizing simultaneously for multiple delayed flights entering a hub, and contrast this with our earlier flight-by-flight analysis. The underlying idea we want to explore in this case study is the value of additional information about other flights in the system, and that impact on our total costs. We first present an illustrative scenario.

We consider a scenario where one-third of the incoming flights into a hub experience delays. The delays range from 45 minutes to an hour. Table 4.5, presented in a similar format as Table 4.2, shows the impacts over a day of operations of using different flight planning strategies, for the flights involved in the disruption.

As in the case of single inbound flight delays, the disruption management strategies that include flight planning decrease cost, on the order of $44,000 - $66,000 per day. Our approaches reduce passenger misconnects from 45% to 63%. By taking advantage of the speed-up possibilities of flights, these approaches all consume more fuel compared to the baseline recovery approaches. Some of the additional fuel cost can be decreased by allowing downstream flights to be held, requiring less speed-up of the long-haul flight. Compared to the case where airlines

	Recovery without flight planning	Naive speed up (airline's rule-of-thumb to CI 300)	Enhanced recovery: don't hold connecting flights	Enhanced recovery: hold connecting flights up to 10 min	Enhanced recovery: hold connecting flights up to 15 min
Fraction of flights delayed into hub = 1/3					
Fuel burn (lb)	0	15228	3547	904	626
Fuel cost ($)	0	13818.84	3218.77	1727.8	568.07
Pax mis-connects	144	82	78	61	53
Delayed pax cost ($)	150773.02	111457.1	103403.89	89578.71	82613.22
Total cost ($)	150773.02	125275.94	106622.66	91306.51	83181.29
Total cost savings (%)	-	16.91	29.28	39.44	44.83

Table 4.5: One-third of inbound flights delayed into hub: Incorporating information about multiple disrupted flights simultaneously

include speeding up to the maximum allowable CI as part of recovery, fuel cost is seen to be lower by $10,000 - $13,000.

	Recovery without flight planning	Naive speed up (airline's rule-of-thumb to CI 300)	Enhanced recovery: don't hold connecting flights	Enhanced recovery: hold connecting flights up to 10 min	Enhanced recovery: hold connecting flights up to 15 min
Fraction of flights delayed into hub = 1/3					
Fuel burn (lb)	0	15228	3547	999	626
Fuel cost ($)	0	13818.84	3218.77	906.55	568.07
Pax mis-connects	144	82	78	61	53
Delayed pax cost ($)	150773.02	111457.1	103403.89	89578.71	82613.22
Total cost ($)	150773.02	125275.94	106622.66	90485.26	83181.29
Total cost savings (%)	-	16.91	29.28	39.98	44.83

Table 4.6: One-third of inbound flights delayed into hub: optimizing flight plans for individual flight disruptions

If the same problem were to be solved by optimizing for each disrupted flight individually, instead of considering the flight network as a whole and optimizing considering all disruptions simultaneously, the solutions are as shown in Table 4.6. When downstream flights are not allowed to be held for passengers, both the strategies of flight-by-flight optimization and network-based optimization will yield the same results (see columns 3 of Tables 4.5 and 4.6). When downstream flights are allowed to be held, then we observe interactions between the downstream networks of these disrupted flights. For example, if long-haul flights f_1 and f_2 into the hub are both delayed at departure, and both share several downstream connections, then f_1 may not be sped up as much (compared to when it is optimized individually). This results when, for example, f_2 requires shared downstream flights to be delayed by 10 minutes and f_1 requires only 5 minutes, then the optimal solution is to delay shared downstream flights by 10 minutes and to operate f_1 at a (slower) speed to take advantage of the newly available slack of 5 minutes. Solutions in column 4 of Table 4.5 and Table 4.6 both indicate the same level of

passenger connectivity available, however, Table 4.5 shows less fuel burn because one flight is not sped up to the extent it would be if optimized for individually. Thus the network-optimal course of action for a disrupted flight and the downstream flights is a function of the status of other flights arriving into the hub.

The key insight from this study is that as additional information about the state of the system becomes available, incorporating such information can add value. Detailed experiments over several scenarios are required to assess the benefits of added information, both pre-departure and enroute, at a network-wide level, which we propose in future work in Chapter 5.

