Distributed Mechanisms for Determining NAS-Wide Service Level Expectations: Year 3 Report

By

Michael Ball
Cynthia Barnhart
Mark Hansen
Yi Liu
Prem Swaroop
Vikrant Vaze
Chiwei Yan

November 12, 2013

The NEXTOR-II team continued its work on a mechanism to support air traffic flow management strategic planning on the day-of-operations. Specifically, the system under design seeks to determine service level expectations that represent a consensus among impacted flight operators. During Phase 1, several alternative approaches were explored: this work led to agreement on an iterative approach whose central engine was based on the majority-judgment voting mechanism. During Phase 2, the detailed system processes were designed and analysed via simulation. During Phase 3, the various system concepts were refined, a benefits analysis was carried out, concept evaluation software was developed and user outreach activities were initiated.

This report contains two chapters and an appendix. Chapter 1 is an executive summary of the entire system. Chapter 2 contains an intuitive description of the various processes underlying it. The appendix provides the technical descriptions of the research topics.

CHAPTER 1:

Distributed Mechanisms for

Determining NAS-Wide Service Level Expectations:

Concept Description

Distributed Mechanisms for Determining NAS-Wide Service Level Expectations: Concept Description

1.1 Introduction

Over the past 15 or more years collaborative air traffic management (CATM) has become a fundamental principle underlying all new air traffic management (ATM) system development in the U.S. In fact, this trend goes beyond the U.S. to Europe and increasingly the rest of the world. Its origins go back to the deployment of new information exchange and resource allocation mechanisms for planning and controlling ground delay programs (GDPs) in the U.S. CATM originally, and for most its life, was knows as collaborative decision making (CDM). GDP decision support tools evolved and became more sophisticated and the underlying GDP ideas were transferred to the enroute environment with the development of airspace flow programs. A variety of other tools, based on CATM paradigms, have been developed and adopted or are on their way to adoption. In Europe, where a somewhat different set of ATM challenges exist, CATM has been most vigorously adopted in the context of the *airport CDM* decision support tools.

It is probably safe to say that the bulk of the CATM-based development has focused on tools and processes to support very specific ATM operational decisions, e.g. assigning ground delay to a specific flight during a GDP. At the same time there is a very important strategic planning aspect to the daily execution of ATM. Specifically, FAA traffic managers consult with airline/flight operator operational personnel at both the local and national levels in planning operational strategies for the day. These take the form of strategic planning teleconns (SPTs). To be sure, the SPTs should be considered a very important part of the general trend toward the widespread use of CATM . However, while the various CATM-based decision support tools employ novel resource allocation and information exchange principles, the SPTs do not employ any new collaborative principles or technologies and are, by-and-large, highly unstructured.

It should be emphasized that the SPTs perform a very legitimate and even vital function in the overall traffic management process. Specifically, flight operators have key information not known by the FAA, including air carrier business objectives and economic tradeoffs and the status of aircraft and personnel, just to name a few. However, while the CATM initiative has produced a host of innovations in the manner in which specific traffic management initiatives (TMIs) are planned and controlled, very little innovation has been directed toward the

operation of SPTs. While this per se may not necessarily be bad, there are several concerns and issues related to SPTs and more generally strategic planning on the day-of-operations that merit research attention:

1. The SPTs are free form and highly unstructured and so, at times, can devote an inordinate amount of time to unimportant topics.
2. Again due to their free-form nature, the SPTs do not attempt to assign priority to the various flight operators based on objective measures. Thus, the more persistent and/or "loudest" flight operators tend to have the most influence.
3. The operational concept for the Next Generation Air Transportation System (NextGen) calls for a performance-based ATM system. One embodiment of this concept calls for the separation of strategic ATM planning into i) service level expectation setting and the ii) planning of an operational response. Flight operator input should be provided in i) and the air navigation service provider (ANSP) should then optimize ii) based on the output of i). In fact, today's SPT's totally focus in ii).

The proposed system – COuNSEL, CONsensus Service Level Expectations -- described in this white paper addresses the service level expectation (SLE) setting problem and, in so doing, seeks to eliminate the deficiencies discussed in 1) and 2).

1.2 Desirable Properties

The starting point for the COuNSEL design process undertaken was to lay out a set of desirable properties that a good design should have. These are listed below.

1) ***Consensus Building:*** the system should take into account the input of all involved flight operators and should generate an output that represents a consensus of those flight operators.
2) ***Equitability:*** the system should treat the involved flight operators equitably. It should be noted, however, that the notion of equitability should take into account appropriate measure of flight operator interests. For example, it can be expected that the influence of a flight operator over the consensus solution should increase as the number of impacted flight operations belonging to that flight operator grows.
3) ***Practicality:*** the system should be easy to administer and efficient in terms of effort and time commitment required by both the ANSP and the flight operators . It should also be efficient in terms of required computing resources.
4) ***Confidentiality:*** the system should seek to minimize input information required from the flight operators and should keep confidential any information provided.

5) ***Strategy Proofness:*** To the extent possible the system should encourage truth-telling on the part of the flight operators and prevent or minimize the opportunities for "strategic behavior", i.e. system gaming.
6) ***Single Winner Determination:*** the system should provide a single recommended consensus output.

1.3 System Concepts and Operation

The basic mode of operation for COuNSEL is fairly straight-forward, however, it is quite different from the SPTs because its basic output is different. As discussed above the system seeks to set service level expectations. Another process, not the subject of this white paper, takes the further step of converting the service level expectations into planned TMIs. Note that SPTs talk directly in terms of TMIs, e.g. discussing which TMIs should be run and which parameter setting should be used. The SLE problem can be viewed as one of setting constraints or guidelines to be used to determining those TMIs and their parameters. A basic question then is what form should the output of a SLE setting process take.

The TMI planning process is viewed as a design problem that requires performance goals. The output of COuNSEL is a set of such goals. In the follow-on step, traffic management specialists carrying out the design process are faced with decisions that require trading off one performance criterion with another. The performance goals provide the designers with the necessary information to do this. As discussed above, this second step is not discussed in this white paper: only the first goal-setting step is addressed.

1.3.1 Performance Metrics

Before describing the exact nature of the output, it is worthwhile to consider an appropriate set of performance criteria. The global ATM community working through ICAO has agreed upon a set of eleven service expectation categories. These were considered carefully in the context of the SLE setting problem and a set of three was chosen to be relevant to the TMI design and control problem. These are discussed below.

Capacity measures the number of flight operations that the overall system or constituent subsystems can process safely over a specified time period. In the context of a GDP, an important capacity metric is the number of arrivals that can be accepted by an airport per hour. Capacity is perhaps the most visible and important performance category as it directly relates to flight delays and more generally the ability of an air carrier to maintain its schedule.

Predictability has multiple interpretations depending on the time frame in question. For planning specific TMIs, predictability refers to the degree to which flight operators know in advance resources available to them and, more generally, the intentions and planned actions of the ANSP. An ANSP could increase predictability by announcing farther in advance its intention to carry out specific TMIs, and giving earlier indications of the ground delays assigned to flights, earlier announcements of the open/closed status of airways, etc.

Efficiency refers to the cost-effectiveness of individual flight operations from the perspective of the flight operator. During a GDP, a policy that leads to high amounts of airborne holding would be less efficient than one that converted that airborne delay into less costly ground delay.

TMI design strategies very often trade off these performance criteria either explicitly or implicitly. For example, one GDP strategy might limit the amount of assigned ground delay, when compared to others. Such a strategy would tend to send a larger number of flights to the airport earlier in hopes that the weather would clear earlier than expected or that a slightly higher than planned acceptance rate could be accommodated. Such a strategy on the average would lead to higher rates of arrival throughput, increasing the capacity metric, but larger amounts of airborne holding, decreasing the efficiency metric.

Another strategy might announce and implement early in the day certain TMI actions, such as ground delays and reroutes. These would provide ample time for the flight operators to plan for the day's operations, allowing them, for example, to cancel strategic flights and to take early steps to re-accommodate passengers. On the other hand, such a strategy would have a tendency to impose unnecessary ground delays or reroutes. Thus, it would tend to have a higher level of predictability but lower levels of capacity and efficiency.

The SLE setting problem is to provide guidance to TFM specialist on how to trade off TMI performance in the three performance categories given above. The approach chosen to do this involves choosing a specific metric for each of the three categories and specifying a goal for each of those metrics. Thus, the output of COuNSEL is a vector of size three that contains a goal for each of the metrics. The metrics chosen are normalized to be between 0 and 1, with 1 being the best possible value and 0 the worst. One can view a value of 1 indicating the best performance level for that performance category on a perfect-weather day. Of course, a very simplistic solution to this goal setting problem would be to choose a goal of 1 for each metric. However, a vector of three 1's provides little insight or tradeoff guidance. Rather one should view the process as starting with an assessment of the weather and traffic conditions. This in turn implies constraints on the set of feasible goal vectors. For example, it would generally be the case that

on a poor weather day, it would be impossible to achieve a vector of three 1's. In general, the constraints implied by the day's conditions would generate an *efficient frontier* of possible vector values. Conceptually any such vector could be achieved on the day given an appropriate TMI. In fact, the choice between these vectors represents the choice among TMI strategies and provides exactly the tradeoff information that is sought. For example, suppose that the SLE vector was ordered as follows:

(capacity metric, predictability metric, efficiency metric)

Consider the following possible vectors chosen from the efficient frontier:

A: (.95, .90, .91), B: (.90, .94, .89), C: (.97, .87, .89)

Suppose particular flight operator had a very heavy emphasis on capacity. That flight operator when given the choice between A and B might choose A, indicating a willingness to increase capacity and to a less extent efficiency, while sacrificing predictability. That flight operator might further be given the choice between A and C and choose C again in order to increase capacity while further sacrificing predictability and efficiency. In this way, by choosing a particular vector, a flight operator is forced to make key performance tradeoffs.

Given this choice of three performance categories one is still left with the problem of choosing three specific metrics. Obviously, the choice of specific metrics is very fundamental and a key driver to the effectiveness of the system. However, the development and/or choice of metrics is not a focus of the research activity summarized here and so specific metric definitions will not be provided in this white paper. Henceforth, it is assumed that metrics for capacity (C), predictability (P) and efficiency (E) have been provided. The output of COuNSEL is a vector of metric values: (m_C, m_P, m_E), where each of m_C, m_P, and m_E are between 0 and 1.

The output vector represents goals for the metric values that the ANSP should seek on the day in question. The "feasible" values for (m_C, m_P, m_E) depend on the conditions of the day so that on poorer weather days, the possible values will tend to be lower (closer to 0) than on better weather days.

1.3.2 Feasible Metric Vectors

COuNSEL seeks to generate a consensus among the flight operators. As such an iterative process is required where each flight operator evaluates and compares possible vectors. Flight operators are also given the opportunity to generate candidate vectors. Figure 1 illustrates the domain of feasible vectors, flight operator preferences and a consensus vector, in the case where there are two (rather than three) metrics.

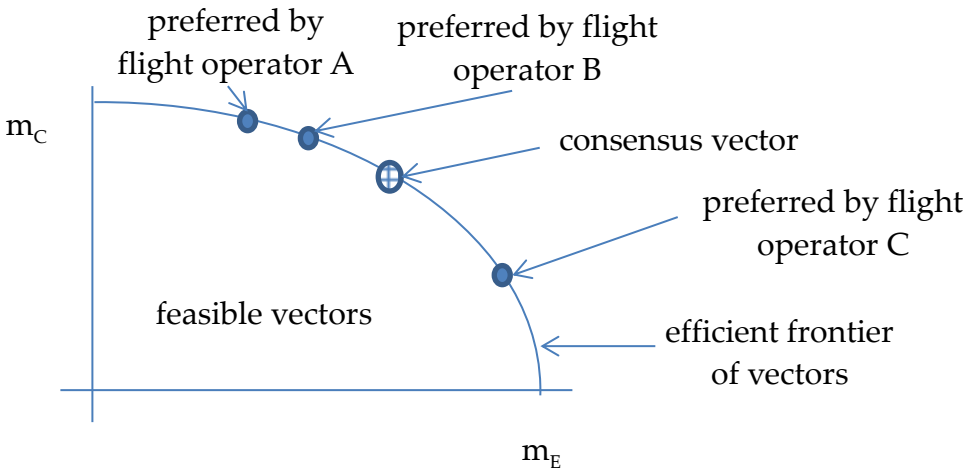


Figure 1.1

For any given day of operations there would be many feasible vectors. However, the flight operators and the ANSP should only consider vectors on the efficient frontier. These dominate the others in the sense that for a vector on the efficient frontier, it is not possible to increase one metric value without decreasing another. Generally, it is the case that each flight operator would have a preferred vector. The consensus vector would tend to represent a compromise among the vectors preferred by each flight operator.

1.3.3 System Operation

Figure 2 illustrates the basic operation of COuNSEL. It proceeds using a process in which candidate vectors are generated and evaluated by the flight operators until one is found that represents a consensus. The "evaluation" of a metric vector on the part of a flight operator involves assigning a "grade" to the vector. A grade is a value between 0 and 100, 100 being the best possible and 0 the worst. The flight operators are free to interpret and assign grades as they see fit. However, in concept, grades should vary in proportion to the value, or inverse of cost, that a vector brings to the flight operator. The flight operators will be asked to grade many vectors and given the opportunity to generate new vectors, including providing their most preferred vector. Over time, it is expected that flight operators will develop formal approaches

for grading vectors and eventually automated systems for grading. The use of flight-operator-assigned grades is very robust in the sense that it can evolve over time as all stakeholders become more familiar and expert with COuNSEL concepts. For example, when the system is first released flight operators might use a manual process driven by judgment and heuristic rules; however, in the long run it is anticipated that the grading process would become fully automated, making it a low-overhead and very fast process.

Similarly, the use of flight-operator-generated candidate vectors would likely evolve. The ability for a flight operator to generate a candidate vector will require that the ANSP provide some additional information or capabilities to the flight operators. As discussed, the feasibility of a vector will depend on the conditions of the NAS on the day-of-operations and so the ANSP must provide to the flight operators either some representation of the constraints defining the feasible region of vectors or the capability to generate feasible vectors, e.g. via a web-based application. It is not necessary for the flight operators to generate vectors, as the ANSP will do this in any event. However, a flight operator may find it advantageous to generate vector(s) as this may make it more likely that the consensus vector will be closer to its most preferred vector. Thus, a possible scenario might be that in the initial deployment of COuNSEL no flight operators would generate vectors but over time flight operators would develop the capability and use it effectively.

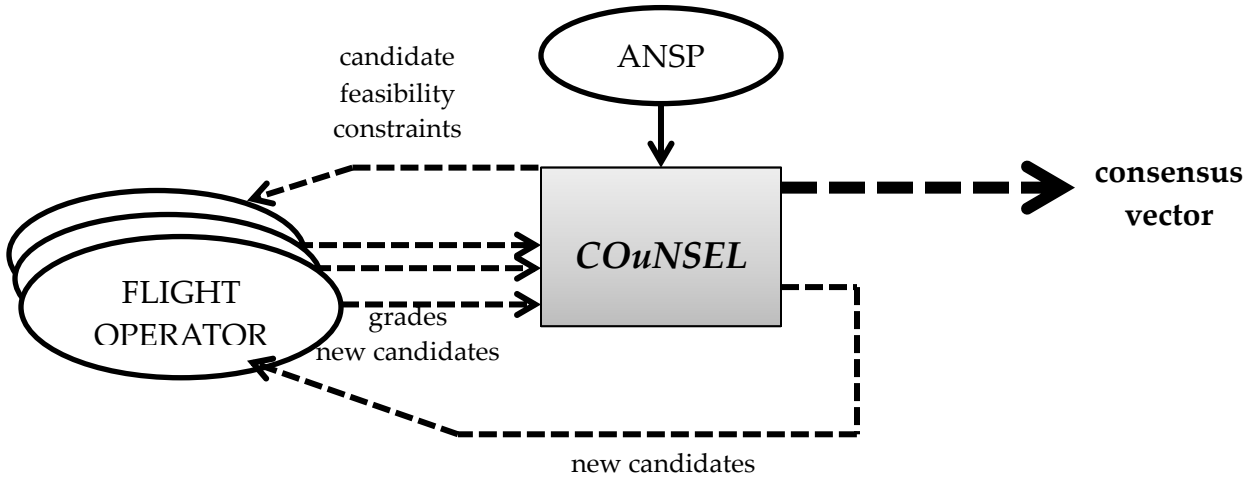


Figure 1.2

Since the COuNSEL process produces the equivalent of a consensus strategic plan on the day-of-operation it is important that it have very fast response time. As discussed above, in the long run, it should be the case that all processes both on the ANSP and flight operator sides should be automated so that the entire process, including multiple iterations should be completed in a

matter of minutes, if not seconds. In the initial stages of implementation, there may be human involvement in the grading so that response time should be slower. However, there is no reason that total response time of less than 30 minutes should not be possible.

1.3.4 Consensus Vector Definition

A fundamental question to ask is what is the definition of a consensus vector. The theory that underlies COuNSEL is the majority judgment voting procedure. The winner is the candidate vector that has the highest *majority grade*. The majority grade of a vector is the grade G, such that a majority of flight operators have assigned a grade of G or higher. Figure 3 illustrates three candidate vectors together with the grades assigned by seven flight operators. The assigned grades are ordered from lowest to highest so that the majority grade for each is the one that appears in the third column.

	GRADE[1]	GRADE[2]	GRADE[3]	GRADE[4]	GRADE[5]	GRADE[6]	GRADE[7]
Vector 1	55	60	76	***78***	88	90	95
Vector 2	50	59	65	***70***	70	85	91
Vector 3	60	60	70	***75***	84	87	89

Figure 1.3

Thus, the majority grades for Vectors 1, 2 and 3 are 78, 70 and 75 respectively and the winning vector is Vector 1. It is possible for ties to occur and there are a set of tie-breaking rules, which will not be discussed here.

1.3.5 Flight Operator Weights: Geographic and Temporal Aspects

Under the process as described, all flight operators have equal influence over the outcome. In fact, few would disagree that it is appropriate that flight operators with larger numbers of impacted operations should have more influence than those with smaller numbers of operations. A simple modification of the definition of consensus vector could be made to take into account flight operator weights. That is, flight operators with larger numbers of operations would be given a high weight (more influence) than those with small numbers of operations. Now rather than a majority of flight operators being required, the condition would be that a set of flight operators whose combined weight was greater than ½ the total. If one views the process as voting, then this approach could be interpreted as giving larger flight operators a larger number of votes. Probably the most natural weighting would be to simply use the number of involved operations as the weight. This could give larger operators excessive influence so a more moderate approach, e.g. weights equal to the square root of the number of operations, could be appropriate. It seems clear that weight should increase with the number of operations, but the best function to be used remains an open question.

Consideration of the flight operator weighting question leads one naturally to notion that influence / weight should vary by geography. For example a carrier with a large hub operation at EWR should have more influence over the strategy or part of the strategy that applies to EWR while a carrier with a large hub operation at ORD should wield more influence at ORD. This consideration in turn leads to the question of how COuNSEL should be applied across the large and complex NAS. Certainly the target application is to develop a daily national NAS strategy. To the extent that the strategy has components that specifically apply to ORD or EWR the carriers with large operations at one or the other airport should have a greater influence over that portion of the strategy.

It is also anticipated that COuNSEL could be applied on a regional basis. For example, it could be used to develop a TMI strategy for the Northeast. In that case, the flight operator with larger numbers of operations in the Northeast would have the most influence.

Another related issue is when and how often COuNSEL might be executed. Probably the "classic" application would be to develop a strategic plan at the beginning of the day. However, given the uncertainly of weather and the dynamic nature of the ATM, it is certainly the case that it would be desirable to modify the strategy over the course of the day. Such modifications would probably occur a few times each day. As discussed above, COuNSEL could be used to develop a regional plan so there might also be multiple executions related to regional planning needs, e.g. the development of plans for different regions of the country.

1.4 Perspectives and System Impact

COuNSEL represents a very different approach to strategic planning for NAS operations. As such much work will be required to move it toward implementation. For example, the concept of providing strategic advice by prioritizing performance criteria and grading performance vectors represents a very different way of doing business for flight operators. Work is on-going to provide intuitive mechanisms for flight operators to provide the necessary information. Human-in-the-loop simulations are also planned. Also, as discussed above, using the output of COuNSEL requires new TMI planning tools. A parallel research activity is addressing this challenging problem.

While challenges remain, the benefits of this new approach could be quite significant. The most obvious benefit will be a reduction in the significant time expended on the SPTs. In fact, this is probably a relatively minor benefit compared to the improvement in overall NAS operations. Specifically, by using NAS strategies that represent a consensus of operator preferences, TMIs will be developed that lead to better overall flight operator performance leading to an overall reduction in flight operator costs. Further, by balancing flight operator input in a formal way,

more equitable strategies and TMIs will result. Historically, equitable treatment of NAS users had led to greater cooperation in CATM, e.g. through high quality information exchanges. Work is ongoing to quantify potential benefits.

CHAPTER 2:

Overview of Research

2.1 Models Underlying Use of Majority Judgment within COuNSEL and Experimental Validation

As discussed in the previous chapter, the central component of COuNSEL is the Majority Judgment voting method. Typically Majority Judgment would be used to vote on, or grade, a small set of candidates. In the COuNSEL setting, the set of candidates is very large (in fact infinite). Any possible (feasible) performance metric vector is a potential candidate. Thus, our key challenge is to apply Majority Judgment in a setting where the set of candidates is so large that it is impractical to ask a player to vote on all candidates. Two other non-standard aspects of our use of Majority Judgment are i) that there is a continuous range of possible grades (any number between 0 and1) and ii) that the players are not equally weighted (each player has a (different) number of votes he or she can cast). In Appendix I, we provide a detailed descriptions of the various models used and also results of experimental testing. Here we provide a summary of this material.

Our approach to the first challenge involves developing a set of models that require voting on only a small set of candidates while implicitly considering the (very large) entire set of candidates. We start by generating a small set of candidates and/or asking the players to suggest candidates: these are used in a first round of voting. We use statistical methods to estimate player value functions based on voting behavior. Using these estimates within mathematical optimization models we have developed, we can find the vector that would (theoretically) win if players were allowed to vote on all possible candidate vectors. Of course, this winning vector is only an estimate of the true winner since it is based on estimated value functions. However, this vector certainly represents a very good new candidate vector. Thus, the process of estimating a winning vector, becomes the engine to generate new candidates that can be added to the candidate set and voted on in new voting rounds. We should note that the models developed can handle continuous grades and unequal player weights.

The models we developed actually represent a theory that allows us to find the "winner" that COuNSEL should choose assuming perfect knowledge of airline value functions. Of course, this Perfect Information winner cannot be found in real-life. However, using this theory we can generate sample data and set up experiments to test the effectiveness of the overall COuNSEL procedure. We conducted near real-life simulations to ascertain the quality of the winner determined by COuNSEL. We followed some intuitive guidelines to mimic airline behavior when generating sample value functions. Broadly, we consider the size of the airline (large / medium / small), type of origin (hub / non-hub), type of destination (domestic / international), and type of operations (hub and spoke / point-to-point / charter / cargo). The Perfect

Information winner was compared against the winner generated by a simulated application of COuNSEL. Overall, we find that the majority grades of the two winners are typically within 0.2% of each other. This small error gap is encouraging.

This research thus presents a sound theoretical method for determining the consensus view among multiple airlines with differing views over infinitely many candidates. This is of course a first step, as many challenges remain in making it operational.

Our experiments assume airlines grade in a way that is consistent with the value functions and that they do not attempt to "game the system". In a separate set of experiments, we investigate the degree to which it is profitable for players to engage in such "strategic behavior". This work is discussed in Section 2.4 and Appendix IV. Of course, over time we will also conduct Human-In-The-Loop experiments to further understand how users will interact with the system.

In developing our basic approach to the problem including the use of Majority Judgment, we explored several other approaches and did analyses that provided an overall understanding of the problem. A description of some of this analysis is included in Appendix V.

2.2 Benefits Assessment

A number of benefits are expected of COuNSEL. The most obvious benefit will be a reduction in the significant time expended on Strategic Planning Teleconns (SPTs). Other benefits, including better overall flight operator performance leading to an overall reduction in flight operator costs, and more equitable strategies and traffic management initiatives, could also be achieved through COuNSEL. In this section, using a Ground Delay Program (GDP) at San Francisco International Airport (SFO) as an example, we provide a rigorous assessment of the potential benefits of COuNSEL compared to the state-of-the-practice and the state-of-the-research in air traffic flow management. Our assessment mainly focuses on the savings in NAS-wide operating costs and the improvement of system-wide performance.

2.2.1 Evaluation Methodology

There are two important building blocks comprising our benefits assessment system: 1) understanding how airlines' preferences differ with characteristics of ground delay programs; 2) assessing NAS-wide performance under different GDP designs. To achieve these two objectives and to facilitate our evaluation, we have built an integrated simulation platform to mimic FAA – airline interactions during a GDP day. Given different GDP designs, our simulator can reveal the associated costs for all stakeholders (airlines, passengers, FAA) and assess NAS-wide performance. Through this machinery, we are able to show the impact different airlines might face under a particular GDP design.

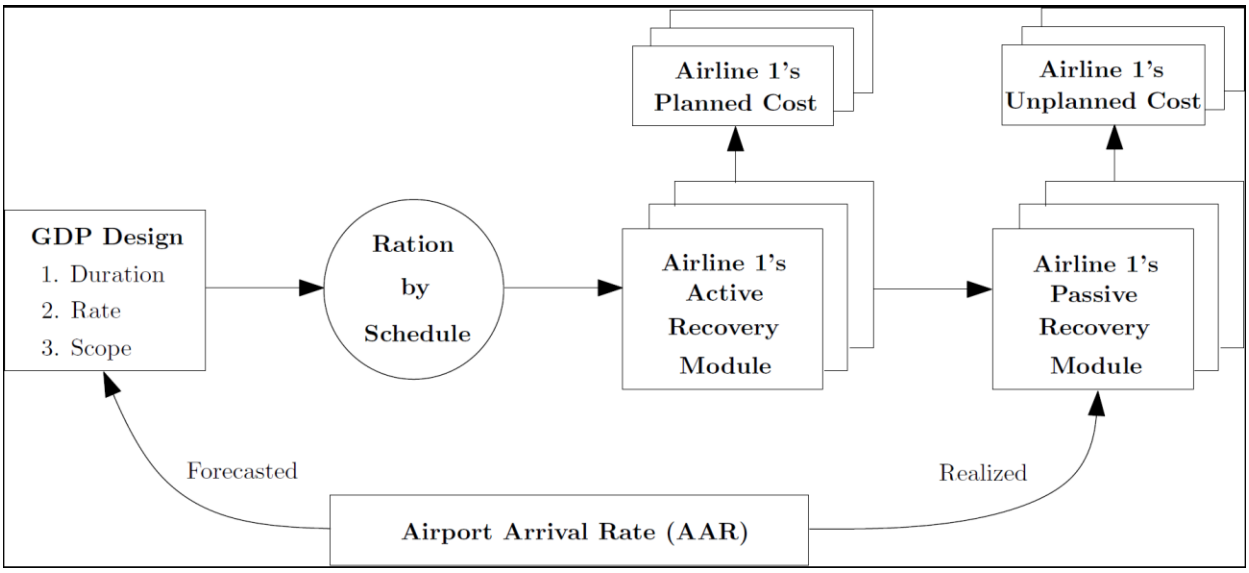


Figure 2.2.1: Evaluation Flow Chart

Figure 2.2.1 briefly depicts our main evaluation procedures. At the beginning of a GDP, the FAA has a weather forecast for use in designing their GDP. Based on the forecast, the FAA might design an aggressive GDP by setting a shorter duration, higher-rate program, thus maximizing overall throughput but potentially inducing higher airborne delays and less predictable delay information to airlines. Alternatively, the FAA can instead choose to be conservative (their typical practice) by setting a longer-duration, lower-rate program, thus providing airlines more accurate delay information, reduced airborne delays, and potentially lower expected throughput. Based on their GDP design, the FAA runs a Ration-by-Schedule (RBS) algorithm to inform airlines about delay information. Then each airline runs its own recovery module, marked as "active recovery module" in figure 2.2.1, in order to try to reduce their own adverse impacts through recovery operations. This will cause them to adjust their schedules, and re-route their fleet. Unfortunately, due to capacity uncertainty, these recovery operations are not the end of the story. At the time when weather conditions are realized, some of the planned recovery operations might again become infeasible. For instance, if it turns out that the FAA underestimated the impact of bad weather (which is usually the case when the GDP design is aggressive), then some of the flights will have additional airborne delays. This may cause fleet and passenger connections that were originally feasible in the recovery plan to get disrupted again. In this case, airlines need some additional recovery tools to get their schedules back on track. We refer to these as the "passive recovery module". We call it "passive" because, such disruptions caused by inaccurate delay information often require urgent fixes. Airlines usually don't have enough time to generate a sophisticated recovery plan. The passive recovery plan needs to be very simple and thus, it is likely to be less effective compared to the recovery plans generated by the active recovery module. Due to this, the

passive recovery module, as we model it, simply propagates all delay when an aircraft connection in the original recovery plan gets disrupted. We don't consider ways to avoid additional passenger disruptions. We denote the total delay cost coming from the active recovery module, based on the delay information provided by the FAA, as the *planned total cost*; and the additional delay cost estimated by the passive recovery module, based on the realized delay information, as *unplanned total cost*. Their sum is the *realized total cost* airlines incur under this particular GDP design. The more aggressive a GDP is, the lower the planned total cost will be, but the higher the unplanned total cost will be. Thus by grading different GDP designs, airlines essentially are making trade-offs between the unplanned total cost and the planned total cost.

2.2.2 Evaluation Results

Flight Schedules for a representative day in the summer of 2007 are obtained at the San Francisco International Airport (SFO). We set up 14 hypothetical GDPs. The planned duration of each GDP, the difference between the planned start and end times, is varied from 3 through 9.5 hours in steps of one half-hour each (i.e., 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9, and 9.5 hours). Each GDP has several possible actual end times, that is, the time after which the airport arrival capacity returns to the VFR level. Therefore, for each possible actual end time value, there is a corresponding unique capacity profile. In our setup, each GDP has seven capacity scenarios, corresponding to actual GDP durations varying between 3 and 9 in steps of 1-hour each (i.e., 3, 4, 5, 6, 7, 8 and 9 hours). All scenarios are assumed to be equally likely to occur (with probability 1/7). Hence, a GDP with a planned duration of 3-hours is considered to be a highly aggressive design in the sense that with high probability (6/7), the airport capacity is over-estimated by the GDP design. On the other hand, a planned duration of 9.5 hours is considered to be highly conservative because with probability 1.0, capacity is under-estimated by the GDP design.

Table 2.2.1 summarizes the total realized cost for each airline under different GDP designs. The italicized number in each row is the minimum value among all the numbers in that row. Based on the trend in total realized costs, we categorize the airlines into 3 broad groups: 1) those that prefer an aggressive (shorter) GDP design, 2) those that prefer a moderate (intermediate duration) GDP design, and 3) those that prefer a conservative (longer) GDP design. Note that these categories are not meant to serve as rigid categorizations for these airlines. Instead, they are relevant only for the specific case of the SFO airport, for the day of operations being considered in our experiment, and under the assumptions made in this experiment about the GDP scenarios, capacity distributions, and other parameters. The first category—airlines preferring aggressive GDP designs—includes US Airways, Frontier Airlines, Northwest Airlines, and Continental/ExpressJet. The second category—airlines preferring moderate GDP

designs—includes Delta Airlines, American/American Eagle, Alaska Airlines and JetBlue Airways. The third category—airlines preferring conservative GDP designs—includes United/SkyWest and AirTran Airways.

Table 2.2.1 also provides the NAS-wide total cost calculated by summing up the expected values of all the airlines' total realized costs. The last row in the table lists the centralized objective value under different designs. Here the centralized objective refers to the summation of airborne delay and ground delay costs from all the controlled flights heading into GDP airport. This objective function is used extensively in centralized GDP decision-making approaches, such as those of Richetta and Odoni, 1993, and Mukherjee and Hansen, 2009, just to name a few. The GDP design having the least centralized cost value (corresponding to 8 hours planned duration, marked by the blue rectangle) serves as our first baseline for comparison and represents the state-of-the-research design. From the analysis of all the GDP advisories issued in the year 2007, we find that 94% of the historical GDP advisories were too long, that is, the planned duration is longer than the actual duration. This means that practical GDP decision-making is typically conservative. Thus, we set the most conservative GDP design (with 9.5 hours planned duration, marked by the red rectangle) as our second baseline and represents the state-of-the-practice design. Finally, the design with 7 hours planned duration (marked by the black rectangle) is the consensus winner produced by COuNSEL, through a single-round majority judgment voting procedure. Interestingly, this coincides with the system optimal design, the one with the lowest NAS-wide cost. In comparing NAS-wide total realized cost, we conclude that COuNSEL reduce NAS-wide cost by 4.2% compared to the design generated by the current state-of-the-practice, and by 2.0% compared to the state-of-the-research design.

Airline - GDP Cost Matrix	Aggressive Design ← GDP Planned Duration (hours) → Conservative Design													
	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5
American & American Eagle	141340	124389	123490	117142	***112998***	125420	115407	128040	127462	126762	128174	130014	129585	134946
Frontier	***60362***	76148	66946	81898	82396	83363	85580	88988	91783	89850	101507	94825	106871	105891
US Airways	***83058***	84186	85994	87735	90695	90418	96115	89400	95663	95637	101711	104089	105107	106905
Continental & ExpressJet	34152	37247	37844	***33511***	36526	33968	39176	37459	39935	40162	41300	43174	44005	47296
JetBlue	9705	9849	10766	8939	8252	7983	***7577***	8367	7707	8563	9446	10863	13090	15468
Delta	36256	35408	34897	34846	34860	***34132***	35880	34732	35531	35773	38467	39139	41918	43874
AirTran	16600	15049	15050	13499	13363	11954	11280	11651	10338	10645	***9592***	10268	9864	12001
Northwest	***22247***	36705	32657	31738	31265	34185	34704	32411	36074	36831	36690	40855	40764	40228
United & SkyWest	489250	448340	426198	408230	402122	386515	357885	354516	330232	330824	322038	309187	304852	***300218***
Alaska	41167	35758	35713	***32724***	35337	37002	34810	36539	34573	36305	36882	37731	38215	38301
NAS wide	934137	903079	869554	850262	847815	844941	818413	822104	***809297***	811352	825808	820144	834271	845128
Centralized Objective	244986	235343	226604	221638	214614	210056	202624	201951	196292	189613	***188450***	195662	209204	220389

Table 2.2.1: Expected value of airlines' total realized costs under different GDP designs

Aside from costs in Table 2.2.2, we also calculate total ground delay, total airborne delay, the number of disrupted passengers, and total passenger delay under different GDP designs. Comparing the COuNSEL design with the centralized design we find that the COuNSEL design lowers total ground delay by 9.8%, and total passenger delay by 3.3%. It increases total airborne delay from 214 minutes to 318 minutes. However, on a per flight basis, it is an increase of just 0.51 minutes/flight from 0.95 minutes/flight to 1.46 minutes/flight, which is a small absolute increment. The benefits are even greater when comparing COuNSEL design to the most conservative design. We find a 22.8% reduction in total ground delay, and a 13.7% reduction in total passenger delay, while inducing only a 1.46 minutes/flight airborne delay. In summary, in this specific SFO case study, the entire National Airspace System is improved if it is controlled under a slightly more aggressive design than what both the state-of-the-research (centralized) and the state-of-the-practice approaches would suggest.

GDP Planned Duration (hours)	NAS wide Total Cost ($)	Total Ground Delay (minutes)	Total Airborne Delay (minutes)	Average Ground Delay per Flight (minutes)	Average Airborne Delay per Flight (minutes)	# Disrupted Passengers	Total Passenger Delay (minutes)
3	934137	2337	422	10.40	1.87	674	480618
3.5	903079	2835	432	12.62	1.92	725	469141
4	869554	3012	403	13.41	1.79	679	467044
4.5	850262	3269	386	14.56	1.71	681	456964
5	847815	3589	382	15.98	1.70	685	453177
5.5	844941	3857	376	17.18	1.67	674	458074
6	818413	4090	371	18.21	1.65	671	459613
6.5	822104	4429	358	19.72	1.59	673	472420
7	809297	4607	328	20.52	1.46	668	480232
7.5	811352	4748	257	21.14	1.14	678	485759
8	825808	5105	214	22.73	0.95	673	496769
8.5	820144	5546	123	24.70	0.54	646	520816
9	834271	5690	85	25.34	0.37	667	531846
9.5	845128	5970	0	26.59	0	662	556366

Table 2.2.2: NAS performance under different GDP designs

2.3 Performance Metrics Tradeoff Models

One of the inputs that is required to reach the consensus performance vector using COuNSEL is a set of feasible performance vectors. In Appendix III, two methodologies are proposed to generate these vectors: theoretical models based on continuum approximation and data-driven models based on historical information. Using the theoretical models, we present the relationship between performance vectors and GDP decision variables. Different flight operators may have different preferences regarding performance goals. Each flight operator may prefer a different point on the trade-off curves, and correspondingly opt for different GDP plans. Using a linear utility function, we further illustrate how a flight operator may select its optimal GDP plan. The idea of data-driven models is to generate a set of possible capacity

scenarios for a given day-of-operation from historical similar days, and calculate feasible performance vectors based on these capacity scenarios and the given demand on that day. This piece is current work and results are expected later. For each methodology, we also present ways to define performance metrics. However, both methodologies are quite general in the sense that expected performance vectors can always be estimated independent from how the performance metrics are defined.

2.4 User voting behavior and incentives

Our research has shown that *COuNSEL* satisfies the principal requirements we set forth for a consensus service-level expectation setting system. Of course, many practical challenges still remain before it may be deployed in operations. One general area of concern involves how the flight operators (players) will interact with the system and whether they will be motivated to vote (grade vectors) in a way that leads effective system operation. Specifically, the players should be motivated to truthfully express their nuances of preferences when providing grades to a variety of candidates.

Ideally, players should employ "value-based" grading. Each player first evaluates the value that each candidate may bring to her. She then provides a grade to the candidate that is proportional to its expected value. A candidate with higher value is understood to be more preferable than a lower valued candidate. In our case, an airline could compute the value for a given candidate as the reciprocal of the expected cost of degradation in service represented by the candidate. Value-based grading thus ensures that a more preferred candidate gets a higher grade than a less preferred one; and that equally preferred candidates get the same grade. Also, the most preferred candidate gets the highest grade possible.

For instance, suppose there are four candidates (C_1, C_2, C_3, C_4) to be graded. Say a particular player computes her value to each candidate as ($7,000, $4,000, $10,000, $4,000) respectively. Her first inclination might be to the highest grade of say 1 to her favorite candidate C_3, and the lowest grade of 0 to all the others. While broadly reflective of her preferences, such a voting strategy would not provide the richness of information the system requires and would likely result in an unsatisfactory outcome for the entire mechanism. Under value-based grading strategy, her grades should be: (.7, .4, 1, .4). It is easy to see that this is a more expressive and better reflective of her preferences. She is able to show that she prefers C_2 and C_4 equally well, and that her most preferred candidate is C_3. Furthermore, players may look for opportunities to "game" the system: instead of reporting their true grades for the candidates, they may adjust the grades in the hope that one of their most favored candidate. For example in the above example, a vector of grades (.9, 0, 1, 0) might seem to be represent a reasonable strategy. Another type of strategic grade might take into account the impact on competitors. For

example, in the above example, knowing that a competitor strongly prefers candidate C_1, the player may submit grades: (.2, .4, 1, .4). Of course, such behavior may also hurt the player – that is, a lesser preferred candidate may win.

Impossibility results due to Arrow, Gibbard and Satterthwaite, have shown that "perfect" voting mechanisms are not possible. That is, under any voting mechanism there will always be some circumstances where strategic voting (gaming the system) may pay off. On the other hand, the inventors of Majority Judgment have shown that, generally, it is difficult for player to profit from such behavior. We have designed and executed a set of experiments to investigate whether the use of Majority Judgment for the service level expectation setting problem has such resistance.

Untruthful or strategic grading by a player may take several forms. She may increase the grade of one or more candidates, and / or decrease the grade of one or more candidates, possibly leaving grades on some candidates unchanged. Strategic grading is beneficial to a player only if the Majority Judgment winner is replaced by a candidate that she regards more preferable to it. Indeed, strategic grading can hurt the player if the new winner is less preferred by her than the existing winner. Or, it may not yield any change to the existing winner.

Our framework is as follows. We assume each player is provided an opportunity to unilaterally change her grade after observing everyone else's grades for a given consideration set of candidates. In practice, such opportunity would not exist -- and the likelihood of hurting oneself would deter the players from strategic grading. Thus, this analysis provides the worst-case strategy proneness of the procedure.

We say that a candidate is manipulable if it is possible for some player to alter the majority grade it would receive from truthful grading. Certain sample sets of candidates may inherently be more manipulable than others -- depending on the number of players, their grades, and number of candidates. The proportion of manipulable candidates to the total number of candidates is one measure of strategy-proneness. However, that does not imply that each such candidate can be manipulated by all the players. Some players may not have any candidate that they prefer over the current winner -- these players will not have an incentive to deviate unilaterally. Among the remaining players, there may be some for whom there are no beneficial opportunities for the candidates that they prefer more than the current winner. These players too would not deviate unilaterally and benefit themselves. The proportion of the players that have any opportunity to benefit from strategic grading is a second measure of strategy-proneness. Another measure of strategy-proneness is the proportion of the total number of such beneficial player-candidate combinations.

2.4.1 Illustration with Equally Weighted Players

We now provide an example, by starting with the set of grades that would be assigned by five players (of equal weight) who used value-based, truthful grading. Note that grades should be between 0 and 1and there are three candidates m_1, m_2, m_3. The (truthful) grades are summarized in Table 2.4.1. For each candidate, the majority grade is highlighted. m_3 is the winning candidate as it has the highest majority grade.

Player	m_1	m_2	m_3
1	0.6	0.3	0.2
2	0.1	0.3	**0.5**
3	0.1	0.6	0.6
4	**0.2**	0.7	0.3
5	0.8	**0.4**	0.7

Table 2.4.1. Players' Grades

We now ask the question whether any player can improve his or her position by deviating from truthful grading. Making the (unrealistic) assumption of full knowledge of all other players' grades, each player can identify a range of grades for every candidate that would lead to a change in the majority grade of the candidate. These are called manipulable ranges, and are listed in Table 2.4.2.

Player	m_1	m_2	m_3
1	[0.1 ... 0.2]	[0.4 ... 0.6]	[0.5 ... 0.6]
2	[0.2 ... 0.6]	[0.4 ... 0.6]	[0.3 ... 0.6]
3	[0.2 ... 0.6]	[0.3 ... 0.4]	[0.3 ... 0.5]
4	[0.1 ... 0.6]	[0.3 ... 0.4]	[0.5 ... 0.6]
5	[0.1 ... 0.2]	[0.3 ... 0.6]	[0.3 ... 0.5]

Table 2.4.2. Manipulable Ranges

Note that not all players have an incentive to deviate, as the consideration set does not have better candidate for them. In the example, players 2 and 3 are such players.

Player 1 has no opportunity to make her most preferred candidate m_1 as the winner. However, she can choose to make her next preferred candidate, m_2 as the winner: she can increase its majority grade within 0.4 and 0.6, which has an overlap with her manipulable range of m_3 over [0.5 … 0.6]. Thus, player 1 can provide say 0.55 to m_2 and make it the new winner.

The candidate-level measure of strategy-proneness is 100%, as all of the candidates in the consideration set are potentially manipulable. However, that does not mean that each player can unilaterally manipulate the grades to benefit. This is thus a broad, naïve measure.

At player-level, we already identified that player 1 can benefit by manipulating m_2 and / or m_3. Also, we noted that the players 2 and 3 already have their most-preferred candidate in the current winner m_3 – and hence do not have incentive to manipulate. Player 4 has an overlap between the manipulable ranges for m_1 and m_3 -- but its preference for m_1 being lesser, it has no incentive to manipulate these. There is no overlap for its most preferred candidate m_2 with m_3. Thus, player 4 actually has no opportunity to strategically grade in a way that might benefit her. Similarly, player 5 has only an opportunity with m_2, but since whe prefers it less than m_3, she cannot benefit by manipulating her grades. Thus, of the five players, only one – player 1 – has a beneficial strategic opportunity. Hence, player-level measure of strategy-proneness is 20%.

Finally, among the 15 player-candidate opportunities, only two are beneficial to any player. Hence, the measure for strategy-proneness at player-candidate level is about 13%.

2.4.2 Simulation Results

The observations made for the illustration with equal-weighted players have been formalized, and extended to scenarios with differential weights. To get a sense of strategy resistance of the procedure, we conducted a number of simulations systematically varying some key parameters. The design of experiments is summarized in Table 2.4.3.

	Preference structure	
Relative weights of the players	P1: None ("unrestricted domain")	P2: Convex function
R1: Equal weights ("unweighted")	P1R1	P2R1
R2: Differential weights (5 players)	P1R2	P2R2
R3: Differential weights (25 players)	P1R3	P2R3

Table 2.4.3. Design of Experiments

The intent behind this design has been to contrast the proposed *COuNSEL* procedure with several other plausible implementations. At the simplest extreme, P1R1 is the basic Majority Judgment, as laid out by its authors. At the other extreme lies P2R3, this is closest to the real-life scenarios that *COuNSEL* may be deployed for.

The progression in the two directions from P1R1 to P2R3 is instructive. R2 and R3 address the proportional representation aspect of *COuNSEL*, which is a key design element that adds equitability. R2 is a very small setup, and might represent the initial deployment phase of *COuNSEL*, in which fewer airlines may participate. R3 is a more likely setup reflecting the later phases of deployment. P2, on the other hand, addresses the key assumption in structuring of the grade functions. An unrestricted domain would easily lead to inconsistent grading over rounds, which is highly undesirable.

We found that the naïve measure for strategy proneness, based on proportion of manipulable candidates, consistently reports very high likelihood of manipulation, typically upwards of 50%. However, the likelihood of an individual player to find a beneficial strategic opportunity drops in the regions of 10% or less. Moreover, as the specific candidates via which the individual players may benefit are also brought into consideration, the likelihood drops to 1-2% levels.

A surprising, though useful, observation has been the rather insignificant impact of attaching weights to the players. Weights are a significant design element in *COuNSEL*, wherein unlike the democratic "one-person one-vote" scenario, it is essential to provide the airlines differential weight in the overall decision-making, for equity reasons.

Another key observation has been the drastic reduction in strategy proneness when the unrestricted domain of grades is replaced with a convex preference structure. Convexity, continuity, and monotonicity have been standard assumptions in the literature. These are also reasonable in our application area, whereby players would more likely have a possibly "single-peaked" preference structure over the feasible candidate space.

