OPTIMALLY AND EQUITABLY DISTRIBUTING DELAYS WITH THE AGGREGATE FLOW MODEL

Michael Bloem, NASA Ames Research Center, Moffett Field, CA
Banavar Sridhar, NASA Ames Research Center, Moffett Field, CA

Abstract

The aggregate flow model is used to determine how to distribute predeparture delays among air traffic control Centers and across time to optimally satisfy constraints on airspace capacity and departure rates. To do so, a quadratic cost on cumulative departure delays is introduced, resulting in an optimization problem that can be quickly solved using convex optimization tools. Simulations using the model demonstrate the behavior of the National Airspace System (NAS) when implementing optimal departure delays for a particular constraint scenario. These results show that capacity-constrained air traffic control Centers suffer the highest delays. Three approaches for increasing the equity of the distribution of delays across the NAS are investigated. The first involves setting an upper bound on the Gini coefficient, a quasi-convex measure of inequality. Another is to make delays in some Centers more costly than in others. The last approach is to put an upper bound on the delay per departure for each Center. Simulation results demonstrate that bounding delay per departure effectively reduces the delays for the constrained Center. Enforcing an upper bound on the Gini coefficient and increasing the weight on delays in some Centers may impose large delays on other Centers when reducing the delays in the constrained Center.

Introduction

Sub-optimal traffic flow management (TFM) initiatives are responsible for some of the delays observed in the NAS. For example, more than 215,000 hours of delay between January 2003 and October 2004 can be attributed to the FAA's Traffic Management System. These delays cost airlines alone around $700 million [1]. A NAS-level model for which optimal TFM actions could be derived would help the FAA Air Traffic Control System Command Center (ATCSCC) make NAS-wide TFM decisions. However, surprisingly few TFM models have been developed for which optimal solutions can be computed for the entire NAS in real-time. NAS-wide TFM is so large and complex that most proposed models and optimization techniques are too computationally intensive to allow for real time NAS-wide optimization. Some of these models are discussed below in the "Models for Traffic Flow Management" subsection.

One model for which optimal solutions can be computed for the entire NAS in real-time is the aggregate flow model [2]. Note that while many models use aggregation of one kind of another, one particular model will be referred to as *the* aggregate flow model. This model describes the behavior of the NAS with around 20 states that evolve in a time-varying linear dynamical system. The stability and response characteristics of the model have been studied [3], and it has been used to manage congestion in a small sample problem [1]. However, there has been no attempt to utilize this model to find TFM actions for the entire NAS.

While the two projects using the aggregate flow model [1, 3] used different approaches, they both found that the most effective way to handle constraints in a part of the NAS was to implement predeparture delays in that part of the NAS. Indeed, research on the impact of weather on delays has found that weather-induced delays are not distributed evenly across air traffic control Centers [4]. This distribution of delays may not be desirable when high-delay or high-priority areas are assigned large amounts of delay. In fact, more recent research has looked at how to take TFM actions specifically to alleviate delays in high-delay parts of the nation, such as New York [5-6]. The aggregate flow model can be used to examine the impact of prioritizing equality in the distribution of delays around the NAS.

In this paper the aggregate flow model is used to determine how to distribute predeparture delays across the NAS to minimize a quadratic cost on

cumulative delays while meeting future capacity constraints. Three methods for incorporating equality concerns into this optimization approach are also analyzed.

In the next sub-section, models for TFM will be classified, discussed, and evaluated. The aggregate flow model will be described in detail, and its strengths and weaknesses will be itemized in the "Aggregate Flow Model" section. In the "Optimization Approach" section, the cost function and constraints will be introduced. Next a simulation scenario and results will be presented and discussed in the "Results and Discussion" section. Finally, in the "Conclusions and Future Work" section, what was learned will be summarized and future work will be proposed.

Models for Traffic Flow Management

Several TFM models have been developed over the past few decades. These models possess a large variety of properties and can be classified in several ways – deterministic versus stochastic, according to the control inputs they allow, regional versus national, etc. However, the most insightful way to classify these models is according to how aircraft are aggregated.

At one extreme, models may not aggregate aircraft at all. Such models that maintain information about each aircraft separately are known as Lagrangian models. The most well-known and widely used Lagrangian TFM model was developed by Bertsimas and Stock [7]. In this model, the decision variables are whether or not a certain flight arrives at a certain sector by a certain time. There also may exist a discrete number of re-routing options for each flight. Minimizing the sum of ground and airborne delay leads to a 0-1 integer programming problem. Even the linear program relaxation of this problem is computationally too demanding to be solved for the entire NAS in real time.

Bayen et al. developed another Lagrangian model [8]. In this model, aircraft were modeled with enough detail that air traffic control commands such as vectoring for spacing and speed changes could be implemented. The completion time for moving aircraft from initial to final states was minimized with a mixed integer linear program. Again this model is too complex to be solved in real-time at scales any larger than a Center.

Other models, broadly known as Eulerian models, do not keep track of each aircraft individually but rather keep track of the number of aircraft that share some characteristic. The most obvious way to aggregate aircraft is according to geographic location. The first such model used as its state the number of aircraft in each air traffic control Center [9]. The dynamics of this model were stochastic in nature and based on Poisson processes. The aggregate flow model is closely related to this model; both models use the same state variables. The models differ in that the aggregate flow model dynamics are based on time-varying linear dynamical equations with noise terms rather than stochastic processes [2]. These models involve more dramatic assumptions than other models, but they also have relatively few state variables that are related in a convenient linear form. This means that this model can easily be used to simulate and optimize over the entire NAS in real-time.

Another possibility is to aggregate flights according to flight status. Ball et al. define the number of aircraft that are held on the ground and bound for a particular airport as control inputs. Aircraft bound for an airport and held in the air constitute another state in the system. The "planned airport arrival rate" is a probabilistic constraint that determines the number of aircraft that can transition out of the airborne state each time step by landing [10]. This is an integer program but its linear program relaxation is guaranteed to give integer results, and it can be solved quickly. Several extensions to this model have allowed for the control of some rerouting possibilities and also time-dependent and probabilistic capacity constraints on airspace. While the control inputs for this model are nicely suited to current collaborative traffic flow management practices, it is not easily extended beyond a single airport.

Several other Eulerian TFM models are based on aircraft aggregation according to network flows. The first network-based model defined a network by first overlaying a grid on the NAS [11]. Eight flows were defined in each grid cell (one from each side to the opposite side and one from each corner to the opposite corner), and flights were aggregated

accordingly. A linear discrete-time dynamical system based on conservation of aircraft describes the movement of aircraft from one flow to another. Model-predictive control then finds appropriate control actions, which involve holding aircraft back in flows rather than allowing them to move from one flow to the next. Determining divergence parameters, which describe how aircraft change direction at the intersection of flows, can be difficult. These parameters could be updated continuously as new information becomes available [12].