Chapter 5

Conclusions and Future Directions

Inherent uncertainty in airline operations guarantees that delays and disruptions are inevitable in the air transportation system. Delays and disruptions cost airlines and passengers billions of dollars a year, as seen by the growing costs in the US from $6B in 2006 to $12B in 2007. The delays and disruptions seen in 2007, the fuel price hikes in 2008 and the economic crisis in 2009 exposed the vulnerabilities of the system. In addition, the growing demand for air travel, combined with the fact that the infrastructure of the airspace system cannot be scaled easily, results in higher expected levels of congestion and delays. In this thesis, we present strategic and operational approaches that can be adopted by airlines to reduce delays and associated costs.

5.1 Summary

5.1.1 Strategic Approaches

Robust Aircraft Routing

We study the application of three types of models to aircraft routing - extreme-value model based on the Bertsimas and Sim [BS04] robust optimization approach; the probabilistic model based on Charnes and Cooper's [CC59] chance-constrained programming (CCP) approach; and the tailored robust aircraft routing model of Lan, Clarke and Barnhart (LCB)[LCB06]. These three robustness mechanisms lead to different models and solutions that have different robustness performances with respect to various metrics of interest.

Increased complexity and solution times are associated with the basic extreme value and probabilistic models, when compared to a deterministic model, because the models have to be solved several times for different values of the robustness parameters. To avoid iterative re-solving, we developed the following extensions to these models: the Delta (extended extreme-value) model for the extreme-value approach, and the Extended CCP for the probabilistic approaches. Our extended models can be solved in a single iteration, with run times equivalent or lower than those of a single iteration each of the basic models.

We evaluate solutions to the different models through simulation, and measure performance via total aircraft delay, on-time performance metrics, and passenger disruption metrics. The extended extreme value and probabilistic approaches can consistently lead to the generation of more robust solutions (compared to the basic models, and the solution currently operated by the airline), as defined by the metrics of interest. These models are also generally applicable, as described in §2.2.2 and §2.2.4.

To understand the differences between the modeling paradigms, we compare the extended extreme value (Delta) model, the extended chance-constrained (ECCP) model and the tailored approach (LCB). The extreme value models' dependence on extreme delay values, ignoring probabilistic information, leads in some cases to large variability in the performance of alternative optimal solutions to the models. In such cases, extra care should be taken in evaluating alternative optimal solutions to these approaches. From this, we conclude that it is not effective to drive the solution process with extreme values that are rare. Probabilistic approaches (such as ECCP) focus on higher-probability delay events, and produce improved solutions according to our metrics. These approaches capture more information about the system and focus on more likely delay events, and thus are more in-line with our metrics of interest, which relate to decreasing total delay. In addition, the tailored LCB approach is seen to be a special case of the probabilistic chance-constrained approach. Though the tailored approach in itself does not explicitly capture knowledge of probability distributions, by simplistically incorporating the 'right' delay quantile in its objective (guided by the Chance-constrained approach), it can achieve improved results through a less complex model.

Our work underscores the importance of choosing an approach that aligns well with the underlying data distributions and the evaluation metrics of interest to the various stakeholders,

including the DoT, the airline and passengers.

Insights from Multiple Applications

In collaboration with IBM Research's Zurich Research Laboratory, we further examine the applicability of the general models applied in the aircraft routing context in Chapter 2, to other problems in pharmaceutical supply chain management and corporate portfolio optimization. We apply the extreme-value based Bertsimas and Sim and Delta approaches, the probabilistic Chance-Constrained Programming (CCP) and Extended Chance-Constrained Programming (ECCP) approaches, and the Conditional-Value-at-Risk(CVaR) approach.

We see that the CVaR approach requires the ability to sample from the joint probability distribution of the uncertain parameters, which can be a considerable challenge in most real-world problems. Also, we quickly run into tractability issues for CVaR, even for medium size problem instances, and certainly for large-scale instances. Thus, this approach proves to be impractical for many of our problem instances.

