

EQUITABLE AND EFFICIENT COORDINATION OF TRAFFIC FLOW MANAGEMENT PROGRAMS

DOUGLAS FEARING

Massachusetts Institute of Technology, Operations Research Center

FACULTY ADVISORS:

CYNTHIA BARNHART

Massachusetts Institute of Technology, Department of Civil and Environmental Engineering

DIMITRIS BERTSIMAS

Massachusetts Institute of Technology, Sloan School of Management

CONSTANTINE CARAMANIS

The University of Texas at Austin, Department of Electrical and Computer Engineering

Abstract: When air traffic demand is projected to exceed capacity, the FAA implements Traffic Flow Management programs. Independently, these programs maintain a first-scheduled, first-served invariant, which is the accepted standard of fairness within the industry. Coordinating multiple programs requires a careful balance between equity and efficiency. In our work, we first develop a fairness metric to measure deviation from first-scheduled, first-served. Next, we develop an IP formulation that attempts to directly minimize this metric. We further develop an exponential penalty approach and show that its computational performance is far superior and its trade-off between delay and fairness compares favorably. In our results, we demonstrate the effectiveness of these models using regional and national scenarios. Additionally, we demonstrate that the exponential penalty approach exhibits exceptional computational performance, implying practical viability. Our tests suggest that this approach could lead to system-wide savings on the order of \$100 million per year.

1 Introduction

The Federal Aviation Administration (FAA) and the airline industry face tremendous challenges due to unexpected weather-induced reductions in system capacity and resulting delays. During the 12-month period ending in September of 2008, 138 million minutes of system delay led to an estimated \$10 billion in costs for US airlines [1]. This figure does not include passenger delay costs, estimated to increase total costs by \$4.5 billion. To put this in perspective, the Air Transportation Association's 2008 Annual Report lists total profits for US airlines of approximately \$3 billion for the 2006 operating year and \$5 billion for the 2007 operating year [2].

Delays over the National Air Space are projected to out-pace increases in overall traffic significantly. A highly referenced article in *The Economist* projects as much as a 5% increase in delay

for each corresponding 1% increase in demand at 2000 demand levels [4]. Increasing capacity by building additional runways and airports is logistically complex due to cost, space limitations, and environmental regulations. Additionally, projects of this type often take a decade or more to plan and complete. Thus, in concert, it is critical to consider systematic tools to improve operational efficiency. In Section 1.4 we detail the contributions of this paper; but to put it in an appropriate context, we now turn to a brief discussion of the existing TFM tools and a non-exhaustive literature review.

1.1 Traffic Flow Management

Traffic Flow Management (TFM) refers to a set of strategic practices utilized by the FAA to ensure safe operations while attempting to minimize costs associated with delay. TFM activities occur on the day of operations and generally impact a significant subset of airline traffic (e.g. all flights into a major airport). According to data received from Metron Aviation, TFM activities account for approximately 20% of all air transportation delays (see calculations in Section 4.5 for details). Based on factors such as number of runways, runway configuration, scheduled personnel coverage, and weather forecasts, the FAA determines maximum capacities for resources in the US air transportation system. These resources include arrival runways, departure runways, and air sectors in the National Airspace System (NAS). TFM programs are initiated only when there are expected to be significant imbalances between demands and capacities, such as in the midst of a severe storm. Minor to moderate inconsistencies between capacity and demand are otherwise resolved through localized Air Traffic Control (ATC) techniques. Since the air traffic controllers' strike in 1981, the primary tool the FAA has used for TFM has been the Ground-Delay Program (GDP). In a GDP, the FAA controls the arrival rate into a reduced-capacity airport by coordinating the departure times for impacted flights. The goal is to allow each aircraft to proceed safely to its destination with minimal airborne delay. The recently introduced Airspace-Flow Program (AFP) is similar to a GDP. The FAA uses an AFP to control the arrival rate into a Flow Constrained Airspace (FCA), e.g. a reduced-capacity air segment of the NAS. To understand the prevalence of these programs, in Figure 1 we provide a table listing the number of days from April 2007 to April 2008 where the corresponding number of GDPs and AFPs were enacted. References [18] and [5] provide further details regarding the TFM problem and its extensions.

Number of AFPs	Number of GDPs											Total
	1	2	3	4	5	6	7	8	9	10	11	
1	11	14	7	5	7	8	4	2	2	0	1	61
2	6	10	11	12	10	6	6	2	2	3	0	68
3	3	2	3	0	1	0	1	0	0	0	0	10
4	0	0	0	0	0	0	0	1	0	0	0	1
Total	20	26	21	17	18	14	11	5	4	3	1	140

Table 1: Number of days from April 2007 to April 2008 with the corresponding number of TFM programs of each type

These tools, namely, GDPs and AFPs, are used in concert with a three-stage, collaborative approach to decision-making. In the first stage, the FAA allocates arrival slots to airlines by applying the *Ration-By-Schedule* (RBS) method for each TFM program. In RBS, arrival slots are allocated according to the original schedule ordering, as is described in detail in the following section. Although fairness is a subjective criterion, the RBS approach is generally considered fair within the airline industry because it maintains a first-scheduled, first-served invariant. In the second stage, airlines undertake a process known as airline recovery. Each airline is allowed to make changes to the schedule within the context of the slots allocated to it. For instance, an airline can swap arrival slots for two of its own flights as long as the swap is consistent with the scheduled departure times. Additionally, an airline can choose to cancel flights due to operational constraints on aircraft routing, crew assignments, etc. In the third stage, the FAA accepts the changes proposed by all airlines. These changes, when merged together, constitute a capacity-feasible schedule because each airline is only allowed to make changes within the set of slots allocated to it. Subsequently, the FAA attempts to improve the schedule by filling in any gaps created by cancellations. This procedure is known as compression and is described in detail in [26]. After compression, the new schedule proposal is sent out to the airlines and the process is repeated as necessary.

1.2 Coordinating Multiple Programs

In RBS, arrival slots for a single resource, either an airport in a GDP or flow constrained airspace (FCA) in an AFP, are allocated to flights according to the original schedule order. For FCAs, the scheduled arrival order is calculated based on the scheduled departure plus the expected travel time to reach the FCA. Once the controlled arrival slots have been allocated for a resource, each affected flight receives a corresponding Controlled Time of Departure (CTD) from its origin, converting the

allocated arrival slot into departure delay at the departure airport.

When multiple TFM programs are implemented concurrently, applying RBS for each one independently may lead to a single flight receiving conflicting CTDs (e.g. from a GDP and one or more AFPs). In order to resolve these conflicts, the FAA prioritizes one of the assigned CTDs for each flight. If the flight intersects a GDP, the GDP-based CTD will take precedence, otherwise the first initiated AFP-based CTD will be used. We refer to this conflict resolution approach as multi-resource RBS.

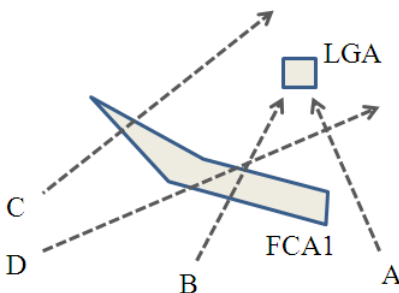


Figure 1: Visual representation of flight routes for flights A, B, C, and D

Consider the following example based on the four flight routes displayed in Figure 1 with planned schedule details listed in Table 2. At 17:00, an AFP is initiated for FCA1 with a controlled arrival rate of 1 flight every 5 minutes from 18:40 until 19:00. Similarly, at 17:00, a GDP is initiated for LGA, with arrivals into LGA restricted to 1 flight every 10 minutes from 18:55 until 19:15. Note that the time a TFM program is initiated determines which flights are impacted, since flights already in the air at the time of initiation are exempted from the program. Performing RBS for each resource independently leads to the CTDs listed in Tables 3 and 4. In this case, flight B receives conflicting CTDs (17:15 from the AFP at FCA1 and 17:25 from the GDP at LGA). Thus, according to multi-resource RBS, flight B will be given a CTD of 17:25 (since the GDP at LGA takes precedence). This leads to the controlled schedule listed in Table 5.

There is one important thing to note regarding this example. In the example it is impossible to simultaneously satisfy first-scheduled, first-served for each resource *and* minimize system delay.

Flight	Departure	FCA1	LGA
A	17:45	—	18:55
B	17:15	18:40	18:55
C	18:00	18:45	—
D	18:15	18:45	—

Table 2: Planned departure and arrival times for flights A, B, C, and D

Flight	Slot	CTD
B	18:40	17:15
C	18:45	18:00
D	18:50	18:20

Flight	Slot	CTD
A	18:55	17:45
B	19:05	17:25

Table 3: RBS CTDs for FCA1

Table 4: RBS CTDs for LGA

That is, the controlled schedule must either deviate from first-scheduled, first-served or incur excess delay. This simple example illustrates the general trade-off that exists between fairness and efficiency in the multi-resource or network setting. For additional examples of this type, the reader is encouraged to review [16].

1.3 Literature Review

The first thorough review of the TFM problem is provided by Odoni (1987 [19]). Over time, two leading research paths have emerged. The first has focused on single resource approaches. These approaches are applicable for a single TFM program or multiple non-conflicting TFM programs. The second research path has focused on network-wide, or multi-resource, approaches to TFM, where typically all airports and air sectors are placed under the FAA’s control. Research into single resource approaches has gained more traction within the industry due primarily to two reasons: the inclusion of collaboration and equity considerations, and also computational tractability (i.e.,

Flight	CTD	FCA1	LGA
A	17:45	—	18:55
B	17:25	18:50	19:05
C	18:00	18:45	—
D	18:20pm	18:50	—

Table 5: Controlled departure and arrival times for flights A, B, C, and D

computations involving a full day of flights for a single resource run quickly). Arguably, it is because of failures in these areas that research into network approaches has gained less traction. Indeed, few network formulations have been able to effectively consider equity or collaboration. Thus far, network research has primarily focused on computational efficiency, due to the inherent complexity of a network-wide TFM model.