The results in themselves are quite encouraging. Even with complete knowledge of everyone's grades, and then being provided with an opportunity to benefit oneself, the likelihood of a particular player to find a beneficial opportunity via a candidate is in the region of 2% or below. In real-life, such opportunity would of course not exist. Moreover, untruthful reporting has a good possibility of hurting the player, as it may result with a new winner that is less preferred than the current winner.

This study provides a theoretical support for the proposed mechanism in terms of its strategy-resistance. In the next steps, Human-In-The-Loop experiments are being planned, whereby inputs from real airlines would be sought in simulated scenarios.

2.5 *COuNSEL* Concept Evaluation Software

Concept evaluation software has been developed to aid in illustrating COuNSEL principles and also to solicit feedback from the user community. The initial version of the software was completed in August of 2013 and demonstrated to the FAA at that time. In the past two months, the NEXTOR-II team has conducted internal gaming exercises with the software in order to provide feedback to the development team. Input for a new software version is now been generated.

COuNSEL features include:

- Browser/web based: can be used on any computer with a standard web browser.
- Distributed operation: as discussed in Chapter 1, COuNSEL users include the FAA/central controller and several flight operator/agents. Individual users can be located at geographically dispersed sites.
- FAA/central controller functions include: i) generating a new traffic management initiative and setting the parameters of that initiative (one initiative corresponds to one application of COuNSEL), ii) inputting candidates, iii) accepting or rejecting user supplied candidates iv) making decision on whether to continue or terminate voting rounds.
- Flight operator/agent functions include: i) generating new candidates (this is implemented using an intuitive GUI function), ii) voting on candidates.
- Distribution of results of each round of voting.

Future versions will include features such as automatic candidate generation, intelligent support for flight operator vote/grade generation, as well as others.

2.6 Summary of Outreach Activities and User Testing

The research team has begun efforts to present COuNSEL to the flight operator community. The aims of these outreach activities are fourfold. First, we seek to establish a baseline of flight operator views about the current process of TMI planning. The second goal is to familiarize flight operators with the COuNSEL concept. Next, we want to get feedback about COuNSEL, including both assessments of its advantages and the concerns that flight operators may have about adopting this system. Finally, at a later stage we intend to engage in user testing of the actual COuNSEL software.

The first step in the process was to identify the appropriate flight operator personnel to contact. Through informal communication with industry contacts, we determined that the Airlines for America (A4A) Air Traffic Control (ATC) council would be an appropriate forum to initiate this process. The ATC council has 20 members representing 12 airlines, as summarized in Table 2.6.1. These individuals hold a variety of titles, but are generally managers who mediate interactions between their companies and FAA ATC.

Members of the research team attended an A4A meeting in Washington DC on July 15, 2013. Prior to the meeting, we prepared and circulated to ATC Council members a White Paper overviewing COuNSEL (Chapter 1 of this report.) At the meeting we presented COuNSEL, took questions, and asked Council members for their support in the outreach effort. Specifically, we requested a 1-hour interview with management at each operator, and that the companies make

potential end-users of COuNSEL available for a 30-minute interview. While flight operators had a number of questions about the concept, they were generally supportive and willing to assist in the outreach. However, they also indicated that it would be better to conduct the effort after the end of the convective weather season.

The outreach protocols have been developed with close cooperation of Delta Airlines. We present the protocol as it was conducted at Delta, along with results obtained. It should be emphasized that these results are based on a handful of respondents from a single carrier, are intended to simply illustrate the conduct of the protocol. Table 1 in Appendix VI summarizes the questions and answers for Delta management. The first part of the Table pertains to the portion of the interview that took place before the respondent was briefed on COuNSEL, while the second part was administered after the COuNSEL briefing.

In the second part of the outreach effort, a survey is administered to individuals who would actually be COuNSEL users. The survey is web-based and uses the Qualtrics platform. The COuNSEL user survey has been designed to collect flight operators' viewpoints concerning current Traffic Management Initiative (TMI) planning and decision making, and their opinions on COuNSEL. We pilot-tested the survey using hard copies that were completed by two Delta Sector Managers. Their responses to the survey are summarized in Table 2 of Appendix VI, which also includes respondents' feedback on the questions themselves. The survey questionnaire has been revised based on the feedback. The revised survey appears in Appendix V. With the end of the convective weather season, we are currently contacting other airlines to arrange for similar interviews with a management, as well as administration of the web-based survey to potential COuNSEL users including users from Delta Airlines.

Flight Operator	**Number of Members**
Air Canada	1
Alaska Airlines	1
American Airlines	2
Atlas Air	1
Delta Airlines	2
FedEx	1
Hawaiian Airlines	2
JetBlue Airways	2
Southwest Airlines	2
United Airlines	2
UPS	2
US Airways	2

Table 2.6.1: ATC Council Membership

APPENDIX:

Technical Research Chapters

APPENDIX I

Designing the Noah's Ark: A Multi-objective Multi-stakeholder Consensus Building Method

Prem Swaroop, Michael O. Ball
Robert H. Smith School of Business and Institute for Systems Research, University of Maryland, College Park, MD 20742.
pswaroop@rhsmith.umd.edu mball@rhsmith.umd.edu

A significant challenge of effective air traffic flow management (ATFM) is to allow for various competing airlines to collaborate with an air navigation service provider (ANSP) in determining flow management initiatives. In this paper, we describe a mechanism whereby the airlines provide "consensus" advice to an ANSP using a voting mechanism. It is based on the recently developed Majority Judgment voting procedure. The result of the procedure is a consensus real-valued vector, that must satisfy a set of constraints imposed by the weather and traffic conditions of the day in question. While we developed and modeled this problem based on specific ATFM features, it appears to be highly generic and amenable to a much broader set of applications. Our analysis of this problem involves several interesting subproblems, including a type of column generation process that creates candidate vectors for input to the voting process.

1. Introduction

A shared perception of a common, imminent, unavoidable, impactful threat or opportunity oftentimes leads even fierce competitors to seek consensus solutions. The mythical Noah's Ark is indeed witnessed in the real-world of business. For example, technology standards bodies have been the foundation for inter-operability of the products and services offerings of firms competing for the same or similar customers. American National Standards Institute (ANSI), a consortium of industry and researchers, performed this key function during the entire Industrial Age; while the more recent Internet Society serves similarly in the Information Age. In the highly competitive airline industry, we see examples of airline alliances which have helped airlines maximize their offerings and reach through collaborating with other competing airlines. At another level, whenever there is bad weather, the airlines come together with the Air Navigation Service Provider – namely, the Federal Aviation Administration (FAA) in the US – to devise effective means to handle the constrained system resources.

Future visions of Air Traffic Flow Management (ATFM) – both in the U.S. and Europe – support a

"performance-based" approach that employs collaboration between the air navigation service providers (ANSPs) and the airlines (ICAO 2005, JPDO 2007, SESAR 2006). A key feature of this outlook is to support the airline operators' business objectives in the ANSP's traffic management initiatives (TMIs), subject only to system-level constraints like safety and security. Our focus is on a performance-based framework that addresses the strategic level planning in advance of the implementation of a TMI. These overarching system performance expectations may then serve as the basis for design and operation of a specific TMI (or a coordinated set of TMIs) that aim to meet the stated expectations.

The framework must (a) be founded upon commonly agreed definitions of service expectations among the several stakeholders, and (b) result in a consensus on the service expectations over independent stakeholders, with possibly conflicting business objectives. We use the Global Air Traffic Management Operational Concept (ICAO 2005) to address the former requirement. Unanimously approved by the U.S. and 187 other States in the eleventh global Air Navigation Conference, it dedicates a section on "expectations of the AT[F]M community." Among 11 performance expectations, three are more specific to the airline operators' business objectives, while the others are more generic to the entire framework – predictability, capacity-utilization, and efficiency. Our focus in this work is on the latter aspect of the framework: given that there is intent to collaborate among the stakeholders, how to design an effective consensus solution that encompasses multiple inter-related objectives.

We postulate six properties as highly desirable for any effective solution for the stated problem: (i) consensus-building, (ii) single solution determination, (iii) practicality, (iv) equitability, (v) confidentiality, and (vi) strategy-proof. These are consistent with the principles of mechanism design (Maskin 2008), and also take into account some specific needs of our application environment.

The first three are desirable for purely pragmatic reasons: it is our stated wish that the method determines an acceptable solution among the multiple stakeholders; the method would be most effective if the method results in an unambiguous solution; and that the method does not take inordinate time and / or effort on the part of the decision makers to yield its solution.

The latter three are higher-level properties. Given that we are dealing with possibly competing stakeholders, we shall like the method to adhere to well-accepted notions of equitability, specifically we shall like the voice of each stakeholder to be fairly represented in the decision making process. Further, as we are likely dealing with independently operating businesses, the method should not require information that may be deemed confidential. Finally, we shall like the method to discourage any strategic behavior among the decision makers.

A recently proposed voting scheme called "Majority Judgment" has many desirable properties. Of primary interest to us has been its high strategy-resistance – while it does not preclude gaming of the system, the probability of a single player to significantly game the system is severely restricted in this design. We therefore base our proposal on Majority Judgment.

In section 2, we describe the problem and present related literature. Section 3 focuses on the mathematical models that underpin the proposed mechanism. After reasonably structuring the underlying information, it presents efficient solution methods. Validation is provided in Section 4 through simulation experiments on a large dataset motivated from real-life. Section 5 concludes.

2. General Problem Statement and Related Work

The general context for the problem we address involves a group of n stakeholders, N, who jointly seek to make a decision. It is not necessarily the case that these stakeholders are cooperative or have common goals: in our application, the stakeholders are the flight operators who in fact are competitors. The form of the decision we seek is a numeric vector $\mathbf{m}$ that is subject to a set of feasibility conditions μ so that $\mathbf{m} \in \mu$. The $\mathbf{m}$ we seek should represent a consensus among a majority of the stakeholders. Each stakeholder $i \in N$ has a value or value function $V_i()$ that maps each $\mathbf{m} \in \mu$ to a real number that represents the value of $\mathbf{m}$ to i. The problem we address is to design a mechanism in which a coordinator exchanges information with the stakeholders and produces the desired $\mathbf{m}$. Of course, this is hardly a well-defined problem yet, as in particular, we have not precisely defined a majority consensus $\mathbf{m}$. Nonetheless, this description does allow us to place our problem in a broader context and to discuss the nature of our contributions. In particular, attacking this problem would seem to require key elements from two large bodies of literature: Voting, and Multi-criteria decision making (MCDM).

The case where $\mathbf{m}$ is of dimension 2 and $\mu = \{(1,0),(0,1)\}$ can be viewed as a classic election among two contenders. Each stakeholder would "vote" for either $(1,0)$, expressing a preference for the first candidate or $(0,1)$, expressing a preference for the second candidate. The vector output would indicate the winning candidate. The case of higher dimensional $\mathbf{m}$ with μ consisting of all unit vectors would correspond to an election with several candidates where one must be chosen. The instant runoff voting mechanism and majority judgment represent mechanisms that would produce a single winner candidate/vector.

Voting in particular, and social choice in general, is concerned with aggregating evaluations over

a multitude of voters, in ways that the final outcome has appeal to a large section of the decision-makers. Over centuries, investigators devising a fool-proof voting system have been riddled by a result – famously known in social choice theory as Arrow's Impossibility Theorem (Arrow 1951). It states: "when voters have three or more distinct alternatives, no voting system can convert the ranked preferences of individuals into a community-wide (complete and transitive) ranking while also meeting a certain set of criteria, namely: unrestricted domain, non-dictatorship, Pareto efficiency, and independence of irrelevant alternatives." Majority Judgment is a recently proposed procedure (Balinski and Laraki 2011), that "bypasses" this result. And hence, its authors claim it to be "a better alternative to all other known voting methods, in theory and in practice."

Majority Judgment involves grading – instead of preference rankings – of each candidate, by all voters, in a common language. It is a natural, rich preference elicitation method, already being practiced in spirit in many contests and juries around the world, as well as a few political elections. It has many good properties; among them, high resistance to strategic voting – which makes it appealing for our work. An outline of the procedure with an example follows later.

A key feature of our problem is that μ is very large; in fact, the μ we employ is a polyhedron and so has the structure of the feasible region of a linear programming problem. Our consideration of, and modeling of, this large space of feasible candidate vectors represents the most essential contribution of this paper.

The decision-making framework in the general MCDM involves a decision maker evaluating a set of candidates on the basis of multiple criteria or attributes (Wallenius et al. 2008). A common assumption about the decision maker's or group's actions is consistency with maximization of a utility or value function that depend on the attributes (Raiffa and Keeney 1976). Wallenius et al. (2008) characterize the distinctions between the discrete and the continuous candidate space versions of the MCDM problem. Our work is related to both the versions. Like the continuous version, we iteratively search a continuous candidate space. However, like the discrete version, we do specify a functional form of the decision makers' value functions, and estimate its parameters over several candidates over the iterations.

We generally assume that $V_i()$ is known in some way to each stakeholder i. Thus, we do not devote attention to methods to "discovering" $V_i()$. We note a significant body of research that focuses on this aspect of the problem. In this literature it is generally assumed the stakeholders can provide some preference information, e.g. the ability to choose between pairs of alternatives. The stakeholders are

then asked to make various preference decisions to elicit functional forms, e.g. the $V_i()$'s, that allow a decision on a complete option to be made. We note in particular the Analytic Hierarchy Process (AHP), which is a well-regarded tool for multi-criteria decision making (Saaty and Vargas 2012). It relies on pairwise comparisons over a set of alternatives, eliciting preference rankings on several criteria organized in a hierarchy on a nine-point scale. The group version of AHP aggregates the individual scores into group scores using their geometric means – similar to the way it aggregates the scores over the hierarchy of criteria. Any mean is less resistant to extremes (and thus strategic behavior) than median – which is used by Majority Judgment.

Green and Rao (1971) introduced conjoint analysis into marketing literature – which has enjoyed considerable success in marketing applications (Green et al. 2001). A decompositional technique, the method presents respondents with descriptions of alternatives with differing levels on a number of attributes, and records their preference order over the alternatives. For reasons just discussed we do not use these methods to determine utility functions but we do use the functional forms from this literature as part of our estimation process.

As the research progresses, there will be need to approximate the efficient frontier of the feasible candidate space using historical or simulated data on candidate realizations. Hence, on the computational side, research dealing with the problem of approximating the efficient frontier of the continuous candidate space is also relevant to our work, e.g. Data Envelopment Analysis (see Charnes et al. (1978), Cook and Seiford (2009)) and potentially, multi-objective linear programming, (see Ruzika and Wiecek (2005), including methods to approximate the efficient frontier (see Sayın (2000) and Karasakal and Köksalan (2009)).

A multi-criteria decision analysis based approach was adopted in a strategic decision making context by Eurocontrol (Grushka-Cockayne et al. 2008). Similar to our setting, the problem involved the ANSP and the airlines collaboratively arriving at a common decision for selecting operational improvements. Further, the decision was subject to constraints like safety and environmental impact, and was expected to improve on objectives like predictability and efficiency. However, unlike our problem that seeks to evaluate at a day-of-operations level, the Eurocontrol was faced with a one-time strategic decision.

2.1. Majority Judgment

Majority Judgment is defined as a social decision function. It involves grading – instead of preference rankings – of each candidate, by all voters, in a common language. It is a natural, rich preference

Candidates:	C_1	C_2	C_3	Maj. Gr.
Worst Grade:	Passable	Reject	Reject	MG-5
.	Passable	Passable	Reject	MG-3
.	Good	Passable	Good	MG
.	Very Good	Good	Very Good	MG-2
.	Very Good	Good	Very Good	MG-4
Best Grade:	Excellent	Good	Very Good	MG-6

Table 1 Majority Judgment example

elicitation method, already being practiced in spirit in many contests and juries around the world, as well as a few political elections.

It takes as input the Grades given by the voters, and produces "Majority Grade" of each candidate as an output. These can be used to compute rank-orderings as well (called "Majority Ranking"). Majority Grade of a candidate is the highest grade approved by an absolute majority of the voters. In case of an odd number of voters, it is the median of the grades; if there are even number of voters, then it is the lower middlemost of the grades. Its high resistance to strategic voting primarily results due to this median-seeking property.

Suppose there are six voters, voting on three candidates: C_1, C_2, C_3. They assign one of these five grades to each candidate: Excellent, Very Good, Good, Passable, Reject. The grades thus obtained by voting are then sorted from worst to best, as given in Table 1.

The majority grade for each candidate (marked "MG" in the last column) is the top fourth grade, as majority (four of six) would give at least that grade to the candidate. Row "MG-2" is found similarly after hiding the "MG" row, and so on; these are useful for tie-breaks when ranking candidates. Majority Ranking for the example is: $C_1 \succ C_3 \succ C_2$.

Majority Judgment requires a common language accepted by all the voters for grading the candidates. Grades may be either continuous or discrete (like above). A continuous grading language could be: $\{0, \ldots, 100\}$, where 0 is commonly understood as "unacceptable", and 100 as "most favorable".

The aspect of common language, while being very intuitive and simple to express, is critical to the overall procedure. Any practical implementation has to carefully come up with the common language that is accepted by all the voters. For some applications, the common language is easier to identify as it forms part of the trade, e.g, tea or wine tasting within a company, or assignment grading in classes. In new applications however, specific focus groups with the voters are sometimes conducted to establish the common language. Furthermore, special training and communication procedures are developed to ensure that new entrants to the system are well-conversed with the common language.

3. Mechanism Design and Underlying Models

As discussed in the preceding sections, Majority Judgment provides a solution to the challenge we have outlined. However, Majority Judgment cannot be directly applied due to the very large – in fact, infinite – size of the set of "candidates". In this section, we develop an analytical framework and set of models for addressing this issue.

3.1. Majority Judgment Winner

Suppose N is the set of n stakeholders – hereafter referred to as players – faced with a potentially infinite set of feasible candidates μ. As discussed in Section 2, each player has a value function V_i that assigns each candidate $\mathbf{m} \in \mu$ a value. In this section we make use of a grade function g_i, which is similarly a function defined on μ. The grade function will be employed by player i to assign a grade to each $\mathbf{m}$ as part of the Majority Judgment process. While g_i is clearly closely related to V_i (and higher V_i values would generally induce higher g_i values), it is possible that a player may consider various strategies for setting g_i based on V_i. However, at this time, we will assume that a simple linear transform is used to produce g_i based on V_i and we will refer to $g_i(\mathbf{m})$ as the value of $\mathbf{m}$ to i. In fact, the only reason that we do not use V_i directly is that we require all grades to use a "common language". Here this implies $0 \leq g_i(\mathbf{m}) \leq 1$ for all $\mathbf{m}$.

Let b denote a *minimal* majority-forming subset of N, and let β denote the set of all possible minimal majority-forming subsets of N. In a "one-person, one-vote" situation, a majority-forming subset is any set of size $\frac{n}{2}+1$ for n even and $\left\lceil \frac{n}{2} \right\rceil$ for n odd. In the weighted case addressed here, each player i is given a weight w_i; the total weight of a minimal majority-forming subset b just exceeds half the total weight of all players:

$$\overline{W} < \sum_{i \in b} w_i; \text{ where } \overline{W} \equiv \frac{\sum_{j \in N} w_j}{2}. \tag{1}$$

Requiring the set to be minimal implies that if any element is deleted from b then equation (1) would no longer hold. Note that the complement $N - b$ clearly does not form a majority.

The *min grade* for a specific b and candidate $\mathbf{m}$ is $u(\mathbf{m}, b) = \min_{i \in b} g_i(\mathbf{m})$. The Majority Grade $v(\mathbf{m})$ for a candidate $\mathbf{m}$ is the highest grade a majority of players is agreeable to assign it, i.e.

$$v(\mathbf{m}) = \max_{b \in \beta} u(\mathbf{m}, b)$$

A Majority Judgment winner is the candidate $\mathbf{m}^*$ with the highest Majority Grade v^*:

$$v(\mathbf{m}^*) \equiv v^* = \max_{\mathbf{m} \in \mu} v(\mathbf{m}).$$

A Majority Judgment winner $\mathbf{m}^*$ thus guarantees a majority of the players a grade of at least v^*.

Determining a winner over a "small" set of candidates is straight-forward in the presence of a trusted, benign "central planner". The players submit their grades for each candidate to the central planner. The planner then sorts the grades for each candidate, and identifies the median grade for each (lower median in case of even number of players) – this is the *majority grade*. The candidate with the highest majority grade is deemed the winner.

Our challenge is to determine such a winner when the size of μ is very large – perhaps infinite. In fact, the proceeding discussion already implicitly associated a subset of players with the winning candidate. This in turn provides a potential approach to making the candidate search finite in the sense that we could search for the winning minimal majority forming subset rather than the winning candidate. Specifically, if we define for any $b \in \beta$

$$\hat{v}(b) = \max_{\mathbf{m} \in \mu} u(\mathbf{m}, b)$$

then it is easy to see that

$$v^* = \max_{b \in \beta} \hat{v}(b). \tag{2}$$

While we have now made a search over a potentially infinite set finite, this reduction depends on the ability to efficiently find $\hat{v}(b)$. The following optimization model can accomplish this:

Subset_Opt(b)

$$\begin{aligned} \hat{v}(b) = \max \quad & z & & \\ s.t. \quad & z \leq x_i & \forall \quad & i \in b \\ & x_i = g_i(\mathbf{m}) & \forall \quad & i \in b \\ & \mathbf{m} \in \mu & & \end{aligned}$$

We will later show that for applications of interest to us this model can be cast as a linear program.

A special type of minimal majority-forming subset is relevant in Majority Judgment: a *majoritarian set* is a minimal majority forming subset that gives the highest grade to some candidate $\mathbf{m}$. That is,

a $b' \in \beta$ is a *majoritarian set* if there exists an $\mathbf{m} \in \mu$ such that

$$u(\mathbf{m}, b') = \max_{b \in \beta} u(\mathbf{m}, b)$$

	C_1	C_2	C_3	C_4
g_1	1.00	.70	.40	.50
g_2	1.00	.90	.70	.85
g_3	.80	1.00	.80	.90
g_4	.60	.75	1.00	.80
g_5	.40	.50	.60	.30

Table 2 Sample grade functions for four candidates

To illustrate these concepts consider the example provided in Table 2.

Assuming all weights are one, there are $\binom{5}{3} = 10$ minimal majority forming subsets but only two majoritarian sets: $\{1,2,3\}$ and $\{2,3,4\}$ ($\{2,3,4\}$ produces the highest grade for each of candidates C_1, C_2, C_3). Note that player 5 is in no majoritarian set since this player tends to give all candidates a low grade. While the grade functions prevent player 5 from being in any majoritarian set, in the weighted case it is possible that a player could be in no majoritarian set because that player was not in any minimal majority-forming subset. An extreme example could occur if the weight of a single player $\hat{i}$ was greater than $\overline{W}$. In such a case, $\{\hat{i}\}$ would be the only minimal majority forming subset and by necessity the only majoritarian set. All other players could be in no majoritarian set irrespective of how they graded. Of course, we may wish to impose rules or assumptions that prevent some of these extreme cases. For example, we will only consider weighting schemes that do not make a single player a majority and we may also require that each player give at least one candidate a grade of one.

The concept of a majoritarian set can potentially allow us to reduce the search space size since if we define β' to be the set of all majoritarian sets then we can replace Equation (2) with:

$$v^* = \max_{b \in \beta'} \hat{v}(b).$$

However, we can in fact reduce the search even more. It should be clear from the preceding discussion that for any $b \in \beta$ there is an $\mathbf{m} \in \mu$ and an $i \in b$ such that $g_i(\mathbf{m}) = u(\mathbf{m}, b) = \hat{v}(b)$, i.e. i is the element of b that assigns $\mathbf{m}$ its minimum grade. In general, a given player i might play such a role for several sets b. We can thus define an optimization problem that determines the highest value of $\hat{v}(b)$ achievable where $i \in b$ and i defines the minimum grade, i.e.

$$\tilde{v}_i = \max\{g_i(\mathbf{m}) : g_i(\mathbf{m}) = u(\mathbf{m}, b), i \in b, \mathbf{m} \in \mu\}$$

We note in general it can be the case that the set optimized over in this expression can be null, in which case $\tilde{v}_i$ is defined to be zero. For example, a player that consistently grades very high could be in many majoritarian sets but might not define the minimum grade for any of them.

We have now developed a new approach to finding v^*, namely:

$$v^* = \max_{i \in N} \tilde{v}_i. \tag{3}$$

We now define an optimization model that determines a value closely related to $\tilde{v}_i$ and will allow us to compute v^* using an equation similar to (3). This optimization model is defined for any $i' \in N$.

Player_Opt(i')

$$\begin{aligned}
\tilde{z}_{i'} = \max \quad & x_{i'} & & \\
s.t. \quad & x_{i'} \leq G^{max}(1 - I_i) + x_i & \forall \quad i \in N & \quad (4)\\
& \sum_{i \in N} w_i I_i \geq \overline{W}' & & \quad (5)\\
& I_i \in \mathbb{B} & \forall \quad i \in N & \\
& x_i = g_i(\mathbf{m}) & \forall \quad i \in N & \\
& \mathbf{m} \in \mu & &
\end{aligned}$$

Here, G^{max} is the maximum grade value and $\overline{W}'$ is the smallest number greater than $\overline{W}$ that can be achieved as the sum of the weights of a subset of players. Note there are two sets of variables. The continuous x_i variables define the grades assigned by each player. The binary I_i variables define the players in the majority forming subset; specifically, $I_i = 1$ implies player i is in the majority forming subset and $I_i = 0$ implies it is not. Constraint (4) insures that $x_{i'}$ is the minimum grade in the set. Constraint (5) insures that the set has total weight larger than $\overline{W}$.

Proposition 1. *The following hold true:*

1. $v^* = \max_{i \in N} \tilde{z}_i$,
2. *any* i^* *that solves* $\max_{i \in N} \tilde{v}_i$ *also solves* $\max_{i \in N} \tilde{z}_i$,
3. *for any* i^* *that solves* $\max_{i \in N} \tilde{v}_i$*, the corresponding majoritarian set* b *when converted to an* I *vector and the corresponding grade vector when expressed as an* x *vector are an optimal solution to* **Player_Opt**(i^*).

Proof All three results follow from two observations. First, consider any optimal solution to **Player_Opt**(i) for some i and let b^* be the set corresponding to the optimal I vector. Constraint set (5) implies that b^* is a majority forming subset. If b^* is not minimal then there is a minimal $b' \subset b^*$ with $\hat{v}(b') \leq \hat{v}(b^*)$. In particular, if $\hat{v}(b') < \hat{v}(b^*)$ then there exists an $i' \in b'$ such that $\tilde{z}_{i'} < \tilde{z}_i$. Second, any majoritarian set b together with a grade minimizing $i \in b$ generates feasible solution to **Player_Opt**(i).

□

3.2. Structural Assumptions and Efficient Modeling of Feasible Set of Candidates and Grade Functions

We now describe some assumptions regarding the structure of the set of feasible candidates and the grade functions. These are appropriate for our target applications (as well as many others) and also aid in the tractability and modeling of the problem.

ASSUMPTION 1. *The feasible candidate space $\mu \subset \mathbb{R}_+$ is continuous and has a concave "efficient frontier".*

The concave efficient frontier is a reasonable assumption if: (a) larger values of each individual metric are desirable, and (b) there is a tradeoff required among the metrics – that is, increasing the value of one metric comes at the expense of other(s). The first requirement can be met by suitable transformations if smaller values are more desirable than larger. Tools like Data Envelopment Analysis are dedicated to finding such efficient frontiers among a miscellany of metrics.

ASSUMPTION 2. *Each player's value function $V_i(\mathbf{m})$ is non-negative, continuous, non-decreasing and concave.*

The non-negativity assumption could be resolved by transformation if the original value function did not have this property. Continuity would seem to be a reasonable assumption (for many applications): very small changes in candidate component values should not induce jumps in value. Non-decreasing is related to the discussion above: higher component values are better. The concavity might perhaps fail in certain settings but in many it could be quite reasonable – expressing a type of diminishing returns property.

ASSUMPTION 3. *The common grading language allows for continuous grades in $\mathbb{G} \equiv [0,1]$, where a higher grade implies better acceptability by a player.*

This assumption defines a common voting language, which is necessary in Majority Judgment.

ASSUMPTION 4. *Each player derives its grade function by a simple linear transformation of its value function. Specifically, define $V_i^{max} = \max_{\mathbf{m}\in\mu} V_i(\mathbf{m})$; then $g_i(\mathbf{m}) = V_i(\mathbf{m})/V_i^{max}$.*

We also might consider slightly more general transformations. However, in general it is possible (and perhaps profitable) for a player to consider a variety of strategies to set the grade function based on their own value function and also knowledge or assumptions regarding the value functions and/or strategies

of the other players. Reducing or eliminating the gain that could be achieved by such "strategic" voting is a very important design consideration. We will address it in future research, currently relying on the strategy-resistance claims of Majority Judgment by its authors.

3.2.1. Linear Representation of Feasible Candidate Set and Grade Functions. The assumptions just described allow us to produce an efficient form of the optimization models previously described. Specifically, Assumption 1 allows us to use a piecewise linear approximation to represent the space of feasible candidates and we can replace $\mathbf{m} \in \mu$ with:

$$\mathbf{c^1} m_1 + \mathbf{c^2} m_2 + \cdots + \mathbf{c^p} m_p \leq \mathbf{c^0} \tag{6}$$

where $\mathbf{c}$'s are appropriately defined coefficient vectors.

Assumptions 2 and 3 allow us to use similar piecewise linear approximations in place of the grade functions. We approximate $x_i = g_i(\mathbf{m})$ with

$$\mathbf{d^1}_i m_1 + \mathbf{d^2}_i m_2 + \cdots + \mathbf{d^p}_i m_p + x_i \leq \mathbf{d^0}_i$$

where $\mathbf{d}_i$'s are appropriately defined coefficient vectors. The fact that higher grades are always preferred allows us to replace each equality constraint with a set of inequalities that approximate the grade functions.

3.2.2. Grade Function Model. In the prior Section we showed how to represent the grade functions using linear constraints. However, doing this requires knowledge of the grade functions. In fact, the central planner will only observe the players' voting behavior. Our candidate generation process requires that we approximate player grade function based on these observations. We will do this using statistical models that assume a particular functional form for the grade functions. The functional form we assume is based on well-accepted notions developed by economists and marketing researchers in the fields of choice modeling and multi-attribute valuation (e.g. Meyer and Johnson (1995)).

Each player takes three steps to determine the grade of a given candidate. The first two involve the value function (V_i) and the last converts the value function approximation into the grade function (g_i). First, she determines the value of each individual component of the candidate – holding the other components at constant levels. Second, she integrates the individual valuations of the components into an overall value of the entire candidate. Third, she normalizes the value of the candidate into its grade.

Specific models are now proposed for each step. First, the value of an individual component m_r to i is modeled as a non-decreasing concave function $\nu_{r_i}(m_r)$. The value can be visualized as net profitability gain as the metric value is increased, holding other metrics at constant levels. The concavity assumption models diminishing marginal returns as the metric value increases. Second, the integration step combines the individual value functions as a multiplicative-multilinear function of $\nu_{r_i}(m_r)$'s, modeling complementarities among the valuations over the different component metrics:

$$\tilde{V}_i(\mathbf{m}) = r_{1_i}\nu_{1_i}(m_1) + r_{2_i}\nu_{2_i}(m_2) + r_{12_i}\nu_{1_i}(m_1)\nu_{2_i}(m_2) + \ldots,$$

with non-negative coefficients $r_{1_i}, r_{2_i}, r_{12_i}, \cdots \geq 0$. The non-negativity of the constants implies that higher values are better, and that the individual components are not substitutes to each other. For more than two components, pair-wise interaction terms are added; higher-order interaction terms are ignored. Finally, the normalization step converts the integrated value into a grade, using a simple linear scaling based on the maximum value $\tilde{V}_i^{max}$. The grade function for player i is thus specified as:

$$g_i(\mathbf{m}) = \frac{r_{1_i}\nu_{1_i}(m_1) + r_{2_i}\nu_{2_i}(m_2) + r_{12_i}\nu_{1_i}(m_1)\nu_{2_i}(m_2) + \ldots}{\tilde{V}_i^{max}},$$

Appendix A provides further implementation details.

3.3. Iterative Procedure

In practice, the true grade functions $g_i(\mathbf{m})$ will be confidential to the players. We use the functional form just described in a procedure that statistically approximates the grade functions based on each player's observed grades, denoted $\hat{g}_i(\mathbf{m})$. Appendix B provides details on the estimation procedure.

The optimization problems **Subset_Opt**(b) and / or **Player_Opt**(i) can be solved with the estimated grade functions $\hat{g}_i(\mathbf{m})$, for some or all $b \in \beta$ or $i \in N$ respectively. The resultant candidates will be an approximation to those computed with the true grade functions. All or a subset of these "generated" candidates are put to vote by the players. This cycle of estimation, new candidate generation, voting is repeated until a stopping criterion is met. Algorithm 1 summarizes the entire mechanism:

Algorithm 1 Algorithm for Proposed Mechanism

```
Initialize consideration set of feasible candidates
repeat
    Obtain players' grades on the consideration set
    Estimate players' grade function coefficients
    Generate new feasible candidates and / or ask players for new feasible candidates
    Introduce some or all new candidates into consideration set
until stopping criteria met
```

3.4. Evaluation

The "optimal" candidate $\mathbf{m}^*$ uses the "true" grade functions $g_i(\mathbf{m})$, while the "winning" candidate $\widehat{\mathbf{m}}^*$ emerges after the mechanism run using estimated grade functions $\hat{g}_i(\mathbf{m})$. The two are compared to evaluate accuracy of the procedure.

Deviation between candidates is determined as the Euclidean distance between the two. For p-dimensional candidate space:

$$d_v = \sigma \sqrt{\sum_{s=1}^{p} (\widehat{m}_s^* - m_s^*)^2}.$$

$\sigma = \pm 1$ assigns a sign to differentiate outcomes with negative versus positive deviation.

Recall the majority grade of the optimal candidate is $v(\mathbf{m}^*)$, or v^*. The "true" majority grade of the winning candidate is computed with the true grade functions for the players, and denoted $v(\widehat{\mathbf{m}}^*)$. Deviation in majority grades is determined as:

$$d_g = \left(\frac{v(\widehat{\mathbf{m}}^*)}{v(\mathbf{m}^*)} - 1 \right) \times 100.$$

By definition, d_g can not exceed 0; however, errors in the piecewise linear approximations of the grade functions may lead to violations.

d_v is an absolute measure, useful in comparing several variants of the mechanism. d_g is relative – akin to "optimality gap", it can be used to assess the overall quality of the mechanism itself.

4. Experimental Results

A large simulation experiment was conducted to validate the proposed mechanism using data from real-life operations. The data selection and preparation is explained first. Instead of randomly fixing the "true" grade functions for the different airlines, some judgment was exercised to mimic reality. This intuition was vetted within the research team which has expertise in air traffic flow management. The procedure to draw the coefficients for grade function with quadratic functional form is detailed in the appendix. Determination of each airline's weight is also a practical challenge. Multiple weighting schemes are explained.

4.1. Data

October 10, 2007 was selected as the sample date. It was a mid-week (Wednesday), with no exceptional events like holidays or expectations of severe weather. The entire day's scheduled departures were included in the dataset.

In terms of geographical scope, the Chicago area airports – ORD (O'Hare) and MDW (Midway) – were included. Operations of feeder airlines were merged into their main airlines' operations. OAG schedule data was used for calculating the number of flights impacted, the left panel of Table 3 sums up results. The setup is representative of real-life: impact of the weather on a part of the National Air Space spanning multiple airports of differing sizes, dominance of a few larger airlines, and a long-tail of smaller airlines.

Heterogeneity in airline operations is evident. The final dataset comprises of 47 airlines, totaling 1603 operations. Six hub-and-spoke airlines make up more than $^3/_4$-th of the operations – 1243 in total. Eight point-to-point airlines make up the next largest group, with 292 operations. 25 international airlines have total 50 operations, three charter airlines have 11 operations, and five cargo airlines have seven operations.

At the airline-level operations, four groups emerge. The first group has three large airlines with large presence: United, American, and Southwest. With over 100 operations each, these make up over 85% of total operations. The second group has five large airlines with small presence: Northwest, Delta, US airways, Continental, and Airtran. With operations between 10 and 100 each, these make up about 8% of total. The third group has between 2 and 9 operations, and comprises of 20 airlines. The fourth group has 19 airlines with a single operation.

4.2. Feasible Candidate Space

An adversely impacted day-of-operations will suffer loss in the service performance metrics as compared to a normal day-of-operations. The metrics are inter-related, requiring trade-offs amongst them. For instance, an "aggressive" approach might yield a high capacity-utilization, but at the expense of delaying the time when final decisions on releasing flights are made, thus reducing predictability. On the other hand, a "conservative" strategy may release fewer flights that are closely tracked by the air traffic controllers; thereby yielding a high predictability, but low capacity-utilization. Infinitely many "moderate" strategies can be proposed in the intervening space.

Research conducted by other members of our research team has shown a concave relationship among representative metrics for three performance categories: efficiency, predictability and capacity (Ball et al. 2011). The relationships are developed for a single airport, by varying the time-period during which the airport suffers a reduced capacity due to bad weather. The metrics are normalized to lie between 0 and 1; the infeasible "ideal point" (1,1,1) represents a normal day-of-operations where all

Airline	MDW	ORD	Characteristics	Profile	nops	log.2	root.10
United (UA) [a]	10	625	Large hub & spoke, large presence	HL	635 (39.6)	9.31 (10.8)	8.85 (10)
American (AA) [b]		500	Large hub & spoke, large presence	HL	500 (31.2)	8.97 (10.4)	8.16 (9.2)
Southwest (WN) [c]	242		Large point-to-point	LH	242 (15.1)	7.92 (9.2)	6.39 (7.2)
Northwest (NW) [d]	11	23	Large hub & spoke, small presence	SS	34 (2.1)	5.09 (5.9)	3.29 (3.7)
Delta (DL) [e]	6	22	Large hub & spoke, small presence	SS	28 (1.7)	4.81 (5.6)	3.08 (3.5)
US Air (US)		27	Large hub & spoke, small presence	SS	27 (1.7)	4.75 (5.5)	3.04 (3.4)
Continental (CO)	2	17	Large hub & spoke, small presence	SS	19 (1.2)	4.25 (4.9)	2.70 (3.1)
Airtran (FL)	18		Large point-to-point	LH	18 (1.1)	4.17 (4.8)	2.65 (3)
Air Canada		8	International, neighboring regions	LH	8 (0.5)	3 (3.5)	2.02 (2.3)
ExpressJet	3	4	Small point-to-point	LH	7 (0.4)	2.81 (3.3)	1.93 (2.2)
Jetblue		7	Small point-to-point	LH	7 (0.4)	2.81 (3.3)	1.93 (2.2)
Chautauqua		6	Small point-to-point	LH	6 (0.4)	2.58 (3)	1.83 (2.1)
Frontier	6		Small point-to-point	LH	6 (0.4)	2.58 (3)	1.83 (2.1)
Mexicana		6	International, neighboring regions	LH	6 (0.4)	2.58 (3)	1.83 (2.1)
Lufthansa		5	International, business-dominant	HL	5 (0.3)	2.32 (2.7)	1.72 (1.9)
Primaris	5		Small charter	HL	5 (0.3)	2.32 (2.7)	1.72 (1.9)
Alaska		4	Small point-to-point	LH	4 (0.2)	2 (2.3)	1.60 (1.8)
Air Midwest	4		Small charter	HL	4 (0.2)	2 (2.3)	1.60 (1.8)
Aeromexico		3	International, neighboring regions	LH	3 (0.2)	1.58 (1.8)	1.45 (1.6)
British Airways		3	International, business-dominant	HL	3 (0.2)	1.58 (1.8)	1.45 (1.6)
Polar Air Cargo		3	Cargo	SS	3 (0.2)	1.58 (1.8)	1.45 (1.6)
Spirit		2	Small point-to-point	LH	2 (0.1)	1 (1.2)	1.26 (1.4)
Aer Lingus		2	International	SS	2 (0.1)	1 (1.2)	1.26 (1.4)
Air Canada Jazz		2	International	SS	2 (0.1)	1 (1.2)	1.26 (1.4)
Lot - Polish		2	International	SS	2 (0.1)	1 (1.2)	1.26 (1.4)
SAS Scandinavian		2	International	SS	2 (0.1)	1 (1.2)	1.26 (1.4)
Singapore		2	International	SS	2 (0.1)	1 (1.2)	1.26 (1.4)
USA 3000		2	Small charter	HL	2 (0.1)	1 (1.2)	1.26 (1.4)
Air France		1	International	SS	1 (0.1)	–	1 (1.1)
Air India		1	International, economy-dominant	LH	1 (0.1)	–	1 (1.1)
Air Jamaica		1	International	SS	1 (0.1)	–	1 (1.1)
Alitalia		1	International	SS	1 (0.1)	–	1 (1.1)
All Nippon		1	International	SS	1 (0.1)	–	1 (1.1)
British Midland		1	International	SS	1 (0.1)	–	1 (1.1)
Iberia		1	International	SS	1 (0.1)	–	1 (1.1)
Japan International		1	International	SS	1 (0.1)	–	1 (1.1)
KLM-Royal Dutch		1	International	SS	1 (0.1)	–	1 (1.1)
Korean		1	International	SS	1 (0.1)	–	1 (1.1)
Martinair Holland		1	International	SS	1 (0.1)	–	1 (1.1)
Pakistan International		1	International, economy-dominant	LH	1 (0.1)	–	1 (1.1)
Swiss		1	International	SS	1 (0.1)	–	1 (1.1)
Turkish		1	International	SS	1 (0.1)	–	1 (1.1)
Virgin Atlantic		1	International	SS	1 (0.1)	–	1 (1.1)
ABX		1	Cargo	SS	1 (0.1)	–	1 (1.1)
Cargoitalia		1	Cargo	SS	1 (0.1)	–	1 (1.1)
Custom Air		1	Cargo	SS	1 (0.1)	–	1 (1.1)
Kalitta		1	Cargo	SS	1 (0.1)	–	1 (1.1)
TOTAL	307	1296			1603 (100)	86.03 (100)	88.38 (100)

[a]includes several United feeders like Go Jet, YV, Shuttle America, United / Skywest, Trans Air; [b]includes American Eagle; [c]includes ATA; [d]includes Mesaba; [e]includes Skywest, Comair, Atlantic Southeast.

Table 3 Data for MDW and ORD airline-wise scheduled departures on 10 Oct, 2007.

the performance metrics are realized at 100% levels. The envelope forms the efficient frontier, while all the interior points serve as feasible region. Two metrics – capacity-utilization and predictability – are used for illustrative purposes here, the proposed procedures extend to any number of metrics.

4.3. "True" Grade Functions

Airlines can be broadly classified along several dimensions. (i) number of operations: large, medium, or small airline. (ii) type of network: hub-and-spoke airline, or point-to-point. (iii) type of operations: cargo

or passengers. (iv) customer focus: business-dominant, or economy-dominant, or type-independent. (v) distance of markets: long-haul, or short-haul. (vi) political markets served: domestic, or international.

To make the setup realistic, these differentiating factors should be reflected in the grade function of the airlines. Some judgment was exercised in modeling the airline grading behavior; it was vetted within the extended research team, which has expertise in air traffic flow management.

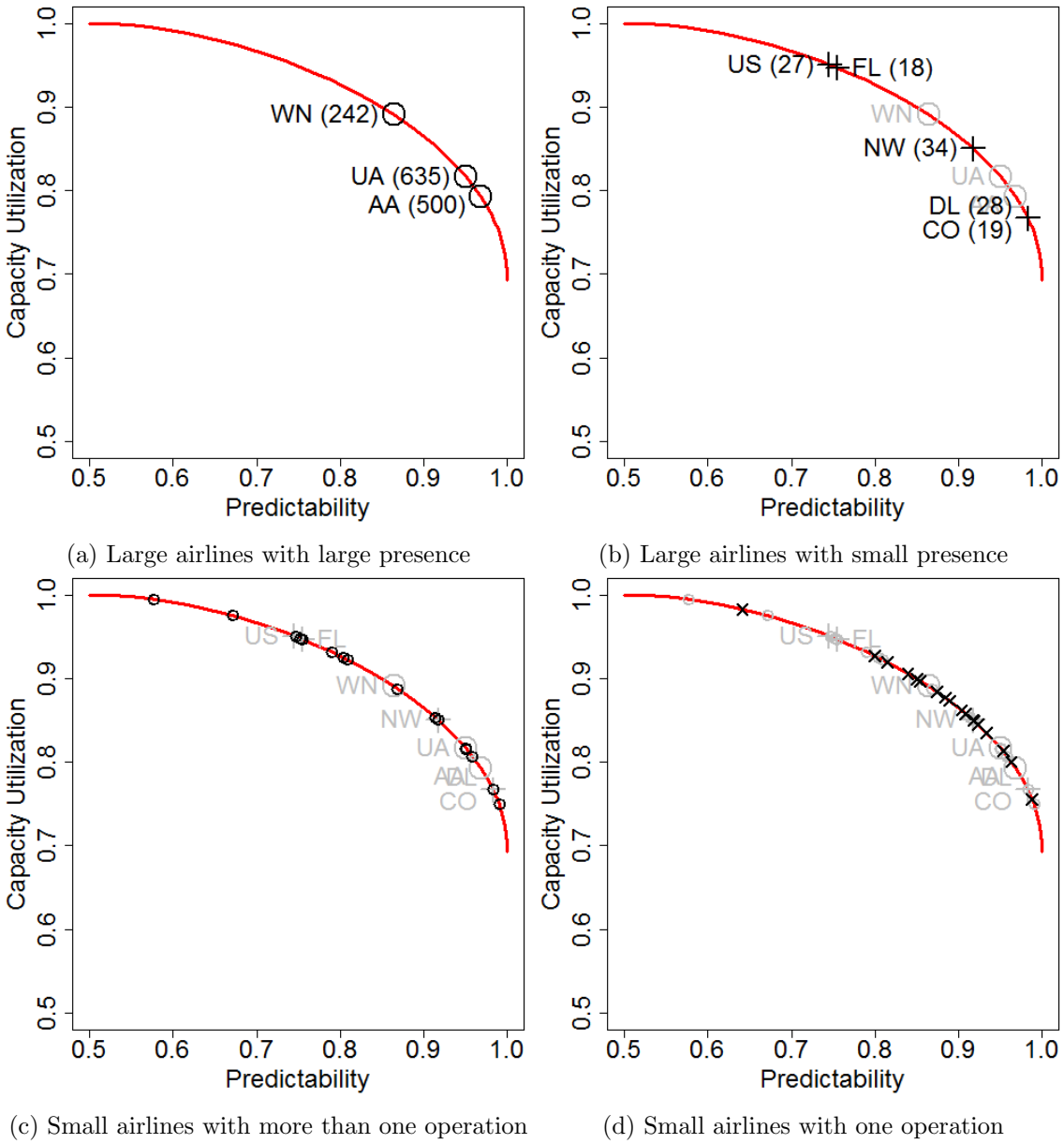
Between the two metrics, we first assessed how each airline would value the two relatively. The possibilities are: "HL", "LH", "SS", where H indicates High, L Low, and S Same; the letters pertaining to predictability and capacity utilization respectively. It does not matter if absolute levels are both H or both L, as the normalization process would not differentiate between the two. Airline characterizations and their posited profiles are summarized in the middle panel of Table 3.