Extreme-value based approaches - the Bertsimas and Sim and Delta approaches - exhibit a high degree of conservatism because they are guided by worst-case realizations. This makes them more applicable when near-extreme conditions are more frequent in the underlying data (heavy-tailed distributions), that is, when extreme realizations are not rare; or when worst-case realizations must be avoided. We show empirically that when extreme values are rare, this approach can lead to unnecessarily conservative solutions. We also show that when the type of distribution is known with very little certainty, this approach becomes more effective.

Probabilistic CCP and ECCP approaches, on the other hand, need more information than the extreme-value based approaches about the distribution, at least in the form of quantiles, to be effective. When little information is available, they are simply equivalent to a mean-value or a worst-case approach. These approaches, which weight more heavily probable data realizations in the optimization, do not give a great deal of importance to worst-case realizations. They result, then, in solutions geared towards optimizing average-case performance metrics.

Because the robust approach's robustness metric can be different from those of the decision-maker, and multiple robustness metrics can be defined for a system, it is difficult to assess the robustness of a solution simply based on the model's robustness parameters and its objective

153

function. This points to the importance of simulation, and evaluation based on multiple criteria to choose robust solutions. In addition, we show that knowledge of underlying data distributions, even if empirically derived, can be used to modify the input parameters of the robust models and lead to improved solutions. In conclusion, the efficacy of any given robust approach is determined not by the approach or model alone, but by the interaction between the model, data and evaluation metrics.

5.1.2 Operational Approaches

Integrated Disruption Management And Flight Planning

We show in this chapter that considerable benefits can be obtained by integrating flight planning into disruption management. In current practice, flight planning and disruption management are treated separately. Flight planning enables pre-departure decisions about flight speeds, routes and fuel burn. Through flight planning, we re-allocate slack in block and ground times in the network, and add additional flexibility into disruption management by changing flight speeds and/or re-timing downstream flights to preserve aircraft, crew and passenger connections.

We glean information from multiple airlines about the state-of-the-practice in flight planning. We show that the consequences of using a static (pre-determined) Cost Index value as is used in practice can result in non-optimal flight plans that burn more fuel than needed or do not ensure network connectivity. An optimization-based decision making tool that optimizes the trade-off of schedule delay costs and fuel burn is required to make the best decisions during operations.

We present models that enable aircraft recovery and passenger recovery by integrating flight planning decisions and disruption management. In the interest of tractability, we also present models for aircraft recovery with approximate passenger connectivity, and solve these models within 1 second.

Through experiments using data from a hub-and-spoke carrier, we show that propagation of inbound hub delays and their associated costs can be decreased significantly using our models. In fact, as the level of delay into the hub increases, disruption management enhanced using flight planning mechanisms provides higher cost savings. We also show that as the number of

154

disrupted flights increases, the savings from integrated flight planning and disruption management increase. In comparison with traditional disruption management approaches, our models decrease passenger misconnections by 47.2% - 53.3%, resulting in a corresponding decrease in passenger-related operating costs of the airline.

5.2 Extensions and Future Directions

Dynamic robustness. Robust optimization is motivated by the fact that the ability to recover from disruptions and uncertainty is closely correlated with the original schedule and network design. To fundamentally address the sequential nature of transportation network operations, where new system information is continuously revealed, a *dynamic* notion of robustness is needed. The new notion of robustness therefore is to incorporate flexibility allowing us to reconfigure easily and inexpensively the system (matching operations to external system state) as information is revealed. The goal in designing such systems is to incorporate new information and enable continual modification of operations to match the system state.

Planning and operations synergies. An open question of interest is the interaction between strategic and operational delay management approaches. Ideally, we would like to have synergistic interactions between the planning and operations stages. Thus far, few approaches have been developed with the express purpose of facilitating this synergy. It is of value to explore if (i) specific robust planning approaches are synergistic with specific dynamic recovery operations; (ii) the added flexibility for individual components of the system are synergistic as more information is revealed in real-time.

Emissions taxes. An interesting extension of our models in Chapter 4 is to capture emissions taxes in the fuel costs. In order to constrain the negative environmental impacts of aviation emissions and associated radiative forcing, market measures such as taxation may be imposed (possibly as part of a cap-and-trade scheme). Estimates of emissions cost incurred during the flight can be obtained from the flight plan or by using the value of emissions index [CT09]. Estimating CO_2 is a function of fuel burn alone, and NO_x estimates can be obtained using the thrust settings of the engines used in the fleet. A new trade-off frontier between fuel costs and delay costs is generated when taxes on emissions are introduced. An understanding of

the optimal trade-off points would help estimate the effect of the taxation measures on airline operating costs.