For the single-resource TFM problem, deterministic, static-stochastic, and dynamic-stochastic versions of the problem were first formulated in the early 1990s (see [24], [21], and [22]). More recent research has extended these models to incorporate collaboration and equity (see [6], [26], and [15]). Vossen et al. define a measure of equity for the single-resource TFM problem and calculate the inequity associated with flight exemptions (2003 [27]). Chang et al. describes the collaborative decision-making (CDM) approach with updated equity considerations that was incorporated into the FAA’s GDP in the late 1990s (2001 [11]). In a recent paper, Brennan describes how the CDM-enhanced GDP approach has been extended to the AFP (2007 [10]).

On the multi-resource side, Vranas, Bertsimas, and Odoni develop the first integer programming formulation for the multi-airport GDP (1994 [28]). Bertsimas and Stock Patterson extend this formulation to the full air traffic system using a novel variable definition (1998 [8]). Subsequent research has primarily focused on computational efficiency and the incorporation of rerouting constraints (see [13], [3], and [7]). Lulli and Odoni discuss the inequities inherent in a network formulation of the TFM problem (2007 [16]), which provides a critical backdrop for our work.

In this paper, we develop integer programming formulations for the multi-resource TFM problem that incorporate fairness considerations. Unlike other network approaches, instead of including all airports and sectors, we choose to restrict the problem to the coordination of multiple, conflicting TFM programs. We believe that by considering this restricted problem, our work will help bridge the gap between the two divergent research paths described above.

1.4 Contributions

The contributions of this paper fall into three categories: (i) fairness modeling, including developing a fairness metric, and analysis of the resulting fairness properties and the relationship to current

industry standards; (ii) two optimization formulations minimizing a delay metric developed in (i), that are computationally tractable for national-scale TFM problems; (iii) computational results and analysis on large-scale (regional and national) instances. The structure of the paper follows these three main points. In Section 2 we discuss inherent fairness properties and develop our fairness metric. In Section 3 we develop two integer programming formulations; and finally in Section 4 we provide and discuss our computational results.

The starting point for our formulations is the model developed in [8] and the first-scheduled, first-served concept of fairness inherent in RBS as described in Section 1.2. RBS has three salient features. First, it is algorithmically trivial to implement and has a linear running time with respect to the number of flight steps. Thus, the approach can be scaled to arbitrarily large problems. Second, for an isolated GDP or AFP, the RBS method always leads to a solution that minimizes the minutes of system delay [26]. Third, it is the industry accepted notion of fairness, endorsed by the primary stakeholders, i.e., the FAA and the airlines. Significantly, to the best of our knowledge, no one has developed an optimization-based extension of RBS for network or multi-resource problems. In particular, as should be apparent based on the example in Section 1.2, any simple extension of RBS will fail on a very important front: it will no longer provide delay-optimality guarantees as in the single-resource case. This is to be expected, since fairness may in general come at the expense of increased aggregate delays. The main modeling contribution of this paper is precisely to make up for this deficiency. Specifically, we desire a formulation for fairness that has the following properties:

- (1) In the single resource setting, it should reduce to (the accepted standard) RBS, which as discussed above, is delay-optimal in this case;
- (2) Since there will typically be a trade-off between aggregate system delay and any flight-based fairness criterion, the formulation should essentially consider a bi-criterion approach, enabling the efficient study of the trade-off curve between the two;
- (3) The formulation should compare favorably to the approach currently utilized in practice for the multi-resource setting.

2 Ration-by-Schedule and Fairness

As discussed in the introduction, understanding and incorporating industry-accepted views of fairness has been a significant road block to the implementation of optimization-based techniques for managing TFM programs. One of the more significant challenges is that the first-scheduled, first-served concept of fairness underlying RBS does not directly extend to the setting where a single flight may interact with multiple TFM programs (e.g. a GDP plus one or more AFPs). With this in mind, we turn our attention to developing a measure of overall schedule fairness that i) is consistent with first-scheduled, first-served in a single resource environment, and ii) naturally extends to the setting where there are interactions between TFM programs.

To provide additional context, we first illustrate problems with the multi-resource RBS approach utilized in practice to resolve conflicts between conflicting TFM programs (i.e. GDPs and AFPs). The main advantage to this approach is that it is a simple extension of RBS in the single-resource setting and thus the resulting schedule is similar to the single-resource RBS schedules. Unfortunately, this simplicity can also lead to significant costs in terms of efficiency and therefore total delays. Next, we describe the properties that we feel should underly any measure of schedule fairness in a multi-resource setting. We use simple examples to demonstrate the importance and significance of the properties we outline. Last, we develop a robust measure of schedule fairness that incorporates these properties. The purpose of this metric is to evaluate the relative fairness of competing scheduling approaches.

2.1 Problems With Multi-Resource Ration-by-Schedule

One downside of the current approach is that AFP capacities, specified in terms of controlled arrival rates, may be (and often are) violated. By examining the controlled schedule from the example in Section 1.2 (Table 5) we see that two flights (B and D) are scheduled to arrive at FCA1 simultaneously even though the controlled arrival rate was established at 1 flight every 5 minutes. It is difficult to measure how much this impacts efficiency because in practice, AFPs are constructed in a trial-and-error fashion. That is, the parameters of each AFP, such as duration and arrival rate, are tweaked until the end result satisfies subjective criteria for safety. Additionally, with an AFP,

traffic flow is controlled through a line or region of air space which may be hundreds of miles long, thus two flights that arrive at the same time may be very far apart geographically. Nonetheless, the multi-resource RBS approach makes it difficult, if not impossible, to precisely control traffic flow through the air. Additionally, as air traffic congestion continues to increase, airspace controls are expected to become more common, only exacerbating this problem.

Another significant issue with the current approach is that it may lead to inefficient resource utilization. By construction, GDP-based arrival slots will always be fully utilized, but conflicting CTDs may lead to gaps in FCA utilization. As described above, if a flight receives conflicting CTDs from a GDP and AFP, the GDP-based CTD will take precedence. If the GDP-based CTD is later than the AFP-based CTD, this could lead to a gap in the FCA schedule. Some of these gaps may be filled by the FAA through a subsequent scheduling step called compression (see [26] for details), but often inefficiencies remain. Again, based on the the example in Section 1.2, we see that the 18:40 arrival slot for FCA1 is unused. This slot is unable to be filled by compression since i) the schedule for flights A and B into LGA is fixed, and ii) flights C and D cannot be released earlier than their planned departure times. Note that if we swap flights A and B into LGA, flight B then uses the 18:40 slot into FCA1 which frees up an 18:50 slot into FCA1. This slot could then be used by a later flight.

The last issue with the FAA's current approach is that the expected RBS order for FCAs is violated based on the resolution of conflicting CTDs. In the example above, flight B was originally scheduled to arrive at FCA1 first, but was instead scheduled second after resolution of the conflicting CTDs. Though the RBS order is violated in this case, it is likely not a fairness issue since LGA is a more restricted resource along flight B's route. On the other hand, consider two flights, the first of which passes through a severely constrained FCA en route to a more mildly constrained arrival airport, and the second of which just passes through the FCA. Because the GDP-based CTD will take precedence for the first flight, the flight is able to avoid the impact of the more severe AFP. The second flight will be impacted solely by the AFP and thus, receive significantly greater, and therefore inequitable, delays. As should be apparent from this example, we can construct scenarios wherein the resulting FCA schedule is arbitrarily unfair.

2.2 Principles for Measuring Fairness

The challenge with incorporating fairness into the multi-resource setting is that the link between original schedule order and delay optimality breaks down when one or more flights are included in multiple TFM programs. Thus, in a multi-resource setting we need to make a trade-off between fairness relative to the original schedule order and efficiency in terms of total system delay. In order to find the appropriate trade-off, we need a method to measure the relative (un)fairness of competing schedules.

The concept of fairness is by nature subjective and often domain-specific. Even within air traffic, there are many plausible ways to measure schedule fairness, each leading to different results. For example, in a single-resource setting, one measure of fairness implemented in practice is the number of slots a flight deviates from its initial order position (e.g. if a flight scheduled to arrive 4th instead is allocated the 12th arrival slot, we would say that flight's schedule was unfair by 8 positions). Unfortunately, in the multi-resource setting, using position-based metrics without considering delay can lead to imbalances in the fairness penalty incurred between resources. Other proposals include measuring schedule fairness by comparing average or maximum flight delays between airlines. This type of measure ignores variation in congestion along flight routes, and thus is also problematic in the multi-resource setting. In this section, we describe properties that we feel are critical for measuring fairness in the multi-resource setting. These properties are motivated primarily as extensions of the successful properties of RBS in the single-resource environment. In the following section, we use these properties to obtain a multi-resource fairness deviation metric.

Property 1: the measure of schedule fairness should be determined relative to the original schedule ordering. The success of RBS in the single-resource setting has led the concept of first-scheduled, first-served to be widely accepted by airlines and the FAA.

Property 2: the measure of schedule fairness should be applicable to a single flight as well as the overall schedule. That is, the measure should be able to determine the amount each flight's schedule varies from first-scheduled, first-served.

Property 3: the unit of fairness deviation and its relative magnitude should be consistent between resources. In a single-resource setting, position-based deviation is an accepted measure of fairness deviation. In the multi-resource setting, this is confounded due to varying congestion levels between resources. An 8-position delay (going from 4th to 12th) could mean 30 minutes of delay in a low-capacity airport, but only 10 minutes of delay in a higher capacity one.

Property 4: there should be no fairness penalty for a flight receiving as much delay as its original schedule order would indicate for any resource along its route. Loosely speaking, this means that a flight should never expect to receive less delay than that caused by the most congested resource along its route.