Another network flow-based modeling approach defines network flows by starting from airspace sectors rather than a grid [13]. Flight trajectories through sectors are clustered into various flows, defining a network. Each network link is then sub-divided into cells, which require one time step to traverse. The system state is the number of aircraft in each cell for each origin-destination pair in the system. Decoupling the system state according to origin-destination pairs avoids the issues related to divergence parameters described earlier. Again the model dynamics are defined by a linear dynamical system, and the control input is to hold aircraft back in a cell rather than allowing them to continue to the next cell. This model is used to minimize total travel time through the network using mixed-integer linear programming optimization. The number of state variables can become large for this model and can easily exceed the number of aircraft in the system, which makes optimizing for the whole NAS difficult. However, techniques such as linear programming approximation and dual decomposition help with these computational issues, and computation times may be low enough for real-time computation of NAS-wide solutions.

A similar approach that is in some ways more realistic uses a finite impulse response filter model [14]. The control inputs in this model are the fraction of aircraft that are allowed to cross each network link at each of a finite number of velocities. This mechanism for assigning delays aligns more closely with TFM practice than holding en-route aircraft. This model can be used with quadratic programming optimization to find optimal control actions, but no attempt has yet been made to apply this approach to NAS-wide problems. Depending on the resolution of the model network, this approach may or may not be fast enough for real-time NAS-wide problems.

Network flows can also be described with partial differential equations (PDEs) rather than linear dynamical systems [15]. The main idea in these models is that aircraft density along network links is modeled with hyperbolic PDEs from the Lighthill-Whitham-Richards traffic model. The objective is to maximize aircraft throughput at a destination airport while maintaining traffic densities below some upper bound, and the control inputs are routing policies and speed assignments. Optimization problems with PDEs are difficult to solve, but this problem has been solved by using adjoint-based methods. A discretized version of a linearized PDE model is referred to as the supply chain model [14]. While the supply chain model is not as accurate as the PDE model, it can be used in a convex optimization formulation and so routing and velocity TFM actions can be found quickly. It is unclear if it would be possible to scale the PDE or supply chain approaches to NAS-wide problems and find real-time solutions.

A final network-flow based approach mentioned here is based on dynamic queuing theory [16]. In this model, airport demand and capacity vary over time, and dynamic queuing theory determines the probability that queues (in the air or on the ground) will be certain lengths at certain times. This methodology has been expanded to include around 60 interconnected airports. One issue with this approach is that while it works well for predictions and studies of the impact of changing relevant parameters (such as airport capacities), it has not been utilized to perform TFM optimization.

Aggregate Flow Model

The state variables for the aggregate flow model are simply the count of aircraft in each air traffic control Center at each time step. In this model the NAS can be viewed as a graph where each node is an air traffic control Center and edges represent borders between Centers, as shown in Figure 1.

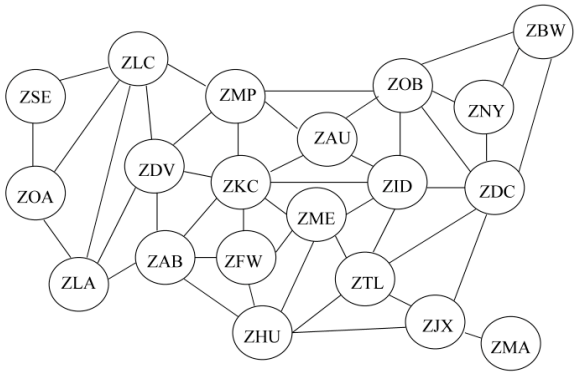


Figure 1. Graph Representation of National Airspace System for Aggregate Flow Model

These states change over time according to linear time-varying equations that capture how aircraft move from one Center to another in each time step and how many aircraft arrive and depart in each Center at each time step. These inflows and outflows are depicted in Figure 2.

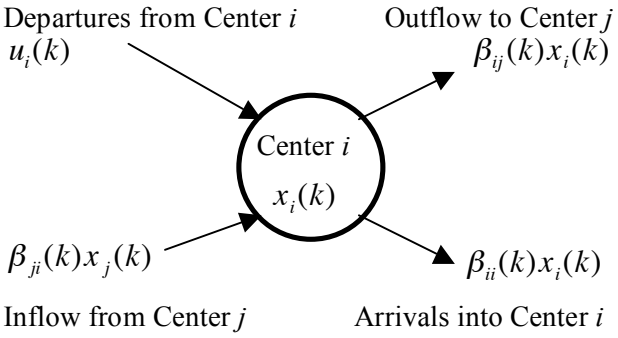


Figure 2. Inflows and Outflows of Air Traffic from a Center

There are four main flows into and out of each Center. One inflow is the number of departures from Center *i* itself. This is the variable that can be used to control the system and it is denoted by $u_i(k)$. The other inflow comes from neighboring Centers and is assumed to be proportional to the number of aircraft in those neighboring Centers. For a neighboring Center *j* this inflow can be expressed as $\beta_{ji}(k)x_j(k)$, where $\beta_{ji}(k)$ is a time-varying parameter between zero and one that denotes the fraction of aircraft in Center *j* that move to Center *i* during time step *k*. Similarly, $\beta_{ij}(k)x_i(k)$ is the outflow from Center *i* to Center *j* during time step *k*. The fourth and final flow is the arrivals in Center *i* during time step *k*, which is also assumed to be proportional to the number of aircraft in Center *i* at the start of time step *k* and can be expressed as $\beta_{ii}(k)x_i(k)$.

Given the β parameters, the dynamics each of the N state variables can be expressed as

$$x_i(k+1) = x_i(k) + \sum_{\substack{j=1 \\ j\neq i}}^{N} \beta_{ji}(k)x_j(k) - \sum_{j=1}^{N} \beta_{ij}(k)x_i(k) + u_i(k) \quad (1)$$

By building the appropriate $\mathbf{A}(k)$ matrix [2], all of these equations can be expressed as a single matrix equation:

$$\mathbf{x}(k+1) = \mathbf{A}(k)\mathbf{x}(k) + \mathbf{u}(k). \quad (2)$$

In this paper, a cost on the cumulative delays in each Center is used, so it is notationally convenient to define $\hat{\mathbf{u}}(K) = \sum_{k=1}^{K} \mathbf{u}(k)$. This means that $\hat{\mathbf{u}}(K)$ is simply a vector containing the cumulative number of departures from each Center at time step *K*. Equation (2) can then be expressed as

$$\mathbf{x}(k+1) = \mathbf{A}(k)\mathbf{x}(k) + \hat{\mathbf{u}}(k) - \hat{\mathbf{u}}(k-1). \quad (3)$$

Aggregate Flow Model Characteristics

The main attractive feature of the aggregate flow model is its ability to model the behavior of the entire NAS with linear equations involving about 20 state variables and 20 control variables. The linear form of the equations means that control and optimization theory can be readily applied to assist in choosing control inputs. The small number of state and control variables means that control inputs for the entire NAS can be readily computed in real time when using many control and optimization techniques. However, the ability of this model to predict the state of the NAS when departure rates differ from their scheduled levels has not been demonstrated.