Enroute speed changes. Enroute flight planning can add to the operational flexibility provided by before-departure flight planning. If information is available about delays to flights to which passengers will connect, or about delays to flights whose passengers share a downstream flight, then we can further enhance airborne recovery and decrease costs. If we are able to incorporate enroute flight plan changes, especially as information about other flights in the system becomes available, we may be able to exploit the synergy demonstrated in Case Study 2 (§4.6.2) to further decrease delay costs.

Robust flight planning. An additional problem of interest is to build 'robust' flight plans that can protect against uncertainty. Our model works with assumptions on flight departure time from the origin airport, and arrival time at the destination airport; assuming that additional delay is not incurred due to taxi delays, en-route weather and congestion. One of the biggest sources of uncertainty in flight operations, however, is the wheels-up time, as described by Altus [Alta], followed by weather and enroute/arrival/ATC issues. Models to estimate the wheels-up time accurately would yield better estimates of outbound delay and in choosing the most cost-effective flight plan.

Flight planning and Air traffic flow management. A further avenue of research is to integrate flight planning with air traffic flow management. For example, because flight planning provides avenues for slack reallocation, we can ask the question - what is the best choice between a longer route without changing the departure time versus a later departure time and a shorter route? Answering such questions might provide further opportunities to alleviate congestion at slot-controlled airports.

Bibliography

[ABW02] A. Armacost, C. Barnhart, and K. Ware. Composite variable formulations for express shipment service network design. *Transportation Science*, 36:1–20, 2002.

[ACL10] S. AhmadBeygi, A. Cohn, and M. Lapp. Decreasing airline delay propagation by re-allocating scheduled slack. *IIE Transactions*, 42:478–489, 2010.

[Age00] Y. Ageeva. Approaches to incorporating robustness into airline scheduling. Masters thesis, Massachusetts Institute of Technology, 2000.

[Air] Air Transport Association. http://www.airlines.org/Economics/DataAnalysis/Pages/AnnualResultsWorldAirlines.aspx.

[Air08] Air Transport Association. Annual and Per-Minute Cost of Delays to U.S. Airlines. http://www.airlines.org/Economics/DataAnalysis/Pages/CostofDelays.aspx, 2008.

[Air10] Airliners.net. Define Long Haul. http://www.airliners.net/aviation-forums/general_aviation/read.main/2260348/, Viewed 10 Aug 2010.

[Alta] S. Altus. Dynamic cost index management in flight planning and flight following. http://www.agifors.org/studygrp/opsctl/2010/program.html, Presented at AGIFORS Airline Operations, 2010.

[Altb] S. Altus. Flight planning: The forgotten field in airline operations. http://www.agifors.org/studygrp/opsctl/2007/, Presented at AGIFORS Airline Operations, 2007.

[Ass08a] Associated Press. http://www.msnbc.msn.com/id/24410809/, May 1, 2008.

[Ass08b] Associated Press. http://www.usatoday.com/travel/flights/2008-05-02-slow-fuel_N.htm, May 2, 2008.

[AV04] T. Andersson and P. Varbrand. The flight perturbation problem. *Transportation Planning and Technology*, 27(2):91–117, 2004.

[Bar09a] C. Barnhart. Airline schedule optimization. In P. Belobaba, A. Odoni, and C. Barnhart, editors, *The Global Airline Industry*, chapter 7. John Wiley and Sons, United Kingdom, 2009.

[Bar09b] C. Barnhart. Irregular operations: Schedule recovery and robustness. In P. Belobaba, A. Odoni, and C. Barnhart, editors, *The Global Airline Industry*, chapter 9. John Wiley and Sons, United Kingdom, 2009.

[Baw73] V. S. Bawa. On chance constrained programming problems with joint constraints. *SIAM Journal of Optimization*, 19(11):1326–1331, 1973.