Property 5: the measure of a flight’s deviation from the original schedule should be calculated relative to the total delay assigned to the flight (ground delay plus air delay), not intermediate arrival times into controlled resources. This property is relevant if the scheduling approach allows both ground delays (by assigning CTDs) and en route delays (by mandating air speed reductions or arrival queuing) to be assigned. In practice, the schedule created by the FAA using RBS assumes that a flight will receive no delays en route and only assigns ground-delay through CTDs. En route delays are subsequently managed by air traffic controllers en route or at the arrival airport. Network TFM models, such as the one described in [8], consider both of these problems simultaneously in order to improve efficiency and predictability.

2.3 Time-Order Deviation Metric

With these properties in mind, we now develop a measure for evaluating fairness of a controlled schedule. We will refer to this measure as the time-order deviation metric.

First, we define a flight’s expected delay relative to a controlled resource along its route as the delay the flight would expect to incur if there were no other controlled resources along the route. For example, if flight A is originally scheduled to arrive 4th into LGA at 19:00 and in the controlled schedule the 4th flight arrives into LGA at 19:30, we would say that flight A has a 30-minute expected delay into LGA. Note that in the controlled schedule we describe the 4th flight might or might not be the same as flight A, the 4th flight in the original schedule.

For each flight, we define its time-order deviation as the amount its total delay exceeds the maximum expected delay along the its route. The maximum expected delay is the maximum expected delay over all controlled resources in the flight’s route. In the case that the maximum expected delay exceeds the flight’s total delay, we set the time-order deviation equal to zero. That is, a schedule is not more fair if a flight arrives earlier than expected, even though this might reduce the overall delay.

Example: Consider a flight scheduled to depart from BOS at 18:00, arrive at the boundary of FCA1 at 18:45 and land at LGA at 19:15. We construct an AFP for FCA1 and a GDP for LGA such that pre-disruption, the flight is scheduled to be the 4th controlled flight into FCA1 and the 3rd controlled flight into LGA. In the resulting schedule, the flight is given a CTD of 18:25 (i.e. 25 minutes of ground delay). In order to calculate the time-order deviation of this flight, we need to know the order of flights into FCA1 and LGA based on the controlled schedule. Based on the partial controlled schedule orders listed in Tables 6 and 7, we can calculate the time-order deviation as follows. First, we calculate the flight’s expected delay into FCA1 as the arrival time of the 4th flight into FCA1 (18:55) minus the flight’s original scheduled arrival time into FCA1 (18:45), which equals 10 minutes. Next, we calculate the flight’s expected delay into LGA as the arrival time of the 3rd flight into LGA (19:20) minus the flight’s original scheduled arrival time into LGA (19:15), which equals 5 minutes. The referenced arrival times in Tables 6 and 7 are highlighted in *bold italics*. Thus, the maximum expected delay for the flight is 10 minutes from FCA1. The total delay for the flight is 25 minutes, so the time-order deviation for this flight is 15 minutes. In Tables 6 and 7 the rows corresponding to the controlled schedule for the original flight have been marked with an *, though they are not used directly in the calculation of the flight’s time-order deviation.

Order	FCA1 Arrival
1	18:35
2	18:45
3	18:50
4	18:55
5	19:00
6*	19:05

Order	LGA Arrival
1	19:00
2	19:10
3	19:20
4	19:30
5*	19:40

Table 6: Controlled flight order for FCA1

Table 7: Controlled flight order for LGA

We define the time-order deviation for a controlled schedule as the sum of the time-order deviations for each flight represented in the schedule. As expected, this measure of fairness satisfies all of the principles laid out in the previous section. That is, i) time-order deviation is calculated relative to the original schedule order, ii) the measure can be applied for each flight in the controlled schedule, iii) the unit of measure (i.e. time) is consistent between resources, iv) the measure is calculated relative to the most restricted resource along each flight’s route (i.e. relative to the maximum expected delay), and v) the measure is based on the total delay and not intermediate arrival times. Note that for a single controlled resource, or for a set of independent controlled resources (such as multiple GDPs), the time-order deviation metric achieves 0 if the controlled schedule matches the schedule resulting from performing RBS independently for each controlled resource.

3 Optimization Approaches

In this section, we describe two integer programming formulations whose solutions describe the air and ground delay that should be assigned to each flight. Each formulation allows for the flexible trade-off between a delay term and a fairness term in the minimization objective. In the first model, the fairness term is a convex approximation of the fairness metric developed in the previous section. We call this the Time-Order Deviation Approximation (TODA) model. Next, we use an exponentially growing delay penalty to enforce fairness. We see that this approach has considerable computational advantages, yet sacrifices little in terms of fairness achieved according to time-order deviation. We refer to this model as the Ration-by-Schedule Exponential Penalty (RBS-EP) model.

In Section 3.1, we develop the common notation as well as define the input data used in both our formulations. Then in Section 3.2, we provide the portion of the optimization formulation that is common to both our TODA and RBS-EP models. Section 3.3 provides the formulation for the TODA model, and Section 3.4 the formulation for the RBS-EP model. Finally we discuss some network issues relating to aircraft connectivity in Section 3.5.

3.1 Data and Notation

We consider a set of discretized time intervals $\mathcal{T} = \{0, \dots, T - 1\}$, where T represents the end of the day, and each interval is defined to have equal duration, typically either 5 minutes or 15 minutes. We consider a set of controlled resources, \mathcal{R} , which will typically include arrival airports (for GDPs) and FCAs (for AFPs). All system resources that are not capacity-controlled provide no binding constraints on the system and are excluded from \mathcal{R} . For each resource, $r \in \mathcal{R}$, and each time interval, $t \in \mathcal{T}$, we specify a capacity of b_{rt} , which can be thought of as either an allowable arrival rate or as a maximum occupancy over the interval. For GDPs and AFPs, resource capacities are specified in terms of an allowable arrival rate.

Additionally, we consider a set of flight legs, \mathcal{F} . For each flight leg, $f \in \mathcal{F}$, we define its controlled flight plan to be the sequence of controlled resources it is scheduled to utilize over the course of the flight. For instance, consider the flight from Boston Logan International Airport (BOS) to New York John F. Kennedy Airport (JFK) in Figure 2 with TFM programs in place at FCA1 and JFK. For this flight, the controlled flight plan would be a sequence containing FCA1 followed by the arrival resource for JFK. Notationally, we let $|f|$ represent the number of steps in the controlled flight plan for flight f , and we use the shorthand $\mathcal{I}(f)$ to represent the set of step indices $\{1, \dots, |f|\}$. For each step in the controlled flight plan, in addition to the resource, r , we must specify the earliest start time, α , the processing time, δ , and the supported types of delay preceding the step, $\psi \subset \{\text{G}, \text{A}\}$. That is, $\alpha \in \mathcal{T}$ represents the first time interval at which the step can be scheduled and $\delta \in \mathbb{N}^+$ the number of time intervals the step needs to be processed (i.e. landing time at an arrival airport or dwell time in an occupancy-controlled FCA). We let $\{\text{G}\}$ represent ground delay and $\{\text{A}\}$ represent airborne delay. In the case $\psi = \emptyset$, we do not allow any delay prior to the specified step. To maintain consistency with current practice, we would let $\psi = \{\text{G}\}$ for the first step in a controlled flight plan and $\psi = \emptyset$ for subsequent steps. In this case, as with a GDP or AFP, we only assign ground delay prior to departure and assign no further delay en route. Notationally, we let $r(f, i)$, $\alpha(f, i)$, $\delta(f, i)$, and $\psi(f, i)$ refer to the appropriate values for step i of the flight plan for flight f . In our formulation, $\alpha(f, i + 1) - \alpha(f, i)$ represents the minimum number of time intervals between the starts of steps i and $i + 1$. Thus, we require $\alpha(f, i) + \delta(f, i)$ to be less than or equal, not equal, to $\alpha(f, i + 1)$. For example, if the resources for two sequential

steps are not geographically adjacent $\alpha(f, i + 1) - \alpha(f, i) - \delta(f, i)$ would represent the travel time between boundaries of the two resources. In Table 8, we provide sample values for these fields based on the example described above (see Figure 2) with 5 minute time intervals starting at 05:00.

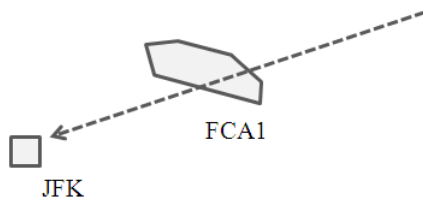


Figure 2: BOS \rightarrow JFK flight path intersecting two controlled resources, FCA1 and JFK

Scheduled Time	i	r	α	δ	ψ
07:35	1	FCA1	31	2	{G}
08:15	2	LGA	39	1	\emptyset

Table 8: Data values for BOS to JFK controlled flight plan based on a 07:00 initial departure

For each resource r , we assume there is a preferred ordering of tasks (i.e. flight steps) corresponding to the original schedule. That is, for resource r we would prefer to start the task indexed by j before the task indexed by $j + 1$, where each task corresponds to a flight step, (f, i) . Using this notation, we let $j(f, i)$ represent the task index of flight step (f, i) for the corresponding resource, $r(f, i)$. Additionally, we let $RBS(r, j)$ represent the time interval task j would be assigned based on performing single-resource RBS for r .

For the aggregate delay cost term in our model we assume a ground delay cost of 1, and let our airborne delay cost be $c_{a/g}$, where $c_{a/g}$ represents the ratio of airborne to ground delay costs. We assume that $c_{a/g}$ is larger than 1, due to fuel costs, depreciation, maintenance, and safety related issues.