The principal assumption in the aggregate flow model is that the number of flights that move to

another Center or arrive during a time step is proportional to the number of aircraft in the Center at the start of that time step. The proportionality constants ($\beta_{ij}(k)$) change over time and are found from historical data. This model performs well when predicting NAS behavior under nominal operating conditions [2]. Using proportionality constants derived from historical data may be less valid when traffic patterns differ from nominal operations and when flights are selectively delayed during a ground delay program. Recent research has considered adapting the proportionality constants in real time and in response to flight plans for the current day in response to this issue [12]. Moreover, deriving proportionality constants from historical data precludes using rerouting for TFM, thereby artificially limiting the space of possible TFM solutions.

The small number of control inputs for this model also means that the controls available are relatively crude. This is appropriate for a NAS-wide model, but controls derived with this model would need to be used in conjunction with a lower-level algorithm to assign departure slots to available flights. This lower-level method may impact the validity of the proportionality constants and also the effectiveness of the resulting TFM initiatives.

Optimization Approach

The aggregate flow model was used in conjunction with optimization techniques to find optimal NAS-wide departure rates in response to expected restrictions on airspace and departure rate capacities. Other approaches were considered, such as a model-following adaptive control approach and an optimal control approach. The model-following adaptive control approach was not selected because of the difficulty in finding an appropriate and theoretically tractable model to follow. Moreover, upper and lower bounds on state variables are not easily implemented when using these types of control theory. Lastly, these approaches did not allow for future expected constraints to be considered when selecting control inputs. Model predictive control (MPC) would overcome many of these difficulties, and in fact the actual implementation of the optimization approach utilized here would likely follow the MPC paradigm of implementing the first control action in a sequence, measuring the result, and then re-optimizing over a relevant time horizon.

Cost Function

The cost function is a quadratic cost on cumulative departure delays over some time period. Previous research applying the aggregate flow model used instantaneous delays in the cost function [1]. However, cumulative delays take into account that when flights are delayed in one time step, actual departures must exceed planned departures in some future time step(s) to get back on schedule. In fact, after a time step where actual departures are below scheduled departures, some flights will be delayed until actual departures exceed scheduled departures enough to catch up with the schedule.

Similarly, using a quadratic cost on delays is appropriate because the marginal cost of a delay for airlines and passengers increases with delay time. As delay time increases, passengers, crews, and aircraft start missing connections and ultimately flights must be cancelled. Therefore the cost of increasing a delay from 5 to 10 minutes is much less than increasing a delay from 55 to 60 minutes, justifying the use of a quadratic cost function.

To express the cost function in a precise form that can be used in an optimization problem, a schedule variable must be defined. Let $\hat{\mathbf{s}}(k)$ be a vector containing the scheduled cumulative departures from all Centers at time step *k*. If everything is running on schedule, $\hat{\mathbf{s}}(k) = \hat{\mathbf{u}}(k)$ for all *k*.

The optimization problem considers some time horizon from $k = 1$ to $k = k_f$. If there are N air traffic control Centers, then the total number of control inputs for this time horizon is Nk_f. For a more concise problem formulation, let

$$\hat{\mathbf{u}} = \left[\hat{\mathbf{u}}(1)^{\mathbf{T}} \quad \hat{\mathbf{u}}(2)^{\mathbf{T}} \quad \dots \quad \hat{\mathbf{u}}(k_f)^{\mathbf{T}}\right]^{\mathbf{T}} \qquad (4)$$

and

$$\hat{\mathbf{s}} = \left[\hat{\mathbf{s}}(1)^{\mathbf{T}} \quad \hat{\mathbf{s}}(2)^{\mathbf{T}} \quad \dots \quad \hat{\mathbf{s}}(k_f)^{\mathbf{T}}\right]^{\mathbf{T}}. \qquad (5)$$

With these variables defined, the quadratic cost to be minimized can be expressed as

$$J(\hat{\mathbf{u}}) = \left\|\hat{\mathbf{s}} - \hat{\mathbf{u}}\right\|_2^2. \quad (6)$$

This cost is the sum of the squared cumulative delay in each Center at each time step.

Physical Constraints

There are several constraints for this optimization problem. Many of them are imposed by the underlying reality of the airspace system. The first and most obvious constraint is that the system follows the system dynamical equations (3).

Assuming that flights do not depart before their scheduled departure time and considering that cumulative departures cannot be negative leads to the constraint

$$\mathbf{0} \le \hat{\mathbf{u}} \le \hat{\mathbf{s}}. \quad (7)$$

Of course cumulative departures should be non-decreasing. The magnitude of the increase in cumulative departures from one time step to the next is bounded above by the possible departure rates at the airports in each Center. These two constraints can be expressed together as

$$\mathbf{0} \le \hat{\mathbf{u}}(k) - \hat{\mathbf{u}}(k-1) \le \mathbf{u}_{\max}(k) \quad \forall\, k = 2 \ldots k_f. \quad (8)$$

Similarly, the number of aircraft in each Center is bounded above by the capacity of the airspace in the Center and below by zero, so

$$\mathbf{0} \le \mathbf{x}(k) \le \mathbf{x}_{\max}(k) \quad \forall\, k = 1 \ldots k_f. \quad (9)$$

Here $\mathbf{x}_{\max}(k)$ is a vector that denotes the capacity of the airspace in each Center at time step k, measured in number of aircraft. Likewise $\mathbf{u}_{\max}(k)$ is a vector that denotes the departure rate capacity for each Center at time step k. For this paper, these vectors are assumed to be known exactly for some time horizon. In reality there would be considerable uncertainty regarding the exact value of $\mathbf{x}_{\max}(k)$ and $\mathbf{u}_{\max}(k)$, particularly for large values of k, and future research should consider this uncertainty explicitly.

Constraints for Equity

Other constraints considered for this problem are imposed not due to physical realities of the NAS but rather out of a desire to distribute air traffic delays across Centers in a particular way. More specifically, each of these constraints imposes some requirements on the equity with which delays are distributed across Centers.