We posit large hub-and-spoke airlines with a significant presence, United and American in this instance, to have HL profile, as they have a large pool of aircrafts to re-balance the impacted passengers – so long as they know the impact adequately in advance. Hence, they would care a lot more about predictability than capacity utilization. However, this can not be said of the other large hub-and-spoke airlines with a small presence (Northwest, Delta, US Air, Continental), hence we assign them the neutral SS profile.

The low-cost point-to-point airlines – of any size – are hypothesized to prefer capacity utilization than predictability. Their predominantly economy passengers are likely interested in completing their itinerary, without a significant time-sensitivity. Hence, we assign LH to large point-to-point airlines (Southwest and Airtran), as well as the smaller ones (ExpressJet, Jetblue, Chautauqua, Frontier, Alaska, and Spirit). We posit the opposite should hold for luxury or time-sensitive passenger focused Charter airlines. Primaris, Air Midwest, USA 3000 are, therefore, assigned HL profile.

We treat the international airlines serving the neighboring countries to be similar to the point-to-point operators, and assign Air Canada, Mexicana De Aviacion, and Aeromexico LH profile too. Lufthansa and British Airways are posited to cater to more time-sensitive passengers, hence assigned HL profile, while Air India and Air Pakistan are treated as opposite and therefore assigned LH profile. All the remaining international airlines are assigned the neutral SS profile . Finally, cargo carriers are also posed to value the two metrics similarly – and are assigned SS profile.

Next, we assessed the degree of curvature for the value function of each individual metric. The possibilities are: small curvature (straight-line like) and large curvature (more concave). We posit that



(a) Large airlines with large presence

(b) Large airlines with small presence

(c) Small airlines with more than one operation

(d) Small airlines with one operation

Figure 1 **Grade-maximizing candidates for different groups of airlines**

the airlines with smaller operations would have a straight-line like curvature, as they would not have as much degree of freedom than the airlines with larger presence. The latter are more likely to observe increasingly diminishing returns, and hence, would have a more concave shape.

Appendix C explains implementation of this intuition using quadratic functional form for the airlines' grade functions. The grade-maximizing candidates are plotted in Figure 1 for the various groups of airlines. The diversity shows the effectiveness of the procedure.

4.4. Weights

The democratic "one-person, one-vote" assigns a weight of one to all the airlines ("eqwt"). This may be perceived as inequitable in many practical decision-making contexts though. E.g., in our case, it implies that airlines with a single operation get same representation as those with hundreds of operations. Nonetheless, this is a benchmark for evaluating other weighting schemes.

Proportional representation can be achieved by replicating each voter's grade as many times as her weight. A basis is needed for determining the weights. To keep matters simple, practical, and minimal in private information, we use publicly available data on total operations impacted as the basis. It is also a very relevant measure to use in the current context.

The weights are traditionally seen as integers, with the interpretation as given above. In our case though, weights can be fractional. A majoritarian set is formed by a set of voters if the proportion of their combined weight is strictly above 0.5.

A simple scheme could use the number of operations as weights ("nops"). However, few voters may get significantly high influence. In this instance, United alone has about 40% operations, the top two airlines make over 70% of total operations. Thus, it may be beneficial to balance the influence of the larger voters.

Logarithmic and power-root transforms on the number of operations would reign in the large positive numbers. However, the choice of base is an open decision. We tried logarithms to three well-known bases: 10, e, and 2, and selected the base 2 for our experiment ("log.2"). The other two bases had lesser differentiation among the airline weights – for the two largest airlines: $\log_{10}(635) = 2.8; \log_{10}(500) = 2.7$ and $\ln(635) = 6.45; \ln(500) = 6.21$. As $\log(1) = 0$, the log transform assigns weight of zero to the airlines with a single operation – which may or may not be desirable. In this instance, the airlines with single operations are mostly international and cargo airlines. If eliminating these is seen as inequitable, a $\log(.) + 1$ would ensure that all airlines have some say in the mechanism.

Alternately, fix the largest airline's proportion of total weight at some desired level, say π_{max}. Power-root transforms can accomplish this. To get π_{max} of 30%, 20%, and 10% ("root.30", "root.20", "root.10") in our example, these are respectively: 1.32585, 1.80390, 2.96015.

While all of these are valid choices, the exact decision of which one to choose would not be taken at the time of each mechanism run. This decision should be made experimentally, and then left unchanged for a relatively long period of time, until there are reasons to reconsider.

We will evaluate results with four weighting schemes: eqwt, nops, log.2, and root.10. log.2 eliminates the 19 airlines with a single operation. root.10 has similar proportional weight for United as log.2.

4.5. Mechanism Design Choices

At this stage, all the inputs for running the procedure are ready. There are a few design choices still to be made though.

4.5.1. Initial consideration set. To initiate the mechanism, the ANSP could provide the airlines a set of candidates. The airlines may heuristically arrive at the grades, through possibly comparing the candidates among themselves.

Alternately, it could communicate the feasible candidate space, and request the airlines to provide their grade-maximizing candidates – to be graded 1. This may be perceived as equitable as the airlines get to submit their most preferred candidates upfront. It also addresses the scaling problem, as the grade of 1 is clearly established for each airline at the outset. However, it does need the airlines to solve a type of profit-maximization problem with feasibility constraints.

Our initial experiments found the former approach converging faster than the latter. Hence, we initialize the consideration set with five or more equally spaced candidates, as there are five coefficients to be estimated for the quadratic value functions.

4.5.2. Extent of agreement. Majority Judgment is a median-seeking procedure. The median has the desirable property that it exactly balances the number of votes that find a candidate's grade too high with those that find its grade too low (Galton 1907). This property will be lost in seeking a non-median based solution, and may encourage strategic behavior.

Having said that, the procedure can be easily extended to allow for any higher (or lower) level of agreement. When seeking a higher (lower) agreement, the Majority Grade of the final candidate could be smaller (higher). Alternate criteria may be explored, for instance, one that seeks a minimum number of airlines to be in the majoritarian set. Any deviations should be subjected to a strategic behavior analysis. In the experiment, the extent of agreement is set at the original, 50% of total weight.

4.5.3. Voter input. At the end of any round, the ANSP may ask for the grade-maximizing candidates from the airlines (if not already done). Alternatively, the ANSP may choose not to ask the airlines for their input. In our experiments, we adopted the latter.

Variants of this alternative may be adopted in practice. For instance, it may be made optional for certain airlines – e.g. those with smaller number of operations, who may possibly not have sufficient infrastructure, and / or stake in the current decision-making context. Furthermore, smaller subsets of airlines may be requested after each round. This would ensure that the consideration set is kept manageable over the rounds. For maintaining equity though, the selection of airlines may be made random, or through a preset procedure.

4.5.4. Consideration set update. At the end of each round, the voter input and the ANSP-generated new candidates are available. A balance has to be made between the size of the consideration set and its quality. Among the new candidates, one could select few candidates with the highest Majority Grades. In our initial experiments, we found that this strategy led to inferior final winners. The inherent error in the estimation of the grade functions is likely the cause.

On the other hand, adding all the new candidates would lead to very large consideration sets. The ANSP may select few diverse candidates among the voter input and new candidates – or it could randomize the selection.

In our experiments, we added all the new candidates generated at the end of each round into the consideration set. This was so we could learn about convergence of the overall procedure with a large input. Results from this experiment would serve to benchmark other strategies in future.

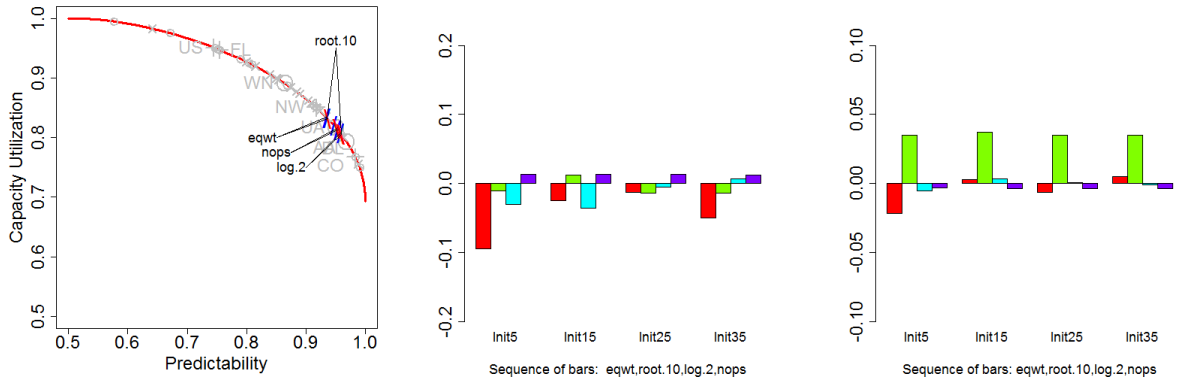
4.5.5. Consistency in grading. We assume the airlines grade every candidate precisely, and report the grades truthfully. In real-life, one or both of the assumptions may not hold, necessitating establishment of consistency rules. In our experiments though, no consistency checks are required. This experiment establishes a benchmark to evaluate the results with different consistency rules.

4.5.6. Stopping criterion. We chose a simple stopping criterion of six rounds for the experiments – in the interest of convergence. More sophisticated stopping criteria should be evaluated against the benchmark established herein.

4.6. Mechanism Evaluation

Several runs of the mechanism were conducted with varying parameters. This section reports evaluations in terms of accuracy and technical performance measures.

Figure 2a shows the optimal and the winning candidates for different weighting schemes, for one of the runs. In this run, apart from root.10, all the other weighting schemes produced winners very close to the optimal candidates.



(a) Optimal ('/') and winning ('\') candidates for Init15

(b) Percentage deviation of the majority grade

(c) Signed Euclidean distance

Figure 2 Evaluation results over several initial consideration set sizes and weighting schemes

As just explained, the initial consideration set is a key parameter of interest. We increased the size of the initial consideration set from 5 through 35, in steps of 10 – the respective runs are called "Init5" through "Init35". Experiments with larger sizes did not yield any significant improvements.

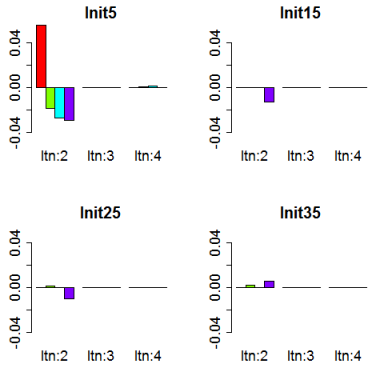
Figure 2b plots the percentage deviation in the majority grades of the winning candidate relative to that of the optimal candidate, d_g. The median absolute percentage error is about 0.013%. By complete enumeration of the majority grades using true grade functions over the entire efficient frontier, we found its range to be (0.88, 0.98) – over all the weighting schemes. Hence, the optimality gap is about $0.013\%/(0.98-0.88) = 0.13\%$, which indicates the high quality of the mechanism outcome.

Figure 2c plots the signed Euclidean distances between the winning and optimal candidate, d_v. A negative sign was ascribed to d_v if the predictability metric of the winning candidate was less than the optimal candidate's (the winning candidate lay to the "left" of the optimal candidate in Figure 2a).

We observe that the winning candidates obtained by the mechanism are quite close to the optimal ones. A larger size does not necessarily mean better solutions consistently – only Init5 seems to suffer in overall quality, but the others are quite similar. Recall this is after six rounds of grading.

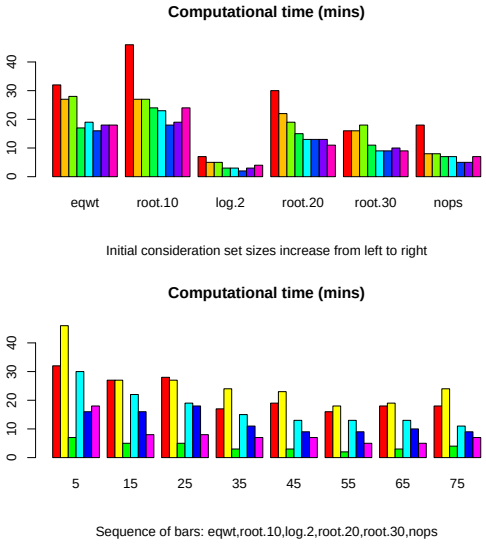
Figure 3a reports on convergence over the rounds. It plots the signed distances for winning candidate in each round over the one in the previous round. We note that except for Init5, all the higher initial consideration set sizes practically converge at the end of the first round itself. However, it may still be beneficial to have at least two rounds.

These experiments were conducted on a personal laptop with Intel Celeron Dual-core CPU (1.8 GHz), having 2 GB RAM, running 32-bit Microsoft Windows 7 Home Premium operating system.



Sequence of bars: eqwt,root.10,log.2,nops

(a) Euclidean (signed) distance of winning candidates for each round over the previous round



(b) Computational results

Figure 3 Evaluation results over several initial consideration set sizes and weighting schemes

Computing environment used was R version 13.0, with API Rcplex to interface with the CPLEX 12.0 solver, obtained through IBM Academic Initiative.

Figure 3b plots the computational times for running six rounds for the respective weighting scheme-initial consideration set size combination. log.2 scheme eliminates the airlines with single operation, hence takes the smallest time. nops gives largest weight to the largest airline, hence takes lesser time than the root-transformed schemes. The computational times increase as the largest airline is apportioned smaller weight: root.10 takes longest, followed by root.20, then root.30, which takes about same time as nops. eqwt interestingly does not take the longest, which gives all airlines equal weight. Finally, an interesting observation is that higher initial consideration set sizes take lesser time to compute.

All the computations were run serially. As each airline's process is independent of other's, there is scope for parallelization. In effect, the computational times could be $\frac{1}{47}$-th of those reported. Moreover, for just two rounds, the computation time should further reduce by 67%.

5. Conclusions

In this paper, we have described a mechanism for generating a consensus vector for use in strategic planning in air traffic flow management. Our approach is based on Majority Judgment but it employs a novel extension: the ability to handle very large sets of candidates. Our experimental results show the methods developed are very effective and can be efficiently carried out.

Several additional steps are required (and currently being carried out) to achieve practicality in the ATFM context. These include developing intuitive mechanisms for the flight operators to understand the performance vectors and to grade them, development of methods to generate the constraints defining the feasible vector space (μ) based on the current weather conditions and air traffic demand, estimation of benefits and human-in-the-loop experiments.

Of particular importance both to the ATFM application and more general applications is the issue of the potential for strategic grading/voting. Our experiments assumed that flight operators graded in a manner that was consistent with their true value functions. While Majority Judgment is generally (somewhat) immune to gaming, this issue deserves further analysis. For example, it could be the case that certain rules should be put in place to help insure reasonable behavior and outcomes, e.g. rules against collusion seem to be warranted.

Finally we note that we are quite excited about the potential application of this mechanism in other areas. There would seem to be a natural fit for many other application contexts.

Acknowledgments

This work was supported by the Federal Aviation Administration through the NEXTOR-II Consortium.

References

Arrow, K. 1951. *Individual values and social choice*. New York: Wiley.

Balinski, M., R. Laraki. 2011. *Majority Judgment: Measuring, Ranking, and Electing*. The MIT Press.

Ball, M. O., C. Barnhart, A. Evans, M. Hansen, Y. Liu, P. Swaroop, V. Vaze. 2011. Distributed Mechanisms for Determining NAS-Wide Service Level Expectations. *NEXTOR Technical Report, July 2011.* **Year 1 Report**. URL `http://www.nextor.org/rep2011.html`.

Charnes, Abraham, William W Cooper, Edwardo Rhodes. 1978. Measuring the efficiency of decision making units. *European journal of operational research* **2**(6) 429–444.

Cook, Wade D, Larry M Seiford. 2009. Data envelopment analysis (dea)–thirty years on. *European Journal of Operational Research* **192**(1) 1–17.

Galton, Francis. 1907. One vote, one value. *Nature* **75** 414.

Green, Paul E, Abba M Krieger, Yoram Wind. 2001. Thirty years of conjoint analysis: Reflections and prospects. *Interfaces* **31**(3 supplement) S56–S73.

Green, Paul E, Vithala R Rao. 1971. Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research* 355–363.

Grushka-Cockayne, Yael, Bert De Reyck, Zeger Degraeve. 2008. An Integrated Decision-Making Approach for Improving European Air Traffic Management. *Management Science* **54**(8) 1395 –1409.

ICAO. 2005. Global Air Traffic Management Operational Concept (Document 9854). *International Civil Aviation Organization* **First Edition**. URL `http://www.icao.int/Meetings/anconf12/Documents/9854_cons_en[1].pdf`.

JPDO. 2007. Concept of Operations for the Next Generation Air Transportation System. *Joint Planning and Development Office, Washington, DC* **Version 2.0**.

Karasakal, Esra, Murat Köksalan. 2009. Generating a representative subset of the nondominated frontier in multiple criteria decision making. *Operations research* **57**(1) 187–199.

Maskin, E.S. 2008. Mechanism design: How to implement social goals. *The American Economic Review* **98**(3) 567–576.

Meyer, R., E.J. Johnson. 1995. Empirical generalizations in the modeling of consumer choice. *Marketing Science* 180–189.

Raiffa, Howard, R Keeney. 1976. *Decisions with multiple objectives: Preferences and value tradeoffs*. John Wiley and Sons.

Ruzika, Stefan, Margaret M Wiecek. 2005. Approximation methods in multiobjective programming. *Journal of optimization theory and applications* **126**(3) 473–501.

Saaty, Thomas L, Luis G Vargas. 2012. *Models, methods, concepts & applications of the analytic hierarchy process*, vol. 175. Springer.

Sayın, Serpil. 2000. Measuring the quality of discrete representations of efficient sets in multiple objective mathematical programming. *Mathematical Programming* **87**(3) 543–560.

SESAR. 2006. Air Transport Framework: The Performance Target. *SESAR Consortium* **2**.

Wallenius, Jyrki, James S Dyer, Peter C Fishburn, Ralph E Steuer, Stanley Zionts, Kalyanmoy Deb. 2008. Multiple criteria decision making, multiattribute utility theory: Recent accomplishments and what lies ahead. *Management Science* **54**(7) 1336–1349.

All appendices deal with individual players; subscript i for player is suppressed.

Appendix A: Grade Function Specification

Without loss of generality, the individual component metrics are normalized to have support in $[0,1]$. Two components are used for explanation, but the specification easily extends to any number of components.

A quadratic form is specified for the value functions of individual components, $\nu_s(m_s) = a_s m_s^2 + b_s m_s$, without an intercept. To obtain the desired increasing function over the range of m_s, the values of a_s and b_s need to be constrained such that $-1 \le a_s < 0; 0 < -2a_s \le b_s \le 1 - a_s$. This yields: $-0.5 \le \frac{a_s}{b_s} < 0$.

Substituting ν's into the grade function, normalizing and renaming the coefficients gives:

$$g(\mathbf{m}) = k_1 m_1 + k_2 m_2 + k_3 m_1^2 + k_4 m_2^2 + k_5 m_1 m_2 + k_6 m_1^2 m_2 + k_7 m_1 m_2^2 + k_8 m_1^2 m_2^2,$$

with following constraints for concavity and the integration rule: $k_1 \ge 0; k_2 \ge 0; -0.5k_1 \le k_3 < 0; -0.5k_2 \le k_4 < 0; k_5 \ge 0$. The renaming yields: $\frac{k_3}{k_1} = \frac{a_1}{b_1}; \frac{k_4}{k_2} = \frac{a_2}{b_2}$, and thus: $-0.5k_1 \le k_3 < 0; -0.5k_2 \le k_4 < 0; -0.5k_5 \le k_6 < 0; -0.5k_5 \le k_7 < 0; 0 \le k_8 \le 0.25k_5$. Note that the normalization involves V_i^{max}, which can be computed using the optimization model provided in **Subset_Opt**(b), as follows. Specify $b = \{i\}$, and replace the constraint $x_i = g_i(\mathbf{m})$ with $x_i = V_i(\mathbf{m})$ – that is, the (un-normalized) value function. This would yield the V_i^{max} for the player i at optimality.

Normalization would only be required if the grade function is specified from the value functions of the individual components of the candidate space. Instead, if the specification with k's is used directly, and the constraints as mentioned above are honored for all k's, then the resulting grade function would automatically have the support in $[0, G^{max}]$. However, it is not possible to recover the original constants a's and b's from the k's. Only global concavity remains to be ensured.

Proposition 2. *Any one of the following constraints is a sufficient condition for global concavity of the grade function as specified above:* $k_1 + 3k_3 m_1 \le 0, k_2 + 3k_4 m_2 \le 0$.

Proof. The Hessian matrix of a function being negative definite in a given region is a necessary and sufficient condition for concavity of the function within it. The region of interest here is: $\mathbf{m} \in ((0,0), \ldots, (1,1)]$:. Denote the Hessian matrix of the grade function as $\mathbf{H_g} = \begin{bmatrix} g_{11} & g_{12} \\ g_{12} & g_{22}, \end{bmatrix}$ where g_{st} is the partial derivative of the $g(\mathbf{m})$ with respect to m_s and m_t. For $\mathbf{m} \ne \mathbf{0}$, non-negative r's and the above relationships for k's:

$$\mathbf{m}^T \mathbf{H_g} \mathbf{m} = \underbrace{2\left[k_3 m_1^2 + k_4 m_2^2\right]}_{<0} + \underbrace{2k_5 m_1 m_2}_{\ge 0} \underbrace{\left[1 + 3\frac{k_3}{k_1} m_1 + 3\frac{k_4}{k_2} m_2 + 6\frac{k_3}{k_1}\frac{k_4}{k_2} m_1 m_2\right]}_{?}$$

The first bracketed term is negative as $k_3, k_4 < 0$, and $\mathbf{m} \neq \mathbf{0}$ by hypothesis. Further, as $k_5 \geq 0$, if the final bracketed term is negative, the entire expression $\mathbf{m}^T\mathbf{H_g}\mathbf{m}$ would be negative, and the Hessian would be negative definite. However, it is not guaranteed to be so, as explained below.

Re-express the final bracketed term as: $\left[1 + 3\frac{k_3}{k_1}m_1 + 3\frac{k_4}{k_2}m_2 + 6\frac{k_3}{k_1}\frac{k_4}{k_2}m_1m_2\right]$

$$= \underbrace{1 + 3\frac{k_3}{k_1}m_1}_{h_1^a} + \underbrace{3\frac{k_4}{k_2}m_2}_{h_2^a}\underbrace{\left(1 + 2\frac{k_3}{k_1}m_1\right)}_{h_3^a} = \underbrace{1 + 3\frac{k_4}{k_2}m_2}_{h_1^b} + \underbrace{3\frac{k_3}{k_1}m_1}_{h_2^b}\underbrace{\left(1 + 2\frac{k_4}{k_2}m_2\right)}_{h_3^b}$$

Recall that $\frac{k_3}{k_1} = \frac{a_1}{b_1}$, hence for $0 < m_1 \leq 1$:

$$-0.5 \leq \frac{k_3}{k_1} < 0 \Rightarrow -1 \leq 2\frac{k_3}{k_1} < 0 \Rightarrow -m_1 \leq 2\frac{k_3}{k_1}m_1\text{`} < 0 \Rightarrow 1 - m_1 \leq h_3^a < 1 \Rightarrow 0 \leq h_3^a < 1.$$

Similarly for $0 < m_2 \leq 1$: $-1.5 \leq h_2^a < 0$. Correspondingly, for $\mathbf{m} \neq \mathbf{0}$: $0 \leq h_3^b < 1; -1.5 \leq h_2^b < 0$. Thus, following hold true for $\mathbf{m} \neq \mathbf{0}$: $-1.5 \leq h_2^a h_3^a \leq 0; -1.5 \leq h_2^b h_3^b \leq 0$. Consider the following two cases for h_1's.

Case 1 $h_1^a \leq 0$ or $h_1^b \leq 0$. This would directly imply that $\mathbf{m}^T\mathbf{H_g}\mathbf{m} \leq 0$, and is thus a sufficient condition for concavity of the grade function.

Case 2 $h_1^a > 0$ and $h_1^b > 0$. It follows then that: $1 + 3\frac{k_3}{k_1}m_1 > 0 \Rightarrow 1 > -3\frac{k_3}{k_1}m_1 \Rightarrow -\frac{1}{3} < \frac{k_3}{k_1}m_1 < 0$, and, similarly: $-\frac{1}{3} < \frac{k_4}{k_2}m_2 < 0$. A feasible range exists for h_2h_3's that allows the bracketed term to be positive. There are other negative terms in the entire expression, which could result in $\mathbf{m}^T\mathbf{H_g}\mathbf{m} > 0$ even in these two Cases. This is why Case 1 conditions are also not necessary; however they do guarantee concavity. Either constraint in the proposition rules out Case 2. □

Appendix B: Grade Function Estimation Procedure

Note from (A) that the grade function $g(\mathbf{m})$ is linear in the parameters k. Further, only five of the eight k's are independent. Treat the observed grade x as the dependent variable, and $m_1, m_2, m_1^2, m_2^2, m_1m_2$ as five explanatory variables. The observational data over h candidates can be represented as: $X = Mk$, where $X_{(h\times 1)}$ is the vector of observations, $M_{(h\times 5)}$ is the matrix with the five columns computed as above from the graded candidates, and $k_{(5\times 1)}$ is the vector of the coefficients.

The sum of squared errors is: $e(k) = (X - Mk)^T(X - Mk) = X^TX - 2X^TMk + k^TM^TMk$. There are additional constraints to be observed on k's, as derived in Appendix A. A constrained least-squares procedure is specified as the following quadratic program:

$$\begin{aligned} \min \quad & -X^TMk + \frac{1}{2}k^TM^TMk \\ \text{s.t.} \quad & A^Tk \geq k_0, \end{aligned}$$

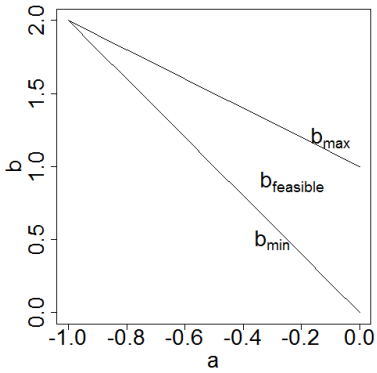


Figure 4 Feasible values for a and b for value functions for individual metrics

where:

$$A^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ -\frac{1}{2} & 0 & 1 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & 1 & 0 \end{bmatrix}$$

and k_0 is vector of seven ϵ's (small positive constant), thus forcing strict inequalities as desired by the constraints.

Appendix C: Airlines' "True" Coefficients for Quadratic Grade Functions

We fix the coefficients for each airline's value function, following the intuition developed in Section 4.3.

C.1. Coefficients for Individual Value Functions

The quadratic value function is: $\nu(m) = am^2 + bm$, where $a \in [-1, 0)$, and $b \in [-2a, 1-a]$ are the coefficients to be fixed. At the higher values of a, that is, near zero, the shape of the value function is similar to a straight line with slope b. On the other hand, the lower end of a's range provides a more concave curvature. Fig 4 shows the feasible values of b over the range of a. Note that b has a larger feasible range at higher values of a. The lower end of a's range allows a much smaller flexibility in choice of b; indeed, at $a = -1, b = 2$.

The highest possible value obtainable by the airline from a metric (i.e. at $m = 1$), is $a + b$. Hence, an airline with profile "HL" would have $a + b$ of the former metric higher than that of the latter. For an "H"-profile metric, high a would yield a straight-line like value function, while low a would yield a more concave one. On the other hand, since low $a + b$ allows only higher values for a, an "L"-profile metric will be straighter. Once $a + b$ is fixed, only one of the two coefficients has to be chosen, say a.

Three ranges within the support of $a + b$ are defined thus: $\{L : (0, {}^1/_2], S : [{}^1/_3, {}^2/_3], H : ({}^1/_2, 1]\}$. Following are repeated for each airline and metric. First, $a + b$ is drawn randomly from the designated ranges in

accordance with the airline-metric profile. Next, a is drawn according to the relative number of operations of the airline, such that larger operations imply smaller a. We employ a acceptance sampling based approach for achieving this, described below and presented in Algorithm 2. This approach accounts for likely errors in our hypotheses, allowing some airlines to have different preference structures than what we posited. Finally, b is computed, and if not feasible, a is drawn again until a feasible b is found. We summarize this procedure in Algorithm 3.

The acceptance sampling algorithm for drawing values of a takes as input the vector of airline-wise operations A_{orig}, the index i_{orig} of the focal airline whose number of operations are reported as i_{orig}-th entry in A_{orig}, and $num.draws$ for number of draws to return for the focal airline. A_{orig} is sorted, and new position of the i_{orig}-th airline is identified – stored as A and i respectively. If there are multiple airlines with exactly same number of operations, any one of those could be designated as i, as the procedure treats similarly sized airlines in a similar fashion.

A proposal probability distribution from which random variables will be drawn is specified as uniform (0,1), such that each draw has mapping onto the desired coefficient a. In this case, $a = -v$. A proposed draw v for the ith airline will be accepted if it falls within its "valid range". If the ith airline has a unique value for number of operations, then its valid range is the width of the ith interval. If multiple airlines have the same number of operations, then the valid range extends to the width of these contiguous intervals. Thus, the ordering of airlines with same number of operations does not matter – which is desirable, as the sorting order for such airlines would be arbitrary.

We wish to allow some probability of accepting a v that happens to fall outside its valid range. Following scheme is adopted. Another iid random variable r is next drawn. v is accepted if r falls in the valid range. Thus, we accept v if either v or r fall within the valid range. Note that the valid range for r need not be the same as that of v; a different range could be used for fine-tuning the acceptance probabilities.

We show the simulation results for a hypothetical set of airline operations: $A = \{1, 1, 4, 4, 4, 7, 9, 10\}$. The first two airlines should predominantly have higher a, followed by the next three, and so on. The last airline should have predominantly lower values of a. We make 1000 draws and plot the histogram in Fig 5. The results are clearly as desired.

C.2. Coefficients for Integration of Individual Value Functions

The integration rule states: $V = r_1\nu_1(m_1) + r_2\nu_2(m_2) + r_{12}\nu_1(m_1)\nu_2(m_2)$. Recall that r's are all non-negative by assumption. That is, the interaction between the two metrics can not *decrease* the overall value to an

Algorithm 2 Acceptance sampling algorithm for drawing a values

sample.a($A_{orig}, i_{orig}, num.draws$)
$A \leftarrow sort(A_{orig}); i \leftarrow \min\{k|A[k] = A_{orig}[i_{orig}]\}$ {sort and identify new position of i_{orig}}
$n \leftarrow |A|; Acc \leftarrow \{\}$ {initialize}
{compute range for valid draws}
$j \leftarrow \min\{k|A[k] = A[i]\}; t_{min} \leftarrow \frac{j-1}{n}$
$j \leftarrow \max\{k|A[k] = A[i]\}; t_{max} \leftarrow \frac{j}{n}$
{make draws}
for $iter \in \{1, \ldots, num.draws\}$ **do**
  while true do
    $v \sim unif(0,1)$ {draw (negative) value for a}
    $r \sim unif(0,1)$ {draw whether to accept v or reject it}
    if $\Big(v \in \{t_{min}, \ldots, t_{max}\}\Big)$**or**$\Big((v \notin \{t_{min}, \ldots, t_{max}\})$**and**$(r \in \{t_{min}, \ldots, t_{max}\})\Big)$ **then**
      $Acc \leftarrow Acc \cup \{-v\}$; break {accept and break out of while loop}
    end if
  end while
end for
return Acc

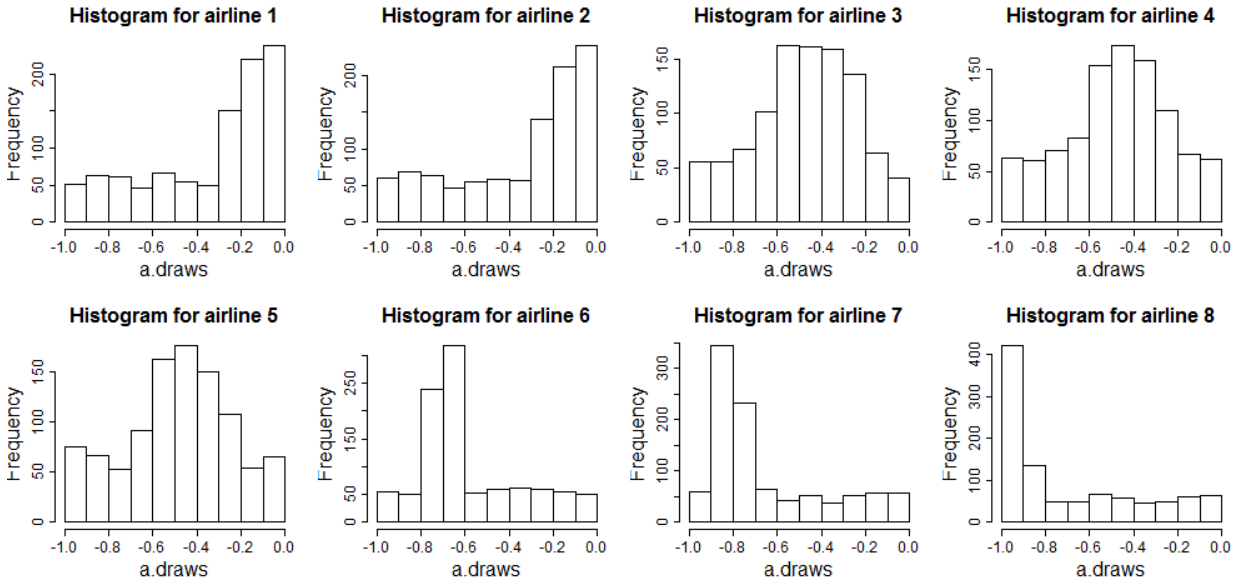


Figure 5 **Acceptance sampling results for hypothetical data of airline operations**

airline. If the value derived from the two metrics are independent of each other to the airline, $r_{12} \to 0$.

r_1, r_2 have to be fixed with respect to the profile for the metric. As these will finally be normalized by V^{max} for each airline, the same positive range can be used for all the airlines without any loss in generality. The following ranges are used: $\{L : [2,4], S : [2,6], H : [4,6)\}$. The interaction effect is constrained to be smaller than the major effects, hence the range for r_{12} is taken as: $[0,2]$.

To ensure global concavity, the drawn values for a, b, r for each airline have to meet the necessary and sufficient condition over the support of (m_1, m_2), as shown in Appendix A:

$$\mathbf{m}^T \mathbf{H_g} \mathbf{m} = 2\left[k_3 m_1^2 + k_4 m_2^2\right] + 2k_5 m_1 m_2 \left[1 + 3\frac{k_3}{k_1} m_1 + 3\frac{k_4}{k_2} m_2 + 6\frac{k_3}{k_1}\frac{k_4}{k_2} m_1 m_2\right] < 0, \quad \text{where}$$

$$k_1 = r_1 b_1 \frac{1}{V^{max}}; k_2 = r_2 b_2 \frac{1}{V^{max}}; k_3 = r_1 a_1 \frac{1}{V^{max}}; k_4 = r_2 a_2 \frac{1}{V^{max}}; k_5 = \frac{r_{12}}{b_1 b_2} \frac{1}{V^{max}}. \quad (7)$$

Since V^{max} is positive by assumption, it has no role in determining the curvature of the grade function. For ensuring concavity, we need to test that the necessary condition below is met at several sample points over the unit square of the metrics:

$$nec(m_1, m_2) = \left[r_1 a_1 m_1^2 + r_2 a_2 m_2^2\right] + \frac{r_{12}}{b_1 b_2} m_1 m_2 \left[1 + 3\frac{a_1}{b_1} m_1 + 3\frac{a_2}{b_2} m_2 + 6\frac{a_1}{b_1}\frac{a_2}{b_2} m_1 m_2\right] < 0.$$

Treating $V^{max} = 1$, un-normalized k's are computed using (7). The LP corresponding to **Subset_Opt**(b) is solved (with individual airline as input) to determine the airline's grade-maximizing candidate. The associated optimal solution is V^{max} for the airline. This is then used to normalize the k coefficients using (7).

C.3. Overall Procedure

The overall algorithm for making the draws is now presented in Algorithm 3. The coefficients a, b and r thus drawn are shown in the left panel of Table 4 – a_1, b_1 are the a, b coefficients for m_1, while a_2, b_2 are for m_2. The grade maximizing candidate and V^{max} for each airline are shown in the middle panel of Table 4. Finally, the normalized k coefficients for each airline are in the right panel. Only $k_1, \cdots, k_5$ are shown, the other three can be directly computed using these five.

Algorithm 3 Algorithm for drawing a and b values

```
gen.true.abr()
for all airlines in A do
  repeat
    for all metrics do
      lookup profile P for the given metric and airline
      repeat
        if P = "H" then
          a.plus.b ~ unif(1/2, 1))
          r ~ unif(4, 6)
        else if P = "L" then
          a.plus.b ~ unif(0, 1/2))
          r ~ unif(2, 4)
        else if P = "S" then
          a.plus.b ~ unif(1/3, 2/3))
          r ~ unif(2, 6)
        end if
        a ← sample.a(A, i, 1)
        b ← a.plus.b − a
      until b ∈ {−2a, ..., 1 − a}
    end for
    r12 ~ unif(0, 2)
  until necessary condition for concavity met at each of several sample (m1, m2) points
  determine grade-maximizing candidate and associated V^max
  normalize coefficients using equations 7
end for
```

Airline	a_1	b_1	a_2	b_2	r_1	r_2	r_{12}	m_1^{max}	m_2^{max}	V^{max}	k_1	k_2	k_3	k_4	k_5
United	-0.55	1.20	-0.06	0.44	5.84	2.79	0.02	0.9500	0.8168	4.67	1.50	0.26	-0.69	-0.03	0.00
American	-0.41	1.04	-0.27	0.70	5.78	3.12	0.30	0.9679	0.7930	4.90	1.23	0.44	-0.48	-0.17	0.04
Southwest	-0.11	0.49	-0.82	1.82	2.69	5.06	1.72	0.8639	0.8906	6.35	0.21	1.45	-0.05	-0.66	0.24
Northwest	-0.45	1.03	-0.25	0.66	3.30	3.10	0.46	0.9172	0.8503	3.15	1.08	0.65	-0.47	-0.24	0.10
Delta	0.00	0.36	-0.59	1.21	5.96	3.86	0.46	0.9829	0.7668	4.43	0.48	1.06	-0.01	-0.51	0.04
US Air	-0.13	0.48	-0.15	0.74	3.55	5.95	0.44	0.7451	0.9506	4.46	0.38	0.99	-0.11	-0.20	0.03
Continental	-0.30	0.88	-0.47	0.97	5.78	3.48	0.94	0.9829	0.7668	5.22	0.97	0.65	-0.33	-0.31	0.15
Airtran	-0.24	0.63	-0.28	1.26	3.37	4.02	1.37	0.7547	0.9468	5.39	0.40	0.94	-0.15	-0.21	0.20
Air Canada	-0.06	0.31	-0.82	1.81	2.62	5.02	1.84	0.8094	0.9219	5.79	0.14	1.57	-0.03	-0.71	0.18
ExpressJet	-0.30	0.77	-0.01	0.87	3.15	5.00	0.50	0.6708	0.9752	5.58	0.43	0.78	-0.17	-0.01	0.06
Jetblue	-0.03	0.11	-0.45	1.19	2.63	5.13	0.18	0.5767	0.9946	3.91	0.07	1.56	-0.02	-0.59	0.01
Chautauqua	-0.15	0.64	-0.71	1.71	2.37	5.42	0.70	0.8045	0.9244	6.55	0.23	1.42	-0.05	-0.59	0.12
Frontier	-0.22	0.59	-0.55	1.47	3.74	5.06	1.73	0.7896	0.9316	6.26	0.35	1.19	-0.13	-0.44	0.24
Mexicana	-0.07	0.29	-0.76	1.56	2.06	5.29	0.02	0.7466	0.9500	4.61	0.13	1.79	-0.03	-0.87	0.00
Lufthansa	-0.69	1.57	-0.17	0.55	5.36	3.00	0.32	0.9501	0.8167	5.80	1.45	0.29	-0.63	-0.09	0.05
Primaris	-0.67	1.58	-0.26	0.57	4.33	3.83	1.86	0.9580	0.8069	5.52	1.24	0.40	-0.53	-0.18	0.30
Alaska	-0.34	0.75	-0.62	1.55	3.22	5.27	1.00	0.7518	0.9480	6.34	0.38	1.29	-0.17	-0.51	0.18
Air Midwest	-0.30	0.98	-0.42	0.91	5.40	3.90	0.06	0.9829	0.7668	5.42	0.98	0.65	-0.30	-0.30	0.01
Aeromexico	-0.25	0.68	-0.61	1.49	2.11	5.93	1.70	0.7552	0.9466	6.44	0.22	1.37	-0.08	-0.56	0.27
British Airways	-0.56	1.19	-0.17	0.43	5.32	3.01	1.72	0.9501	0.8167	4.32	1.47	0.30	-0.69	-0.12	0.21
Polar Air Cargo	-0.58	1.20	-0.29	0.70	2.27	2.84	0.27	0.8688	0.8873	2.56	1.06	0.78	-0.51	-0.32	0.09
Spirit	-0.44	0.93	-0.49	1.00	3.77	5.42	1.23	0.9174	0.8501	4.79	0.73	1.13	-0.35	-0.56	0.24
Aer Lingus	-0.43	1.03	-0.43	1.06	5.27	2.25	0.57	0.9502	0.8166	4.61	1.18	0.52	-0.50	-0.21	0.14
Air Canada Jazz	-0.26	0.91	-0.28	0.76	2.96	3.33	0.59	0.9508	0.8159	3.50	0.77	0.73	-0.22	-0.26	0.12
Lot - Polish	-0.21	0.84	-0.12	0.54	4.08	4.44	0.22	0.9503	0.8164	4.17	0.82	0.58	-0.20	-0.13	0.02
SAS Scandinavian	-0.37	0.88	-0.33	0.72	2.26	2.75	0.73	0.9172	0.8502	2.28	0.87	0.87	-0.37	-0.40	0.20
Singapore	-0.49	1.06	-0.44	0.92	3.06	5.12	0.32	0.9134	0.8536	4.17	0.78	1.13	-0.36	-0.54	0.07
USA 3000	-0.45	1.35	-0.31	0.75	5.13	2.54	0.97	0.9900	0.7494	5.95	1.17	0.32	-0.39	-0.13	0.17
Air France	-0.21	0.80	-0.37	0.87	3.09	5.45	0.17	0.9184	0.8492	4.31	0.57	1.10	-0.15	-0.47	0.03
Air India	-0.23	0.47	-0.25	0.93	2.27	5.85	0.87	0.6411	0.9828	4.50	0.24	1.20	-0.11	-0.33	0.08
Air Jamaica	-0.35	0.86	-0.29	0.84	2.55	4.45	0.55	0.8391	0.9058	3.65	0.60	1.02	-0.25	-0.36	0.11
Alitalia	-0.37	0.78	-0.27	0.72	5.42	5.58	0.92	0.8500	0.8993	4.71	0.89	0.85	-0.42	-0.32	0.11
All Nippon	-0.01	0.57	-0.37	0.96	2.01	3.04	0.16	0.9332	0.8351	2.74	0.42	1.06	-0.01	-0.41	0.03
British Midland	-0.43	0.93	-0.01	0.62	5.77	4.93	0.76	0.7995	0.9268	5.71	0.94	0.53	-0.43	-0.01	0.08
Iberia	-0.31	0.93	-0.40	0.83	3.60	4.68	0.98	0.9530	0.8131	4.36	0.77	0.89	-0.26	-0.43	0.17
Japan Int'l	-0.35	0.98	-0.10	0.65	5.36	2.49	0.07	0.9530	0.8131	4.44	1.18	0.36	-0.42	-0.06	0.01
KLM-Royal Dutch	-0.30	0.91	-0.36	0.75	3.81	2.36	1.24	0.9879	0.7555	3.43	1.01	0.52	-0.33	-0.25	0.25
Korean	-0.49	1.12	-0.18	0.59	4.89	5.01	0.54	0.9035	0.8619	4.99	1.09	0.59	-0.48	-0.18	0.07
Martinair Holland	-0.22	0.82	-0.07	0.50	2.15	3.38	0.04	0.8837	0.8770	2.50	0.71	0.68	-0.19	-0.10	0.01
Pakistan Int'l	-0.42	0.88	-0.52	1.20	2.04	4.26	1.56	0.8144	0.9194	4.14	0.43	1.23	-0.21	-0.54	0.40
Swiss	-0.26	0.74	-0.12	0.54	5.61	4.80	0.58	0.9233	0.8447	4.44	0.93	0.58	-0.32	-0.14	0.05
Turkish	-0.29	0.92	-0.13	0.76	4.46	5.31	0.30	0.8540	0.8969	5.72	0.72	0.70	-0.23	-0.12	0.04
Virgin Atlantic	-0.21	0.80	-0.17	0.62	2.85	5.53	0.21	0.8738	0.8840	3.87	0.59	0.89	-0.16	-0.24	0.03
ABX	-0.37	0.80	-0.39	0.88	4.59	3.98	0.46	0.9084	0.8578	3.92	0.94	0.90	-0.43	-0.39	0.08
Cargoitalia	-0.16	0.65	-0.25	0.83	4.38	4.69	0.06	0.9174	0.8501	4.54	0.63	0.86	-0.15	-0.26	0.01
Custom Air	-0.12	0.67	-0.33	0.67	2.13	4.50	0.23	0.9630	0.8002	2.64	0.54	1.15	-0.10	-0.57	0.04
Kalitta	-0.37	0.98	-0.31	0.87	4.06	5.63	1.75	0.8886	0.8734	5.85	0.68	0.84	-0.26	-0.30	0.26

Table 4 Table with the draws of a, b, r, grade-maximizing candidate and its value, and the normalized k coefficients

APPENDIX II

COuNSEL Benefits Assessment

Cynthia Barnhart and Chiwei Yan, Department of Civil and Environmental Engineering, MIT

Vikrant Vaze, Thayer School of Engineering, Dartmouth College

1. Introduction

FAA traffic managers consult with airline/flight operator operational personnel at both the local and national levels in planning operational strategies for the day. These take the form of strategic planning teleconferences (SPTs). While the SPTs perform a very legitimate and vital function in the overall traffic management process, there are several concerns and issues related to SPTs and more generally, to strategic planning on the day-of-operations. Due to the freeform and unstructured nature of the SPTs, at times, an inordinate amount of time can be devoted to non-critical topics. Furthermore, because priorities are not assigned to the various flight operators based on objective measures, often the more persistent and/or "loudest" flight operators have the most influence. To solve these difficulties and to systematize the decision-making process, several research studies (such as Richetta and Odoni, 1993; Mukherjee and Hansen, 2009, among others) have focused on a centralized decision-making paradigm wherein the objective is to minimize some measure of total cost to the system. Airline preferences, however, are usually not taken into account explicitly in these studies.