Summarizing the above, we have:

- \mathcal{T} = set of discrete time intervals;
- \mathcal{R} = set of capacity-controlled resources;
- b_{rt} = capacity of resource r over time interval t ;
- \mathcal{F} = set of flights;
- $|f|$ = number of steps in controlled flight plan for flight f ;
- $\mathcal{I}(f)$ = set of step indices in controlled flight plan for flight f ;
- $r(f, i)$ = resource required by flight step i for flight f ;
- $\alpha(f, i)$ = earliest start time for flight step i for flight f ;
- $\delta(f, i)$ = processing time of flight step i for flight f ;
- $\psi(f, i)$ = supported delay types preceding flight step i for flight f ;
- $c_{a/g}$ = relative cost per time unit of airborne delay;
- $J(r)$ = number of tasks (i.e. flight steps) assigned to resource r ;
- $\mathcal{J}(r)$ = set of task indices $\{1, \dots, J(r) - 1\}$;
- $j(f, i)$ = the task index of flight step i for flight f ; and
- $RBS(r, j)$ = RBS start interval for task j of resource r ;

3.2 Model Foundation

In this section, we describe the components of the multi-resource TFM formulation that provide the foundation for the two models we develop. This formulation is derived from the Bertsimas, Stock Patterson (BSP) network TFM model [8].

3.2.1 Decision Variables

For both formulations, we use the following variable definitions:

$$y_{fit} = \begin{cases} 1 & \text{if flight plan step } i \text{ for flight } f \text{ has started by time } t; \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

3.2.2 Constraints

We first ensure that the sequence $[y_{fi0} \cdots y_{fi(T-1)}]$, which we refer to as $[y_{fi}]$, is monotonically increasing:

$$y_{fit} \leq y_{fi(t+1)} \quad \forall f \in \mathcal{F}, \forall i \in \mathcal{I}(f), \forall t \in \{0 \dots T-2\}. \quad (1)$$

Next, we guarantee that each flight step is scheduled and that no flight step is scheduled before its minimum start time:

$$y_{fi(T-1)} = 1 \quad \forall f \in \mathcal{F}, \forall i \in \mathcal{I}(f). \quad (2)$$

$$y_{fi(\alpha(f,i)-1)} = 0 \quad \forall f \in \mathcal{F}, \forall i \in \mathcal{I}(F) \text{ s.t. } \alpha(f,i) > 0. \quad (3)$$

We also enforce the appropriate order between flight steps in a controlled flight plan as follows:

$$y_{f(i+1)t} \leq y_{fi(t-\alpha(f,i+1)+\alpha(f,i))} \quad \forall f \in \mathcal{F}, \forall i \in \mathcal{I}(f) \setminus \{|f|\} \text{ s.t. } \psi(f,i+1) \neq \emptyset, \forall t \in \mathcal{T}. \quad (4)$$

$$y_{f(i+1)t} = y_{fi(t-\alpha(f,i+1)+\alpha(f,i))} \quad \forall f \in \mathcal{F}, \forall i \in \mathcal{I}(f) \setminus \{|f|\} \text{ s.t. } \psi(f,i+1) = \emptyset, \forall t \in \mathcal{T}. \quad (5)$$

Constraints (4) allow delay to be introduced between steps i and $i+1$, whereas constraints (5) do not. The last set of constraints is to ensure that resource capacities are not violated:

$$\sum_{\{(f,i):r(f,i)=r\}} (y_{fit} - y_{fi(t-\delta(f,i))}) \leq b_{rt} \quad \forall r \in \mathcal{R}, \forall t \in \mathcal{T}. \quad (6)$$

Note that $y_{fit} - y_{fi(t-\delta(f,i))}$ represents whether flight f is performing flight plan step i at time t .

3.2.3 Objective Function

The delay term in the objective function of each formulation represents the aggregate costs associated with flight delay, which we model as follows. First, we note that the start time of flight plan step i for plane f , $s(f,i)$, can be written as:

$$s(f,i) = T - \sum_{t=0}^{T-1} y_{fit}.$$

Next, we calculate the incremental delay accumulated before flight step i of flight f :

$$\begin{aligned} d(f, 1) &= s(f, 1) - \alpha(f, 1); \text{ and} \\ d(f, i) &= s(f, i) - \alpha(f, i) - (s(f, i - 1) + \alpha(f, i - 1)) \quad \forall i > 1. \end{aligned}$$

In the case where $\psi(f, i) = \emptyset$, we note that $d(f, i)$ is guaranteed to be zero based on constraints (5). Utilizing the above definitions, we can write the delay-part of the objective function as:

$$\min \sum_{\{f \in \mathcal{F}, i \in \mathcal{I}(f): \mathbf{G} \in \psi(f, i)\}} d(f, i) + c_{a/g} \left(\sum_{\{f \in \mathcal{F}, i \in \mathcal{I}(f): \psi(f, i) = \{\mathbf{A}\}\}} d(f, i) \right). \quad (7)$$

The first term represents our ground delay costs. Since our problem is deterministic, when $\psi(p, i) = \{\mathbf{G}, \mathbf{A}\}$, we choose to allocate all of the delay preceding step i as ground delay. The second term represents the airborne delay costs, which thus only accumulate if $\psi(p, i) = \{\mathbf{A}\}$.

3.3 Time-Order Deviation Approximation (TODA) Model

There are two challenges to calculating time-order deviation within a mathematical programming model. The first is that to calculate expected delay for each resource we need the sorted list of scheduled start times. The approach we use to address this is to create auxiliary variables for each of the flight step variables, where these auxiliary variables maintain a fixed relative order. The second challenge is that we need to determine which resource along a flight's route maximizes the expected delays. The resulting problem has a non-convex objective function, as it is the minimum of a collection of linear functions within a minimization objective. Instead of performing this minimization explicitly, we determine for each flight f , which steps i would be assigned the most delay according to single-resource RBS performed for $r(f, i)$. This gives us an estimate of congestion due to capacity-demand imbalances, though it ignores delay introduced due to interactions between resources.

3.3.1 Model Adjustments

We first define the ordered auxiliary variables described above:

$$u_{rjt} = \begin{cases} 1 & \text{if } j \text{ tasks for resource } r \text{ have been scheduled to start by time } t; \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Based on this definition, $(u_{rjt} - u_{rj(t-1)})$ will indicate when task j of resource r starts in the optimized schedule. Note that this may or may not be the same as the start time of the task originally scheduled to occupy position j .

Next, we add the following constraints to the model to ensure that the variables maintain the definition above:

$$u_{rjt} \leq u_{rj(t+1)} \quad \forall r \in \mathcal{R}, \forall j \in \mathcal{J}(r), \forall t \in \{0 \dots T-2\}; \text{ and} \quad (8)$$

$$u_{rj(T-1)} = 1 \quad \forall r \in \mathcal{R}, \forall j \in \{1 \dots J(r)\}. \quad (9)$$

Constraints (8) and (9) ensure that the sequence of ordered auxiliary variables $[u_{rj}]$ maintains the same monotonically increasing form as the sequence of flight step variables $[y_{fi}]$. We also need to ensure that the appropriate order for the auxiliary variables is maintained, that is, task $(j+1)$ cannot start before task j :

$$u_{rjt} \geq u_{r(j+1)t} \quad \forall r \in \mathcal{R}, \forall j \in \mathcal{J}(r), \forall t \in \mathcal{T}. \quad (10)$$

The last, and most important, constraints ensure that by each interval, the number of scheduled flights according to the ordered auxiliary variables and the flight step variables coincides:

$$\sum_{j=1}^{J(r)} u_{rjt} = \sum_{\{(f,i):r(f,i)=r\}} y_{fit} \quad \forall r \in \mathcal{R}, \forall t \in \mathcal{T}. \quad (11)$$

That is, constraints (11) ensure that when a flight step is scheduled within an interval, one of the sequences of ordered auxiliary variables must flip from 0 to 1 in that same interval.

With these definitions in mind, we can calculate the expected delay for flight step (f, i) and resource $r(f, i)$, which we denote $ED(f, i)$:

$$ED(f, i) = \sum_{\alpha(f, i)}^{T-1} (1 - u_{r(f, i)j(f, i)t}).$$

The right-hand side measures the number of intervals from the earliest start time for flight step (f, i) until the $j(f, i)$ task starts for resource $r(f, i)$.

As discussed in the first paragraph of this section, we estimate which resources for each flight f will maximize expected delay. We accomplish this by computing which steps i would be assigned the most delay according to single-resource RBS performed for $r(f, i)$. This gives us an estimate of congestion due to capacity-demand imbalances. For flight f , we denote the set of steps achieving the maximum RBS delay as $\mathcal{I}_{\text{MAX}}(f)$:

$$\begin{aligned} d^{RBS}(f, i) &= RBS(r(f, i), j(f, i)) - \alpha(f, i); \\ d_{\text{MAX}}^{RBS}(f) &= \max_{\mathcal{I}(f)} \{d^{RBS}(f, i)\}; \text{ and} \\ \mathcal{I}_{\text{MAX}}(f) &= \{i \in \mathcal{I}(f) : d^{RBS}(f, i) = d_{\text{MAX}}^{RBS}(f)\}. \end{aligned}$$

When the set $\mathcal{I}_{\text{MAX}}(f)$ corresponds to the steps that achieve the maximum expected delay in the optimized schedule, our approximate time-order deviation will equal the true time-order deviation as described in Section 2.3.

We now have the tools necessary to describe the fairness term we add to (7) to calculate the approximate time-order deviation in our objective function and complete the TODA model:

$$+ \lambda \sum_{f \in \mathcal{F}} \left[(s(f, |f|) - \alpha(f, |f|)) - \frac{1}{|\mathcal{I}_{\text{MAX}}(f)|} \sum_{i \in \mathcal{I}_{\text{MAX}}(f)} ED(f, i) \right]^+. \quad (12)$$

Within the sum, the first difference represents the total delay for flight f , with $s(f, |f|)$ representing the start time for the last flight step of flight f . The last term within the sum represents the average expected delay across the flight steps that achieved maximum RBS delay. The $[\dots]^+$ ensures that we only add the difference between these terms if the total delay exceeds the average expected delay.

Note that $\lambda > 0$ controls the trade-off between the system delay term (7) and the approximate time-order deviation term (12).