The first equity constraint is an upper bound on the Gini coefficient for Center departure delays [17]. The Gini coefficient is a measure of equity that is equal to zero when there is perfect equality (each Center has the same amount of departure delay) and equal to one when there is perfect inequality (one Center incurs all of the departure delay). If $G_{\max}$ is the maximum allowable Gini coefficient for the distribution of departure delays among Centers, with $0 \le G_{\max} \le 1$, and the vector $\mathbf{d}$ contains the delay per departure for each Center, then this constraint can be expressed as

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\mathbf{d}_i - \mathbf{d}_j\right|}{2n\sum_{i=1}^{n}\mathbf{d}_i} \le G_{\max}. \quad (10)$$

A second approach involves adjusting the cost function rather than adding a constraint. Delays can be reduced in some Centers putting a higher weight on their delays in the cost function. To do so, the cost function (6) is modified to have the form

$$J(\hat{\mathbf{u}}) = \left\|\mathbf{C}(\hat{\mathbf{s}} - \hat{\mathbf{u}})\right\|_2^2. \quad (11)$$

Here $\mathbf{C}$ is a diagonal matrix with diagonal entries that designate the weight assigned to delays in each Center at each time step.

Another possible equity-based constraint puts an upper bound on the departure delay per flight over the time horizon for each Center. This upper bound is simply

$$\left\|\mathbf{d}\right\|_\infty \le d_{\max}, \quad (12)$$

where $d_{\max}$ is the desired maximum possible delay per departure for the time period under consideration.

Convex Optimization Problem

Combining either objective (6) or (11) with the constraints (3), (7)-(9), and optionally (10) and (12) leads to the optimization problem used to find the distribution of departure delays for a particular scenario. This problem is a convex optimization problem, a problem in which a convex function is

minimized subject to inequality constraints where convex (or quasiconvex) functions are bounded above by zero and subject to affine equality constraints [18]. Minimizing the objective is equivalent to minimizing its root, in which case either objective (6) or (11) is a norm and therefore clearly convex. The constraints (3) and (7)-(9) are just affine expressions of the vector variable $\hat{\mathbf{u}}$, so they are also convex. Constraint (12) is the sublevel set of a norm, which is convex because any norm is convex. Finally, constraint (10) is convex because the Gini coefficient can be shown to be quasi-convex (see Appendix I), ensuring that its sublevel sets are convex.

Verifying that this optimization problem is a convex optimization problem is important because the theory, and more importantly the practical tools, for minimizing convex optimization problems are well established [18]. In particular, open-source software such as `CVX` can easily be utilized to solve this convex optimization problem [19].

Results and Discussion

Scenario

For the analysis presented here, a five-hour period from 7 am to 12 noon EST on Thursday, May 6, 2004 was considered. This is the same day that was studied in other research on the aggregate flow model [3]. Scheduled departure rates ($\hat{\mathbf{s}}(k)$) and proportionality constants ($\beta_{ij}(k)$) were derived from the actual traffic on this day.

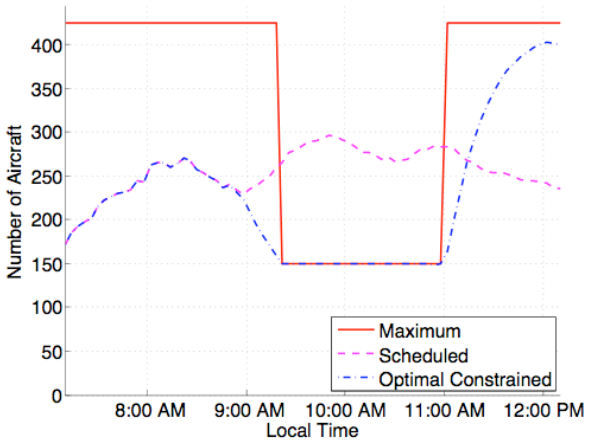
A scenario considered Cleveland Center experiencing a reduction in capacity. The reduction in capacity can be caused by any phenomenon, such as weather or some air traffic control equipment malfunction. Figure 3 depicts the scenario. More specifically, the maximum aircraft count ($\mathbf{x}_{\max}(k)$) in Cleveland was reduced to 150 aircraft for a 100-minute period during this five-hour period. If there were no such restriction on traffic, then the aircraft count in Cleveland Center during this period would peak at just fewer than 300 aircraft. Moreover, during the same 100-minute period, the departure rate in Cleveland Center was reduced to slightly more than 11 departures per 4-minute time step. More than 25 departures were scheduled during some time steps in this period. Such Center-level restrictions are not used in current TFM practice, but they are required for this optimization process. These restrictions could be approximated by summing all of the sector capacities and airport departure rates in a Center.

The optimization problem under consideration is solved using `CVX`, which is called from Matlab. This problem can be solved in less than 10 minutes on a desktop computer, fast enough to be used in a real-time implementation.

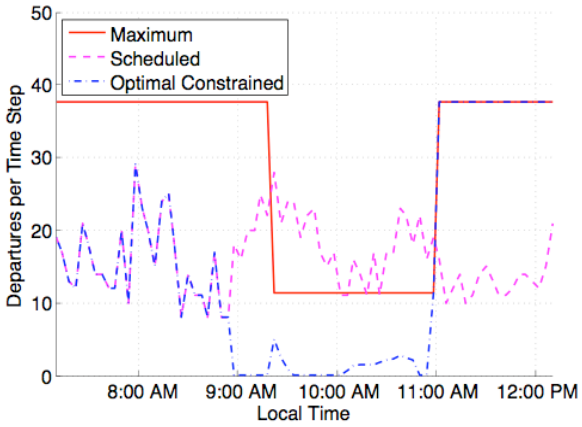
Optimal System Behavior

The first set of simulations demonstrates the optimal departure delays in this scenario when no constraints are implemented for equity. More precisely, the optimization problem discussed above is solved but without constraints (10) and (12).

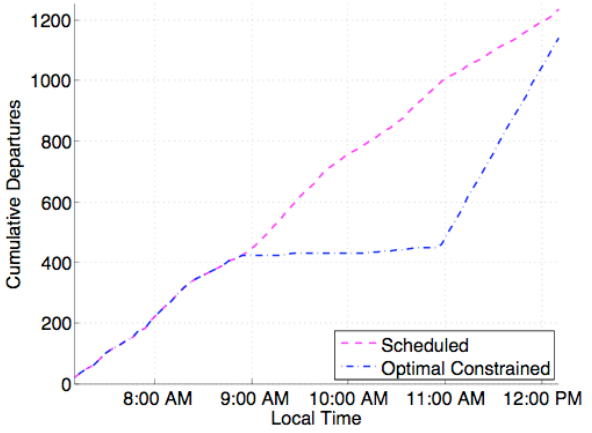
Figure 3 shows the results of the simulation for Cleveland Center. In part (a) of the figure, the restricted aircraft count is plotted over time. About half an hour prior to the activation of the capacity constraints, the aircraft count is reduced dramatically. Immediately after the constraint is lifted, the aircraft count grows quickly as pent-up demand for departures is released. Part (b) shows the departures per time step over time. Before and during the period where the airspace is constrained, the departure rate is even lower than its constrained upper bound. Therefore the airspace constraint is the binding constraint in this situation. As soon as the constraints are lifted, aircraft depart rapidly. Part (c) of this figure demonstrates how aircraft depart at the maximum possible rate following the lifting of the constraints to allow delayed flights to depart. By the end of the simulation, about an hour after the constraints have been lifted, the Cleveland departures are almost back on schedule.