Our proposed framework for setting service level expectations, COuNSEL, provides a distributed mechanism to set service level expectations. Under COuNSEL, the FAA will explicitly collect the individual airline preferences in a standardized format and then derive a consensus service level expectations vector. The FAA will then design the specific traffic management initiatives, such as the Ground Delay Programs (GDPs), according to these expectation levels. A number of benefits are expected of this proposed framework. The most obvious benefit will be a reduction in the significant time expended on SPTs. Other benefits, including better overall flight operator performance leading to an overall reduction in flight operator costs, and more equitable strategies and traffic management initiatives, could also be achieved through COuNSEL. In this chapter, using Ground Delay Program (GDP) at San Francisco International Airport (SFO) as an example, we provide a rigorous assessment of the potential benefits of COuNSEL compared to the state-of-the-practice and the state-of-the-research in air traffic flow management.

2. Evaluation Framework

There are two important building blocks in our benefits assessment system: 1) understanding how airlines' preferences differ with characteristics of ground delay programs; 2) assessing NAS-wide performance under different GDP designs. To achieve these two objectives and facilitate our evaluation, we have built an integrated simulation platform to mimic FAA – airline interactions during a GDP day. Given different GDP designs, our simulator can reveal the associated costs for all stakeholders (airlines, passengers, FAA) and assess NAS-wide performance. Through this machinery, we are able to see what

kinds of impact different airlines might incur under a particular GDP design. In section 2.1, we first give an overview of each component's functionality in our simulator and describe the main evaluation procedures. Following that, in section 2.2, we detail a key component in our simulator—the airline recovery module—and describe our underlying assumptions, models and solution approaches.

2.1 Overview

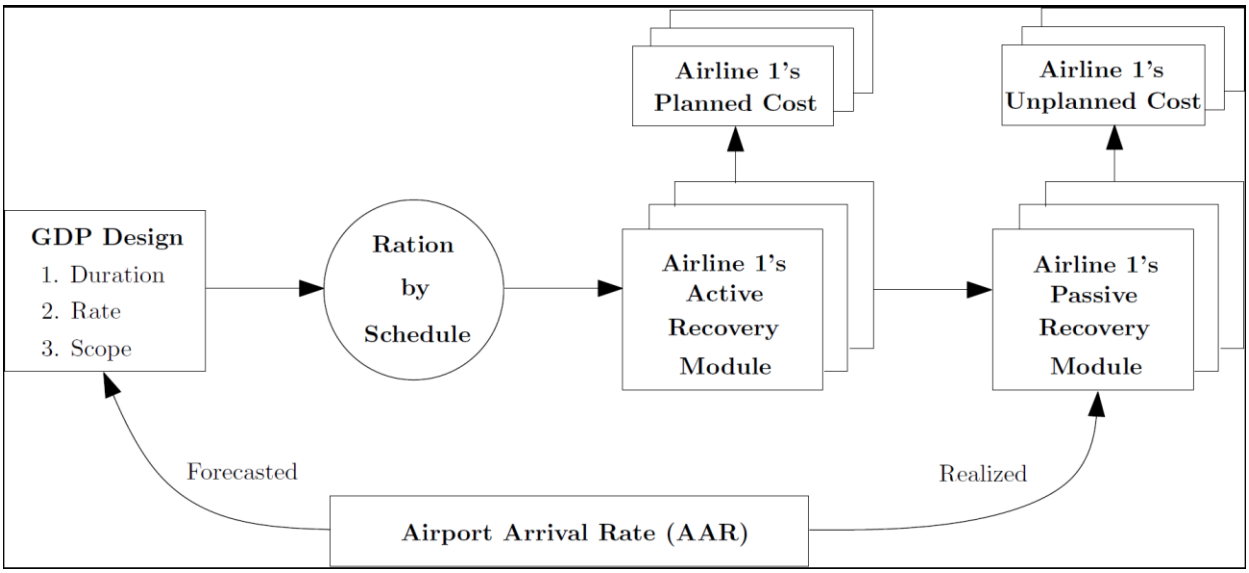


Figure 1: Evaluation Flow Chart

Figure 1 briefly depicts our main evaluation procedures. At the beginning of a GDP, FAA gets some weather forecast to facilitate their GDP design. Based on the forecast, FAA may design an *aggressive* GDP by setting shorter duration, higher rate, thus maximizing overall throughput but potentially inducing higher airborne delay and less predictable delay information to airlines; on the other hand, FAA can also choose to be *conservative* (which is what they usually do in practice) by setting longer duration, lower rate, thus providing more accurate delay information to airlines, reducing airborne delay but potentially losing throughput. Based on their GDP design, FAA runs Ration-by-Schedule (RBS) algorithms to inform airlines about all the delay information. Then each airline runs their recovery module, marked as "*active recovery module*" in figure 1, trying to reduce adverse impacts through recovery operations. This will cause them to adjust their schedules and re-route their fleet. Unfortunately, due to capacity uncertainty, these recovery operations are not the end of the story. At the time when weather conditions finally realize, some of the planned recovery operations may again be infeasible. For instance, if it turns out that FAA underestimates the bad weather impact (usually the case when the GDP design is aggressive), then some of the flights will have additional airborne delay. This may cause fleet and passenger connections, which were originally feasible in the recovery plan, to become disrupted again. In this case, airlines need some additional recovery tools to get their schedules back on track, which we call the "*passive recovery module*". We call it "passive" because such re-disruptions caused by inaccurate delay information often require urgent fixes. Airlines usually don't have enough time to come up with a

sophisticated recovery plan. Such plans need to be simple in nature thus are often less effective compared to the recovery plans designed by the active recovery module. Due to this reason, we model the passive recovery module such that it simply propagates all the delay if the aircraft connection in the original recovery plan is going to be disrupted. Also we don't consider ways to avoid additional passenger disruptions. One could simply regard this passive module as a "simple recourse", which is a common notion in stochastic programming literatures.

We denote the total delay cost coming out of the active recovery module, based on the delay information provided by FAA, as *planned total cost*; the additional delay cost coming out of the passive recovery module, based on the realized delay information, as *unplanned total cost*. They sum up to the *realized total cost* airlines incur due to the GDP. Note that unplanned total cost value might be negative in the case when FAA overestimates GDP impact. In that case, some of the flights may depart earlier than originally controlled by FAA. Thus the unplanned total cost captures the amount of cost reduction brought about by this cancellation. We will use these cost notions extensively in Section 3 where we provide our evaluation results.

2.2 Active Recovery Module: An Integrated Schedule, Route and Passenger Recovery Model

Given delay information generated by Ration-by-Schedule (RBS), the integrated schedule, route and passenger recovery model (SRPRM), formulated in the next section, integrates schedule recovery, fleet re-routing and passenger re-accommodation decisions. In particular, schedule recovery decisions include cancelling flights and re-timing flights through delay and arrival slot substitution. Fleet re-routing decisions mitigate delay through swapping aircraft within the same fleet family. Passenger recovery decisions are captured through explicit modeling of passenger delays and disruptions. We want to point out that due to data unavailability, crew recovery is not explicitly considered in our model. However, by only allowing fleet to be substituted within the same family, we achieve crew compatibility automatically.

2.2.1 Modeling Approach

Instead of a leg-based model, we utilize flight strings, a concept introduced by Barnhart et al. (1998). A flight string is a sequence of flights, with timing decisions, to be operated by the same aircraft. The same sequence of flights might be present in multiple strings, although each sequence must have a unique set of retiming decisions. A string-based model has a number of advantages. Strings are able to capture network effects that individual flight decisions do not. Although the number of strings naturally grows significantly with respect to the number of flights, efficient column generation techniques can be employed.

We begin our detailed model description with the following notation.

Sets:

F: set of flights, indexed by f;
T: set of tails, indexed by t;
$S(t)$: set of flight strings available to tail $t \in T$, indexed by s;

A: set of arrival slots, indexed by a;
P: set of all passenger itineraries, indexed by p;
P^2: set of passenger itineraries containing two flight legs, indexed by p;
$F(p)$: set of all flight legs in itinerary $p \in P$;

Note that in our model formulation, we assume a passenger itinerary cannot contain more than two flight legs. This assumption is justified by Barnhart et al. (2011) where they showed more than 98.5% passengers flew on itineraries with no more than two flight legs in year 2007.

Parameters:

$c_{t,s}$: aircraft delay cost of assigning tail $t \in T$ to string $s \in S(t)$;
n_p: number of passengers in itinerary $p \in P$;
d_p: disruption cost per passenger in itinerary $p \in P$;
c_{pax}: delay cost per passenger minute;
t_{min}^{CT}: minimum passenger connection time;
t_p^{STA}: scheduled time of arrival in itinerary $p \in P$;
t_{fs}^{dep}: departure time of flight f on flight string s;
t_{fs}^{arr}: arrival time of flight f on flight string s;
$\overline{f}(p)$: initial flight leg in itinerary $p \in P$;
$\underline{f}(p)$: final flight leg in itinerary $p \in P$;

Decision Variables:

$$z_f = \begin{cases} 1 & \text{if flight } f \text{ is cancelled,} \\ 0 & \text{otherwise,} \end{cases}$$
$$x_{ts} = \begin{cases} 1 & \text{if tail } t \in T \text{ is assigned to flight string } s \in S, \\ 0 & \text{otherwise,} \end{cases}$$
$$\lambda_p = \begin{cases} 1 & \text{if itinerary } p \in P \text{ is disrupted,} \\ 0 & \text{otherwise,} \end{cases}$$

t_p: delay minutes in itinerary $p \in P$.

Model Formulation:

The SRPRM formulation is given as follows:

$$\min \sum_{t\in T} \sum_{s\in S(t)} c_{t,s} x_{t,s} + \sum_{p\in P} \left(d_p n_p \lambda_p + c_{pax} n_p t_p \right) \quad (1)$$

s.t.:

$$\sum_{t\in T} \sum_{s\in S(t):\, s\ni f} x_{t,s} + z_f = 1, \forall f \in F \quad (2)$$
$$\sum_{t\in T} \sum_{s\in S(t):\, s\ni a} x_{t,s} \leq 1, \forall a \in A \quad (3)$$
$$\sum_{s\in S(t)} x_{t,s} \leq 1, \forall t \in T \quad (4)$$
$$\lambda_p \geq z_f, \forall p \in P, \forall f \in F(p) \quad (5)$$
$$\sum_{t\in T} \sum_{s\in S(t):s\ni \underline{f}(p)} t_{\underline{f}(p),s}^{dep} x_{t,s} - \sum_{t\in T} \sum_{s\in S(t):s\ni \overline{f}(p)} t_{\overline{f}(p),s}^{arr} x_{t,s} \geq t_{min}^{CT} - M\lambda_p, \forall p \in P^2 \quad (6)$$

$t_p \geq \sum_{t \in T} \sum_{s \in S(t): s \ni \underline{f}(p)} t^{arr}_{\underline{f}(p),s} x_{t,s} - t^{STA}_p - M\lambda_p, \forall p \in P$ (7)

$x_{t,s} \in \{0,1\}, \forall t \in T, \forall s \in S(t)$

$z_f \in \{0,1\}, \forall f \in F$

$\lambda_p \in \{0,1\}, \forall p \in P$

$t_p \geq 0, \forall p \in P$

The objective (1) is to minimize the aggregate string assignment cost (aircraft delay cost) and the sum of passenger delay and disruption costs. On the schedule and aircraft side, flight assignment constraints (2) either require a flight to be contained in exactly one string or be cancelled. Slot assignment constraints (3) ensure every arrival slot can be utilized by at most one flight. Similarly, tail assignment constraints (4) restrict a tail to be assigned to at most one flight string. On the passenger side, itineraries with insufficient connection time (constraints (6)) or with one or more canceled flight legs (constraints (5)) are classified as disrupted. Constraints (7) ensure that t_p equals the delay per passenger in minutes on itinerary p if itinerary p is not disrupted; and $t_p = 0$ otherwise.

Underlying Network:

For each fleet family, we construct a connection network $G = (V, A)$, the node set V represents the set of flights and arc set A corresponds to connections between flights. Each flight leg $f \in F$ is represented by a node. Moreover, for each flight f, which is also a controlled flight heading into the GDP impacted airport, we generate its *slot copy nodes*. The timing information for each slot copy node corresponds to one eligible arrival slot to flight f. The arrival time of the slot is no earlier than the scheduled time of arrival of flight f. For each flight that is not a controlled flight in the GDP, we generate *flight delay copy nodes* with 5-minute intervals until a maximum departure delay (and a corresponding arrival delay) of Θ. In our experiment, we let $\Theta = 180$. A connection arc exists from one node i to another node j if the arrival station of i is also the departure station of j and the arrival time of i plus the minimum turn time (30 minutes) is earlier than the departure time of j. Figure 2 illustrates a part of the connection network in which the GDP impacted airport is ORD. Flight 1 (BOS → ORD) is a controlled flight for a GDP that has two slot copy nodes 1' and 1'', each corresponding to an eligible arrival slot time for flight 1. Flight 2 is a non-controlled flight that has one 5-minute flight delay copy node, 2'.

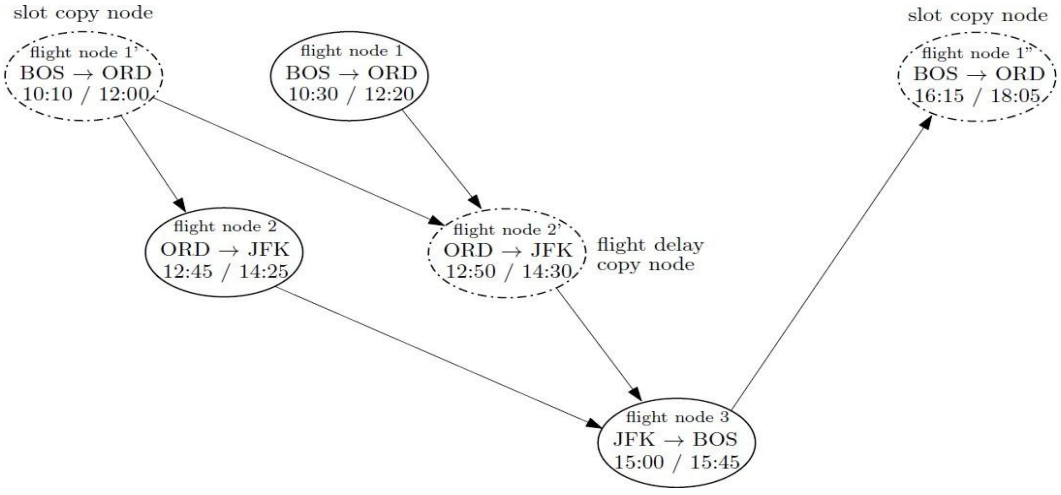


Figure 2: Underlying Connection Network

2.2.2 Solution Approach

Given the large number of flight strings, only a subset of columns is generated in SRPRM. Given a connection network $G = (V, A)$, a dummy source and sink node are added in which a flight string variable corresponds to an $s - t$ path. Paths are constructed by computing the reduced cost for every arc $a \in A$. At each iteration, a path with the most negative reduced cost is generated through the following pricing sub-problem.

Pricing Sub-problem:

The pricing sub-problem is a minimum cost flow problem with an additional side constraint that disallows each feasible path from containing more than one node corresponding to the same flight. As an illustration, in Figure 2, a path cannot be created as flight node 1 → flight node 2' → flight node 3 → flight node 1'', because flight 1 is visited twice along the path. We briefly summarize our pricing algorithm in Algorithm 1.

Algorithm 1: (One-to-One Shortest Path Problem with Node-Visit Restriction)

Initialize the cost of each node to ∞

Initialize the cost of the source to 0

Mark all nodes (except source node) as unknown

WHILE the sink is not marked as known

    Select the unknown node with the lowest cost: n

    Mark n as known

    FOR each node a which is adjacent to n

        IF path to a with n as the penultimate node doesn't contain any node which shares the same flight number as a

            a's cost = min (a's old cost, n's cost + cost of (n, a))

        ENDIF

    ENDFOR

ENDWHILE

The only difference between Algorithm 1 and the classic Dijkstra Algorithm for one-to-one shortest path is as follows. When updating arc cost in each FOR loop, we need to ensure that the path heading to the node a with node n as the penultimate node in the path to be updated doesn't contain any node which shares the same flight number as a. Fortunately, although we have this additional step in each inner iteration, our overall algorithm still has a polynomial complexity, which is shown in the following proposition.

Proposition 1. *Algorithm 1 has polynomial complexity*.

Proof. The step of checking whether the path contains a node which has a specific flight number has complexity $O(|V|)$, where $|V|$ is the number of nodes in the graph. Since Dijkstra's algorithm is a polynomial algorithm, we conclude that Algorithm 1, which only adds a polynomial subroutine to Dijkstra's algorithm, is also polynomial.

Q.E.D.

Integer Solution:

An integer optimal solution to SRPRM can be obtained through a complete branch-and-price routine. Our implementation deploys a heuristic branching rule in the sense that we only explore one particular branch in the B&B tree. Our implementation terminates with a feasible integer solution. A summary of our branch-and-price method is shown in Algorithm 2.

Algorithm 2 (Branch-and-Price with Heuristic Variable Fixing Procedure)

WHILE not reached the maximum number of iterations

Solve the LP relaxation of SRPRM to optimality using column generation

FOR all flight string variables

IF optimal value of the string variable is fractional and greater than or equal to δ

Fix the value of that string variable to 1

ENDIF

ENDFOR

ENDWHILE

Parse the model to a MIP solver to get overall integer solution (string variables, cancellation variables and itinerary variables)

Proposition 2. *Algorithm 2 terminates with a feasible integer solution*.

Proof. We want to show by fixing flight string variables at each iteration, we will not cause infeasibility to the model. We prove this by examining all constraints containing flight string variables. The right hand side of constraints (2), (3) and (4) is 1. Thus at any iteration of algorithm 2, where we get a string variable x whose optimal value is a fractional value, all the other variables in constraints (2), (3) and (4) should not be fixed . So by fixing x to 1, we will not cause any infeasibility in these three set of constraints. Constraints (6) and (7) can always be made feasible by adjusting the value of λ_p, which will not be fixed.

Q.E.D.

The higher the value of δ we choose, the fewer the number of flight strings that will be fixed during the algorithm. So for higher values of δ, our algorithm will be more accurate but computationally more expensive. Thus δ is a parameter controlling the accuracy-speed trade-off. In our implementation, we choose $\delta = 0.6$. We want to point out that our algorithm performs well although it is heuristic in nature. For small instances, we can produce provably optimal solutions, and for medium and large instances, our integer solution is often within 10% of the optimal value of the LP relaxation.

3. Evaluation Results

Using the evaluation framework described in Section 2, we are able to simulate airlines' responses to different GDP designs. In this section, we first discuss in Section 3.1 the details of our experimental setup. Then we present two different types of analyses. In Section 3.2, we highlight how different airlines' preferences differ in prioritizing alternative GDP designs, and provide insights into the underlying operational factors that explain these differences. In Section 3.3, we assess the potential performance benefits to the National Airspace System (NAS) through our decentralized GDP design process, namely COuNSEL. We compare the benefits with the baseline case of a centralized GDP design.

3.1 Experimental Setup

Flight Schedules for a representative day in the summer of 2007 are obtained at San Francisco International Airport (SFO). We assume that an airport operates at its VFR capacity level before a GDP starts; operates at its IFR capacity level during the GDP; and then it returns to its VFR capacity level once the GDP actually ends. The arrival capacities under the VFR and IFR scenarios are obtained from the Airport Capacities Benchmark Report (FAA 2004). We set up 14 hypothetical GDPs. The planned duration of each GDP, the difference between planned start and end times, is varied from 3 through 9.5 hours in steps of one half-hour each (i.e., 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9, and 9.5 hours). We also vary the planned start time of the GDP between 11 am and 5 pm in steps of 1-hour each (i.e., 11am, 12pm, 1pm, 2pm, 3pm, 4pm and 5pm). So we have, in total, 14x7=84 GDP candidates for evaluation. Along with the planned start and end times, each GDP has several possible *actual end times*, that is, the time after which the airport arrival capacity returns to the VFR level. Therefore, for each possible actual end time value, there is a corresponding unique capacity profile. In our setup, each GDP has seven capacity scenarios, corresponding to *actual* GDP durations varying between 3 and 9 in steps of 1-hour each (i.e., 3, 4, 5, 6, 7, 8 and 9 hours). All scenarios are assumed to be equally likely (with probability 1/7) to occur. Although this might not be a valid assumption in many other airports, in the case of SFO it does seem appropriate because marine stratus clouds typically impact capacity. Hence, a GDP with a planned duration of 3-hours is considered to be a highly *aggressive* design in the sense that with high probability (6/7), the airport capacity is over-estimated by the GDP design. On the other hand a GDP, a planned duration of 9.5 hours is considered to be highly *conservative* because with probability 1.0, capacity is under-estimated by the GDP design.

Itinerary data for each airline involved in our evaluation is generated with the algorithm proposed by Barnhart et al. (2011). Itinerary disruption cost is calculated based on the *Passenger Delay Calculator* developed by Bratu (2006). Aircraft delay costs (including airborne and ground delay costs) are estimated separately for different airlines and different fleet types based on Form 41 data (DOT, 2007).

3.2 Revealing Airlines' Preferences

In this section, we highlight how different airlines' have differing preferences in prioritizing alternative GDP designs. We describe the major operational factors that cause these differences. Table 1 provides an overview of some basic statistics about each airline at SFO airport. These include number of operations, fleet composition, number of passengers, percentage of connecting passengers, and the average load factors. Figures 3 through 12 are bar charts describing the total costs (in $) for each airline plotted against the planned GDP duration (in hours) on the x-axis. Each chart consists of 4 data series, namely, total realized cost, total realized cost in the hypothetical scenario assuming no recovery actions, total planned costs and total unplanned costs.

Based on the trend in total realized costs, we categorize the airlines into 3 broad groups: 1) those that prefer an *aggressive (shorter) GDP design*, 2) those that prefer a *moderate (intermediate duration) GDP design*, and 3) those that prefer a *conservative (longer) GDP design*. Note that these categories are not meant to serve as rigid categorizations for these airlines. Instead, they are relevant only for the specific case of the SFO airport, for the day of operations being considered in our experiment, and under the assumptions made in this experiment about the GDP scenarios, capacity distributions, and other parameters. The first category—airlines preferring aggressive GDP designs—includes US Airways, Frontier Airlines, Northwest Airlines, and Continental/ExpressJet. The second category—airlines preferring moderate GDP designs—includes Delta Airlines, American/American Eagle, Alaska Airlines and JetBlue Airways. The third category—airlines preferring conservative GDP designs—includes United/SkyWest and AirTran Airways.

	# Impacted Operations	# Fleet Types (# Aircraft in Each Category)	# Impacted Passengers	% Connecting Passengers	Average Load Factor
United & SkyWest	359	10 (17,4,8,3,9,1,3,7,5,27)	24236	32.33%	75.29%
American & American Eagle	70	5 (4,2,4,3,9)	7678	27.39%	75.53%
US Airways	40	4 (1,4,1,4)	4007	31.57%	80.43%
Continental & ExpressJet	30	5 (1,1,3,1,2)	3244	20.43%	70.15%
Delta Airlines	26	4 (1,1,2,2)	3750	30.29%	80.72%
Alaska Airlines	25	2 (4,3)	2461	9.47%	75.28%
Northwest Airlines	23	4 (2,2,2,1)	3232	25.46%	85.65%
Frontier Airlines	15	2 (2,2)	1351	31.68%	79.35%
JetBlue Airways	9	1 (2)	1180	8.05%	78.46%
AirTran Airways	8	1 (4)	973	32.58%	82.32%

Table 1: Airline Characteristics, for a summer day in year 2007 at SFO airport

Figures 3 through 12 show that the three categories of airlines correspond to three different broad trends in the variation of total realized cost with planned GDP durations. The total realized cost for the airlines in the first category shows a close to monotonic increase in total realized costs (the first data series in the bar charts) as the planned GDP duration increases. Thus, category 1 airlines have a preference for an aggressive, shorter planned GDP duration. The total realized cost for category 2 airlines tend to be comparatively flat and roughly convex, with the point of minimum total realized cost

occurring somewhere in the middle (far from either extreme). Thus, they have a preference for a moderate, intermediate planned GDP duration value. The total realized cost for the category 3 airlines displays a near-monotonic decreasing trend with planned GDP duration, and thus they have a preference for a conservative, longer planned GDP duration.

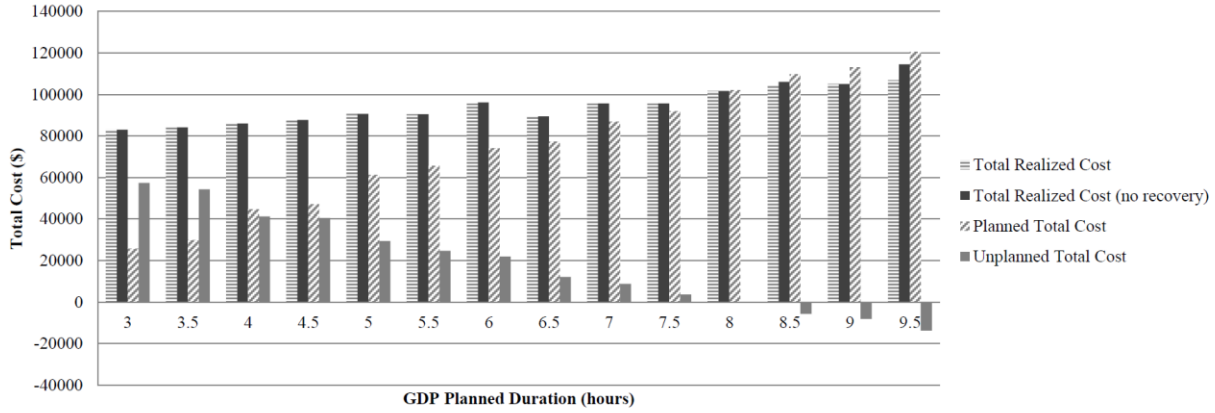


Figure 3: US Airways, Preference for Aggressive Design

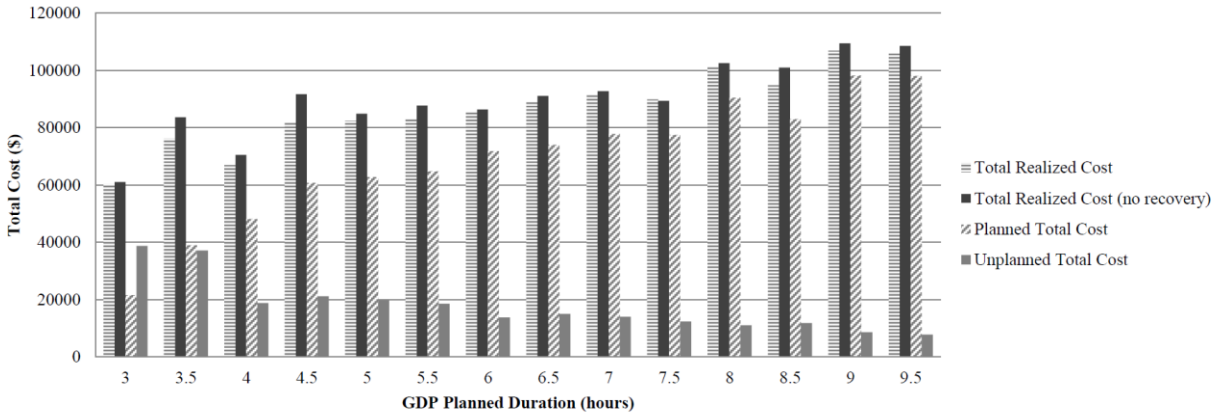


Figure 4: Frontier Airlines, Preference for Aggressive Design

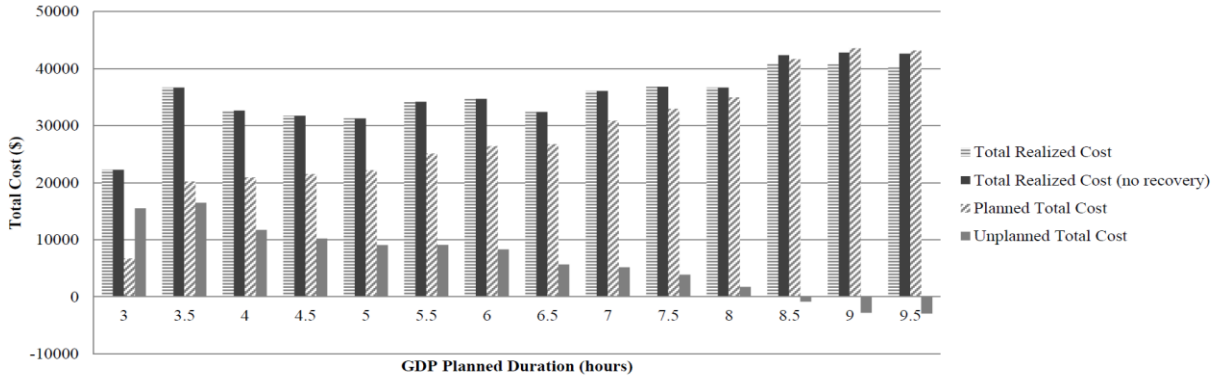


Figure 5: Northwest Airlines, Preference for Aggressive Design

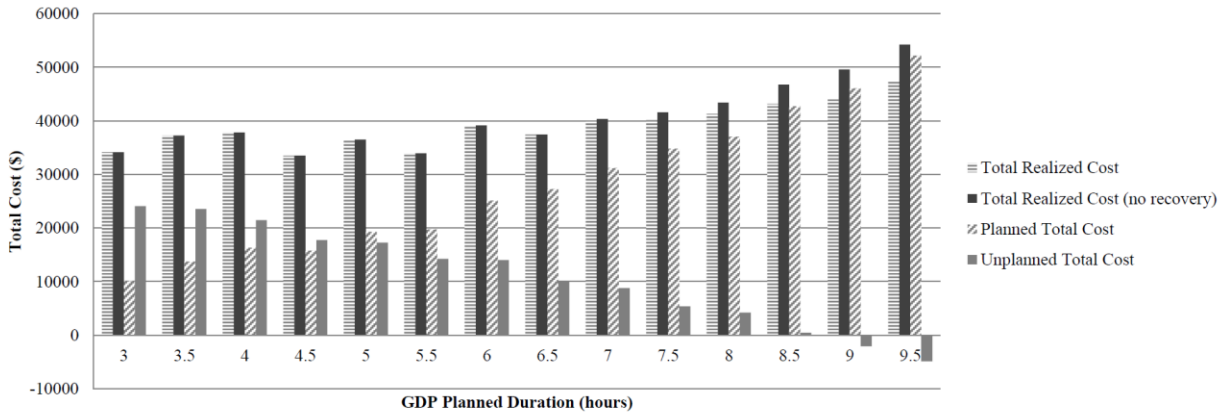


Figure 6: Continental Airlines (and ExpressJet), Preference for Aggressive Design

These differences between airline groups can be explained in several different ways. One prominent difference becomes visible when we focus on the total realized cost in the absence of any recovery actions by the airline. This is the second data series in the bar charts. This data series indicates the total realized cost value for a hypothetical scenario in which the airline takes no recovery actions, and all delays propagate through the network. The longer the planned duration of a GDP, the greater is the extent of disruption to the airline's network and hence, the greater is the total realized cost in the absence of any recovery. In fact, Figures 3 through 12 indicate that the second data series increases almost monotonically with increases in planned GDP duration for all airlines.

When a GDP is announced, however, airlines can plan their recovery actions for the operations that are likely to be affected due to the GDP. The longer the planned GDP duration, the more extensive is the set of possible recovery actions that an airline might take. Therefore, the difference between the first and second data series, indicating the cost reduction realized through the airline's recovery actions, shows a generally increasing trend with increases in the planned GDP duration. Note that some airlines are better than others at taking advantage of advance knowledge of a longer GDP. Such differences are primarily explainable in terms of various operational factors, such as, but not limited to, the number of operations and the consistency of the fleet types for the airline at the impacted airport. As a result of these differences, some airlines are able to reduce their total realized cost considerably through effective recovery actions. This phenomenon appears to be an important determinant in classifying airlines into our three categories.

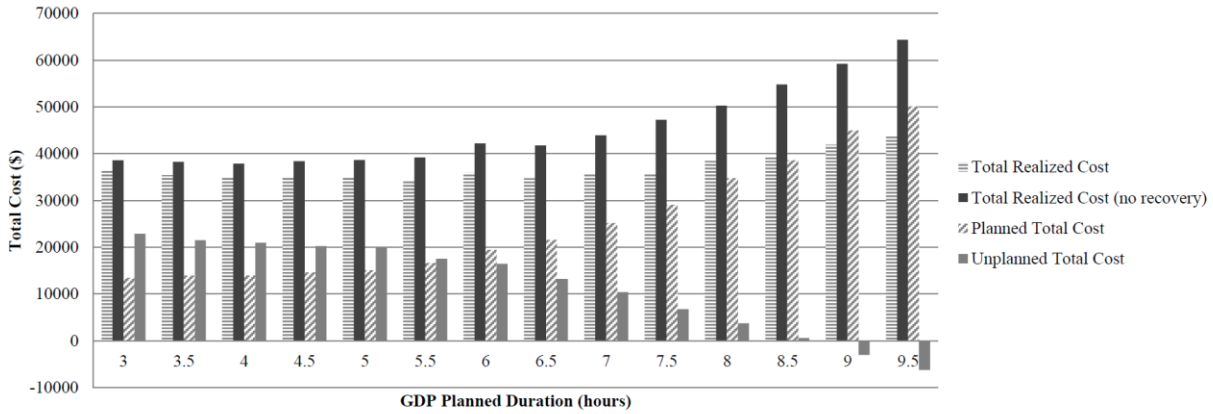


Figure 7: Delta Airlines, Preference for Moderate Design

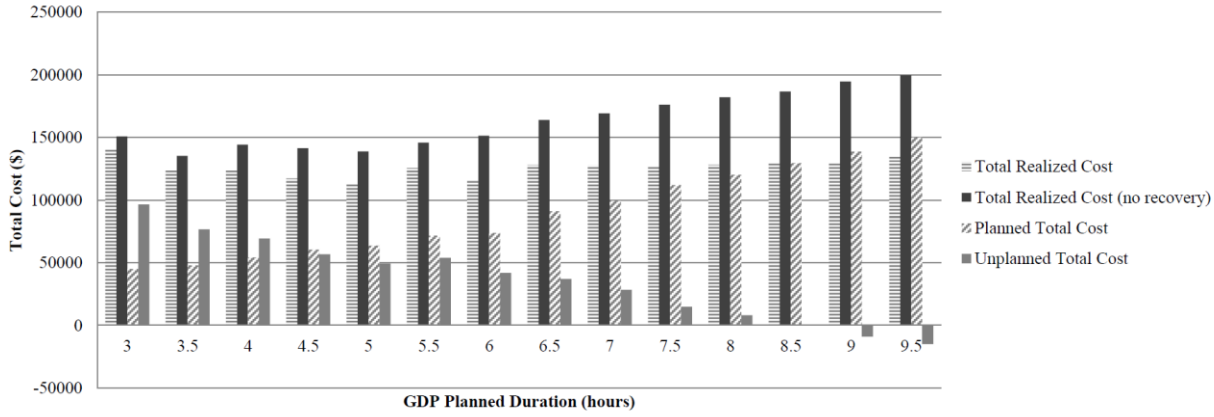


Figure 8: American Airlines (and American Eagle), Preference for Moderate Design

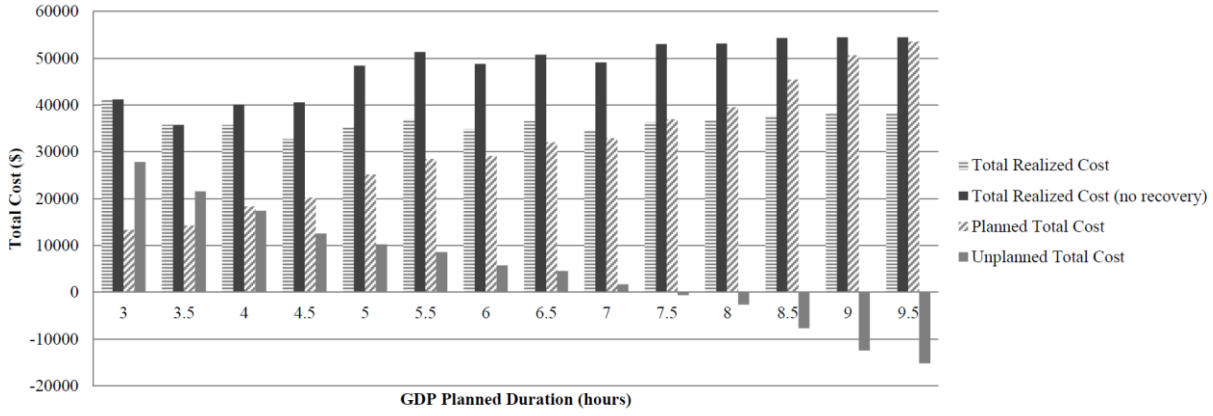


Figure 9: Alaska Airlines, Preference for Moderate Design

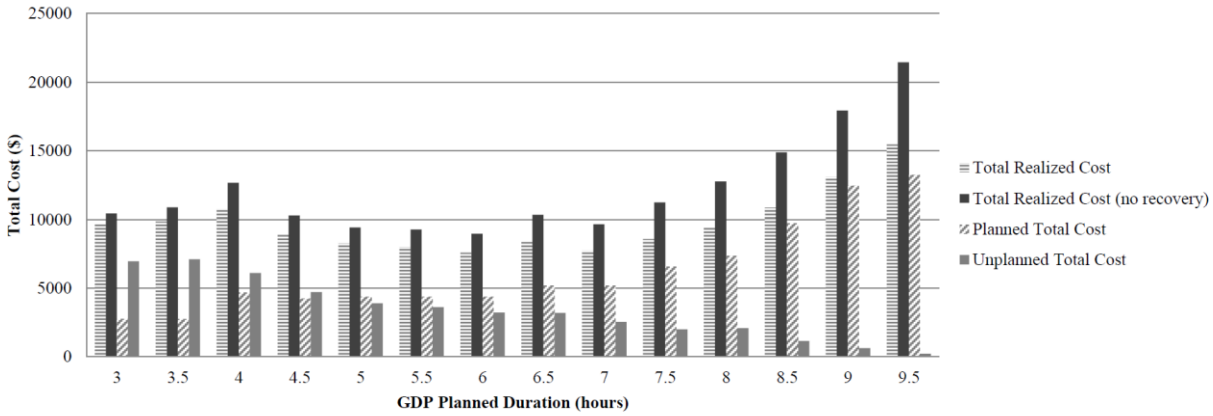


Figure 10: JetBlue Airways, Preference for Moderate Design

For all of the category 1 airlines, across all planned GDP durations, the ratio of total realized cost to total realized cost without recovery does not go below 87%. The recovery potential for these airlines is low even when faced with long planned GDP durations. For all the category 2 airlines, the ratio of total realized cost to total realized cost without recovery is close to 70% for longer GDP durations. This means that these airlines are able to achieve a moderate amount of cost reduction through effective recovery. Finally, for both category 3 airlines, the ratio of total realized cost to total realized cost without recovery is close to (or below) 50% for longer GDP durations. This suggests that these airlines can reduce greatly their realized costs through effective recovery decisions. As a result of these differences, category 1 airlines prefer an aggressive, shorter planned GDP duration because they don't benefit much from recovery decisions when the GDP's are longer. Moreover, with shorter planned durations, there is a chance that the GDP will be lifted as planned and the airline can take full advantage of that. Category 2 airlines prefer a moderate, intermediate planned GDP duration. Category 3 airlines prefer a conservative, longer planned duration primarily because they can choose from a larger set of recovery alternatives and recover a large proportion of the cost if they plan for a longer duration.

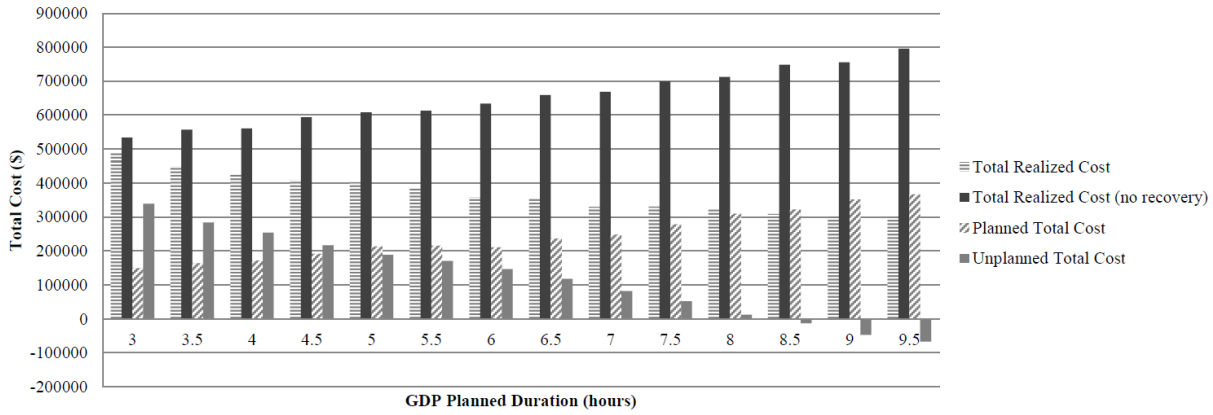


Figure 11: United Airlines (and SkyWest), Preference for Conservative Design

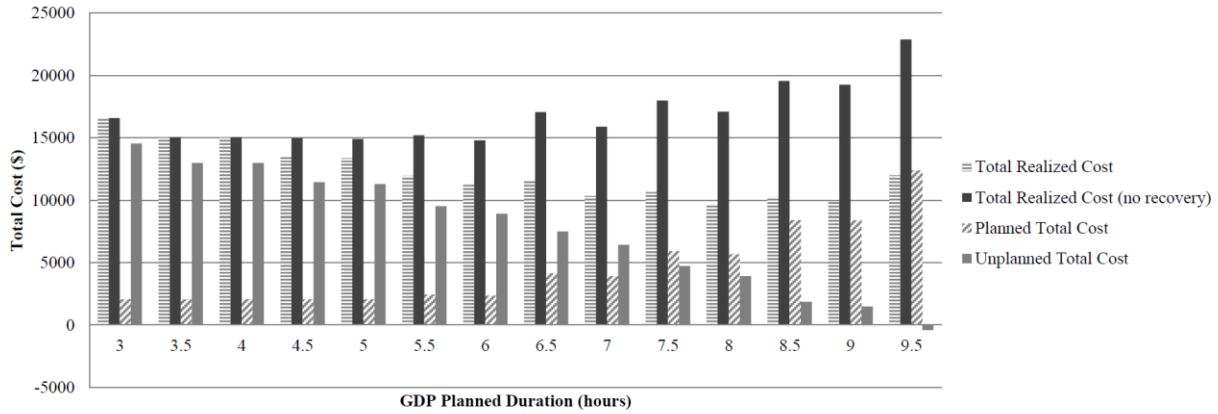


Figure 12: AirTran Airways, Preference for Conservative Design

The third and fourth data series in Figures 3 through 12 show the trends in respectively the planned and the unplanned parts of the total realized costs, plotted against the planned GDP duration. For a particular planned GDP duration, the sum of an airline's planned and unplanned costs equals the airline's total realized cost (the first data series). Note that the planned total costs increase almost monotonically with the planned GDP durations for all airlines. Similarly, the unplanned total costs decrease almost monotonically with the planned GDP durations for all airlines. Variations in the planned cost when compared with the variation in the unplanned costs provide further insights into our airline categorization scheme and the factors leading to the operational differences between different airlines. In particular, the ratio of the difference between the maximum and minimum planned costs to the difference between the maximum and minimum unplanned costs (designated as the *cost ratio*) can provide some insights into the airlines' operational differences.

Comparing the values in Table 1 for the different airline categories provides intuition about why the airlines have such different recovery potential. United/SkyWest Airlines is the airline with by far the largest presence at SFO. It has 359 operations, which is approximately 50% more than the number of operations for the remaining 9 airlines put together. As a result, they have the greatest recovery potential because there are many more opportunities for swapping aircraft, for swapping slots, for rebooking passengers on alternative itineraries, etc. AirTran Airways also has a high recovery potential. Even though it has only 8 operations at SFO, all the 8 operations have the same aircraft type making swapping of aircraft very easy and thus, inducing flexibility in its recovery plans. Also, United/SkyWest and AirTran Airways are the only two airlines that have a cost ratio below 90%. In fact they are much below 90%: 53% for United/SkyWest and 69% for AirTran. This means that for these two airlines, the reduction in unplanned costs is larger compared with the increase in planned costs as the planned GDP duration increases. As a result, these airlines prefer a longer planned duration to ensure that the total cost is minimized. At the other extreme, we have airlines such as Frontier Airlines (247%) and Northwest Airlines (190%), which have the highest cost ratios across all airlines. This means that for these two airlines, the reduction in unplanned costs is smaller compared with the increase in planned costs as the planned GDP duration increases. As a result, these airlines prefer a shorter planned duration to ensure that the total cost is minimized.

3.3 Airport-wise Benefits Assessment

We now add up all the airlines' total realized costs to do an airport-wise performance analysis, in order to evaluate how the National Airspace System, as a whole, performs under different GDP designs. Table 2 summarizes the expected values of the total realized costs for each airline under different GDP designs. The italicized number in each row is the minimum among all the numbers in the same row. Table 2 also provides NAS-wide total cost calculated by summing up all the airlines' total realized costs. The last row in the table lists the *centralized objective value* under different designs. Here centralized objective refers to the *summation of airborne delay and ground delay costs from all the controlled flights heading into GDP airport.* This objective function is used extensively in centralized GDP decision-making approaches such as Richetta and Odoni, 1993, and Mukherjee and Hansen, 2009, just to name a few. The GDP design with the least centralized cost value serves as our *first baseline* which represents the state-of-the-research design. From the analysis of all the GDP advisories issued in year 2007, we find that 94% of the historical GDP advisories are overshot, that is, the planned duration is longer than the actual duration. This means that the practical GDP decision making is often conservative. Thus we believe that it's safe to regard the most conservative GDP design (9.5 hrs planned duration) as our *second baseline* which represents the state-of-the-practice design.