3.4 Ration-by-Schedule Exponential Penalty (RBS-EP) Model

Unlike the TODA model, the RBS-EP model requires the introduction of no new variables or constraints to the foundational model described in Section 3.2. The only change required is adding an additional term to the objective function. The intuition behind the RBS-EP model has two parts. The first is that no flight should expect to receive less delay than its worst-case RBS delay, $d_{\text{MAX}}^{\text{RBS}}(f)$ as defined above. But, due to interactions between resources it is unlikely that each flight will be able to achieve this exactly. So, to provide flexibility, we penalize each interval of delay beyond $d_{\text{MAX}}^{\text{RBS}}(f)$ by an exponentially increasing amount, where the base of the exponent is the parameter that controls the trade-off between aggregate delay and fairness.

3.4.1 Model Adjustments

One of the nice properties of discrete scheduling models is that we can associate different objective coefficients with each possible start time for a task. To achieve an exponentially increasing penalty, we only need to determine the appropriate coefficients for each flight and potential start interval. Thus, we let c_{ft} be the coefficient associated with the last step of flight f starting at time t :

$$c_{ft} = \sum_{\epsilon=1}^{t-\alpha(f,|f|)-d_{\text{MAX}}^{\text{RBS}}(f)} [\lambda^\epsilon - 1] \quad \forall t > \alpha(f, |f|) + d_{\text{MAX}}^{\text{RBS}}(f).$$

Based on the definition above, we have $c_{ft} - c_{f(t-1)} = \lambda^{(t-\alpha(f,|f|)-d_{\text{MAX}}^{\text{RBS}}(f))} - 1$ assuming $t > \alpha(f, |f|) + d_{\text{MAX}}^{\text{RBS}}(f)$. The -1 offsets the linear delay cost from the aggregate delay term defined in (7). With this linear delay cost factored back in, the incremental cost of delaying flight f from time $(t-1)$ to time t is $\lambda^{(t-\alpha(f,|f|)-d_{\text{MAX}}^{\text{RBS}}(f))}$. That is, assuming $\lambda > 1$, the incremental cost of each additional interval of delay increases exponentially beyond $d_{\text{MAX}}^{\text{RBS}}(f)$.

With the cost coefficients c_{ft} defined as above, the fairness term that we add to (7) to complete the RBS-EP model is:

$$+ \sum_{f \in \mathcal{F}} \left[\sum_{t=\alpha(f,|f|)+d_{\text{MAX}}^{\text{RBS}}(f)+1}^{T-1} c_{ft} (y_{f|f|t} - y_{f|f|(t-1)}) \right]. \quad (13)$$

The difference $(y_{f|f|t} - y_{f|f|(t-1)})$ equals 1 if and only if the last step for flight f begins at time t , thus applying a penalty of c_{ft} as desired. Note that in the definition of c_{ft} , the base of the exponent, $\lambda > 1.0$, controls the trade-off between aggregate system delay and fairness.

3.5 Aircraft Connectivity Discussion

As noted, these computational models build off work in [8] and [3]. Beyond the fairness considerations, there is one key difference in the approach we outline. In each of the referenced models, planned aircraft connections between flight legs are maintained in the controlled schedule (i.e. aircraft connections are constrained within the optimization model). In our models, we do not include connectivity constraints between flight legs. Note that both of our models can include these constraints and remain entirely consistent, thus it is an explicit *modeling choice* to omit them. We have made the decision to exclude constraints of this type for the following two reasons. First, this change leads us to an approach that is consistent with current practice. That is, our models are able to utilize the same inputs as existing TFM programs. Second, there is a question as to whether the inclusion of aircraft connectivity constraints would be beneficial to either airlines or passengers. Once schedules are disrupted, there is no guarantee that the aircraft originally scheduled to fly a route (i.e. flight leg to flight leg) will be the one to continue it. Thus constraining the model to enforce planned connections may be too strict. This could be a particularly significant issue for airlines with large hub operations where there is more flexibility to adjust the controlled schedule. A coupled investigation of capacity allocation and its impact on airline recovery is the subject of ongoing research [12] where we plan to address this question.

4 Computational Results

Here we provide computational experiments to demonstrate the practical value of the RBS-EP model. We highlight three key results from our regional and national scenarios. The first is that under a conservative comparison between RBS-EP and current practice, the RBS-EP model improves efficiency, as measured by total delays, while maintaining equivalent levels of equity. The second is that the RBS-EP model closely tracks the tighter TODA approximation of the efficient frontier between aggregate delay and fairness, calculated according to our time-order deviation metric. Finally, the RBS-EP model is computationally efficient, allowing solution of even complex, national-scale problems within reasonable computing times.

4.1 An Apples-to-Apples Comparison

The challenge in comparing our optimization-based approaches to current practice is that the two solve slightly different problems. Our optimization-based approaches ensure that all resource capacity constraints are strictly satisfied. On the other hand, multi-resource RBS allows FCA capacity constraints to be violated, as described in Section 2.1. That is, if the same capacities are utilized as inputs into both procedures, multi-resource RBS would likely perform better because of its ability to arbitrarily exceed FCA capacity constraints (and the inability of our optimization-based approaches to do so).

To level the playing field, we first perform a multi-resource RBS allocation, with one slight modification to the approach utilized in practice. In the multi-resource RBS approach, conflicts between AFPs are resolved by using the CTD associated with the AFP that is initiated first. In our examples, we assume equivalent initiation times for the AFPs, thus we choose the CTD associated with the last FCA in each flight's route. We do not believe this introduces any systematic biases in our comparisons. Additionally, we perform a compression procedure as described in [26] to attempt to fill gaps in FCA resource schedules. For time intervals where the resulting allocation exceeds the initial capacity, we increase the corresponding capacity as an input into each optimization-based approach. By adjusting the capacity, we ensure that our optimization-based approaches do not exceed the initial capacity *any more than* the multi-resource RBS schedule. For instance, based on

the four flight example from Section 2.1 and 5 minute discretization intervals, we would increase the capacity of FCA1 to 2 flights from 18:50 to 18:55, keeping the capacity at 1 for all other intervals. Although this leads to a fairer comparison between the two approaches, the playing field is still tilted toward multi-resource RBS. Because of the inherent limitations of multi-resource RBS, we can only perform a comparison for capacity allocations that correspond directly to a multi-resource RBS schedule. Fortunately, as demonstrated in Section 2.1, this still leaves inefficiencies that optimization-based approaches are capable of exploiting.

4.2 Construction of Regional and National Scenarios

To construct each of our scenarios, we start with flight schedule data that corresponds to a single day of relatively clear weather operations (June 21st 2005). The schedule data was obtained from Flight Schedule Monitor, a decision support tool developed for the FAA by Metron Aviation ([17]). For the purposes of all of our experiments, we will treat this schedule as representing the Official Airline Guide (i.e. the planned airline flight schedules). Thus, the defining characteristics of each scenario are the set of controlled resources and the corresponding capacities.

For arrival resources, we utilize historical GDP data to construct our scenarios. From Metron Aviation, we obtained information on GDP airports and durations from April 2007 through April 2008. We categorize this data into to four regions, specifically Washington D.C. (BWI and IAD), Texas (DFW and IAH), New York (EWR, JFK, LGA and PHL), and Chicago (ORD and MDW). For each region, we choose one day with overlapping GDPs and create two capacity-reduction scenarios, corresponding to moderate or severe disruptions. The execution window for each GDP is taken from the historical data. The arrival capacities for each scenario are based on the FAA's 2004 Airport Capacity Benchmark Report ([25]). This report lists Optimum, Marginal, and IFR (Instrument Flight Rules) capacities in terms of number of operations per hour, which we assume are evenly split between arrivals and departures. IFR operations rates are utilized at an airport when there is reduced pilot visibility, e.g. during fog or a severe storm. For the moderate scenario, we calculate the GDP capacity based on the average of the Marginal and IFR arrival rate. For the severe scenario, we utilize the IFR arrival rate directly.

To control capacity within the regional airspace, we constrain the maximum occupancy for two to four FCA resources corresponding to contiguous air sectors within the region. The execution window for each FCA is set to the intersection of the execution windows of the regional GDPs extended 30 minutes on the front-end. We base our capacities on percentiles of each FCA’s daily utilization split into 15-minute intervals. For the moderate disruption scenarios, we set the capacity to the 90th percentile of utilization. For the severe disruption scenarios, we set the capacity to the 70th percentile of utilization, which typically corresponds to about 60% of the maximum utilization.

Last, we create two national scenarios by combining the capacity controls for each of the regional scenarios. That is, our moderate disruption national scenario is equivalent to the moderate disruption capacity controls for all four regions being implemented on the same day. Thus, in total we develop 10 representative scenarios to compare our optimization-based approaches to multi-resource RBS. For each of the scheduling approaches, including multi-resource RBS, we discretize the flight schedules based on 15 minute intervals. Table 9 provides high-level details for each of the scenarios.

Region	# Impacted Flights	# GDPs	GDP Minutes	# AFPs	AFP Minutes
Washington D.C.	2238	2	795	2	840
Texas	2414	2	555	3	780
New York	3149	4	2740	4	1560
Chicago	5198	2	1400	3	1590
National	10553	10	5490	12	4770

Table 9: Regional and national scenario details

4.3 The Trade-off Between Efficiency and Fairness

In this section, we display ten charts demonstrating the trade-off between efficiency, as measured by aggregate delay, and fairness, as measured by the time-order deviation of the controlled schedule for each of the scenarios described in the previous section. The curves were created by adjusting λ , the parameter that controls this trade-off for each of the optimization-based approaches we developed in Section 3. The less complex scenarios display the trade-off curves for both the TODA and RBS-EP models. The more complex scenarios only display the curve for the RBS-EP model, because the TODA model is not computationally tractable for these scenarios.

In addition to the curves for one or both of our models, each chart displays a point representing the multi-resource RBS schedule. For the purpose of comparison, we consider the first point on the RBS-EP curve with a lower time-order deviation value than the multi-resource RBS schedule. The percentage displayed on the chart indicates the reduction in aggregate delay achieved by this point (i.e. for the corresponding value of λ).