[BB05] S. Bratu and C. Barnhart. An analysis of passenger delays using flight operations and passenger booking data. *Air Traffic Control Quarterly*, 13:1–27, 2005.

[BB06] S. Bratu and C. Barnhart. Flight operations recovery: New approaches considering passenger recovery. *Journal of Scheduling*, 9(3):279–298, 2006.

[BBT$^+$01] G. Burrows, C. A. Brown, T. W. Thom, J. M. C. King, and J. Frearson. Real-time cost management of aircraft operations. *Management Accounting Research*, 12:281–298, 2001.

[BI62] A. Ben-Israel. *On some problems of mathematical programming*. PhD thesis, Northwestern University, 1962.

[BKM02] C. Barnhart, T. Kniker, and Lohatepanont M. Itinerary-based airline fleet assignment. *Transportation Science*, 36:199–217, 2002.

[Bra03] S. Bratu. *Airline Passenger On-Time Schedule Reliability: Analysis, Algorithms and Optimization Decision Models*. PhD thesis, Massachusetts Institute of Technology, 2003.

[Bry06] C. Bryant. Robust planning for effects-based operations. Masters thesis, Massachusetts Institute of Technology, 2006.

[BS03] D. Bertsimas and M. Sim. Robust discrete optimization and network flows. *Mathematical Programming Series B*, 98:49–71, 2003.

[BS04] D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

[BT03] D. Bertsimas and A. Thiele. A robust optimization approach to supply chain management. *Massachusetts Institute of Technology*, 2003.

[BT06] D. Bertsimas and A. Thiele. *M. Johnson, B. Norman and N. Secomandi, editors, Tutorials in Operations Research*. INFORMS, Hanover, MD, 2006.

[BTN99] A. Ben-Tal and A. Nemirovski. Robust solutions to uncertain programs. *Operations Research Letters*, 25:1–13, 1999.

[Bur09a] Bureau of Transportation Statistics. http://www.transtats.bts.gov/HomeDrillChart.asp, 2009.

[Bur09b] Bureau of Transportation Statistics. Airline Service Quality Performance Database. http://aspm.faa.gov/asqp/sys/, 2009.

[Bur09c] Bureau of Transportation Statistics. On-time Performance - Flight Delays at a Glance. http://www.transtats.bts.gov/HomeDrillChart.asp, 2009.

[Bur10] Bureau of Transportation Statistics. http://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp, 2010.

[CC59] A. Charnes and W. W. Cooper. Chance constrained programming. *Management Science*, 6(1):73–79, 1959.

[CC63] A. Charnes and W. W. Cooper. Deterministic equivalents for optimizing and satisficing under chance constraints. *Operations Research*, 11(1):18–39, 1963.

[CCS58] A. Charnes, W. W. Cooper, and G. Y. Symonds. Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil. *Management Science*, 4:235–263, 1958.

[CCT64] A. Charnes, W. W. Cooper, and G. L. Thompson. Critical path analyses via chance constrained and stochastic programming. *Operations Research*, 12(3), 1964.

[CK67] A. Charnes and M. Kirby. Some special p-models in chance-constrained programming. *Management Science*, 14(3):183195, November 1967.

[CT09] A. Cook and G. Tanner, editors. *The challenge of managing airline delay costs.* German Aviation Research Society and University of Belgrade, Faculty of Transport and Traffic Engineering. Conference on Air Traffic Management (ATM) Economics, University of Belgrade, Sep 2009.

[Dan55] G. B. Dantzig. Linear programming under uncertainty. *Management Science*, 1(3/4):197–206, 1955.

[DCT] N. Davis, J. Calusens, and S. Tiourine. Descartes project report, 2002. (Private correspondence).

[Epi07] L. Epiney. Corporate portfolio management: A robust optimzation perspective. Masters thesis, IBM Research, ZRL and Ecole Polytechnique Fdrale de Lausanne, 2007.

[ER02] M. Ehrgott and D. Ryan. Constructing robust crew schedules with bicriteria optimization. *Journal of Multicriteria Decision Analysis*, 11:139–150, 2002.