Airline - GDP Cost Matrix	Aggressive Design ← GDP Planned Duration (hours) → Conservative Design													
	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5
American & American Eagle	141340	124389	123490	117142	***112998***	125420	115407	128040	127462	126762	128174	130014	129585	134946
Frontier	***60362***	76148	66946	81898	82396	83363	85580	88988	91783	89850	101507	94825	106871	105891
US Airways	***83058***	84186	85994	87735	90695	90418	96115	89400	95663	95637	101711	104089	105107	106905
Continental & ExpressJet	34152	37247	37844	***33511***	36526	33968	39176	37459	39935	40162	41300	43174	44005	47296
JetBlue	9705	9849	10766	8939	8252	7983	***7577***	8367	7707	8563	9446	10863	13090	15468
Delta	36256	35408	34897	34846	34860	***34132***	35880	34732	35531	35773	38467	39139	41918	43874
AirTran	16600	15049	15050	13499	13363	11954	11280	11651	10338	10645	***9592***	10268	9864	12001
Northwest	***22247***	36705	32657	31738	31265	34185	34704	32411	36074	36831	36690	40855	40764	40228
United & SkyWest	489250	448340	426198	408230	402122	386515	357885	354516	330232	330824	322038	309187	304852	***300218***
Alaska	41167	35758	35713	***32724***	35337	37002	34810	36539	34573	36305	36882	37731	38215	38301
NAS wide	934137	903079	869554	850262	847815	844941	818413	822104	***809297***	811352	825808	820144	834271	845128
Centralized Objective	244986	235343	226604	221638	214614	210056	202624	201951	196292	189613	***188450***	195662	209204	220389

Table 2: Airlines' costs under different GDP designs

We conclude from Table 2 that the GDP design with planned duration of 8 hrs (marked by the blue rectangle) is the most preferred design from the centralized decision making perspective, with the total NAS wide cost $825,808. The most conservative design (9.5 hrs, marked by the red rectangle) has the total cost of $845,128. The system-optimal design, the one has the lowest NAS wide cost, is has a planned duration of 7 hrs (marked by the black rectangle) and with the total cost $809,297, which is 2.0% less than the centralized design and 4.2% less than the most conservative design. We also want to mention that since each minute of airborne delay is usually much more expensive than each minute of ground delay (roughly in a 3:1 ratio), the centralized decision making approach is also very, if not the most, conservative regardless of what kind of airline composition the airport under consideration has.

Now we show how COuNSEL, a decentralized decision making approach, performs under the same setting. COuNSEL uses a voting procedure to allow each participating airline grade all the candidate designs according to their intrinsic value functions. In the next step, a resolution mechanism called *Majority Judgment* is implemented to select the consensus design based on all the airlines' grades and weights. COuNSEL is originally designed to be a multiple-round mechanism where new candidates can be generated during each round. But for simplicity we only evaluate a single-round version in this analysis where the candidates are the 14 designs explained in Section 3.1. In Table 3, we use a linear transformation to convert all the airlines' costs into grades with a maximum grade of 100. Weight of a particular airline is calculated based on the number of impacted operations it has during the GDP. Here we use a power-root transformation to fix the largest airline's (United's) proportion of total weight at 40%. The italicized numbers are the majority grades of each candidate design. The majority winner, that is, the design having the highest majority grade, turns out to be the one with planned duration of 7 hrs (marked by the black rectangle). Interestingly, this coincides with the system optimal design shown in Table 1. We would like to point out that although the majority judgment procedure doesn't necessarily choose a system optimal design, these results do show the ability of COuNSEL to capture airlines' business objectives better than the centralized approach.

Airline - GDP Grade Matrix	# impacted operation	weights	Aggressive Design ← GDP Planned Duration (hours) → Conservative Design													
			3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5
American & American Eagle	70	17.97	80	91	92	96	100	90	98	88	89	89	88	***87***	***87***	***84***
Frontier	15	6.30	100	79	90	74	73	72	71	68	66	67	59	64	56	57
US Airways	40	12.28	100	99	97	95	92	92	86	93	87	87	82	80	79	78
Continental & ExpressJet	30	10.1	98	90	89	100	92	99	86	89	84	83	81	78	76	71
JetBlue	9	4.45	78	***77***	***70***	85	92	95	100	91	98	88	80	70	58	49
Delta	26	9.16	94	96	98	98	98	100	95	98	96	95	***89***	87	81	78
AirTran	8	4.11	58	64	64	71	72	80	85	82	93	***90***	100	93	97	80
Northwest	23	8.43	100	61	68	70	71	65	64	69	62	60	61	54	55	55
United & SkyWest	359	54.63	61	67	70	***74***	***75***	***78***	***84***	***85***	***91***	91	93	97	98	100
Alaska	25	8.92	***79***	92	92	100	93	88	94	90	95	90	89	87	86	85

Table 3: Airlines' grades under different GDP designs and the majority winner

We conclude this section by taking a look at some detailed NAS performance metrics aside from costs. In Table 4, we calculate total ground delay in minutes, total airborne delay in minutes, number of disrupted passengers, and the total passenger delay in minutes under different GDP designs. COuNSEL design compared with the centralized design, reduces total ground delay by 9.8%, total passenger delay by 3.3%. It increases total airborne delay from 214 minutes to 318 minutes. However, on a per flight basis, it is just a 0.51 minutes/flight increase from 0.95 minutes/flight to 1.46 minutes/flight, which is a small absolute increment. We observe even greater benefits when we compare COuNSEL design to the most conservative design. We find a 22.8% reduction in total ground delay, and a 13.7% reduction in total passenger delay, while only inducing an airborne delay of 1.46 minutes/flight. In summary, in this specific case study at the SFO airport, the entire National Airspace System is better-off being operated

under a slightly more aggressive approach than what both the state-of-the-research (centralized) and the state-of-the-practice approaches would suggest.

GDP Planned Duration (hours)	NAS wide Total Cost ($)	Total Ground Delays (minutes)	Total Airborne Delay (minutes)	Average Ground Delay per Flight (minutes)	Average Airborne Delay per Flight (minutes)	# Disrupted Passengers	Total Passenger Delay (minutes)
3	934137	2337	422	10.40	1.87	674	480618
3.5	903079	2835	432	12.62	1.92	725	469141
4	869554	3012	403	13.41	1.79	679	467044
4.5	850262	3269	386	14.56	1.71	681	456964
5	847815	3589	382	15.98	1.70	685	453177
5.5	844941	3857	376	17.18	1.67	674	458074
6	818413	4090	371	18.21	1.65	671	459613
6.5	822104	4429	358	19.72	1.59	673	472420
7	809297	4607	328	20.52	1.46	668	480232
7.5	811352	4748	257	21.14	1.14	678	485759
8	825808	5105	214	22.73	0.95	673	496769
8.5	820144	5546	123	24.70	0.54	646	520816
9	834271	5690	85	25.34	0.37	667	531846
9.5	845128	5970	0	26.59	0	662	556366

Table 4: Overall NAS performance under different GDP designs

4. Conclusion

By providing a systematic way for information exchange between FAA and airlines in setting the service level expectations, COuNSEL is expected to greatly reduce the time expended in unstructured SPTs and improve NAS-wide performance by incorporating airlines' business objectives. In order to rigorously quantify those benefits, especially those in terms of system operating cost reductions, this study built an integrated simulation platform to mimic FAA – airline interaction on the day of operations. By evaluating different GDP designs through multiple rounds of simulations, we reveal airlines' intrinsic value functions towards prioritizing *aggressive* vs. *conservative* designs. Using these value functions, we applied COuNSEL to generate a consensus design and compared it to both the centralized design informed by previous studies in the literature and the state-of-the-practice design informed by historical data. Taking a representative summer day of 2007 at San Francisco International Airport as our evaluation instance, we showed that COuNSEL reduce NAS-wide costs by 4.2% compared to the current state-of-the-practice, and 2.0% compared to the state-of-the-research design.

Acknowledgement

This work was funded through the National Center of Excellence for Aviation Operations Research (NEXTOR II), a consortium of 8 universities contracted by the FAA to provide research support for a wide variety of aviation issues. Their support is gratefully acknowledged. The authors would also like to thank colleagues on the Service Level Expectations project (Distributed Mechanisms for Determining NAS-Wide Service Level Expectations), including Mike Ball and Prem Swaroop of the University of Maryland, and Mark Hansen and Yi Liu of the University of California, Berkeley for helpful discussions.

APPENDIX III

Performance Metrics Tradeoff Models

Yi Liu, Mark Hansen

University of California, Berkeley

1. Introduction

On the day-of-operation, airport capacity varies and is often reduced due to poor weather, traffic congestion, or other factors. Ground Delay Programs (GDPs) are usually implemented in this case in order to balance the arrivals with the reduced capacity. This is accomplished by delaying take offs bound for the congested airports. As a result, GDPs transfer expensive and unsafe airborne delays to cheaper and safer ground delays. In 2006, 624 GDPs were called at eight airports with the largest numbers of GDPs of that year and affected 167,584 flights from 11 major airlines and their subcarriers (Xiong, 2010). As one of the most common Traffic Management Initiatives (TMI), GDPs are essential to keeping the air traffic efficient and safe. However, the current GDP planning process is ad-hoc and subjective in several respects.

First, different managers may create different plans for the same situation. Depending on their temperament and experience, a manager may set higher or lower capacity rates, and shorter or longer program durations. Clear evaluation criteria to assist managers in designing GDPs are needed. Although TMI performance categories are described in the literature (Bolczak et al., 1997; Bradford et al., 2000; Sridhar et al., 2008), associated performance criteria and day-of-operation performance metrics are not defined for GDPs.

Second, flight operators influence the GDP decisions through planning telecons with the traffic managers, and frequent interaction with the command center personnel. The inputs from the flight operators focus on the decisions on the GDP parameters, and not the underlying performance objectives. It is unclear to both the FAA and airlines how the performance of the program will be influenced by the GDP decisions. A mechanism linking the GDP performance metrics to decision variables is missing.

Third, the vast majority of the existing studies dealing with GDP decision-making, there is a sole objective-- minimizing the delay cost (Odoni, 1987; Hoffman and Ball, 2000; Ball et al., 2003; Mukherjee and Hansen, 2007; Liu et al., 2008; Mukherjee et al., 2009). Little effort has been made to consider other goals, such as predictability and throughput. We therefore lack the ability to evaluate GDP performance using multiple criteria. While delay is an adequate measure of operational effectiveness in some instances, it does not present a complete picture of the many aspects of performance that determine the quality and level of service that Air Traffic Control (ATC) users receive (Bolczak et al., 1997). Different flight operators may have different preferences on performance goals. For instance, low-cost carriers may consider efficiency more important, whereas cargo airlines may consider predictability more valuable. Therefore, an improved decision-making process should be able to measure various dimensions of the GDP performance.

In this chapter, we propose to address these issues by developing GDP performance criteria and assessing the tradeoffs between multiple performance goals in a manner that can inform air traffic management decision-making. We identify four performance criteria for GDPs: capacity utilization, predictability, efficiency, and equity; and specify

the performance metrics for them. We also build performance trade curves between the criteria and associated metrics using proposed GDP models, and illustrate how these curves could be used to assist in GDP decision-making processes when the objective is a linear function of the goal metrics. The research forms a basis for assessing the performance of GDPs using multiple criteria and will ultimately lead to improved GDP decision-making, in which traffic managers and flight operators can make informed tradeoffs based on their assessment of the importance of different performance criteria.

The remainder of the chapter is organized as follows. Section 2 introduces our GDP models with and without consideration to early GDP cancellation. Section 3 specifies GDP performance criteria and the associated performance metrics. Section 4 presents the performance trade-offs and discusses how the trade curves can assist in GDP decision-making. Section 5 summarizes the work on the proposed GDP tradeoff models. Currently, the research team is also working on a data-based methodology to generate feasible performance vectors. This piece of work is presented in Section 6.

2. Ground Delay Program Models

In the literature, Integer Programming (IP) is the predominant technique for GDP performance evaluation (Odoni, 1987; Richetta, 1991; Ball and Lulli, 2004; Mukherjee and Hansen, 2007). IP successfully provides numerical solutions to specific GDP delay cost optimization problems. With this technique, a few efforts have also been made to investigate tradeoffs between performance goal; in particular with respect to equity and efficiency (Ball and Lulli, 2004; Mukherjee and Hansen, 2007). However, there are two disadvantages in the IP technique. First, the computation results are applicable only for a given problem; extra runs are required for a new set of parameters. Therefore, IP fails to characterize the relationship between GDP parameters and multiple dimensions of performance in a generic way. Second, the computational time of IP programs grows with the size of the problem. Considerable effort in improving the IP algorithm is expected before it could be efficient enough to be incorporated in the GDP decision support tool, if multiple resources are involved and multiple performance objectives are considered. In this study, we attempt to address these issues by using the continuum approach based on queueing diagrams (Daganzo, 1997). Most of the notations in this section are also defined in Appendix A.

When the Airport Acceptance Rate (AAR) at the airport is lowered by bad weather, a GDP will be implemented to balance the demand with the reduced capacity. There are three decision variables in the GDP design: the duration of the program, planned airport acceptance rates, and the scope of the area that is subject to GDP. If the duration of the poor weather is underestimated, then there will be airborne delay since more arrivals were planned than could land. Airborne delay could also occur if the AAR is overestimated. On the contrary, capacity may be underutilized if the estimate is too conservative. Compared to the duration of the program, planned airport acceptance levels are more predictable. For example, at SFO, fog could preclude simultaneous arrival operations on its closely spaced parallel runways, reducing the arrival capacity from 60 to 30 flights per hour (Cook and Wood, 2009). However the duration of the fog is hard to predict due to large uncertainty in the weather forecast. In this research, we investigate how the uncertainty in the duration of the program will affect the GDP performance and assume no uncertainty in airport acceptance levels. Different from the duration of the

program, the scope of the GDP is a choice independent of the weather forecast. When GDPs are called, the FAA exempts flights from a GDP by limiting the scope of the GDP to a geographical area surrounding the destination airport (Ball and Lulli, 2004). With a small scope, more flights will be exempted from the GDP, but longer delays will be imposed on the affected flights and thus reduce equity. As discussed later, the decision on the scope has substantial impacts on performances.

The queueing diagram of the arrival traffic at a GDP airport is shown in Figure 1. The brown solid line represents the scheduled cumulative demand curve, which is linear based on the assumption of a constant schedule demand rate τ. Mathematically, it can be formulated as:

$$S(t) = \lambda \cdot t$$

$S(t)$ is the basis for Original Time of Arrival (OTA). The piece-wise blue dash line represents the planned cumulative arrival curve under GDP, which is the basis for allocating Controlled Time of Arrivals (CTA) for the delayed flights. Due to poor weather, AAR drops to a low level, C_L, at time zero. In order to meter flow to the airport, a GDP is initiated and the planned AAR is supposed to switch to the high level, C_H, at time T, when the weather is expected to clear up. Delay is planned to develop in the system until time T and then start to vanish. The planned curve intersects the scheduled curve at time T_2, when there is no delay any more and the two curves overlap with each other afterwards. The duration of the program, T_2, is from the time when the queue starts to grow to the time when there is not more delay in the system. With the above assumptions, we can express the duration of the program as

$$T_2 = \frac{C_H - C_L}{C_H - \lambda} T$$

where, T is the expected weather clearance time. We will consider T as a GDP decision variable that determines the planned duration of the program.

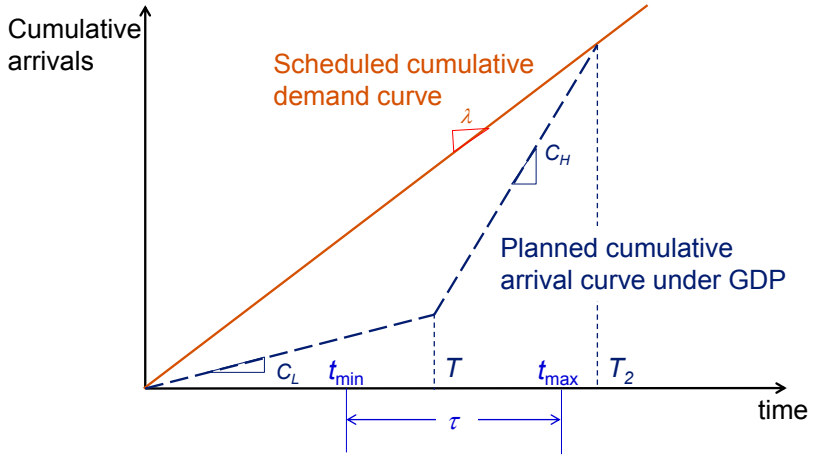


Figure 1: Queueing diagram of the arrival traffic at the GDP affected airport

Given the constant airport acceptance rates, CTA's are determined by the decision on the weather clearance time, T. The formulation of the planned cumulative arrival curve is then

$$N(t|T) = \begin{cases} 0, t \le 0 \\ C_L t, 0 < t \le T \\ C_L T + C_H (t - T), T < t \le T_2 \\ \lambda t, t > T_2 \end{cases}$$

The amount of the planned ground delay in the GDP is

$$D_P(T) = \int_{S(t)>N(t|T)} [S(t) - N(t|T)]dt = \frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot T^2$$

Due to errors in predication, the actual time when the weather will clear up, τ, may be different from the planned clearance time, T. To increase tractability, we assume that the actual clearance time is uniformly distributed between t_{min} and t_{max}. Consequently, the expected clearance time T will be a value between the same bounds. After decisions are made on T, two possible scenarios could occur during the GDP: early clearance, $\tau < T$; and late clearance, $\tau > T$. In the case of early clearance, there is unexpected extra capacity, $C_H - C_L$, between τ and T. We could choose not to revise the program and we will have enough capacity to land the planned arrivals as in the original GDP. Alternatively, we could revise the program by taking advantage of the extra part of capacity. In practice, traffic managers usually cancel the GDP earlier in this case to maximize the arrival throughput. Revision will certainly improve capacity utilization and efficiency, but probably reduce predictability since we are making changes to the original GDP plan. There is widespread consensus in the community that predictability is also important. Therefore, in this research we consider early cancellation as an option, but also allow the option of not revising the GDP. In the case of late clearance, there will be a period when high capacity is planned but the actual capacity is still low. GDP extension is assumed in this case so that some additional delay can be converted to ground delay. In the extension, priority will be given to the flights in the air, and flights on the ground will be held and released when all their delay has been absorbed as ground delay. In modeling the extension, we assume that at the time the extension is made, the actual clearance time is known with certainty.

In Sections 2.1 and 2.2, we will present the GDP models for the cases of early clearance and late clearance respectively. In these two sections, we will assume that all the flights bound for the affected airport are involved in the GDP. In practice, only flights within a certain scope will be subject to GDP. Flights that are geographically farther than the scope will be exempted from the program. Impact of scope on the models will be discussed in Section 2.3

2.1 GDP Models with Early Clearance

When high capacity is available earlier than planned, we are able to land the arrivals as planned in the original GDP, or revise the GDP to allow a higher rate. The queueing diagrams for the arrivals are shown below for the cases of no early cancellation and early cancellation.

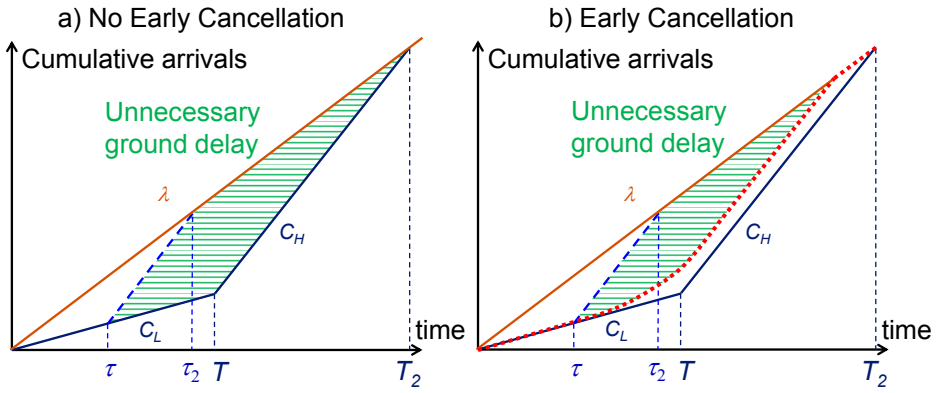


Figure 2: Queueing diagrams for arrivals, early clearance case

If we don't consider GDP cancellation, then the realized cumulative arrival curve will overlap with the planned cumulative arrival curve. The realized throughput, Th, is equal to the planned:

$$Th(t|\tau < T, no\ early\ cancellation) = N(t|T)$$

Realized delay, D_R, is then equal to the planned delay:

$$D_R(\tau < T, no\ early\ cancellation) = D_P(T)$$

If we had perfect information when we were designing the GDP, then we could have allocated the CTA based on the ideal cumulative arrive curve, the slope of which shifts from C_L to C_H at time τ and delay vanishes at time τ_2. More specifically, the ideal cumulative curve is formulated as

$$A(t|\tau) = \begin{cases} 0, t \le 0 \\ C_L t, 0 < t \le \tau \\ C_L T + C_H(t-T), \tau < t \le \tau_2 \\ \lambda t, t > \tau_2 \end{cases}$$

Unnecessary delay, which is defined as the difference between the realized delay and the minimum delay if we had perfect information when designing the GDP, is highlighted by the shaded area in Plot a). If GDP is cancelled earlier to utilize the unexpected part of capacity, then unnecessary delay could be reduced as shown in Plot b). The realized cumulative arrival curve is different from the planned cumulative curve. Conceptually, the realized curve, as depicted as the red dotted curve, should be above the planned curve because we are considering C_H as the acceptance rate starting from τ instead of T in order to accelerate the clearance of delay. This curve begins to deviate from planned cumulative curve not at time τ but sometime after it, because it takes time for flights held on the ground to arrive at the airport.

Without early cancellation, the ideal, planned and realized cumulative arrive curves are just linear piece-wise functions. In the case of early cancellation, if we simply terminate the GDP by releasing flights at the earliest possible take-off time, some flights may encounter airborne delay before landing since we do not have infinite capacity. To avoid this part of airborne delay, we need to release flights by the amount of capacity we have. Here, we assume that GDP is revised at time τ and the new time slots are assigned instantaneously. The model with early cancellation is considerably more complicated because we need to revise the cumulative arrival curve, which will serve as the basis of the new CTA. The revised cumulative curve is jointly determined by the available cumulative arrival demand, D, and the available capacity. Since we are considering early

cancellation, the available capacity is C_L before τ and C_H after τ. The available demands are obtained by releasing flights at their earliest possible take-off time. Before we derive the mathematical expressions for D, we group the affected flights in the original GDP into three groups. D will be the sum of the available cumulative arrival demands from each group. Specifically, at the time when capacity actually increases, τ, flights heading to the congested airport are either:

- Type I: these flights have already departed. They are scheduled to depart before τ and actually have departed before τ under the original GDP. Assigned ground delay in the original GDP has fully occurred for Type I flights and they will therefore arrive at their allocated CTA. We denote the available cumulative demand curve, which is the same as the planned cumulative arrival curve under the original GDP, for this type of flights as D_- and its rate as D'_-.
- Type II: these flights are being held on the ground at τ. Type II flights would have already departed in the original schedule but are waiting on the ground at time τ due to the initiation of the original GDP. These flights can, in principle, depart immediately if capacity permits. If these flights are all released at time τ, the cumulative arrival curve after this action is then the available cumulative demand curve, which is denoted as D_0 with demand rate D'_0.
- Type III: there flights are scheduled to depart after τ. Ground delay assigned in the original GDP has not occurred yet for flights of this type. Therefore, there would be no delay for these flights if they were allowed to take off as scheduled. Assuming they depart as scheduled, they will arrive earlier than the time slots assigned to them under the original GDP. The available cumulative arrival demand curve is the same as the scheduled cumulative arrival curve. The cumulative demand and its rate for this type are denoted as D_+ and D'_+ respectively.

The total available cumulative arrival demand after revision, D, is then the sum of the cumulative demands of each type after the actions. The difference between D and the planned cumulative arrival demand curve in the original GDP reflects the effect of GDP revision, which is affected by the range of the flight time. At this stage, we assume all the flights that are heading to the affected airport will be subject to the GDP. The maximum flight time of these flights is denoted as F_{max}. When F_{max} is small, the delayed flights are concentrated in the vicinity of the affected airport and they could arrive at the airport earlier under revision, which enables the airport to utilize the expected extra capacity efficiently. If the maximum flight time is increased, delay would be absorbed by more flights but the utilization of the unexpected extra capacity would be less efficient. In this analysis, the flight time is assumed to follow a uniform distribution between F_{min} and F_{max}. If the planned clearance time is between τ and the actual clearance time plus F_{min}, then no modification should be made to the GDP. Because in this case the extra demand from the revision first arrives at time $\tau + F_{min}$, there is little we can do if all the available capacity has already been used at this time under the original GDP. Therefore, GDP revision makes a difference only when $\tau + F_{min}$ is less than T. Depending on F_{max}, the available cumulative arrival demand curve is formulated differently. In total, there are three cases for the available cumulative demand curve. Conditions and formulations for each case are discussed in 2.1.1 to 2.1.3, respectively.

2.1.1 Early GDP Cancellation, $\tau + F_{min} < T < T_2 < \tau + F_{max}$

As shown in Figure 3, flights that have taken off before time τ will arrive between τ and $\tau + F_{max}$, which is later than the planned delay clearance time, T_2. In the analysis, we don't consider flights after T_2, because there is no more delay under the original GDP after this time. It is worth mentioning that delay will not vanish earlier than planned after revision even though the system delay will be reduced. One reason is simply because there will be Type I flights arriving at the affected airport from $\tau + F_{min}$ to T_2, and planned delay of these en-route flights has been fully incurred. To generate the revised cumulative arrival curve, we first need the cumulative available demand curves. As discussed before, the total available demand is just the sum of the available demands for each type of flights. The mechanisms of calculating the demands are different and discussed in 2.1.1.1 to 2.1.1.3.

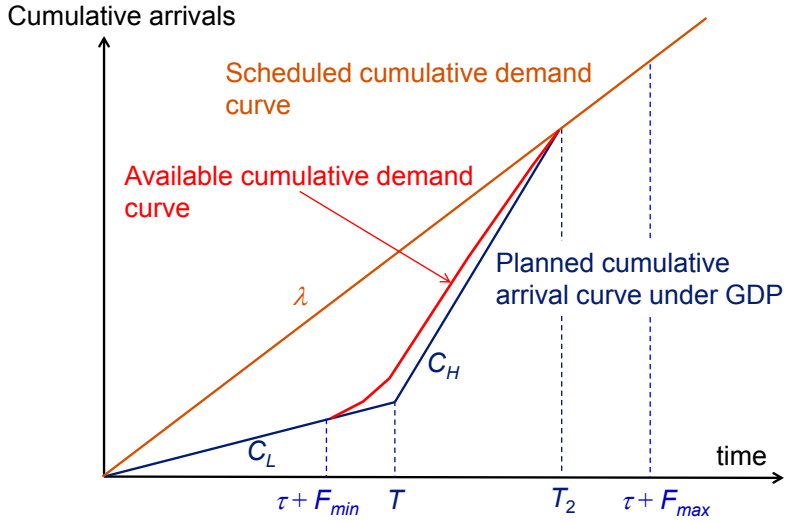


Figure 3. Queueing diagram, $\tau + F_{min} < T < T_2 < \tau + F_{max}$

2.1.1.1 D_-, Type I flights

Type I flights are the flights that have been released from their departure airports when the high capacity is found to be available ahead of time. These flights have arrived at the GDP airport or are in the air. In either case, delay assigned to these flights in the original GDP has already been incurred and they will arrive at their CTA. The principles that are used to derive D_- are:

- Before $\tau + F_{min}$, all planned capacity is utilized for Type I flights. For flights that take off after τ, it is impossible for them to arrive prior to $\tau + F_{min}$ since F_{min} is the minimum flight duration.
- For Type I flights planned to arrive between $\tau + F_{min}$ and T_2, the flight time range at any time t is between $t - \tau$ and F_{max}. If a flight arrives at t with flight time less than $t - \tau$, then this flight must have taken off after τ, so it cannot be a Type I flight. Type I flights are the only flights that have taken off at τ. For a given flight time, type I flights will arrive earlier than other flights which have not taken off yet at τ. Therefore, flights arriving at t with flight time in $[t - \tau, F_{max}]$

will all be Type I flights. The probability that a flights is Type I flight at a given t is then:

$$P_{t,\tau}(D_{-}) = \frac{F_{max} - (t-\tau)}{F_{max} - F_{min}}$$

Since the flight time distribution for all flights is uniform between F_{min} and F_{max}, flight time for Type I flights will be uniformly distributed between $t-\tau$ and F_{max}. Given the capacity rate is C_L before T and C_H afterwards in the GDP, the available demand rate of Type I flights after revision, which is the same as the planned capacity rate for Type I flights under the original GDP, can be expressed as:

$$D'_{-}(t|\tau,T) = \begin{cases} C_L \cdot \dfrac{F_{max} - (t-\tau)}{F_{max} - F_{min}}, & if\ \tau + F_{min} < t < T \\ c_H \cdot \dfrac{F_{max} - (t-\tau)}{F_{max} - F_{min}}, & if\ T < t < T_2 \end{cases}$$

With these, we integrate and express the cumulative arrival demand curve for Type I flights as:

$$D_{-}(t|\tau,T) = \begin{cases} C_L t, & 0 < t < \tau + F_{min} \\ -\dfrac{C_L}{2\cdot\Delta F}[t-(\tau+F_{max})]^2 + C_L\left(\tau + F_{min} + \dfrac{\Delta F}{2}\right), & \tau + F_{min} < t < T \\ -\dfrac{C_H}{2\cdot\Delta F}[t-(\tau+F_{max})]^2 + C_L\left(\tau + F_{min} + \dfrac{\Delta F}{2}\right) + \dfrac{C_H - C_L}{2\cdot\Delta F}[T-(\tau+F_{max})]^2, & T < t < T_2 \end{cases}$$

2.1.1.2 D_0 , Type II flights

Type II flights are the flights that should have taken off at τ if GDP were not initiated, but have not taken off due to the GDP. All these flights have been delayed to some degree at τ. There is no delay planned for the flights arriving after T_2, and thus these flights cannot be Type II flights. Therefore, under original GDP, Type II flights are planned to arrive before T_2. Flight time range of the flights is then between F_{min} and $T_2 - \tau$. Type II flights are held on the ground at τ, and can take off immediately if capacity permits. We can easily plot the cumulative available demands if we know the distribution of flight time since the departure time will be τ for all the Type II flights. Steps for deriving the available demand curve are shown below.

Step 1: $P_\tau(D_0|F)$

$P_\tau(D_0|F)$ is defined as the probability that a flight impacted by the GDP is being held on ground at τ given that its flight time is F. We can write it as

$$P_\tau(D_0|F) = \begin{cases} \dfrac{\lambda\cdot(\tau+F) - N^C(\tau+F)}{\lambda\cdot T_2}, & \tau + F \le T_2 \\ 0, & \tau + F > T_2 \end{cases}$$

where, $N(t)$ is the planned arrival curve under original GDP.

Substitute the formulation of $N(t)$ in the expression of $P_\tau(D_0|F)$:

$$P_\tau(D_0|F) = \begin{cases} \dfrac{(\lambda - C_L)\cdot(\tau + F)}{\lambda \cdot T_2}, F_{min} \le F \le T - \tau \\ \dfrac{(\lambda - C_H)\cdot(\tau + F) + (C_H - C_L)\cdot T}{\lambda \cdot T_2}, T - \tau \le F \le T_2 - \tau \\ 0, otherwise \end{cases}$$

Step 2: $P_\tau(D_0)$

$P_\tau(D_0)$ is the probability that the flight impacted by GDP is being held on ground at τ. It is also the proportion of Type II flights in all the affected flights. Using the total probability theorem, we get

$$P_\tau(D_0) = \int_{F_{min}}^{F_{max}} P_\tau(D_0|F)\, f(F) dF = \int_{F_{min}}^{T_2-\tau} P_\tau(D_0|F)\, f(F) dF$$

Assume F is uniformly distributed over $[F_{min}, F_{max}]$, we get

$$P_\tau(D_0) = \int_{F_{min}}^{T_2-\tau} P_\tau(D_0|F) \frac{1}{F_{max} - F_{min}} dF$$

$$= \frac{1}{\lambda \cdot T_2 \cdot \Delta F} \cdot \left\{ \int_{F_{min}}^{T-\tau} [\lambda \cdot (\tau + F) - C_L \cdot (\tau + F)]\, dF + \int_{T-\tau}^{T_2-\tau} [\lambda \cdot (\tau + F) - C_L \cdot T - C_H \cdot (\tau + F - T)]\, dF \right\}$$

where, $\Delta F = F_{max} - F_{min}$.

Integrating and multiplying bother sides by $\lambda \cdot T_2 \cdot \Delta F$, we get

$$\lambda \cdot T_2 \cdot \Delta F \cdot P_\tau(D_0) = \frac{(\lambda - C_L)}{2}(T - \tau)^2 - \frac{(\lambda - C_H)}{2}(T - \tau)^2 + \frac{(\lambda - C_H)}{2}(T_2 - \tau)^2 - (\lambda - C_L)\cdot\tau^2 - (\lambda - C_L)\cdot\tau\cdot F_{min} - \frac{(\lambda - C_L)}{2}\cdot {F_{min}}^2 + (C_H - C_L)\cdot T\cdot(T_2 - T)$$

Denoting $\mathrm{y}_{D_0} = \lambda \cdot T_2 \cdot \Delta F \cdot P_\tau(D_0)$ and simplifying the right hand side, we obtain

$$\mathrm{y}_{D_0} = \frac{C_H - C_L}{2}(T - \tau)^2 + \frac{\lambda - C_H}{2}(T_2 - \tau)^2 - \frac{\lambda - C_L}{2}(\tau + F_{min})^2 - \frac{\lambda - C_L}{2}\cdot\tau^2 + \frac{C_H - C_L}{C_H - \lambda}\cdot(\lambda - C_L)\cdot T^2$$

The total number of D_0 flights can then be written as

$$D_{0,total} = \lambda \cdot T_2 \cdot P_\tau(D_0) = \frac{\mathrm{y}_{D0}}{\Delta F}$$

Step 3: $f_\tau(F|D_0)$

$f_\tau(F|D_0)$ is defined as the density function of flight time given the flight is Type II flight. Using Bayes' rule, we know

$$f_\tau(F|D_0) = \frac{P_\tau(D_0|F)\cdot f(F)}{P_\tau(D_0)}$$

$$= \begin{cases} \dfrac{(\lambda - C_L)\cdot(\tau + F)}{\lambda \cdot T_2 \cdot \Delta F} \Big/ \dfrac{\mathrm{y}_{D_0}}{\lambda \cdot T_2 \cdot \Delta F}, F_{min} \le F \le T - \tau \\ \dfrac{(\lambda - C_H)\cdot(\tau + F) + (C_H - C_L)\cdot T}{\lambda \cdot T_2 \cdot \Delta F} \Big/ \dfrac{\mathrm{y}_{D_0}}{\lambda \cdot T_2 \cdot \Delta F}, T - \tau \le F \le T_2 - \tau \\ 0, otherwise \end{cases}$$

$$= \begin{cases} \dfrac{(\lambda - C_L)\cdot(\tau + F)}{\mathrm{y}_{D_0}}, F_{min} \le F \le T - \tau \\ \dfrac{(\lambda - C_H)\cdot(\tau + F) + (C_H - C_L)\cdot T}{\mathrm{y}_{D_0}}, T - \tau \le F \le T_2 - \tau \\ 0, otherwise \end{cases}$$

Step 4: $F_\tau(F|D_0)$

Integrating the conditional probability, we can get the flight time distribution for Type II flights.

For $F_{min} \le F \le T - \tau$,

$$F_\tau(F|D_0) = \int_{F_{min}}^{F} \frac{(\lambda - C_L)\cdot(\tau + x)}{y_{D_0}} dx = \frac{(\lambda - C_L)}{y_{D_0}}\left[\left(\tau F + \frac{F^2}{2}\right) - \left(\tau F_{min} + \frac{{F_{min}}^2}{2}\right)\right]$$

For $T - \tau \le F \le T_2 - \tau$

$$F_\tau(F|D_0) = \int_{F_{min}}^{T-\tau} \frac{(\lambda - C_L)\cdot(\tau + x)}{y_{D_0}} dx + \int_{T-\tau}^{F} \frac{(\lambda - C_H)\cdot(\tau + x) + (C_H - C_L)\cdot T}{y_{D_0}} dx$$

$$= \frac{(\lambda - C_L)}{y_{D_0}}\left[\left(\tau(T-\tau) + \frac{(T-\tau)^2}{2}\right) - \left(\tau F_{min} + \frac{{F_{min}}^2}{2}\right)\right] + \frac{(\lambda - C_H)}{y_{D_0}}\left[\left(\tau F + \frac{F^2}{2}\right) - \left(\tau(T-\tau) + \frac{(T-\tau)^2}{2}\right)\right] + \frac{(C_H - C_L)\cdot T}{y_{D_0}}[F - (T-\tau)]$$

Step 5: D_0

The cumulative available demand of Type II flights, D_0, is obtained by assuming all these flights be taking off immediately at time τ. Therefore, if capacity permits, Type II flights arriving at time t after revision are the Type II flights with flight time as $t - \tau$. The cumulative available demand of Type II flights at time t is then equal to the total number of the total Type II flights multiplying the value of the cumulative flight time distribution function at $t - \tau$:

$$D_0(t|\tau, T) = D_{0,total} \cdot F_\tau(t - \tau|D_0)$$

$$= \begin{cases} 0, t < \tau + F_{min} \\ \frac{(\lambda - C_L)}{\Delta F}\cdot\left[\left(\tau(t-\tau) + \frac{(t-\tau)^2}{2}\right) - \left(\tau F_{min} + \frac{{F_{min}}^2}{2}\right)\right], \tau + F_{min} \le t \le T \\ \frac{(\lambda - C_L)}{\Delta F}\left[\left(\tau(T-\tau) + \frac{(T-\tau)^2}{2}\right) - \left(\tau F_{min} + \frac{{F_{min}}^2}{2}\right)\right] + \frac{(\lambda - C_H)}{\Delta F}\left[\left(\tau(t-\tau) + \frac{(t-\tau)^2}{2}\right) - \left(\tau(T-\tau) + \frac{(T-\tau)^2}{2}\right)\right] + \frac{(C_H - C_L)\cdot T}{\Delta F}(t - T), T \le t \le T_2 \end{cases}$$

It should be emphasized that Type II flights are assigned to arrive before T_2 in the GDP plan. Flights scheduled to arrive after T_2 are not involved in the GDP and there is no delay planned for these flights. However, all the Type II flights have been delayed to some degree at τ. Therefore, flights arriving after T_2 cannot be Type II flights and are excluded in the analysis.

2.1.1.3 D_+, Type III flights

Type III flights are scheduled to take off after τ. Delay has not occurred for these flights, and thus there would be no delay for Type III flights if they could take off as scheduled. In the scheduled, Type III flights arrive gradually after $\tau + F_{min}$, and all the scheduled flights after $\tau + F_{max}$ are Type III flights. So the maximum or say possible demand from this group of flights, which is the same as the scheduled cumulative demand, can be formulated as:

$$D_+(t|\tau, T) = \begin{cases} 0, t \le \tau + F_{min} \\ \frac{\lambda}{2\Delta F}(t - \tau - F_{min})^2, \tau + F_{min} < t \le T_2 \end{cases}$$

2.1.1.4 Revised cumulative arrival demand, D

Sum up the three parts; we get the cumulative available arrival demand as:

$$D(t|\tau,T) = \begin{cases} C_L t, & 0 < t < \tau + F_{min} \\ \frac{\lambda - C_L}{\Delta F} \cdot t^2 - \frac{\lambda - C_L}{\Delta F} \cdot (\tau + F_{min}) \cdot t + C_L \cdot t, & \tau + F_{min} \le t < T \\ \frac{\lambda - C_H}{\Delta F} \cdot t^2 - \frac{\lambda - C_H}{\Delta F} \cdot (\tau + F_{min}) \cdot t + C_H \cdot t + \frac{C_H - C_L}{\Delta F} \cdot T \cdot t - \frac{C_H - C_L}{\Delta F} \cdot T \cdot (\tau + F_{max}), & T \le t < T_2 \end{cases}$$

The available demand rate can be calculated as the derivate of $D(t|\tau)$:

$$D'(t|\tau,T) = \begin{cases} C_L, & 0 < t < \tau + \mathrm{F_{min}} \\ 2\frac{\lambda - C_L}{\Delta F} \cdot t - \frac{\lambda - C_L}{\Delta F} \cdot (\tau + \mathrm{F_{min}}) + C_L, & \tau + \mathrm{F_{min}} \le t < T \\ 2\frac{\lambda - C_H}{\Delta F} \cdot t - \frac{\lambda - C_H}{\Delta F} \cdot (\tau + \mathrm{F_{min}}) + C_H + \frac{C_H - C_L}{\Delta F} \cdot T, & \mathrm{T} \le t < T_2 \end{cases}$$

If capacity permits, when a GDP is cancelled early then all the held flights can be released immediately and Type III flights can take off as scheduled when the GDP is cancelled earlier. In other words, D serves as the basis for allocating new CTA's to flights. In practice, 77% of GDPs are cancelled—without the need to assign new CTA's—in this case because of the natural spread of flight time and schedules (Ball et al., 2010). When capacity is insufficient, the throughput rate will be jointly determined by the demand and the capacity:

$$Th'(t|\tau < T, early\ cancellation) = \min\{D'(t|\tau,T), C(t|\tau)\}$$

Where, $C(t|\tau)$ is the capacity at time t:

$$C(t|\tau) = \begin{cases} C_L, t \le \tau \\ C_H, t > \tau \end{cases}$$

The cumulative throughput is then equal to:

$$Th(t|\tau < T, early\ cancellation) = \int_0^t Th'(s|\tau < T, early\ cancellation)ds$$

And the total realized ground delay is:

$$D_R(\tau < T, early\ cancellation) = \int_{S(t) > Th(t)} [S(t) - Th(t|\tau < T, early\ cancellation)]dt$$

Because there are flights that have taken off before τ arriving after T, delay will not vanish in the system until the planned clearance time T. The field of integration is from zero to T for this case.

The algorithm to calculate the revised cumulative curve and the realized delay is the same regardless of the formulation of D. Therefore, for the other two cases of early cancellation, we will just present how to generate the available cumulative demand curve.

2.1.2 Early GDP Cancellation, $\tau + F_{min} < T < \tau + F_{max} < T_2$

In this case, flights that depart before τ can all arrive at the affected airport before the planned delay clearance time T_2. As shown in Figure 4, there will be no more delay in the system after $\tau + F_{max}$, if capacity permits

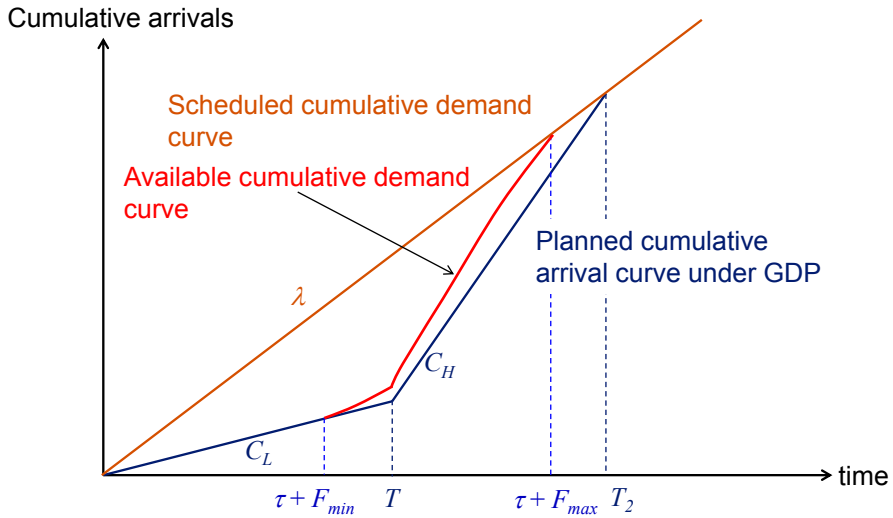


Figure 4. Queueing diagram, $\tau + F_{min} < T < \tau + F_{max} < T_2$

The same methodology is used to generate the available demand curves for the three types of flights. The studied period is still between 0 and T_2. The difference is that the flight time range for Type II flights is $[F_{min}, F_{max}]$ in this case. The available arrival demand curve is found to be

$$D(t|\tau,T) = \begin{cases} C_L t, & 0 < t < \tau + F_{min} \\ \frac{\lambda - C_L}{\Delta F} \cdot t^2 - \frac{\lambda - C_L}{\Delta F} \cdot (\tau + F_{min}) \cdot t + C_L \cdot t, & \tau + F_{min} \le t < T \\ \frac{\lambda - C_H}{\Delta F} \cdot t^2 - \frac{\lambda - C_H}{\Delta F} \cdot (\tau + F_{min}) \cdot t + C_H \cdot t + \frac{C_H - C_L}{\Delta F} \cdot T \cdot t - \frac{C_H - C_L}{\Delta F} \cdot T \cdot (\tau + F_{max}), & T \le t < \tau + F_{max} \\ \lambda t, & \tau + F_{max} \le t \le T_2 \end{cases}$$

2.1.3 Early GDP Cancellation, $\tau + F_{min} < \tau + F_{max} < T$

In this case, the maximum flights time is small and the weather clears up much earlier than the planned time. GDP revision is supposed to be more effective in terms of delay saving. The queueing diagram for arrivals is shown in Figure 5.