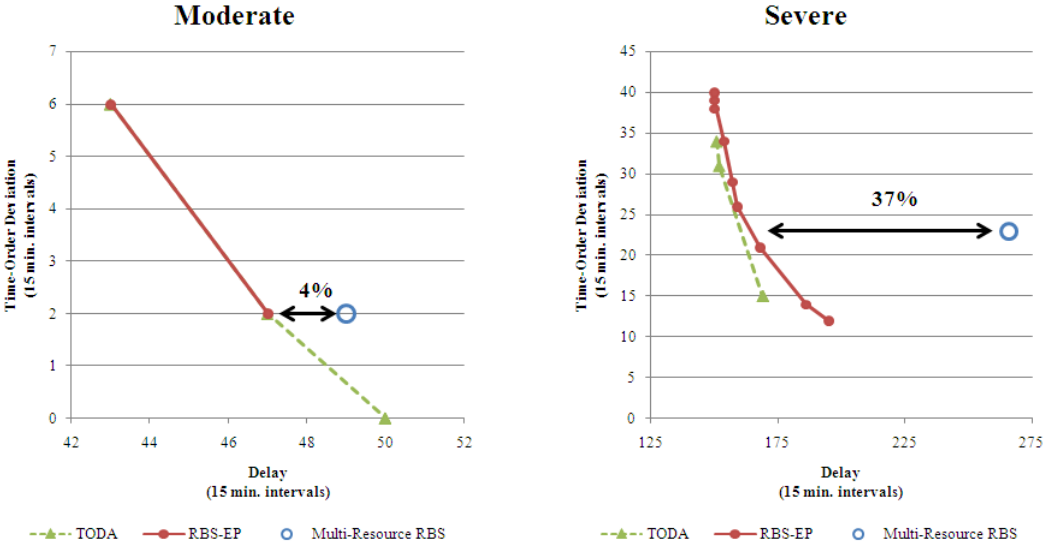


Figure 3: Washington D.C. disruption scenarios

Ideally, we would like the RBS-EP curves to have a monotonically decreasing time-order deviation as λ increases (and therefore delay increases). Unfortunately, this is not always the case (e.g. the moderate scenarios for New York and Chicago). Still, this should be expected because the RBS-EP model is not directly minimizing time-order deviation. We are highly encouraged by the strong trend between an increasing λ and the decreasing time-order deviation of the resulting schedule. That is, by simply adjusting the functional form of the delay term, we have created a model that tracks the much more complex time-order deviation metric.

For the Texas moderate disruption scenario (Figure 4), we list an efficiency improvement of -1% . For this scenario, the indicated values of λ for both the TODA and RBS-EP models lead to an increase from 178 to 179 intervals of delay as compared with the multi-resource RBS schedule. Though both curves pass through the point representing the multi-resource RBS schedule, the

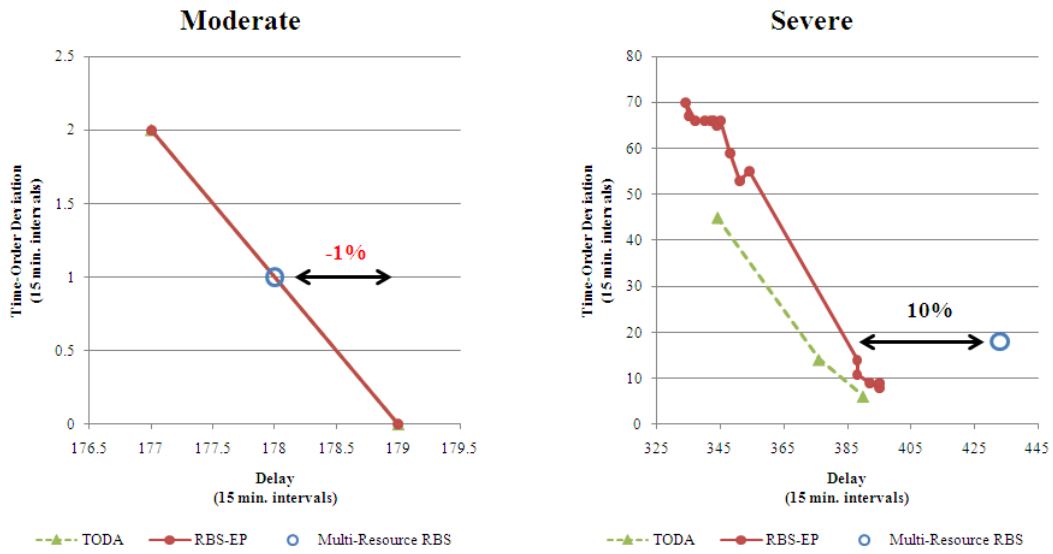


Figure 4: Texas disruption scenarios

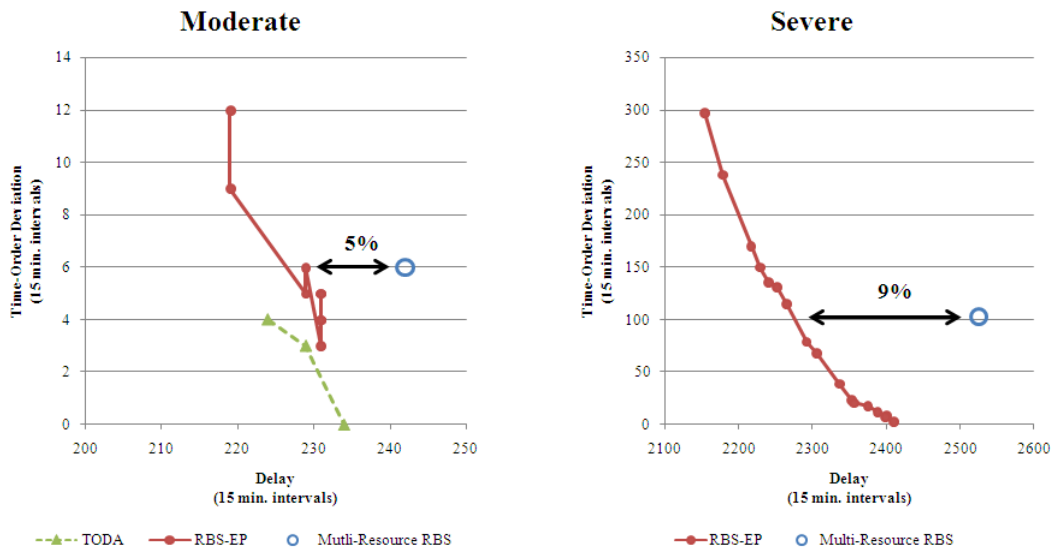


Figure 5: New York disruption scenarios

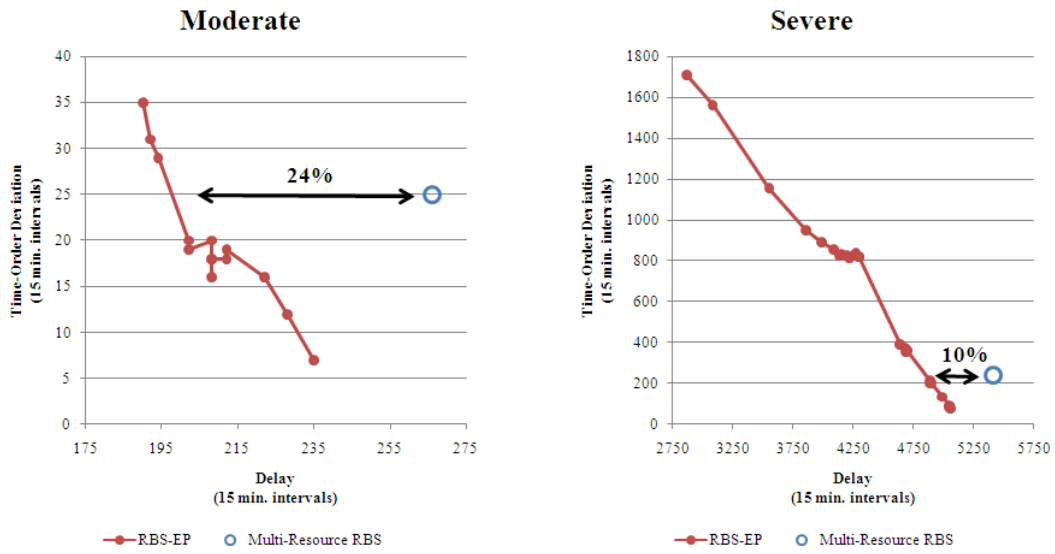


Figure 6: Chicago disruption scenarios

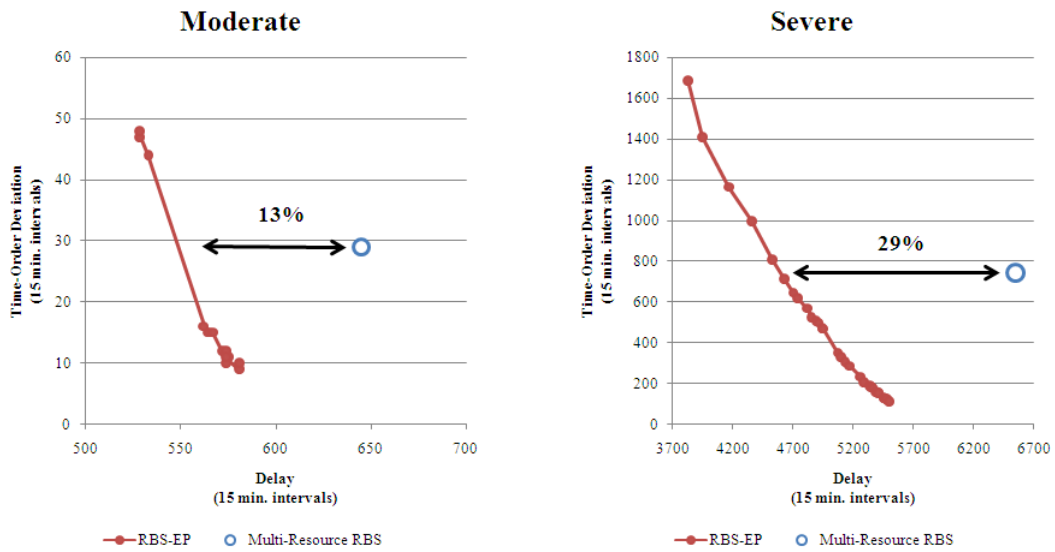


Figure 7: National disruption scenarios

curves jump from 177 intervals of delay with a time-order deviation of 2 intervals to 179 intervals of delay with a time-order deviation value of 0 intervals. Thus, based on our selection criteria, we choose the schedule with 179 intervals of delay since it is the first schedule at least as fair as the multi-resource RBS schedule. The fact that we do not achieve an efficiency improvement is likely due to the limited range of time-order deviation values. For our severe disruption scenarios, the efficiency improvements range from 9% (New York) to 37% (Washington D.C.).