[EUR06] EUROCONTROL Route Availability Document RD Users Manual. https://www.cfmu.eurocontrol.int/j_nip/cfmu/gallery/content/public/library/handbook_supplements/handbook_supplements/docu_rad_users_manual_latest.pdf, 31 July 2006.

[Eur10] European Central Bank. Exchange Rates. http://www.ecb.int/stats/exchange/eurofxref/html/eurofxref-graph-usd.en.html, 2010.

[Fed09] Federal Aviation Administration. http://www.faa.gov/air_traffic/publications/media/FAA_Economic_Impact_Rpt_2009.pdf, December 2009.

[FJ48] M. Friedman and Savage L. J. The utility analysis of choices involving risks. *Journal of Political Economy*, 56:297–304, 1948.

[FM09] M. Fischetti and M. Monaci. Light robustness. In R. K. Ahuja, R. H. Mohring, and C. D. Zaroliagis, editors, *Robust and Online Large-Scale Optimization Models and Techniques for Transportation Systems*, chapter 3. Springer, Germany, 2009.

[Gal05] O. Gallay. Robustness and regulatory risk in supply chains. Masters thesis, IBM ZRL and Ecole Polytechnique Federale de Lausanne, 2005.

[GM08] H. J. Greenberg and T. Morrison. Robust optimization. In A. Ravi Ravindran, editor, *Operations and Management Science Handbook*, chapter 14, pages 14–1–14–33. CRC Press, Florida, 2008.

[GR68] S. K. Gupta and J. Rosenhead. Robustness in sequential investment decisions. *Management Science*, 15(2):B–18–29, 1968.

[Gre07] H. J. Greenberg. Representing uncertainty in decision support. *ORMS Today*, June 2007.

[HPS03] D. M. Hanssens, L. J. Parson, and R. L. Schultz. *Market Response Models: Econometric and Time Series Analyses*. International Series in Quantitative Marketing. Kluwer Academic Publishers, Massachusetts, second edition, 2003.

[Int08] International Air Transport Association. Annual Report. http://www.iata.org/about/ Documents/IATAAnnualReport2008.pdf, 2008.

[Int10a] International Air Transport Association. Jet Fuel Price Development. http://www. iata.org/whatwedo/economics/fuel_monitor/Pages/price_development.aspx, 2010.

[Int10b] International Air Transport Association. Traffic and Capacity Analysis. http://www. iata.org/whatwedo/economics/Pages/traffic_analysis.aspx, 2010.

[Jag74] R. Jagannathan. Chance-constrained programming with joint constraints. *Operations Research*, 22(2):358–372, 1974.

[JEC08] Joint Economic Committee. Your flight has been delayed again. URL:http://jec. senate.gov, 2008.

[Jepa] Jeppesen, a Boeing Company. http://www.jeppesen.com/index.jsp.

[Jepb] Jeppesen Commercial and Military Aviation Operations Services. Jeppesen technology services JetPlan engine user manual. (Private correspondence).

[Jia06] H. Jiang. *Dynamic Airline Scheduling and Robust Airline Schedule De-Peaking*. PhD thesis, Massachusetts Institute of Technology, 2006.

[JR73] R. Jagannathan and M. R. Rao. A class of nonlinear chance-constrained programming models with joint constraints. *Operations Research*, 21(1):360–364, 1973.

[KPU02] P. Krokhmal, J. Palmquist, and S. Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *The Journal of Risk*, 4(2):11–27, 2002.

[KT09]   Ban Kawas and Aurlie Thiele. A log-robust optimization approach to portfolio management. *OR Spectrum*, pages 1–27, 2009. 10.1007/s00291-008-0162-3.

[Kuc09]   S. Kucukyavuz. On mixing sets arising in chance-constrained programming. Submitted, 2009.

[LCB06]   S. Lan, J.P. Clarke, and C. Barnhart. Planning for robust airline operations: Optimizing aircraft routings and flight departure times to minimize passenger disruptions. *Transportation Science*, 40(1):15–28, 2006.

[Lin03]   J. Linderoth. Chance constrained programming. Lehigh University Lecture Notes IE495 Lecture 22, 2003.

[LR02]   G. L. Lilien and A. Rangaswamy. *Marketing Engineering: Computer-Assisted Marketing Analysis and Planning.* International Series in Quantitative Marketing. Prentice Hall, Englewood Cliffs, NJ, second edition, 2002.