(a)



(b)



(c)

Figure 3. Optimal Constrained Response in Terms of (a) Number of Aircraft, (b) Departure Rate, and (c) Cumulative Departures in Cleveland Center

The optimization approach is able to satisfy these constraints with 133,030 minutes of departure delay, or just over 6 minutes of delay per departure in the NAS. Almost half of those minutes of departure delay were absorbed by Cleveland Center. This corresponds to more than 47 minutes per departure from Cleveland Center. This inequity in the distribution of delays among air traffic control Centers is portrayed graphically in the histogram of delays in Figure 4. Such inequality has been observed in real delay data [4] and motivates the analysis of equity constraints presented in the following sub-section.

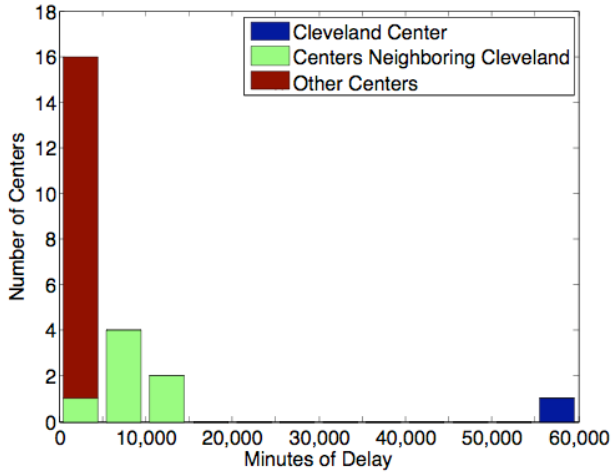


Figure 4. Departure Delay Histogram for Optimal Constrained Response

Delay and Equity Tradeoff

The scope of the aggregate flow model and the flexibility of the convex optimization approach used here allow for some unique NAS-level analyses. Here a study analyzes several methods for trading off delay and equity in the distribution of delay between Centers.

To investigate this tradeoff, we will first enforce constraint (10), an upper bound on the Gini coefficient. When no constraint on inequality is imposed, the Gini coefficient is equal to 0.64, a relatively high value indicating an unequal distribution of delay per departure among Centers. The upper bound on the Gini coefficient ($G_{\max}$) is varied between 0.01 (almost perfect equality in delay per departure distribution) and 0.7 (highly unequal delay per departure distribution). Actually an upper bound on the Gini coefficient above 0.64

will have no effect on the optimization problem, because this is the value of the Gini coefficient at the optimal distribution of delay.

Figure 5 shows the tradeoff between total departure delay and equality, as measured by the Gini coefficient, for this particular scenario. When no constraint on equality is imposed, the delay is about 133,000 minutes. If the upper bound on the Gini coefficient is decreased to 0.2 in order to enforce a more equitable distribution of delays, the total delay increases dramatically to about 200,000 minutes. Enforcing even more equality leads to higher total departure delays.

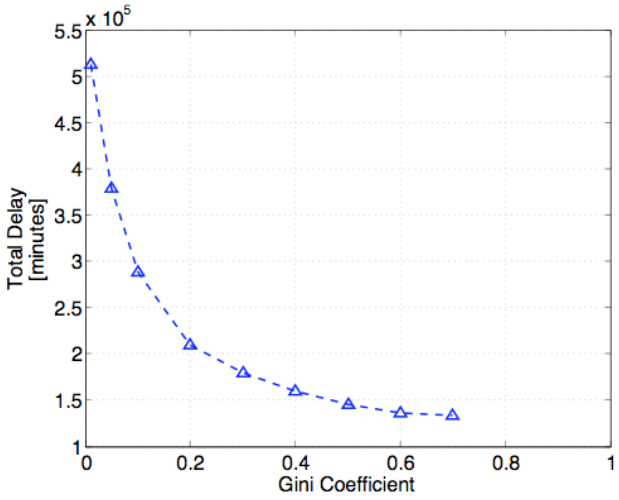


Figure 5. Tradeoff Between Total Delay and Gini Coefficient

This tradeoff can also be investigated on a per departure basis. Figure 6 shows the average departure delay in Cleveland Center, the Center with the most delays in this scenario, and the NAS-wide average departure delay. As the upper bound on inequality decreases, the average delay per departure increases NAS-wide and also decreases in Cleveland.

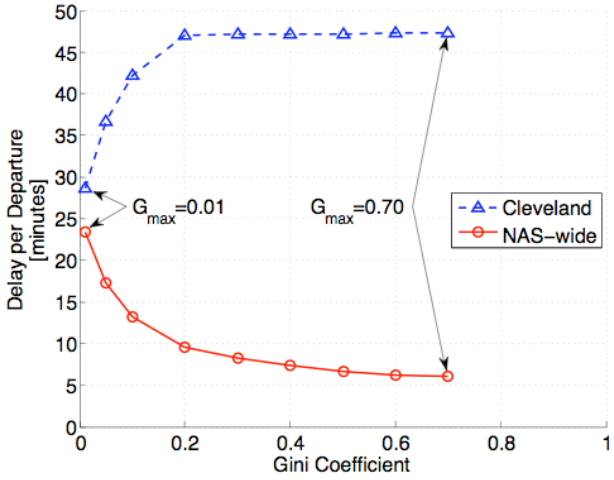


Figure 6. Impact of Gini Coefficient Bound on Delay per departure NAS-wide and in Cleveland

In Figure 6 it can be seen that as the Gini coefficient upper bound decreases from 0.7 to 0.2, there is almost no change in the delay per departure in Cleveland. This means that the gains in equality as the Gini coefficient decreases to 0.2 are essentially totally achieved by increasing delays in other Centers rather than decreasing delays in Cleveland. It is only when an extremely high level of equality is enforced that delays in Cleveland start to decrease. Unfortunately, as can be seen in Figure 5, enforcing such high levels of equality leads to exceedingly large increases in NAS-wide total delay values. Thus the aggregate flow model indicates that enforcing equality in average delay distributions is a poor way of alleviating delay in a particular Center.

Weighted Centers in Cost Function

A second option for shaping the distribution of delays among Centers is simply to put a higher cost on delays in some Centers than on delays in others. For example, in the scenario under consideration, Cleveland Center is constrained and therefore absorbs a high amount of delay when no effort is made to distribute the delays evenly. Therefore it makes sense to put a higher weight on delays in Cleveland when computing TFM actions in this scenario. Very busy or high-priority Centers could always receive a larger weight in the cost function when computing optimal departure delays.