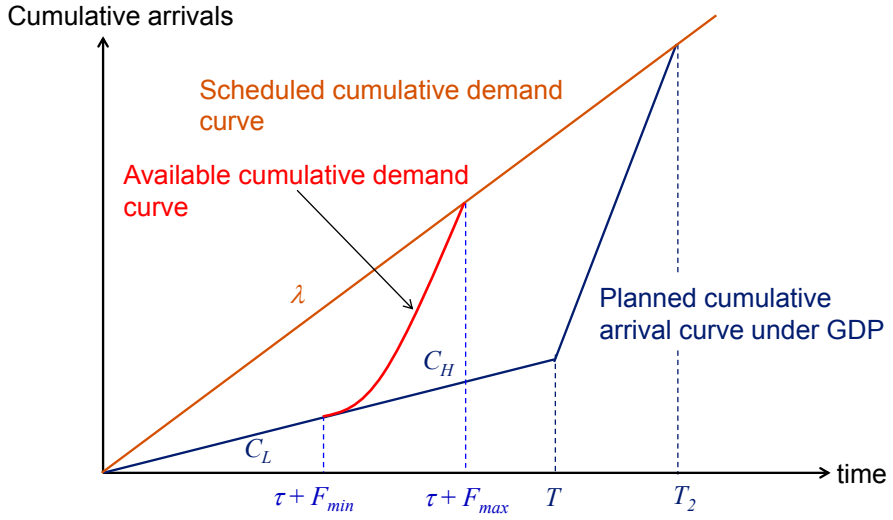


Figure 5. Queueing diagram, $\tau + F_{min} < \tau + F_{max} < T$

The formulation of available demand curve is simple compared to the other two cases:

$$D(t|\tau,T)=\begin{cases} C_L t, & 0\le t<\tau+F_{min} \\ \dfrac{\lambda-C_L}{\Delta F}\cdot t^2-\dfrac{\lambda-C_L}{\Delta F}\cdot(\tau+F_{min})\cdot t+C_L\cdot t, & \tau+F_{min}\le t<\tau+F_{max} \\ \lambda t, & \tau+F_{max}\le t\le T_2 \end{cases}$$

2.2 GDP Extension Models

In the case of late clearance, GDP will be extended. We assume that flight operators are informed of the new CTA at time T. When revising GDP, we assume that τ is known with certainty. Extension is realized by giving priority to flights in the air and further holding flights on the ground if necessary. Plot a) of Figure 6 is the queueing diagram for the late clearance case without GDP extension. The shaded trapezoid represents the total amount of airborne delay due to unexpected late recovery of the capacity. By extending the GDP, we can transform part of this airborne delay to ground delay even though we are not able to reduce the amount of delay in the system, as seen in Plot b).

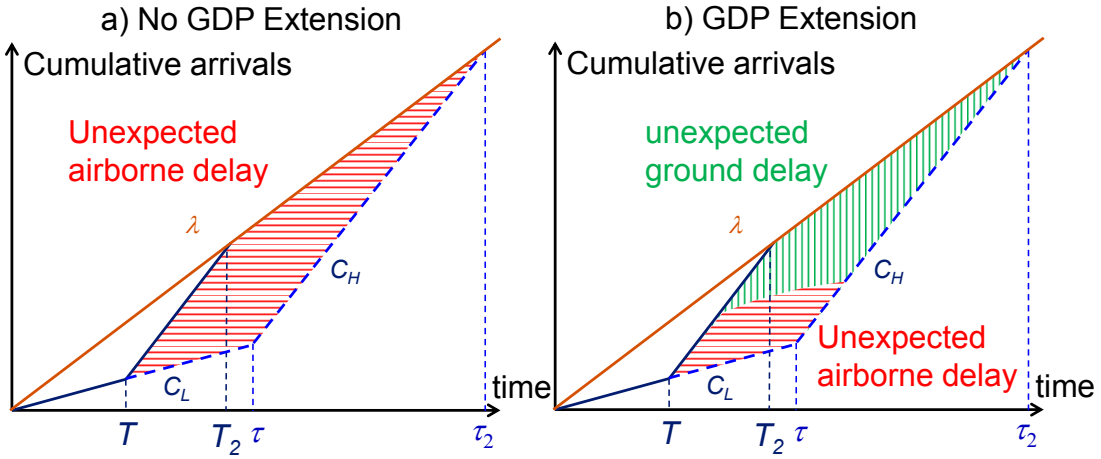


Figure 6. Queueing diagrams for the arrivals, Late Clearance

The cumulative throughput curve is:

$$Th(t|\tau\ge T)=A(t|\tau)$$

where, $A(t|\tau)$ is the cumulative arrival curve if we had perfect information at the beginning of the GDP.

When we extend the GDP, the landing sequence may not be ration-by-schedule anymore because landing priority is given to the en-route flights. An example is illustrated with Table 1, where GDP is implemented due to dense fog. The fog is planned to burn off at 10:00 am in the original plan. At 10:00 am, it does not clear up and the forecast is predicting the weather will be clear at 11:00 am. As a result, there will be another half hour of delay on average for the affected flights. In GDP extension, flights that have not taken off at 10:00 am will be further held on the ground. As seen in Table 1, Controlled Time of Departure (CTD) of Flight 2 is updated to 45 minutes later and it will arrive at the affected airport at 1:45pm instead of 1:00 pm. Delay for Flight 2 increases from 1.5 hours to 2.25 hours, and the extra 0.75 hour is realized as ground delay. On the contrary, flight 1 has already taken off before 10:00 am; therefore, any extra delay will be in the form of airborne delay. Because priority is given to en-route flights in GDP extension, Flight 1 lands at 1:30 pm which is earlier than the actual arrival time of Flight

2. Due to GDP, Flight 1 is delayed for 1.75 hours with 0.25-hour airborne delay. If there were no GDP revision, then Flight 1 would have landed at 1:45 pm with 0.5-hour airborne delay.

Table 1: Example of landing re-sequence

Flight	OTA	Flight Time (hour)	CTD	CTA	New CTD	New CTA
1	11:45 am	5	8:15 am	1:15 pm	8:15 am	1:30 pm
2	11:30 am	1.5	11:30 am	1:00 pm	1:15 pm	1:45 pm

With the GDP extension, we transferred 15-minute airborne delay to ground delay. However, it should be noticed that the total extra delay for these flights is 1 hour, which is the same as if there were no GDP extension. In addition, the arrival sequence is reversed compared to the arrival time in the schedule. Our research focuses the performance at the system level and ignores the impact of GDP extension at the flight level.

Similar to the case of early clearance, we also categorize flights into three groups: Type I, II and III. However, the critical time used to define the groups will be T instead of τ in the previous case. For instance, Type I flights are the flights that have taken off at time T. Denote the planned cumulative arrival curve for Type I flights in this case as $C\_$. We assume that:

- The actual clearance time τ is known at time T, when we extend the GDP.
- All the capacity will be used to land Type I flights first. Type II/III flights will be held on the ground if necessary and released when there are available arrival slots.

To estimate airborne delay in the system, we need to generate the cumulative arrive curve for Type I flights. The difference between this and the actual cumulative arrive curve— with slope shifts to C_H at τ—will be the unexpected airborne delay. The rest of the delay will be the unexpected ground delay. There are four different cases for GDP extension, as illustrated in Figure 7. The unexpected airborne delay is highlighted as in the shaded area.

Plot a) represents the case where all the unexpected delay is airborne delay. This happens because all the unexpected delay is encountered by Type I flights. It is most likely when T is close to τ and C_H is much bigger than C_L. No action should be taken because there is no delay for flights that have not taken off at this time. The airborne delay is

$$AD(\tau \geq T) = \int_{S(t)>A(t|\tau)} [S(t) - A(t|\tau)]dt - D_P = \frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot (T^2 - \tau^2)$$

There will also be ground delay at the end of the program, which is the same as the planned ground delay in this case. So the total realized delay is

$$D_R(\tau \geq T) = AD(\tau \geq T) + D_P = \frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \tau^2$$

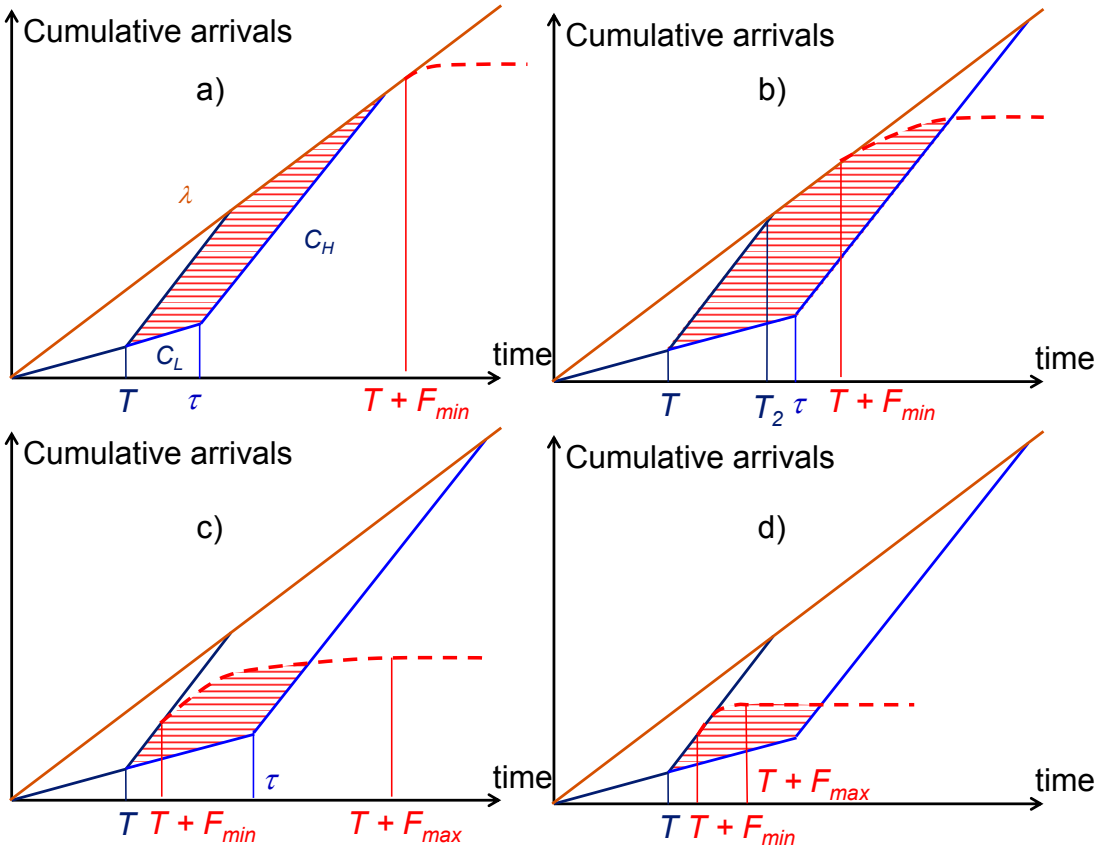


Figure 7. GDP extension models

In the other three plots, GDP extension should be considered and the amount of airborne delay depends on the magnitudes of the minimal flight time, the maximum flight time, and the difference between the planned delay clearance time and the planned capacity recovery time. The expressions of $C\_$, cumulative arrival curve for Type I flights for the three cases are summarized in Appendix B. The algorithm of generating these curves is similar to that shown in Section 2.1.1.1. The area between $C\_$ and the actual cumulative curve is the airborne delay for each case. In other words, airborne delay can be calculated as:

$$AD(\tau \geq T) = \int_{C\_(t|\tau) > A(t|\tau)} [C\_(t|\tau) - A(t|\tau)]dt$$

The total amount of delay is:

$$D_R(\tau \geq T) = \int_{S(t) > A(t|\tau)} [S(t) - A(t|\tau)]dt = \frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \tau^2$$

Therefore, the realized ground delay is

$$GD(\tau \geq T) = D_R(\tau \geq T) - AD(\tau \geq T)$$

2.3 Impact of GDP Scope

So far, we have assumed that all the flights heading to the affected airport with arrival time in the constrained period are involved in the GDP. In practice, only flights within a certain region will be subject to the GDP. Flights that are geographically farther than the scope will be exempted from the program. As mentioned, scope is an important design parameter of the GDP. In this analysis, it is reflected by F_{Scope}, the maximum flight time of the GDP affected flights. Flights with flight time between F_{Scope} and F_{max} will exempted from the program. The demand rate of the exempted flights is denoted by λ_e. By assuming a uniform distribution for flight time, we can obtain

$$\lambda_e = \frac{F_{max} - F_{scope}}{F_{max} - F_{min}} \cdot \lambda$$

All the delay will be absorbed by the non-exempted flights whereas the exempted flights will arrive at the airport on time. The queueing diagrams of the GDP arrivals for the non-exemption case and the exemption case are shown in Figure 8. The non-exemption case is represented with dashed lines and the case with exempted flights is represented with solid lines. Compared to the non-exemption case, both the demand rates and the capacity rates in the exemption case are reduced by λ_e, which should be less than C_L. Denote the delay clearance time in the exemption case as $T_{2,e}$. It is found that $T_{2,e}$ is equal to T_2, when delay clears in the non-exemption case. The planned cumulative arrival curve is shaped the same as that in the non-exemption case:

$$N_e(t|T) = \begin{cases} 0, t \le 0 \\ (C_L - \lambda_e) \cdot t + \lambda_e \cdot t, 0 < t \le T \\ (C_L - \lambda_e) \cdot T + (C_H - \lambda_e) \cdot (t - T) + \lambda_e \cdot t, T < t \le T_2 \\ \lambda t, t > T_2 \end{cases} = N^C(t|T)$$

It should be mentioned that this is not the cumulative curve for the exemption case in Figure 8. Since there is no delay for exempted flights, to estimate the delays for the GDP with exemption, we only need to replace $\lambda/C_H/C_L$ in the previous models with $\lambda - \lambda_e/C_H - \lambda_e/C_L - \lambda_e$. The planned delay will be the same for different exemption rates with the same T because the planned cumulative arrival curve remains unchanged.

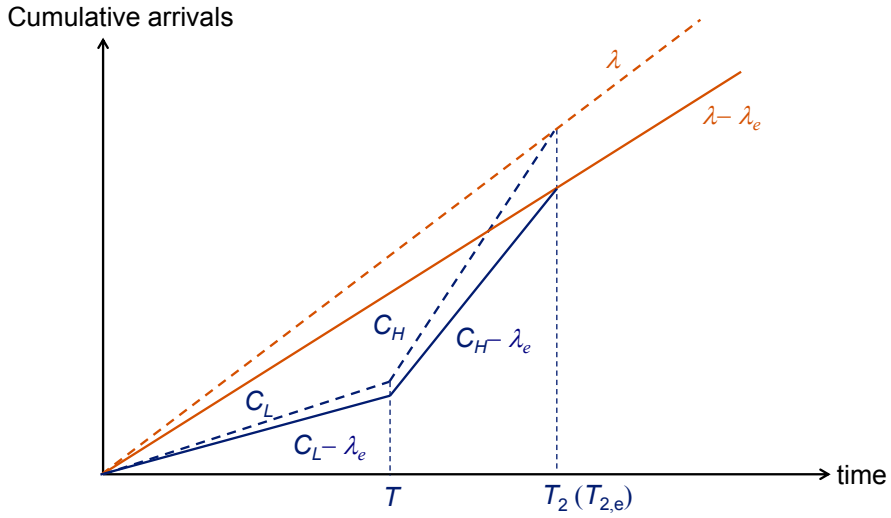


Figure 8. Queuing diagrams of GDP affected arrivals, non-exemption and exemption cases.

3. Performance Metrics

In this Section, we introduce our performance metrics with the proposed GDP models. For each metric, we will first present the general definition based on the performance criteria. After that, we will derive the formulations for the metric under three different situations: early clearance without GDP cancellation; early clearance with GDP cancellation; late clearance with GDP extension. The Section is summarized in 3.5.

3.1 Capacity Utilization

This metric is specified to measure how fully we used our capacity. It is defined as the ratio of throughputs:

$$\alpha_c(\tau, T) = \frac{N_R}{N_I}$$

where,

N_I is ideal throughput under perfect information at the time when queue clears, τ_2;

N_R is realized throughput at this time.

These values are shown in Figure 9, for the cases of early clearance and late clearance respectively. As we see in Plot a), the realized throughputs are less than the idealized throughput at τ_2. Therefore, capacity utilization is less than 1 in the case of early clearance. However, N_R is increased if we consider early GDP cancellation, which benefits capacity utilization. In the case of late clearance, the ideal throughput is the same as the realized throughput since delay could only clear at time τ_2 even if we had perfect information at the beginning. As a result, capacity utilization is equal to 1. In summary, we have

$$\alpha_c(\tau < T|T) = \frac{Th(\tau_2|\tau, T)}{A^C(\tau_2|\tau)} = \begin{cases} \dfrac{N(\tau_2|T)}{\lambda \cdot \tau_2}, \text{no early cancellation} \\ \dfrac{Th(\tau_2|\tau < T, early\ cancellation)}{\lambda \cdot \tau_2}, \text{early cancellation} \end{cases}$$

and

$$\alpha_c(\tau \geq T|T) = 1$$

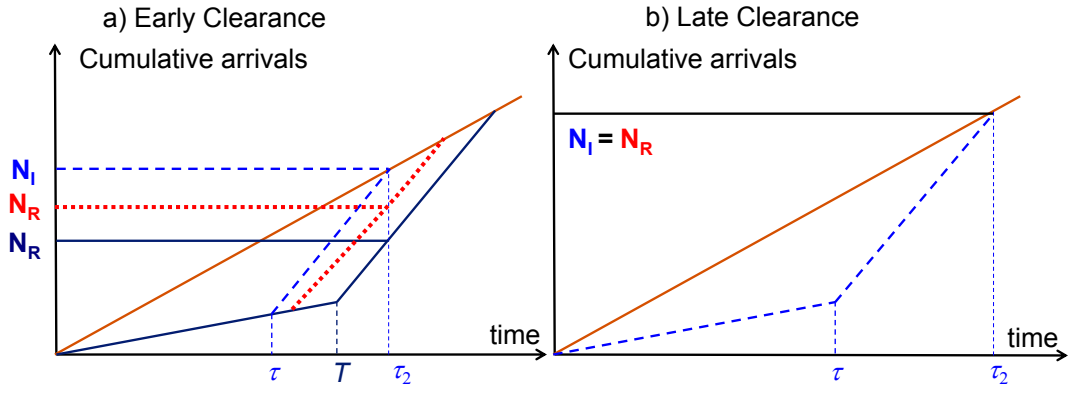


Figure 9. Ideal throughput and realized throughputs

Note: The red dot line represents the cumulative throughput curve with early cancellation

3.2 Predictability

In principle, predictability is defined to measure the amount of information available in advance. In the design of a GDP, a certain amount of delay is planned. Flight operators will be informed of the expected delays before the GDP is implemented. The realized delay will usually be different from the planned delay due to error in prediction. Predictability is then identified to measure how different the realized delay is from the planned delay:

$$\alpha_P(\tau, T) = \frac{min(D_P, D_R)}{max(D_P, D_R)}$$

where,

D_P is flight delay planned at the beginning of the GDP;

D_R is total realized flight delay.

As shown in Figure 10, D_P is determined by the planned weather clearance time at the beginning of the GDP and does not change with the real clearance time. On the contrary, D_R depends on when the weather will clear up and whether we choose to revise the GDP D_R or not. In the case of early clearance, D_R will be equal to D_P if we choose not to revise the GDP. Flight operators could run their operations relying on the CTA allocated in the original GDP and no further adjustment will be needed. If the GDP is revised, as seen in the bottom-left plot, we will be able to save delay in the system and realized delay will be less than the planned delay. In the case of late weather clearance, realized delay is larger than the planned delay due to unexpected late capacity recovery. We can further write our predictability metric as

$$\alpha_P(\tau < T|T) = \begin{cases} 1, \text{ no early cancellation} \\ \dfrac{D_R(\tau < T, early\ cancellation)}{\frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot T^2}, \text{ early cancellation} \end{cases}$$

and

$$\alpha_P(\tau \geq T|T) = \frac{\frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot T^2}{D_R(\tau \geq T)} = \frac{\frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot T^2}{\frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \tau^2} = \frac{T^2}{\tau^2}$$

Predictability will be less than 1 when we make change to the original GDP plan, either by cancelling the program earlier or extending it. Choosing not to revise the GDP when τ is less than T, will cause unnecessary delay in the system but will benefit predictability. By the current practice, GDP is usually cancelled early whenever possible. However, there is widespread consensus in the community that predictability is important. Day-of-operation predictability allows a multitude of benefits, such as reduced communication between command center and airline dispatchers and pilot workload mitigation. In this analysis, we leave early GDP cancellation as an option, which will allow us to examine a GDP design in a more comprehensive way.

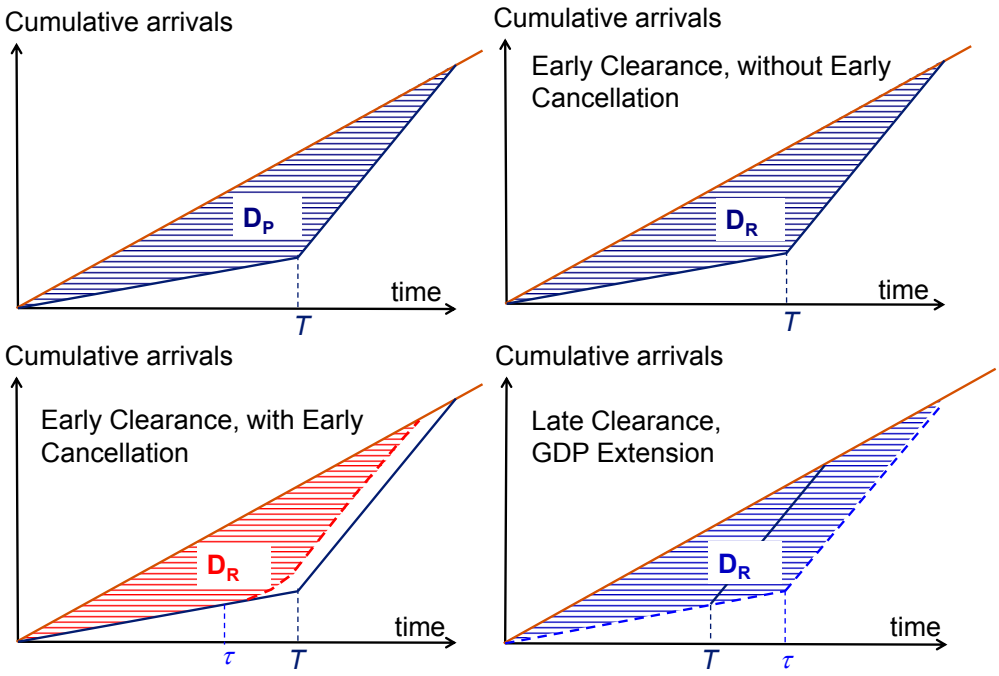


Figure 10. Planned delay and realized delays

3.3 Efficiency

A primary motivation for the GDP is that as long as delay is unavoidable, it is cheaper and safer for flights to absorb delay on the ground before take-off, rather than in the air. The efficiency metric is defined to measure realized delay cost relative to the minimum cost that could be incurred under perfect information. In this metric, we distinguish cost of airborne delay from the cost of ground delay and assume the cost ratio is β (>1). In other words, 1-minute of airborne delay is equivalent to β-minute of ground delay. The efficiency metric is then

$$\alpha_e(\tau, T) = \frac{C_I}{C_R}$$

where,

C_I is minimum cost that would be incurred if perfect information were available about when the capacity will increase:

$$C_I(\tau) = \int_{S(t)>A(t|\tau)} [S(t) - A(t|\tau)]dt = \frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \tau^2$$

C_R is total realized cost;

C_I will always be ground delay cost but C_R could include airborne delay cost. The costs are illustrated in Figure 11. All the costs are ground delay cost except the realized cost in the GDP extension case, as shown in the bottom-right plot. As discussed in Section 2.2, there will be airborne delay-highlighted with dots- for flights that have taken off at time T.

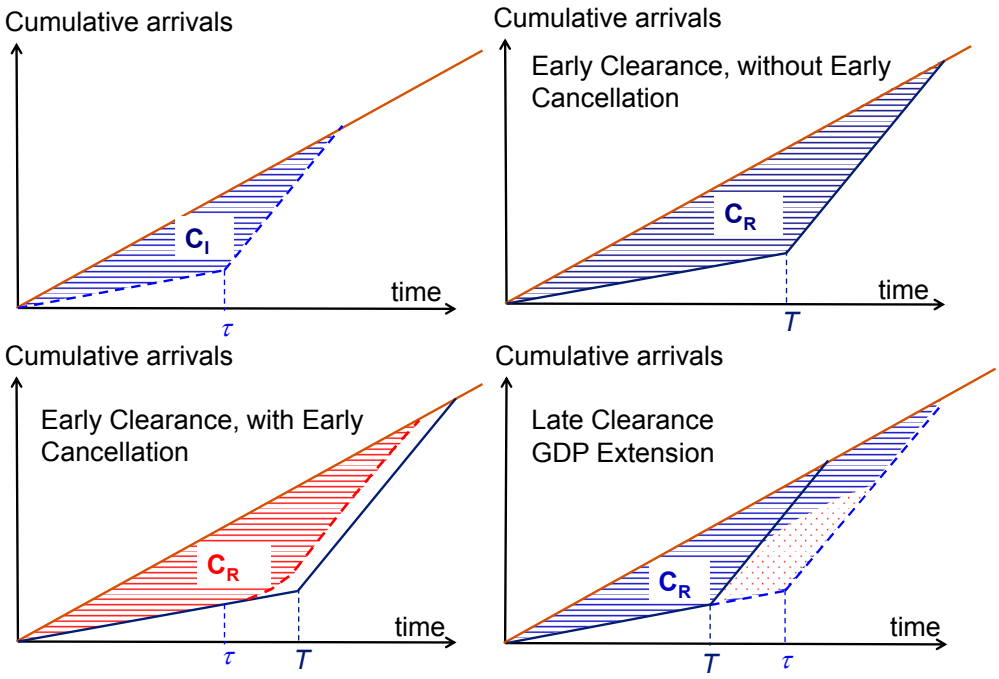


Figure 11. Minimum cost and realized costs

Following the notations in Section 2, we can further write the efficiency metric as:

$$\alpha_e(\tau < T|T) = \begin{cases} \dfrac{\frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \tau^2}{\frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot T^2} = \dfrac{\tau^2}{T^2}, & \text{no early cancellation} \\ \dfrac{\frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \tau^2}{D_R(\tau < T, early\ cancellation)}, & \text{early cancellation} \end{cases}$$

and

$$\alpha_e(\tau \geq T|T) = \frac{\frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \tau^2}{GD(\tau \geq T) + \beta \cdot AD(\tau \geq T)}$$

3.4 Equity

When there are flights exempted from the GDP, total planned delay and delay planned clearance time are still the same as the case without exemption. However, the maximum planned flight delay is different. The maximum planned flight delay in this model can be expressed as:

$$d^e_{P,max} = \frac{\lambda - C_L}{\lambda - \lambda_e} T$$

In the non-exemption case, λ_e is zero and the maximum planned delay is minimized. $d^e_{P,max}$ is illustrated in Figure 12.

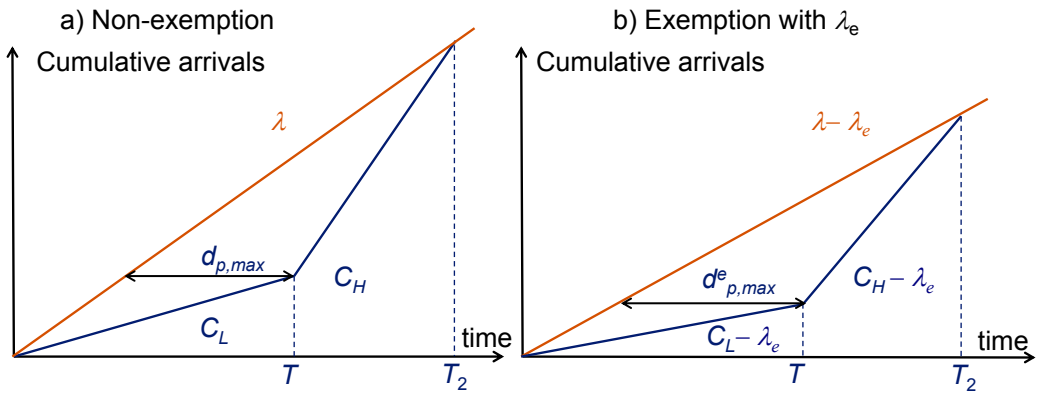


Figure 12. Maximum planned delays, non-exemption and exemption cases

In the exemption case, $d^e_{P,max}$ increases with increasing exempted demand rate. With more flights exempted from the program, more delays are allocated to the affected flights whereas the exempted flights are 'free-riders'. This raises the equity issue in the design of GDPs: how much of the demand should be exempted from the program. In practice, the FAA exempts flights from a GDP by limiting the scope of the GDP to a geographical area surrounding the destination airport (Ball and Lulli, 2004). A flight operator, whose flights are mostly long-haul, will prefer a smaller scope so that more of its flights can arrive on time. On the contrary, flight operators with more short-haul flights may prefer a larger scope, in which case total delay will be absorbed by more flights and delay per flight will be reduced. Preserving equity among competing flight operators is an important goal of the FAA. In this study, the equity metric is defined to measure the maximum planned delay, relative to the maximum delay when no flights are exempted from the GDP (Hoffman et al., 2007; Mukherjee and Hansen, 2007; Ball, et al., 2010). Different from the other performance metrics, equity will only be measured when the GDP is planned. Its value will not be updated upon a GDP revision. The performance metrics of equity is expressed as:

$$\alpha_f = \frac{\min(d^e_{P,max})}{d^e_{P,max}} = \frac{\lambda - \lambda_e}{\lambda} = \frac{F_{scope} - F_{min}}{F_{max} - F_{min}}$$

As seen in the formula, the equity of a GDP is independent of the decision on the clearance time and only affected by the scope. Specifically, equity is an increasing function of scope.

3.5 Expectations of Performance Metrics

We constructed performance metrics for capacity utilization, predictability, efficiency and equity in the previous sections. All the metrics are dimensionless, and between 0 and 1. The expected value of equity metric is determined once we select the scope of the GDP and independent of the prediction errors. However, the values of the other three metrics depend on τ, the real weather clearance time. We need to integrate over τ to get the expected values of the performances:

$$\alpha(T) = E[\alpha(\tau, T)] = \int_{t_{min}}^{t_{max}} \alpha(\tau, T) \cdot f(\tau) d\tau = \int_{t_{min}}^{t_{max}} \alpha(\tau, T) \cdot \frac{1}{t_{max} - t_{min}} d\tau$$

$$= \frac{1}{t_{max} - t_{min}} \cdot [\int_{t_{min}}^{T} \alpha(\tau < T|T) d\tau + \int_{T}^{t_{max}} \alpha(\tau \geq T|T) d\tau]$$

where, α is any of the three metrics: capacity utilization, predictability and efficiency.

Since GDP decisions are made before the real clearance time is known, the program performance should be assessed using the expected values of the defined metrics.

4 Performance Trade-offs and User Optimization

In this section, through a numerical example, we will first present the influence of the GDP decisions on the performance expectations and the trade-offs among multiple performance goals. Then, we will illustrate how the research results could assist decision-making in the design of GDP under capacity uncertainty.

The set of parameter values in the example is shown in Table 2. Capacity values are chosen referring to the airport capacity benchmark report by the FAA (FAA, 2004). The lower and upper bounds for the τ are estimated after reviewing the air traffic control system command center advisories database, which is available in the FAA website. The cost ratio of airborne delay to ground delay is set as 2 (Mukherjee and Hansen, 2007).

Table 2: Parameter values used in the numerical example

Parameter	Notation	Values	Unit
Scheduled demand rate	λ	60	Arrival per hour
High airport acceptance rate	C_H	80	Arrival per hour
Low airport acceptance rate	C_L	40	Arrival per hour
Lower bound for τ	t_{min}	2	Hour
upper bound for τ	t_{max}	6	Hour
Minimum flight time	F_{min}	0.5	Hour
Maximum flight time	F_{max}	7	Hour
Cost ratio	β	2	-

4.1 Performance Metrics and Their Trade-offs

There are three decisions in the design process of a GDP: scope of the GDP, planned clearance time, and whether to cancel the GDP in the case of early clearance or not. As discussed before, equity is determined by the GDP scope only, and it increases monotonically with the scope. Expectations of the other three performance metrics also depend on the assumed clearance time and the policy on early cancellation. The influence of the decisions on the three performance metrics is shown in Figure 13, where the left three plots demonstrate the non-exemption case and the right plots demonstrate the exemption case, with half of the demand exempted from the GDP.

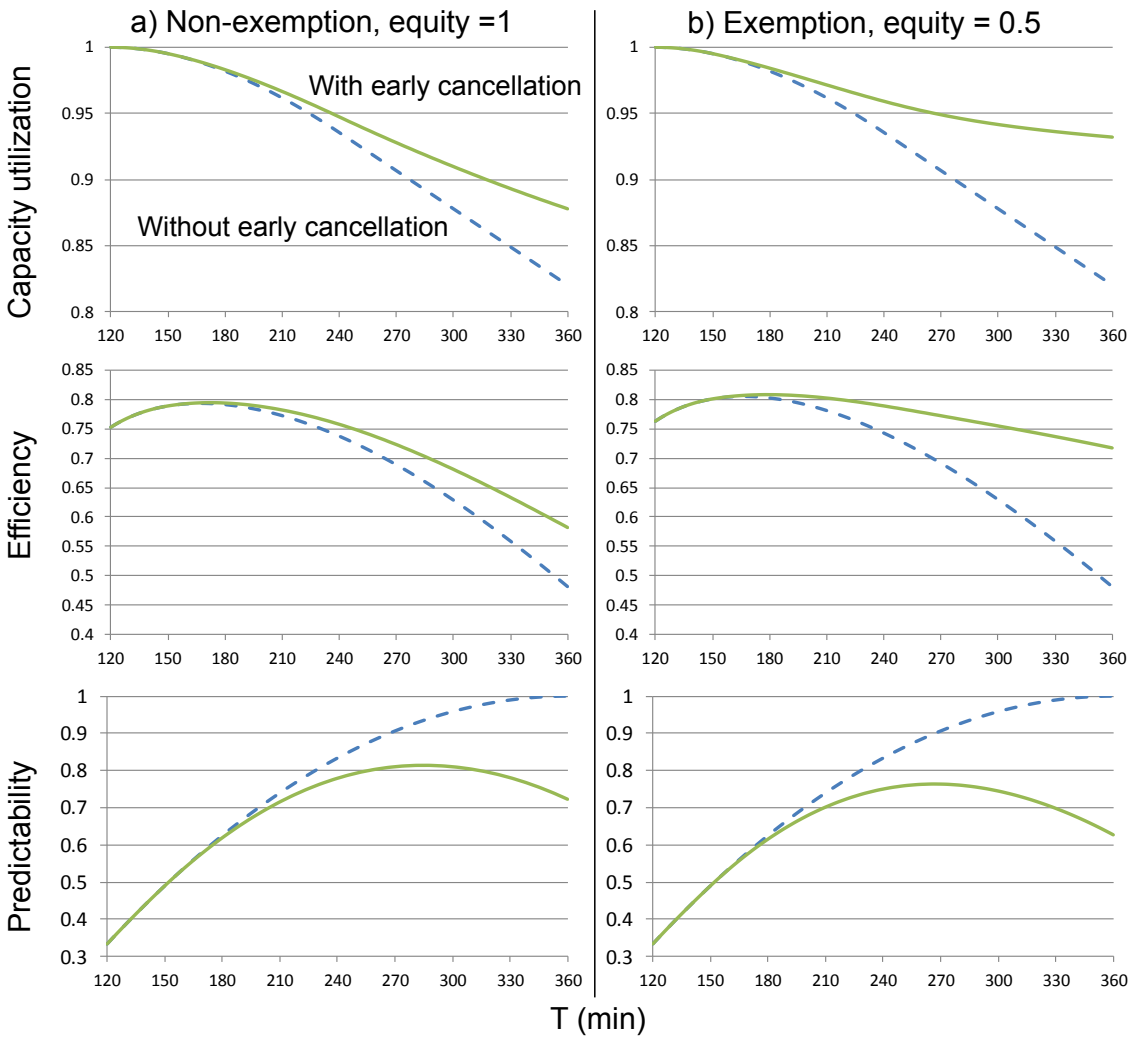


Figure 13. Values of performance metrics as functions of T, with and without early cancellation, with and without GDP exemption

When no flight is exempted from the GDP, equity is equal to one as for the left three plots. Capacity utilization decreases with the planned clearance time because there is a larger chance of early clearance with larger T and part of the high capacity cannot be utilized. Early GDP cancellation will enable us to take advantage of the unexpected high capacity, which benefits capacity utilization. As T increases, efficiency first increases because of reduced chance of expensive airborne delay. After a certain point, efficiency will decrease with T because it is very likely that realized ground delay is much larger than it could be if we had perfect information. Early GDP cancellation saves delay in the system and increases the efficiency. Predictability increases with T without early GDP cancellation. With a larger T, it is very likely that capacity will recover earlier than planned, without early cancellation, and the realized delay will be the same as the planned delay so that predictability approaches 1. In contrast to efficiency and capacity utilization, predictability degrades when we permit early GDP cancellation. Basically, the more adaptable we make the GDP, the less predictability we have. The impact of early cancellation is more obvious with a larger T for all the three metrics.

The conclusions above all also hold when there is exemption, as shown in the right plots. By comparing the early cancellation plots to the non-exemption case (solid lines), we see that exemption will increase capacity utilization and efficiency, but decrease predictability in the system. By exempting long-haul flights from the GDP, the delayed flights are concentrated in the vicinity of the affected airport and they could arrive at the airport earlier if there is early cancellation, which enables the airport to utilize the expected extra capacity earlier. This benefits capacity and efficiency, but reduces predictability. In the case of GDP extension, more flights will still be on the ground if they are closer to the affected airport. As a result, more airborne delay could be transferred to ground delay, which increases efficiency. By comparing the plots without early cancellation for the non-exemption and exemption cases (dashed lines), we find that the plots for capacity and predictability are the same regardless of exemption rate whereas efficiency is slightly improved with exemption. The efficiency gain from reduced scope is because the GDP extension can shift more airborne delay to ground delay.

Performance trade-off curves are shown in Figure 15. Movement toward the right along these curves is associated with earlier planned clearance times. Equity is equal to 1 for the left two plots. The bottom-left plot presents the trade-offs between efficiency and capacity utilization. The dashed blue line is for the case without early cancellation and the solid green line is for the case with early cancellation. Both plots have internal optima, and the points located on the left of the internal peaks are inferior because we can increase efficiency and capacity utilization simultaneously by decreasing T. On the right of the peaks, the line for early cancellation is above that for no early cancellation. Therefore, if only efficiency and capacity utilization are concerned, then we will always choose to terminate the GDP earlier if possible and we tend to pick an earlier planned clearance time. The situation changes when predictability is also taken in to account. Comparing the two dashed plots on the left, we see that a choice of larger T degrades efficiency and capacity utilization but benefits predictability. As a result, flight operators who value predictability more may prefer a larger T. Additionally, early cancellation is not necessarily to be a better choice when predictability is important. For instance, at capacity utilization equal to 0.894, both predictability and efficiency are higher if we choose not to terminate the GDP earlier. It should be pointed out that the planned clearance time of the case with no early cancellation case is smaller than that of the early cancellation case here. The trade-off relationship is similar when equity is reduced to 0.5, as shown in the right plots. Since early cancellation is more effective with long-haul flights exempted, the differences between two cases are more pronounced.

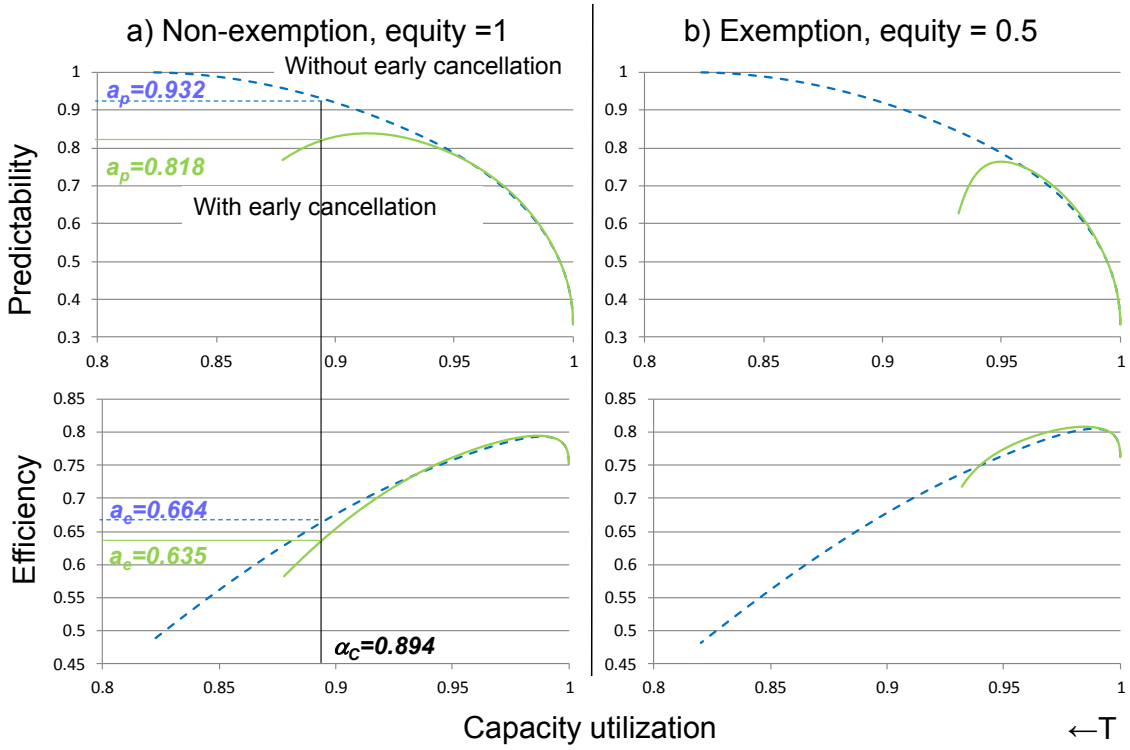


Figure 14. Performance trade-off curves, with *T* increase from the right to the left

4.2 User Optimization

Different flight operators may have different preferences regarding performance goals. Each flight operator may prefer a different point on the trade-off curves, and correspondingly opt for different GDP plans. Rationally, each user will prefer the point that maximizes their utility. The decision-making process could then be formed as a utility optimization problem. In 4.2.1 and 4.2.2, we will present two utility functions. In 4.2.1, the FAA will predetermine the level of equity and the flight operators will only consider the other three metrics in the utility function. In 4.2.2, equity is votable and all the four metrics are involved in the flight operator utility maximization process. We assume linear utility functions to illustrate how the trade-off curves could be used by flight operators to select their preferred GDP decisions. Using the same set up, optimal solutions can always be found for users with concave utility functions.

4.2.1 Predetermined Equity

The constrained optimization problem could be set up as the following. Each system user will choose a set of the performance vectors, $[\alpha_c, \alpha_p, \alpha_e]$, to

maximize: $U(\alpha_c, \alpha_p, \alpha_e)$,

subject to: $F_{early\_cancellation}(\alpha_c, \alpha_p, \alpha_e) = 0$ or $F_{no\_early\_cancellation}(\alpha_c, \alpha_p, \alpha_e) = 0$

where, the constraints are limiting the feasible region to the points on the trade-off curves.

Here, we assume a linear utility function:

$$U(\alpha_c, \alpha_p, \alpha_e) = C_c \cdot \alpha_c + C_p \cdot \alpha_p + C_e \cdot \alpha_e$$

where, the coefficients are the weights of the performance goals.

Implied ideal plans of three users with different preferences on performance goals are compared in Table 3. User 1 is concerned most with capacity utilization and weights predictability and efficiency equally. Predictability has double the importance to User 2 compared to the other metrics, and is an even more of a critical performance goal to User 3. If User 1 is the only user in the system, then the GDP should be planned for 3.8 hours and early cancellation should be performed if possible, as long as all the flights are involved in the system. If half demand is exempted, then the GDP should be planned for a little longer. T increases with exemption rate if early cancellation happens, because early cancellation is more effective with more short-haul flights. With predictability as the most critical performance goal, both User 2 and User 3 choose not to revise the program in the case of early clearance. The preferred planned clearance times of the two users are not affected by the value of equity or say the choices on the GDP scopes. As discussed before, GDP extension only benefits efficiency but has no effect on predictability and capacity utilization. Without early cancellation, exemption makes no difference to the performance metrics in the case of early clearance. This is because realized throughput and realized delay are the same with different GDP exemption rates when no early cancellation is considered, as discussed in Section 2.3. Overall, the preferred T remains unchanged for different values of equity when predictability is dominating.

Table 3: Preferred GDP decisions by different system users, predetermined equity

	Weights			Equity =1		Equity =0.5	
User	C_c	C_P	C_e	T (Hour)	Early cancellation?	T (Hour)	Early cancellation?
1	0.5	0.25	0.25	3.8	Yes	3.88	Yes
2	0.25	0.5	0.25	4.88	No	4.88	No
3	0	0.75	0.25	5.4	No	5.4	No

4.2.1 Endogenous Equity

In this mechanism, equity like the other performance goals, is treated as an argument in the utility function. Thus the problem is:

maximize: $U(\alpha_c, \alpha_p, \alpha_e, \alpha_f)$,

subject to: $F_{no\_early\_cancellation}(\alpha_c, \alpha_p, \alpha_e, \alpha_f) = 0$ or $F_{early\_cancellation}(\alpha_c, \alpha_p, \alpha_e, \alpha_f) = 0$

where, the constraints are limiting the feasible region to the points on the trade-off curves. We consider 6 equity levels, from 0.5 to 1 (most equitable) with increment 0.1. We have different trade-off curves at each equity level. In total, we have 6 pairs of 3-d trade-off curves.

Again, we assume a linear utility function:

$$U(\alpha_c, \alpha_p, \alpha_e, \alpha_f) = C_c \cdot \alpha_c + C_p \cdot \alpha_p + C_e \cdot \alpha_e + C_f \cdot \alpha_f$$

where, the coefficients are the weights of the performance goals.

The optimization results and corresponding GDP decisions are shown in Table 4. The italicized rows assume that equity is unimportant. As seen from the italicized rows, the largest exemption rate, 0.5, is always better regardless of the preferences on the other three performance goals. The unexpected part of capacity can be utilized more effectively and earlier with a larger exemption rate when the GDP is cancelled earlier. Therefore,

when capacity utilization is the dominating metric as in 1*, early cancellation should be allowed, but not when predictability is dominating, as in 2* and 3*.. GDP extension is not affected the performances in terms of capacity utilization and predictability, but benefits efficiency. GDP extension is more efficient when only short-haul flights are involved. Overall, small scopes are always preferred in the design of GDPs when equity is not an issue. It should be mentioned that the scope cannot be smaller than the lower bound, C_L.