[LS82]   J.-C. Larreche and V. Srinivasan. Stratport: a model for the evaluation and formulation of business portfolio strategies. *Management Science*, 28(9):979–1001, 1982.

[Mar52]   H. M. Markowitz. Portfolio selection. *Journal of Finance*, 7:77–91, 1952.

[Mar07]   L. Marla. Robust optimization for network-based resource allocation problems. Masters thesis, Massachusetts Institute of Technology, 2007.

[MB09]   L. Marla and C. Barnhart. Robust optimization: Insights from aircraft routing. Working Paper, Massachusetts Institute of Technology, 2009.

[MPRS09] L. Marla, E. Pratsini, A. Rikun, and G. Stauffer. Robust modeling and planning: Insights into robust optimization methodologies from three industrial applications. Working Paper, IBM Research and Massachusetts Institute of Technology, 2009.

[MVZ81]   J. M. Mulvey, R. J. Vanderbei, and S. A. Zeinos. Robust optimization of large-scale systems. *Operations Research*, 43:34–56, 1981.

[MW65]   B. Miller and H. Wagner. Chance constrained programming with joint constraints. *Operations Research*, 13(6):930–945, 1965.

[NS06a]   A. Nemirovski and A. Shapiro. Convex approximations of chance-constrained programs. *SIAM Journal of Optimization*, 17(4):969–996, 2006.

[NS06b]   A. Nemirovski and A. Shapiro. Scenario approximations of chance constraints. In G. Calafiore and F. Dabbene, editors, *Probabilistic and Randomized Methods for Design under Uncertainty*. Springer, 2006.

[RJN04]   J. M. Rosenberger, E. L. Johnson, and G. L. Nemhauser. A robust fleet assignment model with hub isolation and short cycles. *Transportation Science*, 38:357–368, 2004.

[RU00]   R.T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000.

[Sak06]   P. Sakamoto. UAV Mission Planning Under Uncertainty. Masters thesis, Massachusetts Institute of Technology, 2006.

[Sav51]   L. J. Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46:55–67, 1951.

[Sav54]   L.J. Savage. *The Foundations of Statistics*. New York: Wiley, 1954.

[SJKN05] A. Schaefer, E. Johnson, A. Kleywegt, and G. L. Nemhauser. Airline crew scheduling under uncertainty. *Transportation Science*, 39:340–348, 2005.

[SK06]   S. Shebalov and D. Klabjan. Robust airline crew pairing: Move-up crews. *Transportation Science*, 40(3):300–312, 2006.

[SMW]   C. Saint-Martin and G. Wagner. Optimizing the use of cost index at Air Canada. http://www.agifors.org/studygrp/opsctl/2009/, Presented at AGIFORS Airline Operations, 2009.

[Soy73]   A. L. Soyster. Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations Research*, 21, 1973.

[Tho09]   Thomas Cook Airlines. http://www.thomascook.com/brands/thomas-cook-airlines/airline-faqs/baggage-restrictions/, Viewed July 27 2009.

[TP09]   A. Tomer and R. Puentes. Expect delays: An analysis of air travel trends in the United States. *Washington: Brookings Institution*, 2009.

[TTE09]   A. Thiele, T. Terry, and M. Epelman. Robust linear optimization with recourse. Revision submitted, *Naval Research Logistics*, March 2009.

[TYB00]   Benjamin G. Thengvall, Gang Yu, and Jonathan F. Bard. Balancing user preferences for aircraft schedule recovery during irregular operations. *IIE Transactions*, 32(3):181–193, 2000.

[Vaa]   B. Vaaben. Disruption management passenger re-accommodation strategies. http://www.agifors.org/studygrp/opsctl/2009/, Presented at AGIFORS Airline Operations, 2009.

[YB06]   J. Yen and J. Birge. A stochastic programming approach to the airline crew scheduling problem. *Transportation Science*, 40:3–14, 2006.

[YY96]   S. Yan and A. Young. A decision support framework for multi-fleet routing and multi-stop flight scheduling. *Transportation Research-A*, 30(5):379–398, 1996.