For the analysis of this method, all of the departure delays in Cleveland Center were assigned weights that took values between 1 and 10, while the weights on delays in each other Center remained at 1. Figure 7 shows the impact of increasing the weight on Center delays in Cleveland on delay per departure in Cleveland and NAS-wide, as well as the impact on equality as measured by the Gini coefficient. With this approach, Cleveland delay per departure can be driven below NAS-wide delay per departure; attempting to do so with the other approaches led to infeasible problems. Of course doing so means that other Centers are experiencing higher delay per departure than Cleveland, even though it is the constrained Center. This is why the Gini coefficient remains high as the Cleveland Center delays are reduced.

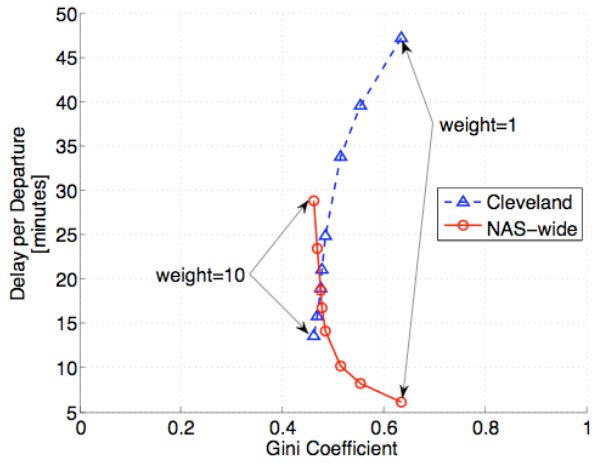


Figure 7. Impact of Increasing Weight on Cleveland Delays on Delay per departure NAS-wide and in Cleveland and on Gini Coefficient

Delay and Center Delay per Departure Tradeoff

A more direct approach to reducing the delays in a high-delay Center is to implement a constraint that bounds the delay per departure in each Center. This is accomplished by using constraint (12) rather than (10) in the optimization problem. In the simulations with this constraint, the scenario described and simulated above was again used.

The tradeoff between total delay and the maximum delay per departure in any given Center is depicted in Figure 8. The y-axis on this figure is identical to that in Figure 5 to facilitate comparison between the results. Reductions in the bound on Center delay per departure induce smaller increases in total delay than reductions in the Gini coefficient bound.

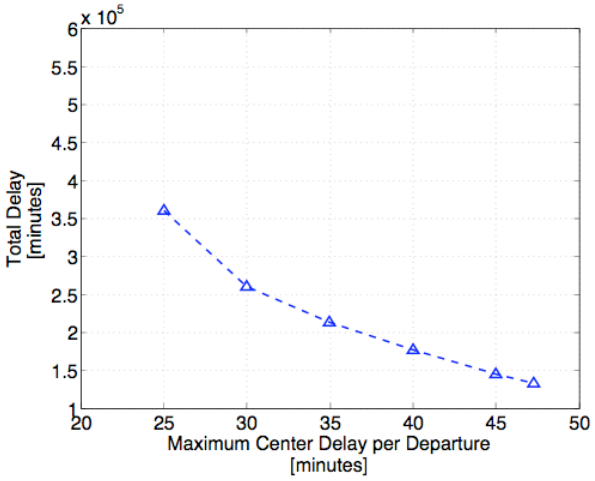


Figure 8. Tradeoff Between Total Delay and Maximum Center Delay per Departure

Figure 8 shows the impact of various Center delay per departure bounds on delay per departure in Cleveland and the entire NAS, as well as on the Gini coefficient. As the delay per departure in Cleveland are reduced by the upper bound, the NAS-wide delay per departure increase. However, comparing Figure 8 with Figure 5 reveals that the bound on Center delay per departure is much more effective at reducing severe delays in a constrained Center without imposing excessive delay demands on other Centers. For example, Figure 5 indicates that bounding the Gini coefficient above by a very low value (around 0.07) will reduce the delay per departure in Cleveland to about 30 minutes by increasing the delay per departure NAS-wide to about 25 minutes. Enforcing an upper bound on Center delay per departure as shown in Figure 9 reduces the delay per departure in Cleveland to 30 minutes while only increasing the delay per departure NAS-wide to about 12 minutes.

Interestingly, Figure 9 also indicates that the Gini coefficient remains relatively high even as the maximum Center delay per departure ($d_{\max}$) is reduced. Even when the delay per departure is bounded above by 25 minutes per departure, the Gini coefficient is still around 0.3. Figure 10 shows

the histogram of the delay distribution in this scenario (not delay per departure). Many Centers have low levels of total delay, while a few have total departure delays as high as or higher than that of the constrained Center. Intuitively this makes sense – the neighbors of the constrained Center would be expected to sustain larger levels of delays than Centers far from the constrained Center. While a few Centers have higher total departure delays in this case, Cleveland Center has the highest delay per departure (25 minutes).

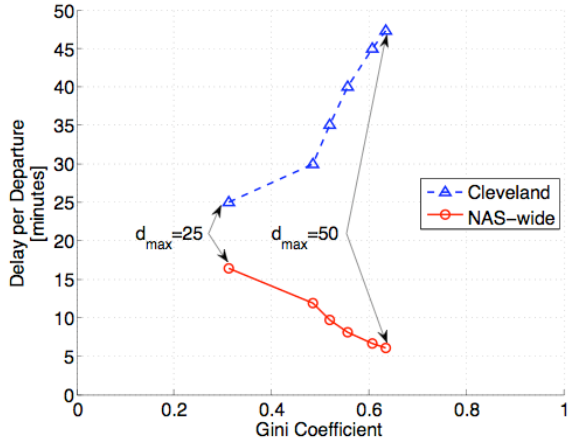


Figure 9. Impact of Center Delay per Departure Bound on Delay per departure NAS-wide and in Cleveland and on Gini Coefficient

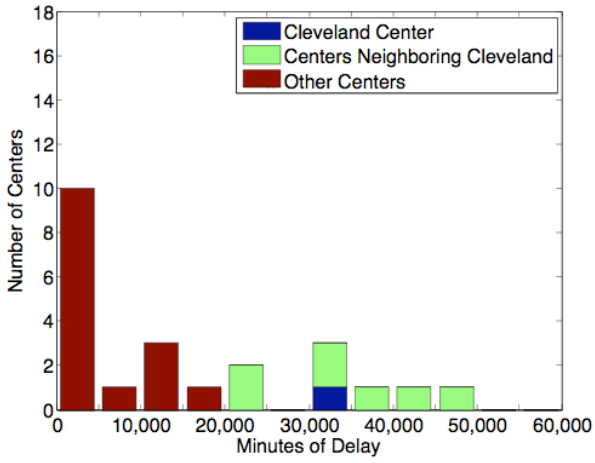


Figure 10. Distribution of Departure Delay When Maximum Center Delay per Departure is 25 minutes

To compare the results of these three methods more directly, curves showing the tradeoff between Cleveland delay per departure and NAS-wide delay per departure for each approach are plotted in Figure 11. Using a bound on the Gini coefficient is clearly an inferior way to reduce delay in a constrained Center, as its tradeoff curve is strictly worse than those generated by the other two options. Bounding Center delay per departure and putting a weight on delays in the constrained Center perform almost identically until the bound on Center delay per departure becomes too low. Therefore, when the goal is to reduce delays in one Center, the best approach is to put a higher weight on delays in that Center.