In Table 10, we summarize the information displayed in the charts above. For each scenario and optimization-based approach, we list the value of λ that leads to the efficiency improvement and the corresponding trade-off between fairness and delay. All of these results were obtained using enhanced computational techniques described in complete detail in [12]. These techniques allow us to significantly improve computational performance at the cost of a slight increase in objective costs, typically resulting in solutions within 1% of the true optimum. We choose to report these values as opposed to the true optimums since we believe these techniques will be an important part of any practical implementation. The efficiency gains listed in Table 10 correspond to the computational performance numbers provided in Section 4.6. Note that for the TODA results, we list an explicit optimality gap. This explicit gap is in addition to any sub-optimality that may occur due to our computational techniques. With all of these caveats in mind, it is important to note that solving these problems to optimality would only *improve* the efficiency gains listed.

Scenario	Multi-Resource RBS		TODA Model				RBS-EP Model		
	Delay	TOD	Gain	TOD	λ	Gap	Gain	TOD	λ
Washington D.C. (I)	49	2	4.1%	2	1.38	0.0%	4.1%	2	2.08
Washington D.C. (II)	266	23	40.2%	15	1.28	2.1%	36.8%	21	2.08
Texas (I)	178	1	-0.6%	0	1.32	0.0%	-0.6%	0	2.08
Texas (II)	433	18	13.2%	14	1.42	2.2%	10.4%	14	2.08
New York (I)	242	6	7.4%	4	1.02	0.8%	5.4%	5	2.08
New York (II)	2525	103	—	—	—	—	9.2%	79	1.55
Chicago (I)	266	25	—	—	—	—	24.1%	20	2.08
Chicago (II)	5411	241	—	—	—	—	9.9%	216	3.07
National (I)	645	29	—	—	—	—	12.6%	15	2.08
National (II)	6545	745	—	—	—	—	29.3%	713	1.28

Table 10: Summary of disruption scenario results. Delay and TOD (Time-Order Deviation) are reported in number of 15-minute intervals.

4.4 Flight Delay Distribution

In addition to the summary statistics listed in Table 10, it is important to consider the distribution of delays for impacted flights. Airlines typically build slack into their flight schedules to preserve connections between aircraft, crew, and passengers. Delay that is less than the planned slack can be absorbed without schedule modifications. Delay that exceeds the planned slack often requires costly recovery operations. Thus, we need to ensure that our approach does not lead to a heavy tail of flight delays (i.e. a larger number of flights receiving a large amount of delay).

Consider the flight delay distributions charted in Figure 8. These figures chart the number of flights receiving at least the specified number of intervals of delay based on the multi-resource RBS approach and the RBS-EP model utilizing the λ values listed in Table 10. For the New York severe disruption scenario, the distribution is similar for flights receiving at least 4 15-minute intervals of delay (i.e. 1 hour), though the RBS-EP model leads to more flights receiving at least 3 intervals of delay. For the National severe disruption scenario, the RBS-EP model has a much longer tail, with 15 flights receiving more than 9 intervals of delay, the maximum allocated by the multi-resource RBS approach. Based on the discussion in the preceding paragraph, this could be a significant issue.

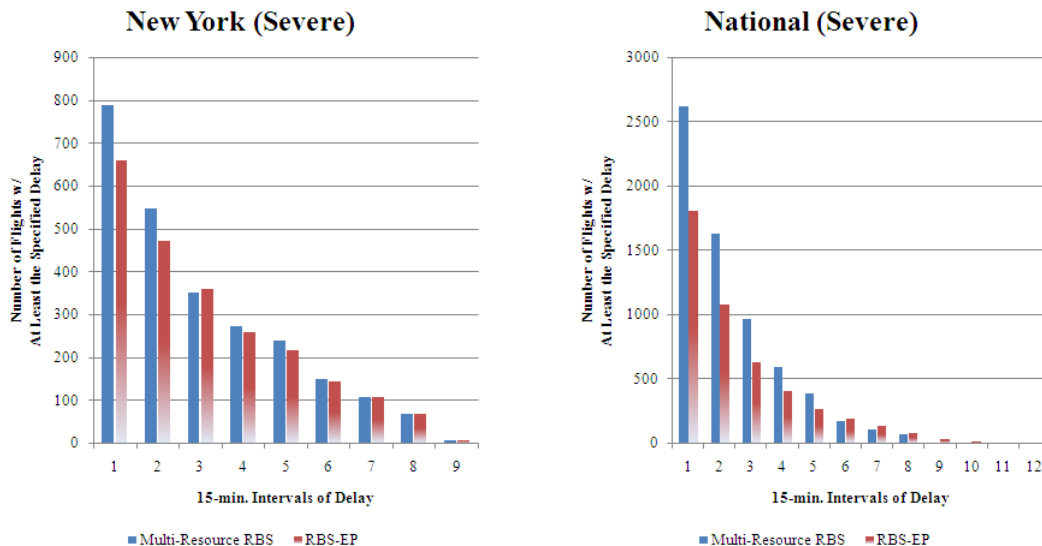


Figure 8: Allocated flight delay distributions with RBS-EP λ values from Table 10

Fortunately, the RBS-EP model provides a mechanism for resolving these types of issues. By increasing the value of λ , the base of the exponential penalty, it puts additional pressure on the tail of the flight delay distribution. For example, consider the updated charts in Figure 9. To create these charts, we utilize λ values of 2.08 for each of the RBS-EP models, as compared to 1.55 for the New York scenario and 1.28 for the National scenario above. By increasing the value of λ we have increased the aggregate delay from 2292 to 2337 intervals of delay in the New York scenario and from 4472 to 5048 in the National scenario. Though, in so doing, we have managed to shrink the tails of the delay distribution, with the resulting schedules still significantly more efficient than the multi-resource RBS schedules. This trade-off between aggregate delay and the distribution of delay is another important consideration for choosing an appropriate value of λ in practice.

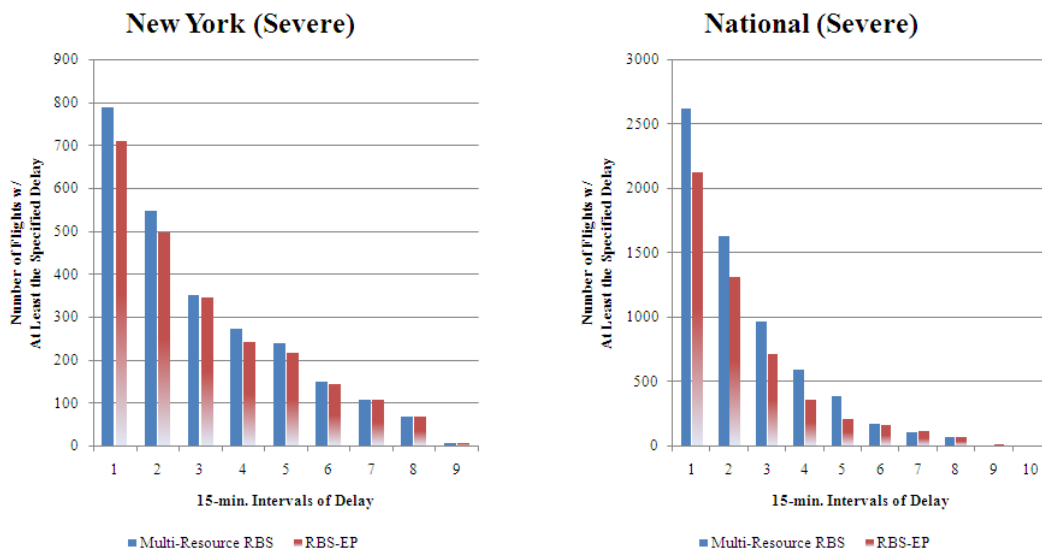


Figure 9: Updated flight delay distributions based on RBS-EP λ value of 2.08

4.5 The Value of Efficiency

In Table 10, we see that the multi-resource RBS approach allocates a total of 16,560 intervals of delay across the 10 scenarios, whereas the RBS-EP model allocates 13,575 intervals of delay. Thus, the RBS-EP model leads to an overall efficiency improvement of 18% in our tests. In general, we prefer a scheduling approach that is more efficient, but we would like to determine how much cost reduction, both for airlines and passengers, can be attributed to this efficiency gain.

As mentioned in the introduction, the Air Transportation Association estimates that 138 million minutes of delay during the 12-month period ending in September of 2008 cost airlines \$10 billion and passengers \$4.5 billion [1]. The Bureau of Transportation Statistics (BTS) estimates that in 2007, 37.7% of flight delays were due to the previous flight arriving late, thus we estimate that the remaining 62.3% of flight delays are due to direct impacts [20]. These direct impacts, to which we attribute the full delay costs, led to 86 million minutes of delay. From the GDP data we received from Metron Aviation, we find that from April 2007 to April 2008, 17.8 million minutes of delay were assigned through GDPs and AFPs. This represents approximately 20.7% of the direct impact delay, thus we attribute 20.7% of the total delay costs to TFM programs (\$2.1 billion in airline costs and \$930 million in passenger costs). Of the 17.8 million minutes of delay assigned through these initiatives, 7 million minutes of delay or 39.4% was assigned on days where both GDPs and AFPs were implemented. We consider these days our baseline for improvement, since the multi-resource RBS schedule is optimal when there are no conflicts between TFM programs. A 1% efficiency improvement on these days would save airlines \$8.1 million and passengers \$3.6 million annually. An 18% efficiency improvement, the average from our tests, would result in a total cost savings of \$212 million annually.