Table 4: Preferred GDP decisions by different system users, votable equity

	Weights					Implied ideal plan	
User	C_c	C_p	C_e	C_f	Equity (Scope)	T (Hour)	Early cancellation?
1	**0.5**	**0.25**	**0.25**	**0.001**	**0.5**	**3.88**	**Yes**
*1**	*0.5*	*0.25*	*0.25*	*0*	*0.5*	*3.88*	*Yes*
2	**0.25**	**0.5**	**0.25**	**0.001**	**1**	**4.88**	**No**
*2**	*0.25*	*0.5*	*0.25*	*0*	*0.5*	*4.88*	*No*
3	**0**	**0.75**	**0.15**	**0.1**	**1**	**5.64**	**No**
*3**	*0*	*0.75*	*0.25*	*0*	*0.5*	*5.4*	*No*

The bolded rows highlight the situations where equity is an argument in the utility function. The preference of User 1 remains unchanged whereas User 2 and 3 update their implied ideal plans. A small weight of equity makes sense since the magnitude of change in equity is big. Comparing 2 to 2*, we find that the preferred planned clearance time is independent of the scope when predictability dominates and early cancellation is not considered. This happens because exemption is not benefitting predictability without early cancellation. Comparing 3 to 3*, the choice on the planned clearance time is more conservative since predictability is more valuable. With a larger T, there is a larger chance that the high capacity level could be available earlier and the realized delay will be equal to the planned delay since early cancellation is not considered.

5. Summary on the Theoretical Performance Tradeoff Models

In this chapter, we model the GDPs using continuum approximation. Two models are developed, based on the policy on early GDP cancellation. In the model with no early cancellation, flights arrive at their planned arrival time as allocated in the GDP. No further change is expected and realized delay is equal to the planned delay at the beginning of the GDP. In the case of early cancellation, we revise the cumulative arrive curve by taking advantage of the unexpected extra capacity. The revised cumulative arrive curve is determined by the available cumulative arrival demand and the available capacities together. Analytical solutions are presented for the available cumulative arrival demands for all the possible situations, and the available capacities. By utilizing the unexpected extra capacity, we save delay in the system and delay may vanish earlier. In the case of late clearance, GDP extension is assumed to further transfer some airborne delay to ground delay. When extending the GDP, landing priority is given to the en-route flights and flights that have not taken off yet are further delayed on the ground if necessary.

In Section 3, we identify criteria and develop day-of-operation metrics in the GDP design for four performance goals: capacity utilization, predictability, efficiency, and equity. All the metrics are dimensionless and have values between 0 and 1. To evaluate service level expectation, we integrate the performance metrics over the actual clearance time to calculate the expected values of the defined metrics. Using the expectations of the performance goals, we represent trade-offs in the design of GDPS and relate these to the GDP decisions on: the GDP planned clearance time, the GDP scope and the choice on GDP early cancellation in Section 4.1. It is found that the equity of GDPs is determined by the choice on the scope, whereas the other metrics are also affected by the other two decisions.

Capacity utilization decreases with the planned clearance time, T, because there is a larger chance of early clearance with larger T and part of the high capacity cannot be utilized. Early GDP cancellation will enable us to take advantage of the unexpected high capacity, which benefits capacity utilization. As T increases, efficiency first increases because of reduced chance of expensive airborne delay. After a certain point, efficiency will decrease with T because it is very likely that realized ground delay is much larger than it could be if we had perfect information. Early GDP cancellation saves delay in the system and increases the efficiency. Predictability increases with T without early GDP cancellation. With a larger T, it is very likely that capacity will recover earlier than planned and the realized delay will be the same as the planned delay without early cancellation so that predictability approaches 1. Different from the other two metrics, predictability degrades when we permit early GDP cancellation. The impact of early cancellation is more obvious with a larger T for all the three metrics. The conclusions above also hold when there is exemption. With exemption, expectations of the capacity utilization and efficiency increase, whereas it decreases predictability since it makes the program more adaptable. In the case of GDP extension, capacity and predictability remain unchanged regardless of exemption rate whereas efficiency is slightly improved with exemption. This is because GDP extension can only transfer expensive airborne delay to cheaper ground delay but cannot reduce the amount of delay or improve throughputs.

If only efficiency and capacity utilization are concerned, then we will always choose to terminate the GDP earlier if possible and we tend to pick a small planned clearance time. The situation changes when predictability is also taken in to account. A choice of larger T or larger scope degrades efficiency and capacity utilization but benefits predictability. Different flight operators may have different preferences on performance goals and prefer different points on the trade-off curves that maximize their utilities.

In Section 4.2, we illustrate how the trades could assist the system users in GDP decision-making use a linear function of the performance metrics as the objective. Equity choice is considered in two ways. In one case it is predetermined, while in the other it is considered an argument in the utility function. When capacity utilization is the dominant factor in the utility function, we cancel the case of early clearance. On the contrary, if predictability is the dominating metric, then we choose not to take advantage of the unexpected high capacity and operate as planned in the original GDP. When equity is not considered, smaller GDP scopes are always preferred for flight operators with difference preferences on performance goals. When equity matters, we may choose a larger GDP depending on the weights assigned by the flight operators to equity.

The work enables us to make GDP decisions using multiple criteria. This capacity will lead to improved decision-making, in which traffic managers and flight operators can make informed trade-offs based on their assessment of the importance of different performance criteria. An obvious problem is that different flight operators may have different utility functions. In that case, there must be a process for taking conflicting inputs and arriving at an acceptable compromising plan. Work on this subject, by other researchers, is currently underway.

6. Current work on Data-based Feasible Performance Vectors Generation

One of the inputs that is required to reach the consensus performance vector using COuNSEL is a set of feasible performance vectors. In the previous sections, theoretical models assuming continuum approximation were presented to generate feasible performance vectors under uncertainty. In this section, we will illustrate a data-based methodology for achieving this purpose. This approach is practical and is not tied to any specific performance metrics. It makes use of information that is available to air traffic managers at the time TMI's are being planned.

Most GDPs are implemented because poor weather diminishes the airport acceptance rate. The arrival capacity is predicted based on the weather forecast, and the GDP duration is determined based on capacity and demand information. Future arrival capacity, however, cannot be predicted with certainty. Recognizing this, we allow multiple capacity scenarios, which are derived from historical data in situations when the forecast was similar. Different capacity scenarios, combined with fixed demand, lead to different GDP plans. All these plans are reasonable given the limited information at the GDP decision time. This motivates the following idea for generating feasible performance vectors. The framework of the idea is illustrated in Figure 15.

For a given day-of-operation, the weather forecast—Terminal Area Forecast (TAF)—will be matched with weather forecasts of the historical days using a technique called Dynamic Time Warping (DTW). Similar historical days are identified when the degree of similarity in the TAF exceeds a given threshold. The exhaustive set of possible capacity scenarios is then defined as the collection of the realized capacity profiles on the similar historical days. Each capacity scenario leads to one identical GDP plan. Whatever plan is implemented, the real capacity scenario could be any of the possible scenarios. If the GDP is planned under capacity scenario j but the real capacity scenario is i, GDP performance vector Vji can be estimated by assigning time slots based on scenario j and setting scenario i as the actual available capacity. The expected performance vector of scenario j can then be estimated as:

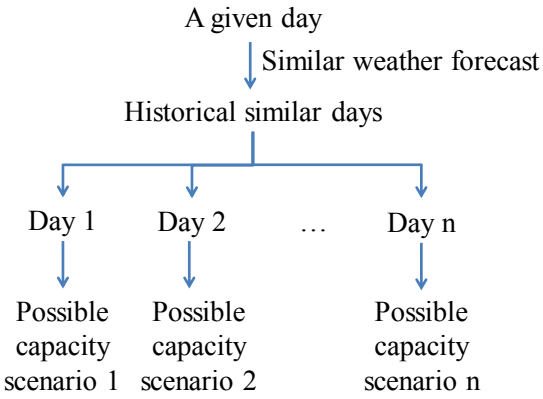
$$E[Vj] = \frac{\sum_i Vji}{n}$$

where n is the number of possible scenarios. This methodology is quite general in the sense that expected performance vectors can always be estimated independent from how the performance metrics are defined.

In the reminder of this section, we present possible definitions for performance metrics based on available data sources in 6.1. After that, we illustrate how these metrics can be evaluated using existing data sources in 6.2. Statistics on the performance metrics are not

presented but can be found in a paper by Liu and Hansen (in press). Different definitions for performance metrics are possible. Our feasible performance vector generation methodology will be capable to deal with any definition of the performance metrics, so long as it is based on planned rates, realized rates, and demand.

Step 1: Possible Capacity Scenarios



Step 2: Expected Performance Vectors given Fixed Demand

Plan\Realization	Capacity scenario 1	Capacity scenario 2	…	Capacity scenario n	Expected performance
Capacity scenario 1	$V11$	$V12$		$V1n$	$\sum_i V1i/n$
Capacity scenario 2	$V21$	$V22$		$V2n$	$\sum_i V2i/n$
…					
Capacity scenario n	$Vn1$	$Vn2$		Vnn	$\sum_i Vni/n$

Figure 15: Framework of Feasible Performance Vectors Generation

6.1 Performance Metrics

This section defines five performance criteria for GDP performance assessment. They are taken from the list of 11 globally endorsed key air transport performance areas. We sought Key Performance Indicators (KPIs) for the identified criteria that were

- Dimensionless and defined on a common scale—the KPIs will allow meaningful performance comparisons across goals;
- simple—the individual KPIs will be easy to understand and calculate using existing data sources;
- and robust— KPIs should not return unreasonable values under extreme conditions. One example is given when predictability metric is introduced later.

None of the KPIs is intended to capture overall GDP performance on its own. However, when put together, the five KPIs provide a fairly complete picture of GDP performance.

6.1.1 Capacity Utilization

In its own performance metrics, the FAA places considerable emphasis on capacity utilization. For example, one of the most widely tracked metrics is the ratio of arrival operations to called rates of arrivals. Called rates are airport arrival acceptance rates planned based on the current runway configuration and expected visibility conditions. In the case of early weather clearance, planned rates in the GDP may be less than the real airport acceptance rate that could have been utilized. In this chapter, capacity utilization for a given GDP is defined as the ratio of actual arrivals between the GDP start time and the GDP end time, to the maximum arrivals that could have been landed assuming perfect information at the beginning of the GDP for the same GDP period. The metric for this criterion is then written as:

$$\alpha_{cu} = \frac{N_A}{N_P}$$

where,

N_A is count of the actual arrivals between GDP start time and GDP end time;

N_P is count of the arrivals that could have been planned for the same period assuming that we had perfect information in airport acceptance rates at the beginning of the GDP. Assuming demand exceeds capacity, it should be the total of real airport acceptance rates during the GDP period.

In theory, the metric should be between zero and one, with one as the best. GDPs are usually implemented due to poor weather at the affected airport. Due to uncertainty in weather forecasts, error in predicting GDP end time is highly likely. After the implementation of a GDP, if the weather condition improves earlier than predicted, then arrival capacity could be restored earlier than planned. In the current practice, GDPs are usually cancelled earlier in this case. At the time when the GDP is cancelled, either all flights could take off assuming no further delay, or new time slots according to the recovered capacity would be assigned to the flights, which allows for early delay clearance in the system. In both cases, capacity may not be fully utilized immediately because it takes time for early released flights to get to the airport. As a result, part of the capacity would be unrecoverable, making capacity utilization less than one.

On the contrary, if capacity degradation persists beyond the planned clearance time, then the GDP will be extended, and further delay will be assigned to flights through new Controlled Times of Arrival (CTAs). Since more flights are approaching the airport than the real airport acceptance rate, there could be airborne delay in the terminal area but available capacity will be well utilized and capacity utilization should be close to one.

6.1.2 Efficiency

To relieve congestion in the terminal area at the affected airport, the GDP is designed to shift the potential airborne delay to ground delay at the departure airports. Studies have shown that a unit airborne delay cost is more than a unit ground delay cost. Estimate of the cost ratio varies widely, but values in the range 2 to 3 are common. Given this, a GDP is viewed as the most efficient if all the delay induced by the GDP occurs on the ground. Accordingly, the efficiency metric is defined as:

$$\alpha_e = \frac{GD}{TD}$$

where,

GD is the amount of GDP induced departure delay of all the GDP-affected flights that are not cancelled;

TD is the amount of GDP induced arrival delay of the same set of flights.

Departure delay is always realized as ground delay at the departure airports, whereas arrival delay is a combination of ground delay occurring before takeoff and airborne delay en route. In our metric, delays are not measured against the scheduled departure/arrival time. Instead, delays are calculated as the differences between the actual departure/arrival time and the estimated time of departure/arrival prior to the flight becoming controlled in the GDP. On a day-of-operation, there could be other reasons for flight delays besides a GDP. By basing delay on estimated times on the day-of-operation, we exclude the impact of non-GDP factors on delays.

Similar to the capacity utilization metric, in theory, the value of this metric should be between zero and one with one as the best. In the case of early weather clearance, real airport acceptance rate is sufficient to land all the planned arrivals. No airborne delay is expected in the terminal area and departure delay should be close to the arrival delay leading to high efficiency. When the GDP is extended due to late weather clearance, aircraft could be delayed in the air before landing which degrades efficiency.

6.1.3 Predictability

When a GDP is implemented, CTA and Controlled Time of Departure (CTD) are issued to flights that are estimated or scheduled to arrive during the GDP time horizon, which is the period between the start and end time of the program. The CTA represents the target arrival time of the flight, but the CTD is the enforcement mechanism—flights must take off close to their CTD. Flight operators will then adjust their operations according to the assigned time slots. This could involve changes in gate assignment, airport surface operations, crew swapping, and substitution of flights between assigned arrival slots. Most of the time, GDPs are planned ahead of time based on weather forecast and demand forecast, so GDP report time—when the GDP is issued—is earlier than the GDP start time. The lead time, difference between report time and start time, allows flight operators to make adjustments in their operations. Due to errors in prediction, GDPs are often revised after the first implementation, via adjustments to the arrival rate or the end time of the program. Every program revision requires adjustments on the part of the flight operators. No matter what the revision is, it requires further change in the flight plans and extra effort from both the flight operator side and the traffic manager side. A program that is revised too frequently can be very disruptive. In the light of this, we assumed that "perfect" predictability is achieved when planned delay at the beginning of the GDP and realized delay at the end of the GDP are equal. We want predictability to be one in this case, and decrease as the planned and realized delays become more different. Accordingly, predictability is measured as:

$$\alpha_p = \frac{min(D_P, D_R)}{max(D_P, D_R)}$$

where,

D_P is the planned arrival delay in the system at the beginning of the GDP;

D_R is the realized arrival delay in the system at the end of the GDP.

As shown in Figure 1, predictability is valued as one only when the realized delay is equal to planned delay. When realized delay is different from the planned delay, either

smaller or larger, predictability will degrade. To preserve the 0-1 scale, the degradation occurs at a diminishing rate when realized delay is larger than planned delay. Put another way, the predictability score is the same—0.5—whether realized delay is twice or half as the planned delay. Other predictability metrics should be considered in future research.

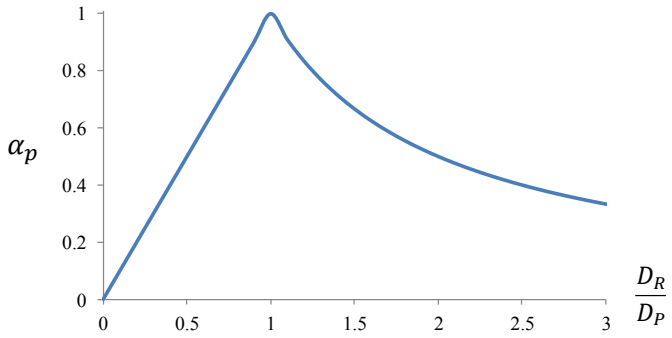


Figure 16: Predictability Metric against the Ratio of Realized Delay to Planned Delay

Delay in the predictability metric is measured with respect to the flight schedule. This is because most of the adaptations to delay which dispatchers must make—such as crew scheduling, misconnections, aircraft rotations—are related to delay against the original schedule. We thus define D_P as the sum of the planned flight arrival delays—differences between CTA and scheduled time of arrival—for all the flights that are affected in the first GDP plan. When a GDP extension is announced with a new set of time slots, additional flights could be affected by the GDP. When estimating the realized system delay, D_R, all the flights that are affected by the GDP—whether in the initial plan or subsequently—are considered. Like D_P, D_R is measured relative to the original flight schedule.

Collaborative Decision Making (CDM) poses further issues for the predictability metric. Since 1998, GDPs have been implemented under CDM, which allows airline substitution and cancellations, and inter-airline compression to fill open slots from substitution and cancellation. Swapping slots does not change the total delay in the system because it shifts the same amount of delay from one flight to another. On the other hand, cancellations relieve congestion and reduce delay at the system level. In an extreme case, if enough flights were cancelled such that demand drops below capacity, then the GDP would be cancelled immediately and realized delay would be zero for the flights that were not cancelled. However, we could not claim that the GDP is unpredictable. If the GDP were planned with perfect foresight, then the realized delay would equal the planned delay if no flights were cancelled, which could have led to perfect predictability. The predictability metric defined here aims to evaluate accuracy of the planned GDP decisions compared to the decision that could have been made with perfect information. Therefore, we need to exclude the cancellation effect from the analysis. To do so, when calculating realized delay, all the cancelled flights are assumed to arrive at the end of the queue in sequence, restoring the delays resulting from cancellations.

6.1.4 Equity

In practice, all the flights estimated or scheduled to arrive during the GDP time horizon are assigned CTA/CTD, but not all of them are assigned delays. The traffic

specialist who constructs the program can apply any combination of the following exemption criteria:

- Airborne, flights in the air are always exempted;
- Geographical scope: tier-based scope or distance based scope;
- Departure status. For instance, one GDP could exempt flights that are taking off within 45 minutes of the GDP report time;
- Flight-specific;
- Airport-specific, to relieve congestion at that airport;
- En route center specific;
- Subject to prior ground stop;

The most common combination is geographical scope and departure status, in addition to exempting airborne flights. The geographical scope, or GDP scope, is an important GDP planning parameter. Flights that are geographically outside the scope are exempted in terms of delay. Research has shown that reducing GDP scope enables airport capacity to be utilized more efficiently under uncertainty. On the other hand, the choice of GDP scope also affects equity. Full scope leads to the most equitable GDP. When the scope is reduced so that certain flights are exempted from the GDP, program equity decreases. For the same reason, the program is more equitable when no flights are exempted by the criteria of departure status. If we give exemption to flights taking off within, say, 30 minutes at the GDP report time, then delay would be absorbed by a reduced number of flights and equity degrades.

Given this, we define equity by comparing how much ground delay an airline is planned to receive under the planned GDP scope to how much delay an airline would have received assuming no flights were exempted from the GDP except flights in the air. Rather than absolute delay quantities, we compare the ratios of delay shares to flight shares. As mentioned earlier, original GDP plans are usually revised later when there is change in weather or demand level. In general, called rates and GDP end time would be updated, whereas GDP scopes are unchanged. Our equity metric measures the equity level of the first GDP plan and is not affected by later actions. Before introducing the equity metric, we will first introduce how we estimate the delay shares in the planned GDP and in the hypothetical GDP assuming no exemption.

The ground delay an airline is planned to receive in the GDP is measured with respect to the arrival traffic that the airline contributes at the GDP-affected airport. A normalized ground delay share of airline k then could be expressed as the ratio of the fraction of ground delay airline k absorbs to the fraction of arrival traffic it contributes:

$$P_p{}^k = \frac{GD_k/GD}{m_f^k/m_f}$$

where,

$P_p{}^k$ is planned traffic-normalized ground delay share for airline k in the original GDP plan;

GD_k is the amount of ground delay assigned to all the flights owned by airline k in the first GDP plan; GD is the total planned system delay, sum of GD_k over all the airlines;

m_f^k is the number of flights involved in the first GDP plan in airline k ; m_f is the total number of affected flights in the original GDP plan. m_f^k/m_f is then the airline's share of all the arrivals, no matter if they are exempted or not.

${P_p}^k$ may be larger than one if airline k's ground delay share exceeds its traffic share. To estimate equity, we compare ${P_p}^k$ to the airlines' hypothetical ground delay share in the baseline—assuming no exemptions by GDP scope or departure status. Define ${P_b}^k$ as normalized percent of delay for airline k, which is the ratio of the fraction of ground delay airline k could have received in the baseline to the fraction of traffic it contributes:

$$ {P_b}^k = \frac{{GD'}_k/GD'}{m_f^k/m_f} $$

where, ${GD'}_k$ is the amount of ground delay airline k could have received assuming no exemption; GD' is the total amount of ground delay. The denominator is the same as in the normalized planned delay share.

The normalized delay shares are defined at the airline level. To evaluate the equity performance at system level, we define the metric as:

$$ \alpha_{eq} = \frac{1}{\exp(\frac{1}{m_a}\sum_k^{m_a} |{P_p}^k - {P_b}^k|)} $$

where, m_a is the number of flight operators affected by the GDP. The exponential function is employed to amplify the minor differences between different GDPs. The average absolute difference in normalized ground delay share could be larger than one, if the exempted flights in the plan are from a small group of flight operators rather than spread more evenly among all operators. The reciprocal of the exponent will bound the equity value between zero and one. When each airline receives the same share of planned ground delay as their share of ground delay under the non-exemption scenario, the program is the most equitable and thus the value of metric is equal to one. If airlines receive different shares of ground delay in the plan and in the baseline, equity will be less than one. Table 1 illustrates the calculation process of equity metric using nominal traffic shares and delay shares. As shown, the normalized ground delay shares, ${P_p}^k$ and ${P_b}^k$, could be larger than one.

Table 5: Illustration of Equity Metric Calculation

Flight operator	Arrival traffic share	Delay share in the initial GDP plan	Delay share in the non-exemption scenario	${P_p}^k$	${P_b}^k$	$\lvert{P_p}^k - {P_b}^k\rvert$
1	0.25	0.25	0.25	1	1	0
2	0.3	0.35	0.2	1.17	0.67	0.5
3	0.25	0.2	0.4	0.8	1.6	0.8
4	0.2	0.2	0.15	1	0.75	0.25
					Sum:	1.55
					m_a:	4
					α_{eq}:	0.68

6.1.5 Flexibility

Flexibility should reflect the ability of the system to permit users to adapt their operations to changing conditions. If all the GDP-affected flights were owned by a single flight operator, then this operator could have maximum flexibility in the CDM process

and spend less time in coordination. When there are multiple flight operators, each operator has few options to adjust their operations to evolving circumstances. Flexibility of GDPs is then positively correlated with the level of operation concentration and could be measured using Herfindahl-Hirschman Index (HHI), which is defined as the sum of the squared operation share of each airline involved in the GDP:

$$\alpha_f = \sum_i (\frac{N_i}{N_T})^2$$

where,

N_i is the number of affected flights from airline i in the first GDP plan;

N_T is the total number of affected flights in the first GDP plan.

During the CDM process, substitution and cancellation decisions are made mainly by the parent carriers rather than the regional affiliate. Therefore, our flexibility metric is estimated with respect to the parent carriers. Like equity, flexibility is estimated based on the initial GDP plan and will not be updated later. When only one flight operator is involved in the program, the value of the metric is one. Such a case should not be common given the competition in the aviation industry. As airline operation shares at an airport are fairly stable, so should be the shares of flights affected by a GDP at the airport. The within-airport variation in the flexibility metric is thus expected to be relatively small compared to the other metrics.

6.2 Data and Metric Evaluation

Using the metrics defined in the previous section, we will assess historical GDP performance at SFO and EWR for 2006 and 2011. We chose these airports because they are typically the top two airports for GDPs. And they are very different in causes for GDPs. At SFO, GDPs are often implemented due to foggy weather; at EWR, convective weather, such as wind and thunderstorm, has been the main weather problem. We selected two years that are 5 years apart to assess the change in performance across time. Two databases are used for this analysis: Traffic Flow Management System (TFMS) Aggregate Demand Lists (ADLs) and Aviation System Performance Metrics (ASPM). The ADL data was provided by Metron Aviation. Main data fields that are referred in the performance analysis are summarized in Table 2.

The ADL data contains two types of information: GDP parameters and individual flight information. One GDP could have several GDP events, each of which corresponds to the issuance of a new or revised GDP plan. GDP parameters include the report time and start and end time for each GDP event, which are essential for performance analysis. For each flight affected by the GDP, we have a snapshot of the flight's status at every GDP implementation or revision, as well as a summary record of the flight's final disposition. Key individual data that we used in our computation of our metrics include OCTA/OCTD, CTA/CTD, ARTA/ARTD, and IGTA/IGTD. Detail descriptions on the data are provided in Table 2. Using the ADL data, we are able to track the status of the program and of each flight as they change over time. It is worth mentioning that the CTA/CTD fields can change at any time during the program due to an airline flight substitution, GDP compression, or a GDP revision. Quarter-hour rates issued in GDP events are also available from the ADL. However, the GDP rates are predictions, and may not equal the arrival acceptance rates that are actually called. The latter are thus obtained from quarter-hour ASPM data.

Table 6: Data Fields Used in the Historical GDP Performance Assessment

Data type	Data field	Description
GDP parameter	Event report time	When the GDP event is reported to flight operators; report time of the first event is GDP report time
GDP parameter	Event start/end time	When the event starts/ends. Flights scheduled to arrive between the event start and end time will be involved in this GDP event. Start time of the first event is GDP start time. End time of the last event is GDP end time. GDP end time is not necessary to be the largest GDP event end time
GDP parameter	GDP purge time	GDP is cancelled at GDP purge time. Most of the time, GDP purge time is earlier than GDP end time, or say the last event end time.
Individual flight	BETA/BETD	Base Estimate Time of Arrival/Departure, a snapshot of the flights ETA/ETD just before it was affected by the GDP
Individual flight	CTA/CTD	Controlled Time of Arrival/Departure, arrival/departure time slot assigned to a flight in the GDP; CTA/CTD gets updated continuously upon GDP revisions
Individual flight	OCTA/OCTD	Original Controlled Time of Arrival/Departure, the first CTA/CTD for a GDP flight for a given GDP event
Individual flight	ARTA/ARTD	Actual Runway Time of Arrival/Departure
Individual flight	IGTA/IGTD	Initial Gate Time of Arrival/Departure; IGTA is used to determine the sequence of time slots in a GDP
ASPM, quarter-hour	ARR_CT	Actual arrivals
ASPM, quarter-hour	ARR_RATE	Actual airport acceptance rate

Capacity utilization is estimated based on ASPM quarter-hour data. For each quarter, actual airport acceptance rate is considered as the count of arrivals that could have been planned for that quarter assuming perfect information, which is the denominator of the capacity utilization metric. The metric value is then calculated as the ratio of the sum of actual arrivals to the sum of actual acceptance rates, over the GDP time horizon between GDP start time and GDP end time.

ADL data is used for estimating the metrics for the other four performance goals. For flights that have ever been affected by the GDP, efficiency is calculated as the ratio of the sum of realized ground delays to the sum of realized total delays. For each flight, ground delay is the difference between ARTD and BETD, and total delay is the difference between ARTA and best estimated time of arrival. After the implementation of the GDP, the flight plan may be further changed due to adverse weather en route and other TMIs. For instance, flight time may increase if the flight is detoured due to a thunderstorm. BETA is the snapshot of estimated time of arrival before the flight becomes active in the GDP and it is not updated upon more information. On the contrary, CTD/CTA fields are frequently updated to reflect operational changes. Therefore, we use the difference between CTA and CTD to estimate the unimpeded flight time after takeoff.

In predictability, flight delay is measured with respect to the original flight schedule. Planned delay is calculated as the sum of the differences between OCTA and IGTA for all the flights that are delayed in the first GDP plan. More flights would be delayed when there is a GDP revision. All the flights that have ever been delayed in the GDP are considered when calculating realized system delay. For flights that are cancelled, the hypothetic ARTA is estimated assuming the flights were put at the end of the queue. Realized delay then is obtained as the sum of the differences between ARTA and IGTA.

In contrast to capacity utilization, predictability, and efficiency, equity and flexibility metrics are based entirely on the initial GDP plan data: neither updates to the program nor realized operational outcomes affect these metrics. Planned ground delay for each flight in

the GDP is assigned as the difference between OCTA and BETA. In the no-exemption scenario, time slots that are assigned in the planned GDP to GDP-affected flights that are still on the ground are reassigned to these flights according to RBS. This assumes all the flights that are scheduled to arrive during the GDP time horizon and are still on the ground at GDP report time would share delay together. The hypothetical ground delay in the non-exemption scenario is then measured as the difference between the new time slot and BETA. Delay shares of each airline are then calculated, and together with airline flight shares, used to calculate the values of $P_p{}^k$ and $P_b{}^k$ in the equity metric equation (see Section 6.1.4). In the calculation of equity and flexibility metrics, we identify a flight operator by its major carrier since decisions on slot swapping and substitution are usually made by the parent carriers who own the slots rather than its regional feeder operators.

References

Andreatta, G., Brunetta, L., Guastalla, G., 2000. From ground holding to free flight: an exact approach, Transportation Science 34(4), 394-401.

Ball, M.O., Hoffman, R., Mukherjee, A., 2010. Ground delay program planning under uncertainty based on the ration-by-distance principle, Transportation Science 44 (1), 1-14.

Ball, M.O., Hoffman, R., Odoni, A., Rifkin, R., 2003. A stochastic integer program with dual network structure and its application to the ground-holding problem, Operations Research 51(1), 167-171.

Ball, M.O., Lulli, G., 2004. Ground delay program: optimizing over the included flight set based on distance, Air Traffic Control Quarterly 12, 1-25.

Ball, M.O., Vossen, T., Hoffman, R., 2001. Analysis of demand uncertainty effects in ground delay programs, 4th US/Europe Air Traffic Management R&D Seminar, Santa Fe, New Mexico.

Bertsimas, D., Patterson, S.S., 2000. The traffic flow management rerouting problem in air traffic control: a dynamic network flow approach, Transportation Science 34(3), 239-255.

Bertsimas, D., Lulli, G., Odoni, A., 2011. An integer optimization approach to large-scale air traffic flow management, Operations Research 59(1), 211-227.

Bolczak, C.N., Hoffman, J.H., Jensen, A.J., Trigeiro, W.W., 1997. National airspace system performance measurement: overview, MITRE Technical Report MTR 97W0000035, Center for Advanced Aviation System Development, McLean, Virginia.

Bradford, S., Knorr, D., Liang, D., 2000. Performance measures for future architecture, Proceedings of 3rd USA/Europe Air Traffic Management R&D Seminar, Napoli, Italy.

Churchill, A.M., 2010. Coordinated and robust aviation network resource allocation, PhD Thesis, University of Maryland, College Park, Maryland.

Clarke, J.B., Solak, S., Chang., Y.H., Ren, L. L., Vela, A.E., 2009. Air traffic flow management in the presence of uncertainty, Proceedings of 8th USA/Europe Air Traffic Management R&D Seminar, Napa, California.

Cook, L.S., Wood, B., 2009. A model for determining ground delay program parameters using a probabilistic forecast of stratus clearing, 8th USA/Europe Air Traffic Management R&D Seminar.

Daganzo, C.F., 1997. Fundamentals of Transportation and Traffic Operations, Elsevier Science Inc., New York.

FAA, 2004. Airport Capacity Benchmarking Report 2004. Washington, D.C..

Gosling, G.D., 1999. Aviation system performance measures, Nextor Working Paper UCB-ITS-WP-99-1.

Grabbe, S., Sridhar, B., Mukherjee, A., 2009. Integrated traffic flow management decision making, AIAA Guidance, Navigation and Control Conference, Chicago.

Hoffman, R., Ball, M.O., 2000. A comparison of formulations for the single-airport ground-holding problem with banking constraints, Operations Research, 48 (4), 578-590.

Kim, A.M., 2011. Collaborative resource allocation strategies for air traffic flow management, PhD Thesis, University of California, Berkeley, California.

Kotnyek, B., Richetta, O., 2006. Equitable models for the stochastic ground holding problem under collaborative decision making, Transportation Science 40 (2), 133-146.

Liu, P.B., Hansen, M., Mukherjee, A., 2008. Scenario-based air traffic flow management: from theory to practice, Transportation Research Part B 42, 685-702

Liu, Y., Hansen, M., in press. Ground Delay Program performance evaluation, Transportation Research Record.

Mukherjee, A., Hansen, M., 2007. A dynamic stochastic model for the single airport ground holding problem, Transportation Science 41(4), 444-456.

Mukherjee, A., Hansen, M., Grabbe, S., 2009. Ground delay program planning under uncertainty in airport capacity, AIAA Guidance, Navigation, and Control Conference Chicago.

Odoni, A.R., 1987. The flow management problem in air traffic control. In: Odoni, A.R., Bianco, L., Szego. G. (Eds.), Flow Control of Congested Networks, Springer-Verlag, Berlin, 269-288.

Richetta, O., 1991. Ground holding strategies for air traffic control under uncertainty, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Mass.

Sridhar. B., Grabbe. S.R., Mukherjee, A., 2008. Modeling and optimization in traffic flow management, Proceedings of the IEEE 96(12), 2060-2080.

Vranas, P., Bertsimas, D., 1994a. The multi-airport ground holding problem for air traffic control, Operations Research 42(2), 249-261.

Vranas, P., Bertsimas, D., Odoni, A.R., 1994b. Dynamic ground-holding policies for a network of airports, Transportation Science 28, 275-291

Xiong, J., 2010. Revealed preference of airline's behavior under air traffic management initiatives, PhD Thesis, University of California, Berkeley, California.

Appendix A: Notations for the Theoretical Models

λ: scheduled arrival rate.
λ_e: exempted demand rate.
$C_H(C_L)$: planned high (low) airport acceptance rate.
T: expected weather clearance time.
T_2: planned delay clearance time.
τ: actual weather clearance time.
τ_2: ideal delay clearance time if perfect information were available at the decision time.
$t_{max}(t_{min})$: upper (lower) bound of the actual clearance time.
$F_{max}(F_{min})$: upper (lower) bound of the flight time.
F_{Scope}: maximum flight time of the GDP affected flights.
$d^e_{P,max}$: maximum planned flight delay with exemption rate as λ_e.
S: scheduled cumulative arrival curve.
N: planned cumulative arrival curve.
N_e: planned cumulative arrival curve with exemption rate as λ_e.
Th: realized cumulative arrival curve.
A: ideal cumulative arrival curve if perfect information were available at the decision time.
$D(D')$: available cumulative arrival demand (demand rate) if capacity permits and GDP is cancelled earlier.
$D_-(D'_-)$: available cumulative arrival demand (demand rate) of Type I flights.
$D_0(D'_0)$: available cumulative arrival demand (demand rate) of Type II flights.
$D_+(D'_+)$: available cumulative arrival demand (demand rate) of Type III flights.
C_-: planned cumulative arrival curve for Type I flights in the case of GDP extension.
D_P: planned delay in the original GDP.
D_R: realized delay at the end of the GDP.
AD: realized airborne delay in the case of GDP extension at the end of the GDP.
GD: realized ground delay at the end of the GDP.

Appendix B: The Cumulative Arrival Curves for Type I flights, Late Clearance

1. For Plot b) in Figure 7:
Condition:

$$T_2 \leq T + F_{min} \leq \tau_2$$

Formulation:

$$C_- = \begin{cases} C_L \cdot t, t \leq T \\ C_L \cdot T + C_H \cdot (t - T), T < t \leq T_2 \\ \lambda \cdot t, T_2 < t \leq T + F_{min} \\ -\frac{\lambda}{2\Delta F} \cdot [t - (T + F_{max})]^2 + \lambda \cdot (T + F_{min}) + \frac{\lambda}{2} \cdot \Delta F, T + F_{min} \leq t \leq T + F_{max} \\ \lambda \cdot (T + F_{min}) + \frac{\lambda}{2} \cdot \Delta F, t > T + F_{max} \end{cases}$$

2. For Plot c) in Figure 7:
Condition:

$$T + F_{min} \leq T_2 \leq T + F_{max}$$

Formulation:

$$C_- = \begin{cases} C_L \cdot t, t \leq T \\ C_L \cdot T + C_H \cdot (t - T), T < t \leq T + F_{min} \\ -\frac{C_H}{2\Delta F} \cdot [t - (T + F_{max})]^2 + C_H \cdot \left(F_{min} + \frac{\Delta F}{2}\right) + C_L \cdot T, T + F_{min} \leq t \leq T_2 \\ -\frac{\lambda}{2\Delta F} \cdot [t - (T + F_{max})]^2 + \frac{\lambda - C_H}{2\Delta F} \cdot [T_2 - (T + F_{max})]^2 + C_H \cdot \left(F_{min} + \frac{\Delta F}{2}\right) + C_L \cdot T, T_2 < t \leq T + F_{max} \\ \frac{\lambda - C_H}{2\Delta F} \cdot [T_2 - (T + F_{max})]^2 + C_H \cdot \left(F_{min} + \frac{\Delta F}{2}\right) + C_L \cdot T, t > T + F_{max} \end{cases}$$

3. For Plot d) in Figure 7:
Condition:

$$T + F_{max} \leq T_2$$

Formulation:

$$C_- = \begin{cases} C_L \cdot t, t \leq T \\ C_L \cdot T + C_H \cdot (t - T), T < t \leq T + F_{min} \\ -\frac{C_H}{2\Delta F} \cdot [t - (T + F_{max})]^2 + C_H \cdot \left(F_{min} + \frac{\Delta F}{2}\right) + C_L \cdot T, T + F_{min} \leq t \leq T + F_{max} \\ C_H \cdot \left(F_{min} + \frac{\Delta F}{2}\right) + C_L \cdot T, t > T + F_{max} \end{cases}$$

APPENDIX IV

Strategic Opportunity Analysis in *COuNSEL* – A Consensus-Building Mechanism for Setting Service Level Expectations

Prem Swaroop, Michael O. Ball
Robert H. Smith School of Business and Institute for Systems Research, University of Maryland, College Park, MD 20742.
pswaroop@rhsmith.umd.edu mball@rhsmith.umd.edu

COuNSEL has been accepted as a technically viable consensus-building mechanism for the stated problem – although many practical challenges still remain before it may be deployed in operations. A key issue worthy of further investigation is its strong strategy-resistance – as claimed by the authors of Majority Judgment, the voting procedure embedded in *COuNSEL*. Using the broad ideas of Nash Equilibria, we characterize the necessary and sufficient conditions for an airline to benefit from unilaterally deviating from truthfully grading one or more candidates. The framework provides the airline with all the other airlines' grades on a set of candidates, and allows it an opportunity to present new grades. The analysis is repeated over multiple instances, and likelihood of beneficial strategic opportunity is presented.

1. Introduction

COuNSEL is a multi-objective multi-stakeholder consensus-building mechanism that has several desirable properties. It is based on Majority Judgment voting procedure, in which players provide a grade for each candidate in the consideration set, in a common language. The procedure uses the input of grades to compute a Majority Grade for each candidate; the candidate with the highest Majority Grade is deemed winner. Majority Judgment is described by its authors as being highly strategy-resistant (Balinski and Laraki 2011). We wish to verify this claim using simulations.

Our framework is as follows. Assume each player is provided an opportunity to unilaterally change her grade *after* observing everyone else's grades for a given consideration set of candidates. In practice, such opportunity would not exist – and the likelihood of hurting oneself would deter the players from strategic grading. Thus, this analysis provides the worst-case strategy proneness of the procedure.

The core idea behind this framework for analysis is similar to Nash Equilibria. It has origins in mechanism design, particularly in implementation theory (Maskin 1999). Gibbard and Satterthwaite's impos-

sibility theorem established that true incentive-compatibility is not attainable if there is no restriction on the players' preference structure, unless a player is dictatorial. This realization led to investigation of weaker notions of strategy-proofness. Many solution concepts have been studied, e.g., Bayesian and sub-game perfect equilibria; however Nash equilibrium and Pareto optimality have been of particular interest. Such mechanisms are termed Nash implementable. Maskin identified two properties that the social choice rule underlying a mechanism with three or more players must possess in order to be Nash implementable: monotonicity and no veto power. These results were tightened later, and extended to two players (Moore and Repullo 1990) – with potential applications in contracting theory, which invariably deals with two-party settings.

The Nash equilibrium solution concept assumes complete information and allows unrestricted domain of preferences – albeit observing convexity, continuity, and monotonicity. Maskin (1985) provides justifications for using such a complete information solution concept for an inherently incomplete information process like these social choice rule mechanisms. First, by definition, Nash equilibrium is a fixed-point among players' strategy spaces. Thus, it represents a stationary point in a process whereby players (with incomplete information) iteratively adjust their preference elicitations, until no unilateral deviation from true preferences benefits any player. Second, Nash equilibrium is a fitting solution concept in cases where the planner has incomplete information (or may not even exist), but the players are well-informed about each others preferences, such as in committee decisions.

Given that complete strategy-proofness is ruled out in any mechanism, it is of interest to quantify the extent of manipulability. This is particularly important in our case, as Majority Judgment is not a traditional voting procedure, and is therefore not as well-studied. Moreover, we intend to use weights for the players, and not the traditional "one person-one vote" setting. Of course, no single player will be given 50% or more of the total weight over all players to disallow dictatorial powers. However, this uneven distribution of decision power is worthy of investigation with regard to strategy-proneness. Finally, while unrestricted domain is of interest in itself, it would be useful to compare against a scenario where the players' preferences are convex.

Untruthful or strategic grading by a player may take several forms. She may increase the grade of one or more candidates, and / or decrease the grade of one or more candidates, possibly leaving grades on some candidates unchanged. Strategic grading is beneficial to a player only if the majority judgment winner is replaced by a candidate that she regards more preferable to it. Indeed, strategic grading can

hurt the player if the new winner is less preferred by her than the existing winner. Or, it may not yield any change to the existing winner.

Some consideration sets may inherently be more manipulable than others – depending on the number of players, their grades, and number of candidates. Proportion of manipulable candidates to the total number of candidates is one measure of strategy-proneness. However, that does not imply that each such candidate can be manipulated by all the players. Some players may not have any candidate that they prefer over the current winner – these players will not have an incentive to deviate unilaterally. Among the remaining players, there may be some for whom there are no beneficial opportunities for the candidates that they prefer more than the current winner. These players too would not deviate unilaterally and benefit themselves. The proportion of the players that have any opportunity to benefit from strategic grading is a second measure of strategy-proneness. Another measure of strategy-proneness is the proportion of the total number of such beneficial player-candidate combinations.

Section 2 intuitively describes the procedure to identify strategic opportunity within this framework, using an illustrative example. Section 3 formalizes the description, and exhaustively identifies the necessary and sufficient conditions for beneficial strategic opportunities for a player. The measures for strategy-proneness, or manipulability, are also formally defined. Results from simulations for six types of scenario configurations are presented in Section 4. The first three allow the players unrestricted domain in grading; that is, no preference structure is imposed on the players. The latter three impose a convex grading function for each player. The three scenarios with these two assumptions on preference structures that were simulated are: players have equal weights, 5 players with differential weights, and 25 players with differential weights. The very first scenario, namely players have equal weights, and are allowed unrestricted domain in grading, is the basic Majority Judgment procedure. The last scenario, namely 25 differentiated players with a convex preference structure, is closer to the proposed *COuNSEL* procedure. The progression from the basic Majority Judgment to the last scenario is instructive. Section 5 concludes.

2. Illustration

Suppose five players (of equal weight) provide grades to three candidates as summarized in Table 1a. The grades are unrestricted, that is, no structure is imposed on the preferences. Of course, the grades should be within the allowable range – in this case in $[0 \dots 1]$. The grades are sorted for each candidate,

Player	$\mathbf{m}_1$	$\mathbf{m}_2$	$\mathbf{m}_3$
1	0.6	0.3	0.2
2	0.1	0.3	**0.5**
3	0.1	0.6	0.6
4	**0.2**	0.7	0.3
5	0.8	**0.4**	0.7

(a) Grades provided by five players to three candidates

	$\mathbf{m}_1$	$\mathbf{m}_2$	$\mathbf{m}_3$
	0.1	0.3	0.2
	0.1	0.3	0.3
M.G.	0.2	0.4	0.5
	0.6	0.6	0.6
	0.8	0.7	0.7

(b) Grades in increasing order for each candidate

Player	$\mathbf{m}_1$	$\mathbf{m}_2$	$\mathbf{m}_3$
1	[0.1 . . . 0.2]	[0.4 . . . 0.6]	[0.5 . . . 0.6]
2	[0.2 . . . 0.6]	[0.4 . . . 0.6]	[0.3 . . . 0.6]
3	[0.2 . . . 0.6]	[0.3 . . . 0.4]	[0.3 . . . 0.5]
4	[0.1 . . . 0.6]	[0.3 . . . 0.4]	[0.5 . . . 0.6]
5	[0.1 . . . 0.2]	[0.3 . . . 0.6]	[0.3 . . . 0.5]

(c) Each players' manipulable range for each candidate.

Table 1 Illustrative example.

and presented in Table 1b. The majority grades are marked as "M.G.". The candidate $\mathbf{m}_3$ is the winner in this example.

We highlight several observations relevant to unilateral grading decisions. First, not all players have an incentive to deviate, as the consideration set does not have better candidate for them. In the example, players 2 and 3 are such players.

Second, to influence the majority grade of any candidate, a player has to grade *towards* its majority grade. In other words, if her grade for a particular candidate is higher (lower) than the current majority grade, then her new grade for it must be smaller (greater) than her current grade to have any chance to change the majority grade. This also implies that if her grade is higher (lower) than the current majority grade, then she can only decrease (increase) the new majority grade. If her grade is same as the majority grade for the candidate, then she can influence it upwards or downwards. Player 1 in this example clearly does not like the current winner, and would rather prefer $\mathbf{m}_1$ as the winner. However, her decreasing the grade on $\mathbf{m}_3$ will not change its majority grade – nor would increasing her grade on $\mathbf{m}_1$. The only way for her to change the new majority grade for $\mathbf{m}_1$ is to decrease her new grade on it, resulting in a lower majority grade; the opposite holds for $\mathbf{m}_3$.

The third observation relates to the extent of strategic grading opportunity available for a given candidate. A player can unilaterally influence the majority grade of a candidate within a specific range determined by the ordering of the grades provided by all the players. If player 1's new grade for $\mathbf{m}_3$

is below the current majority grade of 0.5, the majority grade remains at 0.5. Any grade between 0.5 and 0.6 would become the new majority grade, but any higher than 0.6 would not increase it beyond 0.6. Thus, player 1's "manipulable" range for $\mathbf{m}_3$ is [0.5,0.6]. Similarly for $\mathbf{m}_1$, a new grade by player 1 above the current majority grade of 0.2 will not have any impact. Any grade between 0.1 and 0.2 would become the new majority grade, any lower than 0.1 would keep it 0.1. Player 1's manipulable range for $\mathbf{m}_1$ is [0.1,0.2].