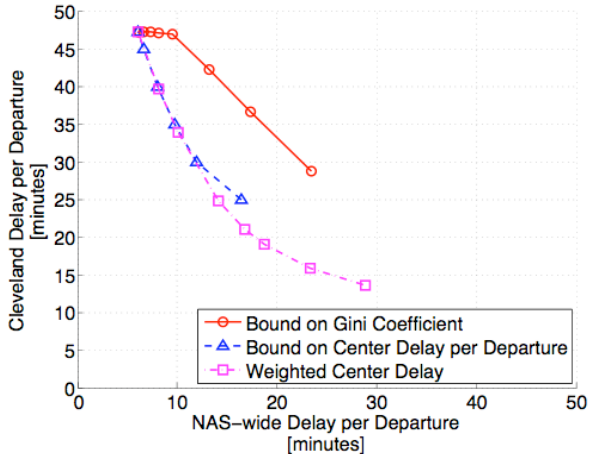


Figure 11. Tradeoff Between Cleveland Delay per Departure and NAS-wide Delay per Departure for Three Approaches

Figure 11 masks the fact that when weighting Center delays, at some point the weight becomes so large that the constrained Center is no longer experiencing the highest delay per departure. Figure 12 shows the delay per departure in the Center with the largest delay per departure on the y-axis and the NAS-wide delay per departure on the x-axis. Putting a higher weight on delays in the constrained Center in the cost function leads to reasonable results, except when the weight becomes too large and induces other Centers to experience excessive delays. Putting an upper bound on the maximum Center delay per departure produces the most intuitive and useful results, largely because the bound applies to all Centers, not just the Center that is constrained.

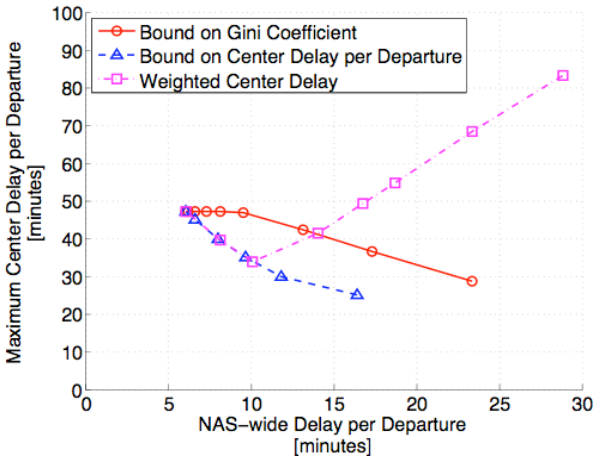


Figure 12. Tradeoff Between Maximum Center Delay per Departure and NAS-wide Delay per Departure for Three Approaches

The three approaches presented above all use the aggregate flow model. While this model allows for finding optimal NAS-wide solutions, these approaches are also hindered by the limitations of the model. In particular, it is not possible in the aggregate flow model to implement ground delay programs only for flights departing for particular Centers or to change aircraft routes. These more precise aircraft-level actions may be useful in easing the delay at a highly constrained Center [5-6].

Conclusions

The aggregate flow model can be used to distribute delays over time and air traffic control Centers to minimize a quadratic cost on delays in response to expected airspace and departure rate constraints. The aggregate flow model has a simple linear structure and significantly fewer state variables and control inputs than other TFM models, so the computational complexity of NAS wide optimization using this model is significantly lower than for optimization approaches using other TFM models. The ability to quickly perform NAS wide TFM optimization of predeparture delays with this approach could provide useful advice to decision makers at the FAA when they are responding to constraints on airspace or departure rates.

This optimization approach tends to allocate delays primarily to the Center that is experiencing capacity constraints. Modifications to the approach can shape the equity of the resulting optimal delay distribution. Bounding the Gini coefficient above enforces a more equitable distribution of delay per departure, but increasing equality does not alleviate delays in a particular Center without severely punishing other Centers. Simulation results indicate that more reasonable approaches are to directly constrain Center delay per departure or to increase the weight on delays in the constrained Center. Both of these approaches work well to decrease the delays in the constrained Center, but applying the upper bound on delay per departure in any Center is a more useful tool because its impact is more obvious. In particular, when applying weights to Center delays it is possible to apply excessive weighting, which leads to other Centers experiencing delays that exceed those initially experienced by the constrained Center.

Future Work

There is significant future work to be done in this area of TFM research. Other optimization formulations that use the aggregate flow model should be considered. Uncertainty, particularly in the timing and severity of airspace constraints, should be considered in this modeling and optimization approach.

Important work remains to validate that the optimal departure rates proposed by this approach impact the NAS as the model predicts they will. This will involve determining how to implement such departure rates. In particular, which flights should be allowed to depart when a departure rate is lower than scheduled? Does this choice affect the validity of the proportionality constants in the model?

Finally, the investigation here into equity in the distribution of delays among air traffic control Centers naturally could be applied to equity in the distribution of delays among airlines. One simple way to research implementing equity constraints in the distribution of delays would be to use several aggregate flow models running in parallel, one for each airline, to model the NAS. This would increase the system state but hopefully not prohibitively. Then the upper bound on the Gini coefficient could enforce equity for airlines. Upper bounds on the Gini coefficient and the infinity norm

could also be used with totally separate models that also make use of convex optimization techniques.

References

[1] Grabbe, S. and Sridhar, B., August 2005, "Congestion Management with an Aggregate Flow Model," AIAA Guidance, Navigation, and Control Conference and Exhibit, San Francisco, CA.

[2] Sridhar, B., Soni, T., Sheth, K., and Chatterji, G., August 2004, "An Aggregate Flow Model for Air Traffic Management," AIAA Guidance, Navigation, and Control Conference, Providence, RI.

[3] Chatterji, G. B. and Sridhar, B., September 2005, "Some Properties of the Aggregate Flow Model of Air Traffic," AIAA Aviation, Technology, Integration and Operations Conference, Arlington, VA.

[4] Sridhar, B. and Swei, S. S. M., September 2007, "Computation of Aggregate Delay Using Center-based Weather Impacted Traffic Index," AIAA Aviation Technology, Integration and Operations Conference, Belfast, Northern Ireland.

[5] Grabbe, S., November 2007, "Impact of Deterministic Constraints on New York Flights," INFORMS Conference, Seattle, WA.

[6] Grabbe, S., Sridhar, B., and Mukherjee, A., "New York Flow Control with Deterministic En route Capacity Constraints," Air Traffic Control Quarterly (submitted).