It is worth noting that the attribution approach utilized above likely underestimates the value in at least three ways. First, the ATA estimates for the total costs associated with delays are conservative. The U.S. Congress Joint Economic Committee estimates delays in calendar year 2007 to have cost airlines \$19 billion, passengers \$12 billion, and other industries \$10 billion [14], which is almost a factor of three larger than the ATA estimates. Second, by focusing our analysis on direct impact delays, we are assuming propagated delay costs are allocated proportionally between different causes. Note that this likely underestimates the costs associated with TFM programs, because TFM programs typically lead to larger magnitudes of delays which are more likely to exceed schedule slack and be propagated. Third, the Air Transportation Association estimates passenger delay costs by multiplying the total passenger delay hours by an average time value of \$35.70 per hour. This approach underestimates the impacts of schedule disruptions and flight cancellations, both of which are more prevalent during TFM initiatives.

4.6 Computational Performance Measurements

In a practical setting, we expect the FAA to choose a value of λ in advance to represent an appropriate trade-off between fairness and aggregate delay (and between aggregate delay and the distribution of delay). With this in mind, we note that in Table 10, many of the λ values for the RBS-EP model are 2.08 or less (the exception being the Chicago severe disruption scenario). Additionally, for the the New York severe and National severe disruption scenarios, it could easily be argued that choosing $\lambda = 2.08$ is appropriate due to the distribution of delay, as discussed in Section 4.4. Thus, for our computational performance measurements, we use the RBS-EP model with $\lambda = 2.08$. We also list the performance measurements with $\lambda = 1.01$ as a reference to indicate the computational benefits of the RBS-EP model. These measurements are based on the enhanced computational techniques mentioned briefly in the previous section and described in full in [12]. The computational tests are performed on a PC with dual Xeon 3220 Quad-Core processors, 8 Gigabytes of RAM, running Ubuntu v8.04 and CPLEX v11.2 through the Java interface.

Scenario	CPLEX Solver Time (sec.)	
	$\lambda = 1.01$	$\lambda = 2.08$
Washington D.C. (I)	0.068	0.011
Washington D.C. (II)	0.038	0.019
Texas (I)	0.023	0.014
Texas (II)	0.108	0.023
New York (I)	0.042	0.025
New York (II)	3.445	0.311
Chicago (I)	0.059	0.069
Chicago (II)	37.404	0.633
National (I)	0.179	2.993
National (II)	113.073	5.215
Total	154.439	9.313

Table 11: CPLEX computation times for RBS-EP model with λ values of 1.01 and 2.08

In general, larger values of λ lead to improved performance for the RBS-EP model. The one scenario where the RBS-EP model with $\lambda = 2.08$ performs significantly worse is the National moderate disruption scenario. In this case, CPLEX finds a solution for $\lambda = 1.01$ during pre-solve, but requires branching to find a $\lambda = 2.08$ solution. An interesting result from our computational tests is that by using $\lambda = 2.08$, the CPLEX solver time is cut dramatically for the most complex

scenarios: the Chicago, New York, and National severe disruption scenarios. This suggests that not only is 2.08 a good choice of λ for balancing aggregate delay against fairness and flight delay distribution, it also provides valuable performance benefits for complex, large-scale scenarios.

5 Conclusion

In this research, we develop an optimization-based formulation that could be readily incorporated in practice by the FAA. Specifically, based on principles that have made RBS successful, we have developed a time-order deviation metric for schedule fairness that extends to the multi-resource setting. This metric allows us to evaluate optimization-based scheduling approaches relative to each other, but more importantly it allows us to compare these approaches to current practice. Using this metric, we have demonstrated that our two formulations, the TODA and RBS-EP models, can improve operational efficiency while maintaining a consistent level of fairness. Additionally, these models allow for precise management of airspace-based capacities, which is not possible with the current multi-resource RBS approach. Last, we have demonstrated that the RBS-EP model is computationally tractable in practice, even for complex regional and national-scale problems.

Introducing optimization into the FAA's practices has been a significant challenge, as should be apparent from the literature review in Section 1.3. The RBS-EP model addresses many of these challenges and should thus provide a strong foundation for future research. Our goal is to have the RBS-EP model represent the first step in an ongoing sequence of practical enhancements to the FAA's TFM procedures.

References

- [1] Air Transportation Association. (2009). Cost of ATC delays. Retrieved April 8, 2009 from <http://www.airlines.org/economics/cost+of+delays>
- [2] Air Transportation Association. (2008). 2008 Annual Report. Retrieved April 8, 2009 from http://www.airlines.org/economics/review_and_outlook/annual+reports.htm
- [3] Andreatta, G., L. Brunetta, G. Guastalla. (2000). From ground holding to free flight: an exact approach. In *Transportation Science*, Vol. 34, 4, 394-401.
- [4] Anonymous. (2000). A jam at 32,000 feet. In *The Economist*, February 5, 2000.
- [5] Ball, M. O., C. Barnhart, G. Nemhauser, A. Odoni. (2007). Air transportation: irregular operations and control. In *Handbook in OR & MS*, Vol. 14, Chapter 1, 23-38.
- [6] Ball, M. O., R. Hoffman, A. R. Odoni, R. Rifkin. (2003). A stochastic integer program with dual network structure and its application to the ground-holding problem. In *Operations Research*, Vol. 51, 1, 167-171.
- [7] Bertsimas, D., Lulli, G., Odoni, A. R. (2008). The air traffic flow management problem: an integer optimization approach. In *Integer Programming and Combinatorial Optimization (34 - 46)*. Springer Berlin / Heidelberg.
- [8] Bertsimas, D. J., S. S. Patterson. (1998). The air traffic flow management problem with enroute capacities. In *Operations Research*, Vol. 46, 3, 406-422.
- [9] Bratu, S., C. Barnhart. (2006). Flight operations recovery: new approaches considering passenger recovery. In *Journal of Scheduling*, Vol. 9, 3, 279-298.
- [10] Brennan, M. (2007). Airspace flow programs - a fast path to deployment. In *The Journal of Air Traffic Control*, Vol. 49, 1, 51 - 55.
- [11] Chang, K., K. Howard, R. Oiesen, L. Shishler, M. Tanino, M. C. Wambsganns. (2001). Enhancements to the FAA ground-delay program under collaborative decision making. In *Transportation Science*, Vol. 34, 1, 57-76.
- [12] Fearing, D. (2010). The equity, efficiency, and passenger impacts of traffic flow management. Massachusetts Institute of Technology, Ph.D. Thesis.
- [13] Hoffman, R., M. O. Ball. (2000). A comparison of formulations for the single-airport ground-holding problem with banking constraints. In *Operations Research*, Vol. 48, 4, 578-590.
- [14] Joint Economic Committee. (2008). Your flight has been delayed again: flight delays cost passengers, airlines, and the U.S. economy billions. Retrieved May 13, 2009 from <http://jec.senate.gov/index.cfm?FuseAction=Reports.Reports>
- [15] Kotnyek, B., O. Richetta. (2006). Equitable models for the stochastic ground-holding problem under collaborative decision-making. In *Transportation Science*, Vol. 40, 2, 133-146.
- [16] Lulli, G., A. R. Odoni. (2007). The European air traffic flow management problem. In *Transportation Science*, Vol. 41, 4, 431-443.

- [17] Metron Aviation. (2009). Flight schedule monitor. Retrieved May 13, 2009 from <http://www.metronaviation.com/fsm.php>
- [18] de Neufville, R., A. R. Odoni. (2003). Air traffic management. In *Airport Systems: Planning, Design, and Management*, Chapter 13, 525-545.
- [19] Odoni, A. R. (1987). The flow management problem in air traffic control. In *Flow Control of Congested Networks*, Springer-Verlag, Berlin, 269-288.
- [20] Research and Innovative Technology Administration. Understanding the reporting causes of flight delays and cancellations. Retrieved April 9, 2008 from the Bureau of Transportation Statistics. <http://www.bts.gov/help/aviation/html/understanding.html>
- [21] Richetta, O., A. R. Odoni. (1993). Solving optimally the static ground-holding policy problem in air traffic control. In *Transportation Science*, Vol. 27, 228-238.
- [22] Richetta, O., A. R. Odoni. (1994). Dynamic solution to the ground-holding policy problem in air traffic control. In *Transportation Research Part A*, Vol. 28, 167-185.
- [23] Patterson, S. S. (1997). Dynamic flow management problems in air transportation. Ph.D. Thesis, Massachusetts Institute of Technology.
- [24] Terrab, M., A. R. Odoni. (1993). Strategic flow management for air traffic control. In *Operations Research*, Vol. 41, 1, 138-152.
- [25] U.S. Department of Transportation, Federal Aviation Administration, The MITRE Corporation, Center for Advanced Aviation System Development. Sep. 2004. Airport Capacity Benchmark Report 2004.
- [26] Vossen, T., M. O. Ball. (2005). Optimization and mediated bartering models for ground delay programs. In *Naval Research Logistics*, Vol. 53, 1, 75-90.
- [27] Vossen, T., M. O. Ball, R. Hoffman, M. Wambsganns. (2003). A general approach to equity in traffic flow management and its application to mitigating exemption bias in ground delay programs. In *Air Traffic Control Quarterly*, Vol. 11, 277-292.
- [28] Vranas, P. B., D. J. Bertsimas, A. R. Odoni. (1994). The multi-airport ground-holding problem in air traffic control. In *Operations Research*, Vol. 42, 2, 249-261.