[7] Bertsimas, D. and Stock, S., August 1994, "The Air Traffic Flow Management Problem with Enroute Capacities," White Paper 3726-94, Massachusetts Institute of Technology, Cambridge, MA.

[8] Bayen, A. M., Grieder, P., Sipma, H., Meyer, G., and Tomlin, C. J., May 2002, "Delay Predictive Models of the National Airspace System Using Hybrid Control Theory," American Control Conference, Anchorage, AK.

[9] Roy, S., Sridhar, B., and Verghese, G. C., June 2003, "An Aggregate Dynamic Stochastic Model for an Air Traffic System," USA/Europe Air Traffic Management Research and Development Seminar, Budapest, Hungary.

[10] Ball, M. O., Hoffman, R., Odoni, A. R., and Rifkin, R., January-February 2003, "A Stochastic Integer Program with Dual Network Structure and Its Applications to the Ground-Holding Problem," Operations Research, No. 1, Vol. 51, pp 167-177.

[11] Menon, P. K., Sweriduk, G. D., Lam, T., Cheng, V. H. L., and Bilimoria, K. D., August 2003, "Air Traffic Flow Modeling, Analysis, and Control," AIAA Guidance, Navigation, and Control Conference and Exhibit, Austin, TX.

[12] Saraf, A. P. and Slater, G. L., September 2007, "Adaptive Eulerian Model for Optimal Air Traffic Management," AIAA Aviation Technology, Integration and Operations Conference, Belfast, Northern Ireland.

[13] Robelin, C., Sun, D., Wu, G., and Bayen, A. M., June 2006, "MILP Control of Aggregate Eulerian Network Airspace Models," American Control Conference, Minneapolis, MN.

[14] Roy, K. and Tomlin, C. J., August 2008, "Traffic Flow Management using Supply Chain and FIR Filter Methods," AIAA Guidance, Navigation, and Control Conference and Exhibit, Honolulu, HI.

[15] Bayen, A. M., Raffard, R. L., and Tomlin, C. J., July 2004, "Adjoint-Based Constrained Control of Eulerian Transportation Networks: Application to Air Traffic Control," American Control Conference, Boston, MA.

[16] Malone, K. M., June 1995, "Dynamic Queuing Systems: Behavior and Approximations for Individual Queues and for Networks," PhD Thesis, Massachusetts Institute of Technology, Cambridge, MA.

[17] Gini, C., 1912, "Variabilitá e mutabilita," Reprinted in Memorie di metodologia statistica (Ed. E. Pizetti and T. Salvemini), Rome: Libreria Eredi Virgilio Veschi, 1955.

[18] Boyd, S. and Vandenberghe, L., 2004, "Convex Optimization," Cambridge, UK: Cambridge University Press, http://www.stanford.edu/~boyd/cvxbook/

[19] Grant, M. and Boyd, S., March 2008, "CVX: Matlab software for disciplined convex programming," Web page and software, http://stanford.edu/~boyd/cvx

Acknowledgements

Dr. P. K. Menon of Optimal Synthesis Incorporated provided helpful insight into previous Eulerian model research in the early stages of this research. Professor Stephen Boyd of Stanford University assisted in evaluating the convexity properties of the Gini coefficient.

Email Addresses

Michael Bloem can be contacted at michael.bloem@nasa.gov and Banavar Sridhar can be contacted at banavar.sridhar@nasa.gov.

Appendix I

A simple proof of the quasiconvexity of the Gini coefficient is given here. This proof is based on showing that the expression for the Gini coefficient is quasiconvex in its variables. These variables are not controlled directly in this case, so composition rules are invoked to show that the Gini coefficient is also quasiconvex in the control variables. For an elaboration on this approach to studying the convexity properties of functions, see chapter 3 of [18].

Recall from equation (10) that the Gini coefficient can be expressed as

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\mathbf{d}_i-\mathbf{d}_j\right|}{2n\sum_{i=1}^{n}\mathbf{d}_i}. \qquad (13)$$

Here $\mathbf{d}_i$ refers to the delay per departure in Center i. First it will be shown that (13) is quasiconvex in $\mathbf{d}$.

To show that (13) is quasiconvex in $\mathbf{d}$, the sublevel sets of (13) will be examined. If these are all convex sets, then the Gini coefficient is quasiconvex in $\mathbf{d}$. The γ-sublevel set can be expressed as

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\mathbf{d}_i-\mathbf{d}_j\right|}{2n\sum_{i=1}^{n}\mathbf{d}_i}\le\gamma. \qquad (14)$$

This can be re-written as

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\mathbf{d}_i-\mathbf{d}_j\right|-2\gamma n\sum_{i=1}^{n}\mathbf{d}_i\le 0, \qquad (15)$$

where the inequality does not change direction because the sum of delays is nonnegative.

Expression (15) can be shown to describe a convex set in $\mathbf{d}$ by verifying that the left hand side of the inequality is a convex function in $\mathbf{d}$ for all values of γ. This is sufficient because any sublevel set of a convex function is a convex set, and here we are studying the 0-sublevel set of the expression on the left hand side. The left hand side can be expressed as the pointwise maximum of 2^{n-1} linear expressions in $\mathbf{d}$, with another linear expression in $\mathbf{d}$ subtracted from the result. For example, if $n=2$, the left hand side can be expressed as:

$$\max\left\{2(\mathbf{d}_1-\mathbf{d}_2),\ 2(\mathbf{d}_2-\mathbf{d}_1)\right\}+\left(-2\gamma n\sum_{i=1}^{n}\mathbf{d}_i\right).$$

The pointwise maximum of linear expressions is convex, so the first term this version of the left hand side is convex. The second term is linear and therefore also convex. Therefore the entire left hand side of (15) is convex, meaning that (15) as a whole describes a convex set for any value of γ. All of its sublevel sets are convex, so the Gini coefficient (13) is quasiconvex in $\mathbf{d}$.

Now that the quasiconvexity of (13) in $\mathbf{d}$ has been established, quasiconvexity-preserving operations will be utilized to show that it is also convex in the control variable $\hat{\mathbf{u}}$. Note that each component of $\mathbf{d}$ can be computed as the sum of the delays in a Center divided the total number of scheduled departures in the Center:

$$\mathbf{d}_i=\frac{\Delta t\sum_{k=1}^{K}\left[\hat{\mathbf{s}}_i(k)-\hat{\mathbf{u}}_i(k)\right]}{\hat{\mathbf{s}}_i(K)}, \qquad (13)$$

where Δt represents the model time step. The delay per departure in each Center is simply an affine expression of the control variable $\hat{\mathbf{u}}$.

Overall, the Gini coefficient is a quasiconvex function in $\mathbf{d}$, which is an affine transformation of $\hat{\mathbf{u}}$. Quasiconvexity is preserved under composition

with an affine transformation, so the Gini coefficient is quasiconvex in $\hat{\mathbf{u}}$.