Clearly, player 1 has no opportunity to make her most preferred candidate $\mathbf{m}_1$ as the winner in this example. The fourth observation is regarding comparative grading over multiple candidates. Following the last two observations for $\mathbf{m}_2$, player 1 can only increase its majority grade, and that increase is bounded between 0.4 and 0.6. The range of grades between 0.5 and 0.6 overlaps with that of her manipulable range of $\mathbf{m}_3$, the current winner. Thus, player 1 can provide new grades for the two vectors $\mathbf{m}_2$ and $\mathbf{m}_3$ within [0.5,0.6] such that the grade for the former is less than that of the latter. This would make $\mathbf{m}_2$ the new winner, which she prefers over the current winner $\mathbf{m}_3$. The manipulable ranges for each candidate for the players who have an opportunity to benefit from strategic grading are reported in Table 1c.

Building on the previous observation, the fifth observation characterizes strategy-proneness of a given candidate for a player. A candidate is prone to (beneficial) strategy only if its manipulable range has an overlap with that of the current winner for any player. $\mathbf{m}_1$'s manipulable ranges for players 1 and 5 have no such overlap, similarly $\mathbf{m}_2$'s manipulable range for player 4 has no such overlap with those of the winner.

The sixth observation is about the relative position of a player's grade for a candidate with respect to its majority grade – in relation to those of the winner. When the player's grade is not same as majority grade for a candidate, its relation to the majority grade should be same as that for the winner. For player 1, the grade (0.2) for the winner $\mathbf{m}_3$ is below the majority grade (0.5). This is also true for $\mathbf{m}_2$: her grade (0.3) is below the majority grade (0.4) – but not for $\mathbf{m}_1$. The former is manipulable, but the latter is therefore not. The converse also holds, though there is no instance in this example. Such an opportunity also exists when a player provides the same grade as the majority grade for a candidate, and grades the winner lower than its majority grade. For example, player 4 grade for $\mathbf{m}_1$ is its majority grade, while she grades lower (0.3) than the majority grade for the winner (0.5). Another case is when a player grades the same as majority grade for the winner, and has a higher grade for a candidate

than its majority grade. There is no instance in this example of this happening. These relationships are established formally in a later section.

Seventh, at an overall level, a candidate would not yield any benefit to any player if no player has an overlap of its manipulable range with that of the winner. In this example, all the candidates have an overlap with the winner's. Consider a candidate whose sorted grades are: {0.1,0.15,0.2,0.25,0.8}. Its manipulable range for any player has to be within [0.15,0.25], while the winner's has to be within [0.3,0.6]. Indeed, any candidate for which the grade just above the majority grade (the second highest grade in this example) is lower than the grade just below the winner's majority grade will not yield any benefit to any player. Each candidate in the consideration set should be pre-screened using this observation before analyzing at player-level.

Measures for Strategy-Proneness. Let us analyze the example with regard to strategy proneness. As just noted, all of the candidates (100%) in the consideration set are potentially manipulable. However, that does not mean that each player can unilaterally manipulate the grades to benefit.

We already identified that player 1 can benefit by manipulating $\mathbf{m}_2$ and / or $\mathbf{m}_3$. Also, we noted that the players 2 and 3 already have their most-preferred candidate in the current winner $\mathbf{m}_3$ – and hence do not have incentive to manipulate. Player 4 has an overlap between the manipulable ranges for $\mathbf{m}_1$ and $\mathbf{m}_3$ – but its preference for $\mathbf{m}_1$ being lesser, it has no incentive to manipulate these. There is no overlap for its most preferred candidate $\mathbf{m}_2$ with $\mathbf{m}_3$. Thus, player 4 actually has no opportunity to strategically grade that might benefit her. Similarly, player 5 has only an opportunity with $\mathbf{m}_2$, but since it prefers it less than the $\mathbf{m}_3$, it cannot benefit by manipulating her grades.

Thus, of the five players, only one – 20% – has a beneficial strategic opportunity. Among the 15 player-candidate opportunities, only two – about 13% – are beneficial to any player.

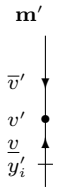
3. Conditions for Beneficial Strategic Grading

We formalize the observations regarding beneficial strategic grading opportunities for a player i with respect to a candidate $\mathbf{m}'$, whose majority grade is v'. For ease of exposition, the analysis and development begins with the equally weighted players case, that is, where all the players have the same weight. We relax this restriction later in the section, and explain the approach for the more general case of differentially weighted players.

3.1. Equally Weighted Players

Sorted in increasing order, the grade just before the majority grade is denoted $\underline{v}'$, and the grade just after the majority grade as $\overline{v}'$. Player i's grade for $\mathbf{m}'$ is denoted y_i'. Denote the winning candidate as $\mathbf{m}^*$, and the notation regarding it replaces the prime ($'$) with asterisk (*) in above.

A simple line diagram is used extensively in this section, it is explained below.



A candidate $\mathbf{m}'$ is depicted with a vertical bar, which represents the allowable grading range as per the common grading language. The majority grade v' is marked with a circle, and the two neighboring grades $\underline{v}'$ and $\overline{v}'$ are marked with upwards and downwards pointing arrowheads. Player i's grade for the candidate is marked with a horizontal tick marks.

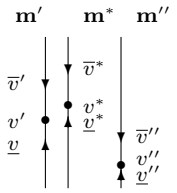
For any strategic grading by i for $\mathbf{m}'$ that changes its majority grade, the manipulable ranges are defined as below.

$$\delta_i' = \begin{cases} [v', \overline{v}'] & \text{if } y_i' < v' \\ [\underline{v}', v'] & \text{if } y_i' > v' \\ [\underline{v}', \overline{v}'] & \text{if } y_i' = v' \end{cases}$$

Looking at each candidate against the winner, the overlap of manipulable ranges between $\mathbf{m}'$ and $\mathbf{m}^*$ is a necessary condition:

$$\overline{v}' > \underline{v}^* \tag{1}$$

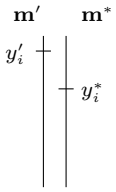
For instance, in the following, $\mathbf{m}'$ is potentially manipulable, but $\mathbf{m}''$ cannot be beneficially manipulated by any player.



The proportion of the candidates in the consideration set that meet the conditions of (1) gives an idea of strategy proneness of the setting at an overall level. A strategy-proof consideration set would

have no candidate with such an overlap – although it would be quite unlikely in practice. Indeed, as the seventh observation in Section 2 implied, this would be an overly strong measure, and an investigation of player-wise opportunities is required for a better and tighter quantification of strategy-proneness.

At a player level, a necessary condition for player i to strategically grade $\mathbf{m}'$ is that she grades it higher than she does the winner: $y'_i > y^*_i$.



This is not sufficient, as noted in the observations. Specific relationships among her grades for $\mathbf{m}'$ and $\mathbf{m}^*$ are required. We examine all possible relationships in Table 2, and summarize the necessary and sufficient conditions.

Combining cases 1 and 9, we see that among candidates that have: $(y'_i \leq v')$ & $(y^*_i < v^*)$, if there exists a candidate with $\overline{v}' \geq v^*$, then player i could increase its grade to anywhere in $(v^*, \overline{v}']$ without changing grades of the rest of the candidates. This is also a sufficient condition for a beneficial strategic grading opportunity for i, as she can only manipulate her grade for a single candidate and benefit herself. Of course, if multiple candidates meet the conditions, then she could only manipulate the candidate that she grades highest amongst these. Hence, one sufficient condition is:

$$(y'_i > y^*_i) \;\&\; (y^*_i < v^*) \;\&\; (y'_i \leq v') \;\&\; (\overline{v}' \geq v^*)$$

Cases 2 and 8 can be combined as: $(y'_i > v')$ & $(y^*_i \geq v^*)$. A candidate could be potentially graded strategically to benefit i if $v' \geq \underline{v}^*$ is also met. However, this is not a sufficient condition. For, the required strategy is to down-grade the winner *as well as* any other candidates whose majority grade lies between v' and v^*, so that their majority grade becomes lower than v'. Such candidates may not be manipulable by the player i. Some more screening conditions need to be added in this case.

First, recall that any candidate with $y''_i < v''$ cannot be manipulated by i so as to reduce its majority grade. Thus, the highest majority grade among such candidates, say v''^{max} forms a bound below which i cannot reduce the majority grade of the other candidates. For example, examine the following consideration set. Player i prefers $\mathbf{m}'$ the most. Candidate $\mathbf{m}^*$ is currently winning. Now, i can reduce its majority grade down to $\underline{v}^*$, but this will make $\mathbf{m}''$ as the new winner, not $\mathbf{m}'$. While $\mathbf{m}''$ is much to her dislike, she cannot influence its majority grade downwards.

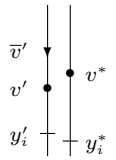
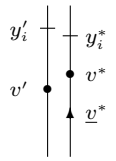
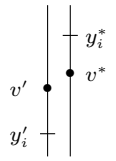
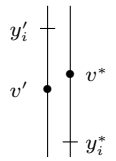
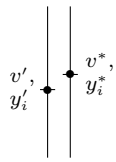
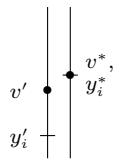
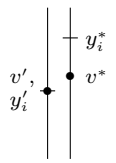
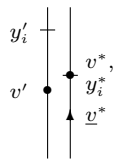
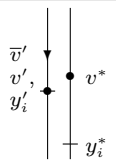
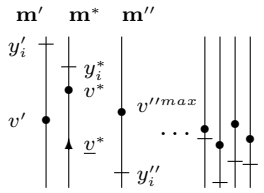
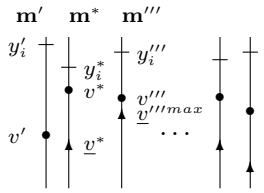
Case	Relative Positions	$\mathbf{m'}$	$\mathbf{m^*}$	Required Condition	Manipulable Range
1.	$(y_i' < v') \;\&\; (y_i^* < v^*)$.			$\overline{v}' \geq v^*$	$[v^*, \overline{v}']$
2.	$(y_i' > v') \;\&\; (y_i^* > v^*)$.			$v' \geq \underline{v}^*$	$[\underline{v}^*, v']$
3.	$(y_i' < v') \;\&\; (y_i^* > v^*)$.			NA	
4.	$(y_i' > v') \;\&\; (y_i^* < v^*)$.			NA	
5.	$(y_i' = v') \;\&\; (y_i^* = v^*)$.			NA	
6.	$(y_i' < v') \;\&\; (y_i^* = v^*)$.			NA	
7.	$(y_i' = v') \;\&\; (y_i^* > v^*)$.			NA	
8.	$(y_i' > v') \;\&\; (y_i^* = v^*)$.			$v' \geq \underline{v}^*$	$[\underline{v}^*, v']$
9.	$(y_i' = v') \;\&\; (y_i^* < v^*)$.			$\overline{v}' \geq v^*$	$[v^*, \overline{v}']$

Table 2 **Examination of relative positions between the majority grade and a player's grade for a non-winner candidate m' and the winner m^*.**



Secondly, for a candidate with $y_i''' \geq v'''$, she could reduce its majority grade to $\underline{v}'''$. If she were to down-grade all of these candidates, the highest majority grade among these, say $\underline{v}'''^{max}$ would form a similar bound as above. Pictorially, examine the following consideration set. Player i likes the $\mathbf{m}'$ over the current winner $\mathbf{m}^*$. She could reduce the majority grade of the winner to lower than v', but she would also need to reduce the majority grade of $\mathbf{m}'''$ and other such candidates to make $\mathbf{m}'$ as the new winner. However, the lowest majority grade she can get for all such candidates is $\underline{v}'''^{max}$.



Finally, the two conditions are combined as follows. To decide whether $\mathbf{m}$ can be made the new winner by i, all the remaining candidates are evaluated. Depending on the relative position of her grade with respect to its majority grade, the candidate is marked as one of $\mathbf{m}''$ or $\mathbf{m}'''$. The bounds v''^{max} and $\underline{v}'''^{max}$ are determined, and the higher of these two is taken as $\underline{w}_i'$:

$$\underline{w}_i' = \max(v''^{max}, \underline{v}'''^{max}).$$

If $v' > \underline{w}_i$, then i can make $\mathbf{m}'$ as the winner, otherwise not. This will form the other sufficient condition for beneficial strategic grading:

$$(y_i' > y_i^*) \;\&\; (y_i^* \geq v^*) \;\&\; (y_i' > v') \;\&\; (v' > \underline{w}_i')$$

Putting it all together, the following is the necessary and sufficient condition that allows beneficial strategic grading opportunity to a player via a non-winner candidate $\mathbf{m}'$:

$$(y_i' > y_i^*) \;\&\; \left\{ \Big((y_i^* < v^*) \;\&\; (y_i' \leq v') \;\&\; (\overline{v}' \geq v^*) \Big) \;\Big|\; \Big((y_i^* \geq v^*) \;\&\; (y_i' > v') \;\&\; (v' > \underline{w}_i') \Big) \right\} \qquad (2)$$

The first term in (2) states simply that the player has to prefer an alternate candidate over the winner. The first two terms of the two groups of conditions within the bracket state the relative positioning of

the player's grade with respect to the majority grade for the alternate candidate and the winner. The two groups are mutually exclusive. Note that the beneficial opportunities are only likely if the player's grade is on the *same* side of the majority grade for *both* the candidates. Depending on which side of the majority grade the player's grades fall, specific conditions are required to be met for her to benefit – as stated in the final condition in the two groups of conditions.

In terms of exact strategies, if the player's grades are below the majority grades for both the candidate and the winner, she could simply raise her grade on the alternate candidate all the way to the maximum possible grade, G^{max} (though the majority grade of the candidate would remain at $\overline{v}'$ by her doing so), while keeping the grade on the winner at the same level. Of course, this is the simplest strategy for her; one can imagine several other strategies that would result beneficially to her. For instance, she could increase the grade of the winner too – while ensuring that her grade on the winner is smaller than the grade on the alternate candidate. Or, she could increase the grade of the alternate candidate barely above the winner's majority grade, or may be at any other level above it.

On the other hand, if her grades are both higher than the majority grades, her simplest strategy would be to keep the grade on the alternate candidate at the same level, and give all the other candidates the lowest possible grade. Unlike the previous case, it is necessary that she down-grades the other candidates as well – as just down-grading the winner does not guarantee that the alternate candidate becomes the winner.

Measures for Strategy-Proneness. Suppose the consideration set comprises of M candidates, and there are N players. We formally define the three measures of interest.

1. Likelihood of manipulability of a candidate, φ^C: the number of non-winner candidates that meet condition (1) $\div$ $(M-1)$.

2. Likelihood of manipulability by a player, φ^P: the number of players for whom any candidate meets condition (2) $\div$ N.

3. Likelihood of manipulability of the consideration set, φ^S: the number of player-candidate pairs that meet condition (2) $\div$ $(M \times N)$.

3.2. Differentially-Weighted Players

In *COuNSEL*, the airlines are assigned different weights which are a function of the impact they suffer from the weather. The equally weighted case explained thus far needs four types of modifications to account for the players' weights.

First, the definitions of the Majority Grade v', and its neighbors $\underline{v}'$ and $\overline{v}'$ are modified. Instead of a simple median, a weighted median is sought for identifying v'.

Table 3a provides an example with six players, whose grades for a candidate and their weights are listed. The players are then sorted in the increasing order of their grades, as shown in Table 3b. In this ordered list, the cumulative weights are computed for each player. π is the proportion of each individual player's weight to the total weight (20 in this example). Π is the cumulative proportional weight in the increasing order of grades. The player whose cumulative weight meets or exceeds half the total weight (20/2=10 in this example) provides the majority grade v' – player B in this example. Coincidentally, if the players had equal weight, the majority grade would have been the same – but this need not be the case, as we shall see shortly. The grades just below and above v' are respectively marked $\underline{v}'$ and $\overline{v}'$, as earlier. The majoritarian set in this example is formed by players B, C, D, and F.

Recall that no player is assigned a weight that is larger than half the total weight, to avoid giving it dictatorial powers. This implies that when the players are ordered in increasing order of their grades for a given candidate $\mathbf{m}'$, the weighted majority grade v' is always flanked by at least one grade on either side. That is, with three or more players, $\underline{v}'$ and $\overline{v}'$ are always defined in the differentially weighted case – just like the equally weighted case.

Aside from this modification, the rest of the procedure for determining the winning candidate over a consideration set remains the same. That is, the weighted majority grade is computed for each candidate in the consideration set, and the candidate with the largest majority grade is declared the winner.

The second modification has to do with manipulability of a candidate $\mathbf{m}'$ by a player i with proportional weight π_i, whose grade for $\mathbf{m}'$ is y_i'. Like in the equally weighted case, to influence the majority grade of $\mathbf{m}'$, i has to provide a new grade *towards* v. However, the equally weighted case ensured that each player could influence v – by grading in this fashion, i could move from majoritarian set to the non-majoritarian set and vice-versa. This was possible due to the fact that in the equally weighted case, the majoritarian set is a minimal majority-forming set: if any player moved out, it no more forms the majority. The converse held true for the non-majoritarian set: if any player moved in, it would now have formed a majority.

As weights are "lumpy", this no more holds true for the differentially weighted case. For instance, consider player F in Table 3b. She is currently in the majoritarian set for the given candidate. Hence, she has to provide a grade below the majority grade of 0.24 to influence it downwards – she cannot

Players	Grades	Weights
A	0.15	6
B	0.24	5
C	0.96	4
D	0.33	3
E	0.18	1
F	0.63	1

(a) Example grades

Players	Ordered Grades	Weights	Cumulative Weights	π	Π	
A	0.15	6	6	0.30	0.30	
E	0.18	1	7	0.05	0.35	$\underline{v}'$
B	0.24	5	12	0.25	0.60	v'
D	0.33	3	15	0.15	0.75	$\overline{v}'$
F	0.63	1	16	0.05	0.80	
C	0.96	4	20	0.20	1.00	

(b) Players ordered by grades

Table 3 Weighted Majority Grade example

Ordered Grades	Players	Weights	Cumulative Weights	π	Π	
A	0.15	6	6	0.30	0.30	
E	0.18	1	7	0.05	0.35	
F	0.20	1	8	0.05	0.40	
B	0.24	5	13	0.25	0.65	v'
D	0.33	3	16	0.15	0.80	
C	0.96	4	20	0.20	1.00	

(a) Player F has provided a reduced grade

Ordered Grades	Players	Weights	Cumulative Weights	π	Π	
A	0.15	6	6	0.30	0.30	
E	0.18	1	7	0.05	0.35	
C	0.20	4	11	0.20	0.55	v'
B	0.24	5	16	0.25	0.80	
D	0.33	3	19	0.15	0.95	
F	0.63	1	20	0.05	1.00	

(b) Player C has provided a reduced grade

Table 4 Manipulation in Differentially-Weighted Case: Downwards Revision

increase it by increasing her grade, and any grade above 0.24 also would not change anything. Suppose she provides 0.20, Table 4a is the amended table. Note that the majority grade remains at 0.24, as player F's weight is insufficient to move the new cumulative proportional weight Π to 0.5 or above.

To formalize this observation, denote the cumulative proportional weight of the player that provided

Ordered Grades	Players	Weights	Cumulative Weights	π	Π	
A	0.15	6	6	0.30	0.30	
B	0.24	5	11	0.25	0.55	v'
E	0.30	1	12	0.05	0.60	
D	0.33	3	15	0.15	0.75	
F	0.63	1	16	0.05	0.80	
C	0.96	4	20	0.20	1.00	

(a) Player E has provided an increased grade

Ordered Grades	Players	Weights	Cumulative Weights	π	Π	
E	0.18	1	1	0.05	0.05	
B	0.24	5	6	0.25	0.30	
A	0.30	6	12	0.30	0.60	v'
D	0.33	3	15	0.15	0.75	
F	0.63	1	16	0.05	0.80	
C	0.96	4	20	0.20	1.00	

(b) Player A has provided an increased grade

Table 5 Manipulation in Differentially-Weighted Case: Upwards Revision

the majority grade v' for the candidate $\mathbf{m}'$ as Π'. For the player that provided $\underline{v}'$, it is denoted as $\underline{\Pi}'$; and for the player that provided $\overline{v}'$, it is denoted as $\overline{\Pi}'$. In Table 3b, $\underline{\Pi}' = 0.35$, $\Pi' = 0.60$, and $\overline{\Pi}' = 0.75$.

So, for a player i whose grade $y_i' > v'$, the only way to influence the majority grade would now be qualified by the additional condition that $\underline{\Pi}' + \pi_i \geq 0.5$. Player B in Table 3b could only get 0.35+0.05=0.40, which being less than 0.5, was not sufficient, as seen in the amended Table 4a. Player C, on the other hand, could manipulate its majority grade: 0.35+0.20=0.55 clearly crossed 0.5, as seen in Table 4b.

Conversely, a player with $y_i' < v'$ can influence the majority grade upwards only if $\Pi' - \pi_i < 0.5$. Table 5a shows player E could not influence, as 0.60-0.05=0.55 exceeded 0.5; while player A could do so, because 0.60-0.30=0.30 was below 0.5.

Finally, for a player with $y_i' = v'$, manipulability is possible in either direction, so long as $\underline{v}' < v' < \overline{v}'$. Indeed, even if strict inequality does not hold, manipulation by such a player is possible due to differential weights – as we shall see next.

The third modification has to with the manipulable ranges. With differential weights, it is possible that a player can manipulate the majority grade beyond $\underline{v}'$ and $\overline{v}'$. For instance, see Table 6. In Table 6a, player C reduced her grade further, below that of E – who in the original Table 3b had provided $\underline{v}'$. This caused the new majority grade to become lower than the original $\underline{v}'$. Conversely, player A in Table 6b effectively changed the majority grade above the original $\overline{v}'$. The manipulable ranges are thus

Ordered Grades	Players	Weights	Cumulative Weights	π	Π	
A	0.15	6	6	0.30	0.30	
C	0.16	4	10	0.20	0.50	v'
E	0.18	1	11	0.05	0.55	
B	0.24	5	16	0.25	0.80	
D	0.33	3	19	0.15	0.95	
F	0.63	1	20	0.05	1.00	

(a) Player C has provided a further reduced grade

Ordered Grades	Players	Weights	Cumulative Weights	π	Π	
E	0.18	1	1	0.05	0.05	
B	0.24	5	6	0.25	0.30	
D	0.33	3	9	0.15	0.45	
A	0.60	6	15	0.30	0.75	v'
F	0.63	1	16	0.05	0.80	
C	0.96	4	20	0.20	1.00	

(b) Player A has provided a further increased grade

Table 6 Manipulation in Differentially-Weighted Case: Larger Revisions

not constrained to be within $\underline{v}'$ and $\overline{v}'$.

For player i with $y_i \geq v'$, the lower bound for the manipulable range is given by the grade of the player j with the smallest Π_j, where $\Pi_j + \pi_i \geq 0.5$. Denote this grade as $\underline{u}'_i$ – note that it depends on the particular player i under consideration. Conversely, for player i with $y_i \leq v'$, the upper bound for the manipulable range is given by the grade of the player j with the smallest Π_j, where $\Pi_j - \pi_i \geq 0.5$. Denote this grade as $\overline{u}'_i$.

For any strategic grading by i for $\mathbf{m}'$ that changes its majority grade, the manipulable ranges are defined as below.

$$\delta'_i = \begin{cases} [v', \overline{u}'_i] & \text{if } y'_i < v' \\ [\underline{u}'_i, v'] & \text{if } y'_i > v' \\ [\underline{u}'_i, \overline{u}'_i] & \text{if } y'_i = v' \end{cases}$$

The fourth and final modification updates the necessary and sufficient conditions for manipulability over multiple candidates in the candidate set. The core necessary condition that $y'_i > y^*_i$ remains – i must grade the alternate candidate higher than she grades the winning candidate. The observations made in the first two columns of Table 2 continue to hold – the only ways to benefit from strategic grading via a candidate $\mathbf{m}'$ require the player's grades for both the winner and $\mathbf{m}'$ to be on the same

side of their respective majority grades. Specifically:

$$\text{a. } (y_i^* < v^*) \ \& \ (y_i' \leq v'), \text{ or}$$

$$\text{b. } (y_i^* \geq v^*) \ \& \ (y_i' > v').$$

However, the latter columns need an update, as described above.

Case (a) requires a simpler manipulation – grade for only $\mathbf{m}'$ needs to be increased to make it a winner. The sufficient condition in this case is that there is room for benefit:

$$(y_i' > y_i^*) \ \& \ (y_i^* < v^*) \ \& \ (y_i' \leq v') \ \& \ (\overline{u}' \geq v^*).$$

Case (b) requires a complex manipulation – grades for multiple candidates need to be decreased, to make them all losers against $\mathbf{m}'$. Using a similar approach as developed in the equally-weighted case, the rest of the consideration set is split into two categories: (i) with $y_i'' < v''$ and (ii) $y_i''' \geq v'''$. Among category (i) candidates, the highest majority grade v''^{max} is the lower bound below which majority grade cannot be decreased by i. This is same as the equally-weighted case. Among category (ii) candidates, there is a modification: the highest $\underline{u}'''^{max}$ forms the lower bound. Thus, $\underline{\underline{w}}_i'$ needs to be updated as:

$$\underline{\underline{w}}_i' = \max(v''^{max}, \underline{u}'''^{max}).$$

Putting it all together, for the differentially-weighted case, the following is the necessary and sufficient condition that allows beneficial strategic grading opportunity to a player via a non-winner candidate $\mathbf{m}'$:

$$(y_i' > y_i^*) \ \& \ \left\{ \Big((y_i^* < v^*) \ \& \ (y_i' \leq v') \ \& \ (\overline{u}' \geq v^*) \Big) \ \Big| \ \Big((y_i^* \geq v^*) \ \& \ (y_i' > v') \ \& \ (v' > \underline{\underline{w}}_i') \Big) \right\}. \tag{3}$$

Measures for Strategy-Proneness. Suppose the consideration set comprises of M candidates, and there are N players. We formally define the three measures of interest.

1. Likelihood of manipulability of a candidate, φ^C: the number of non-winner candidates that meet condition (1) $\div$ $(M-1)$.

2. Likelihood of manipulability by a player, φ^P: the number of players for whom any candidate meets condition (3) $\div$ N.

3. Likelihood of manipulability of the consideration set, φ^S: the number of player-candidate pairs that meet condition (3) $\div$ $(M \times N)$.

φ^C uses the same condition as the equally-weighted case, as it is at the overall consideration set level. φ^P and φ^S are now updated with the modified condition derived in this section.

	Preference structure	
Relative weights of the players	P1: None ("unrestricted domain")	P2: Convex function
R1: Equal weights ("unweighted")	P1R1	P2R1
R2: Differential weights (N=5)	P1R2	P2R2
R3: Differential weights (N=25)	P1R3	P2R3

Table 7 Design of experiments for investigation of strategy resistance

4. Simulation Results

To get a sense of strategy resistance of the procedure, we conducted a number of simulations systematically varying some key parameters. The design of experiments is summarized in Table 7.

The intent behind this design has been to contrast the proposed *COuNSEL* procedure with several other plausible implementations. At the simplest extreme, P1R1 is the basic Majority Judgment, as laid out by its authors. At the other extreme lies P2R3, which is closest to the real-life scenarios that *COuNSEL* may be deployed for. The progression in the two directions from P1R1 to P2R3 is instructive. R2 and R3 address the proportional representation aspect of *COuNSEL*, which is a key design element that adds equitability. R2 is a very small setup, and might represent the initial deployment phase of *COuNSEL*, in which fewer airlines may participate. R3 is a more likely setup reflecting the later phases of deployment. P2, on the other hand, addresses the key assumption in structuring of the grade functions. An unrestricted domain would easily lead to inconsistent grading over rounds, which is highly undesirable.

With this broad overview, the specific details for each scenario are now explained. Consideration set sizes of 5, 10, or 15 candidates are simulated in all the scenarios. The players' grades for the consideration set are generated randomly within the grading range of $\{0 \ldots 1\}$ for the unrestricted domain scenarios (P1). An increasing quadratic function of a randomly generated number that restricts the function maxima to be within the grading range is used for convex preference scenarios (P2).

In the equal weight scenarios (R1), number of players is one of 5, 15, 25, 35, and 45. Five different weighting schemes are simulated in the differential weight scenarios. In the differential weight (N=5) scenarios (R2), the number of players is fixed at 5; while the differential weight (N=25) scenarios (R3) have 25 players.

Table 8 summarizes the weighting schemes for the R2 scenarios. The first scheme gives all players equal weight for comparison. The proportion of largest weight to the total weight is 0.20 in this case. The Herfindahl-Hirschman Index, or HHI, is reported as a measure of the "market concentration". HHI

Weighting Scheme	Player Weights					Total Weight	Largest to Total	HHI
	A	B	C	D	E			
1	1	1	1	1	1	5	0.20	0.20
2	2	1	1	1	1	6	0.33	0.22
3	2	2	1	1	1	7	0.29	0.23
4	3	2	1	1	1	8	0.38	0.25
5	3	3	1	1	1	9	0.33	0.26

Table 8 Weights for the different weighting schemes for R2 scenarios

Weighting Scheme	Player Weights															Total Weight	Largest to Total	HHI
	A	B	C	D	E	F	G	H	I	J	K	L	M	...	Y			
1	1	1	1	1	1	1	1	1	1	1	1	1	1	...	1	25	0.04	0.040
2	2	2	2	2	2	2	2	2	2	2	2	2	1	...	1	37	0.05	0.045
3	4	4	4	4	2	2	2	2	2	2	2	2	1	...	1	45	0.09	0.054
4	8	8	4	4	2	2	2	2	2	2	2	2	1	...	1	53	0.15	0.073
5	16	8	4	4	2	2	2	2	2	2	2	2	1	...	1	61	0.26	0.107

Table 9 Weights for the different weighting schemes for R3 scenarios

is computed as sum of the square of the market shares of all players, where market share of a player is the proportion of her weight to the total weight. From scheme 1 through 5, the HHI keeps increasing, as two players (namely A and B) are given progressively higher weights. Player A has the largest weight; its proportion to total weight never crosses 50%, as that would provide it dictatorial power.

Table 9 summarizes the weighting schemes for the R3 scenarios. These have 25 players each, marked A through Y. The first scheme gives all players equal weight for comparison. The thirteen players marked M through Y have the same weight of 1 for all the schemes. Weights for the initial twelve players are systematically varied, so that the HHI increases as we go down the list. Player A has the largest weight; its proportion to total weight is kept below 50%.

A hundred simulations runs were conducted for each of the four scenarios. The averages of the three metrics for strategy proneness are reported.

4.1. P1R1: Unrestricted domain, Equal weights

In this scenario, the numbers of candidates (M) and players (N) are systematically varied; each player having the same weight as others. This is very similar to the basic Majority Judgment procedure described by its authors. It thus forms a benchmark to which results from the other scenarios are compared.

Number of candidates, M	Number of players, N				
	5	15	25	35	45
5	69.25	43.50	32.00	25.75	22.00
10	69.44	30.44	25.56	16.89	14.67
15	64.21	28.07	22.29	22.29	13.57

(a) P1R1: Mean φ^C (%)

Number of candidates, M	Number of players, N				
	5	15	25	35	45
5	9.60	8.73	6.68	4.43	4.49
10	15.20	10.07	7.52	6.00	6.47
15	19.40	9.67	8.12	7.74	5.87

(b) P1R1: Mean φ^P (%)

Number of candidates, M	Number of players, N				
	5	15	25	35	45
5	2.40	2.01	1.48	1.02	1.00
10	2.46	1.25	0.90	0.65	0.76
15	2.16	0.81	0.66	0.62	0.46

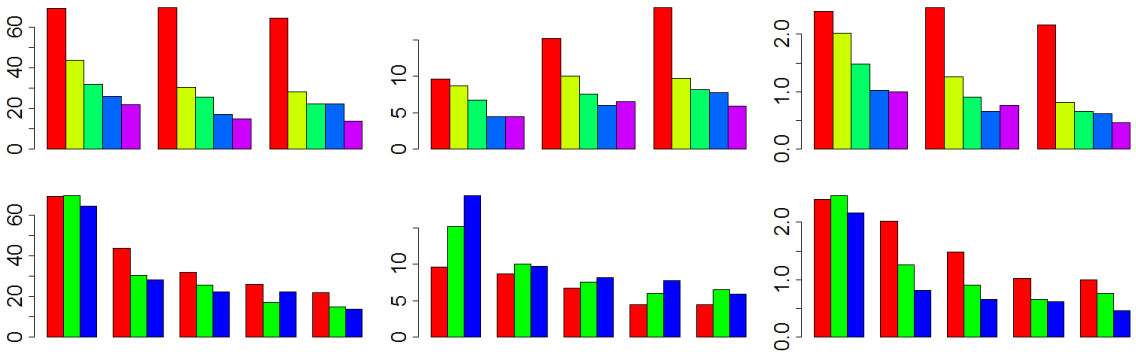
(c) P1R1: Mean φ^S (%)

Table 10 Strategy-proneness Measures for P1R1

At a broad level, many candidates appear to be manipulable overall, as Table 10a reports. However, when it comes to individual players, a much smaller proportion of players actually have beneficial opportunities, as reported in Table 10b. More specifically, as each player evaluates each candidate, the likelihoods are even smaller, as reported in Table 10c. This is as expected.

Figure 1 pictorially depicts the information in the tables. The top row has groups of bars for consideration set sizes, M of 5, 10, 15 respectively. Within each group, the individual bars represent the measures for the different number of players, N of 5,15,25,35,45. The bottom row presents the same information, but groups them in the other way: the broad groups are for various N's, and the individual bars within each group have measures for different M. The patterns are clearly noticeable with this layout.

For a fixed size of consideration set, all the three measures decline as more players are included – as the top row depicts. This effect tapers off as the number of candidates increase. This makes intuitive sense – as more grades are provided by the increased number of players for each candidate, the gap



(a) Mean φ^C (b) Mean φ^P (c) Mean φ^S

Figure 1 Strategy-proneness Measures for P1R1

Measures for proneness to beneficial strategic opportunity in P1R1 scenario. Within each of the three groups of bars in the top row, the consideration set size, M is fixed to one of 5, 10, 15. Across the five groups of bars in the bottom row, the number of candidates, N is fixed to be one of 5,15,2,53,45. Players have equal weight. Players' grades are unrestricted within the allowable range.

between the $\underline{v}'$ and $\overline{v}'$ narrows. This gap is a decisive factor in manipulability of a candidate.

For a fixed number of players, φ^C and φ^S decrease as the consideration set size increases, while the opposite holds true for φ^P. This trend also tapers off with larger consideration set sizes. The decrease in the φ^C and φ^S has the same intuitive explanation as above. More candidates being available increases chance of a player to look forward to strategic grading – thus, φ^P increases with consideration set size. This has design implications – if *COuNSEL* has to be initiated in a region that has smaller number of airlines, then it should force them to grade more candidates in order to minimize strategic grading opportunities.

Equal weight to each airline is clearly ruled out in the implementation of *COuNSEL*. For, it would imply that airlines with large impact due to the weather would have the same voice in the decision-making as other airlines with perhaps a single impacted flight. However, this forms a benchmark for our investigation, as the proposed procedure should retain Majority Judgment's key desirable property of strategy resistance.

4.2. P1R2: Unrestricted domain, Differential weights with 5 players

In this scenario, the number of players is fixed at $N = 5$, and the consideration set size is one of $M = 5, 10, 15$. Players may have different weights; the five weighting schemes reported in Table 8 are simulated. This represents a likely scenario in the initial pilot phase of *COuNSEL*, in which a small number of airlines may be involved. A key difference from *COuNSEL* is the unrestricted domain, as

	Weighting Scheme				
Number of candidates, M	1	2	3	4	5
5	69.25	69.00	69.50	60.25	73.25
10	69.44	63.11	64.44	58.44	63.67
15	64.21	59.07	65.21	58.93	58.14

(a) P1R2: Mean φ^C (%)

	Weighting Scheme				
Number of candidates, M	1	2	3	4	5
5	9.60	7.80	14.40	7.20	11.80
10	15.20	13.40	13.40	11.60	17.20
15	19.40	16.40	18.80	15.20	16.40

(b) P1R2: Mean φ^P (%)

	Weighting Scheme				
Number of candidates, M	1	2	3	4	5
5	2.40	2.04	4.04	1.92	2.92
10	2.46	1.92	2.10	1.50	2.50
15	2.16	1.69	1.97	1.51	1.73

(c) P1R2: Mean φ^S (%)

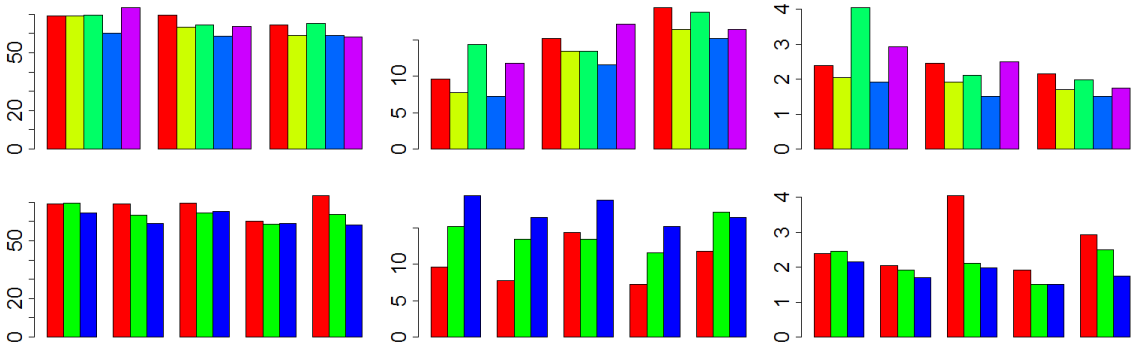
Table 11 Strategy-proneness Measures for P1R2

COuNSEL assumes a structured grade function for each airline.

Tables and figures similar to those in the equal weights scenario are reported for the three measures. Very broadly, the strategy-proneness measures are not significantly different with differential weighting schemes as compared to the equal weighted scheme. A mild systematic pattern is evident from the top row of Figure 2. For a fixed size of consideration set, the weighting scheme appears to have smallest strategy-proneness; this scheme has the largest proportional weight to the player A. No clear patterns are visible with respect to HHI. The bottom row continues the pattern with the equal weights scenario. Given a weighting scheme, φ^C and φ^S decrease as the consideration set size increases, while φ^P increases. The intuition behind this remains the same.

4.3. P1R3: Unrestricted domain, Differential weights with 25 players

In this scenario, the number of players is fixed at $N = 25$, and the consideration set size is one of $M = 5, 10, 15$. Players may have different weights; the five weighting schemes reported in Table 9 are



(a) Mean φ^C (b) Mean φ^P (c) Mean φ^S

Figure 2 Strategy-proneness Measures for P1R2

Measures for proneness to beneficial strategic opportunity in P1R2 scenario. Number of candidates N is fixed at 5. The players have same weight in the first weighting scheme, and different weights in the others. Within each of the three groups of bars in the top row, the consideration set size, M is fixed to one of 5, 10, 15. Across the five groups of bars in the bottom row, the weighting scheme is varied so that HHI increases from left to right. Players' grades are unrestricted within the allowable range.

simulated. This represents a later deployment phase of *COuNSEL*, whereby several airlines are involved in the decision-making process. Again, the unrestricted domain of preferences is a key difference from *COuNSEL*.

Tables and figures are reported for the three measures. A systematic pattern is evident from the top row of Figure 3: for a fixed consideration set size, increasing HHI (which also increases the proportional weight of the largest player in this scenario) tends to reduce strategy-proneness. The bottom row continues the pattern with the equal weights scenario. Given a weighting scheme, φ^C and φ^S decrease as the consideration set size increases, while φ^P increases. The intuition behind this remains the same.

4.4. P2R1: Convex preference structure, Equal weights

In this scenario, the numbers of candidates (M) and players (N) are systematically varied; each player having the same weight as others. The key difference from P1R1 is that the players have a convex grading function of a special type. The mechanics of drawing such convex grades are summarized in Appendix A.

Compared to P1R1 scenario, the strategy-proneness measures are all dramatically lower. The convex structure forces the grades to be more concentrated near the peaks for each player. This potentially reduces the gap between $\underline{v}'$ and $\overline{v}'$ for all the candidates, leading to reduction in strategy proneness.

The general pattern of reductions in all the strategy-proneness measures within a fixed consideration set size continues, as the top row of Figure 4 shows. The tapering off effect is also evident in the top

	Weighting Scheme				
Number of candidates, M	1	2	3	4	5
5	32.00	29.25	25.00	21.75	18.50
10	25.56	22.11	21.44	15.11	9.56
15	22.29	20.43	18.79	13.71	10.71

(a) P1R3: Mean φ^C (%)

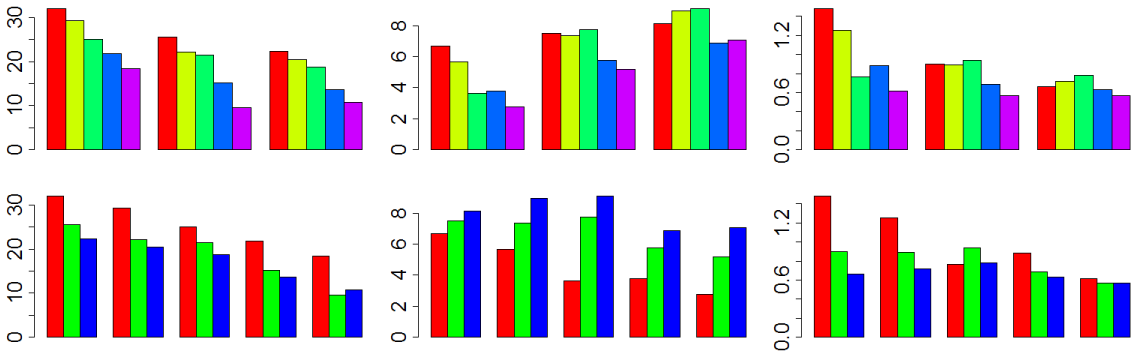
	Weighting Scheme				
Number of candidates, M	1	2	3	4	5
5	6.68	5.12	4.56	4.76	6.20
10	7.52	7.04	10.04	7.48	7.32
15	8.12	9.16	10.36	9.32	11.04

(b) P1R3: Mean φ^P (%)

	Weighting Scheme				
Number of candidates, M	1	2	3	4	5
5	1.48	1.11	0.95	0.99	1.32
10	0.90	0.85	1.16	0.80	0.78
15	0.66	0.71	0.89	0.75	0.88

(c) P1R3: Mean φ^S (%)

Table 12 Strategy-proneness Measures for P1R3



(a) Mean φ^C (b) Mean φ^P (c) Mean φ^S

Figure 3 Strategy-proneness Measures for P1R3

Measures for proneness to beneficial strategic opportunity in P1R3 scenario. Number of candidates N is fixed at 25. The players have same weight in the first weighting scheme, and different weights in the others. Within each of the three groups of bars in the top row, the consideration set size, M is fixed to one of 5, 10, 15. Across the five groups of bars in the bottom row, the weighting scheme is varied so that HHI increases from left to right. Players' grades are unrestricted within the allowable range.

Number of candidates, M	Number of players, N				
	5	15	25	35	45
5	30.00	16.75	9.25	7.50	7.25
10	30.67	13.22	7.89	6.00	6.11
15	36.43	15.29	10.14	5.36	4.93

(a) P2R1: Mean φ^C (%)

Number of candidates, M	Number of players, N				
	5	15	25	35	45
5	3.00	2.80	1.16	0.97	0.91
10	9.20	4.93	4.12	3.26	3.24
15	14.20	8.60	8.40	4.71	4.42

(b) P2R1: Mean φ^P (%)

Number of candidates, M	Number of players, N				
	5	15	25	35	45
5	0.64	0.65	0.27	0.19	0.18
10	1.04	0.61	0.48	0.38	0.32
15	1.55	0.68	0.68	0.37	0.31

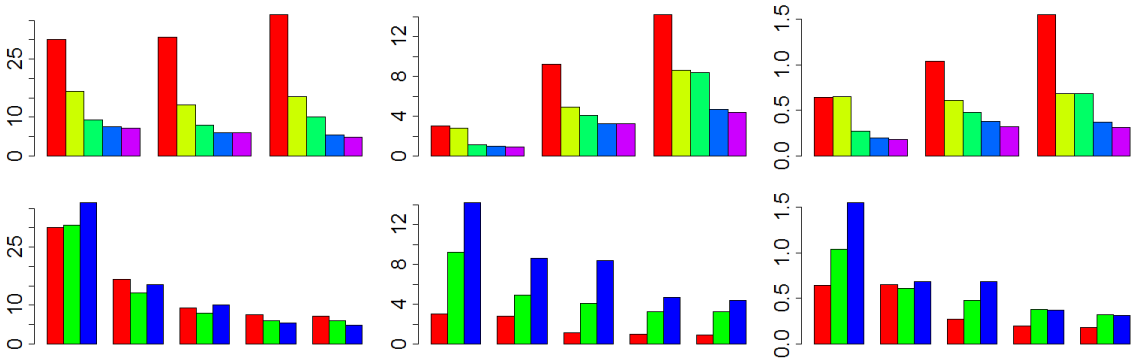
(c) P2R1: Mean φ^S (%)

Table 13 Strategy-proneness Measures for P2R1

row. The bottom row has similar patterns as P1R1 for φ^C and φ^P – the former is more or less similar within each group having the same number of players, while the latter increases within each group. However, the φ^S measure increases as the consideration set size increases, with fixed number of players. Recall φ^P counts a player as potentially manipulative if she has opportunity via even a single candidate, whereas φ^S counts exact player-candidate pairs that are manipulable. Compared to P1R1, this implies that more candidates are manipulable for the players who have an opportunity to manipulate at all, as the number of candidates increase. Note, however, that the overall levels of φ^P and φ^S are both lower than those in P1R1. All the three measures taper off as number of players increases.

4.5. P2R2: Convex preference structure, Differential weights with 5 players

In this scenario, the number of players is fixed at $N = 5$, and the consideration set size is one of $M = 5, 10, 15$. Players may have different weights; the five weighting schemes reported in Table 8 are simulated. The key difference from P1R2 is that the players have a convex grading function of a special



(a) Mean φ^C (b) Mean φ^P (c) Mean φ^S

Figure 4 Strategy-proneness Measures for P2R1

Measures for proneness to beneficial strategic opportunity in P2R1 scenario. Within each of the three groups of bars in the top row, the consideration set size, M is fixed to one of 5, 10, 15. Across the five groups of bars in the bottom row, the number of candidates, N is fixed to be one of 5,15,2,53,45. Players have equal weight. Players' grades are convex within the allowable range.

type – as specified for the P2R1 scenario.

As with P2R1 versus P1R1, there is a dramatic reduction in the strategy-proneness measures compared to its analogous unrestricted domain, namely P1R2. The main observation continues from P1R2: compared to the equal-weighted scenario with convex grading functions, the measures do not change dramatically due to introduction of weights – especially for φ^C. The patterns for the other two remain similar to those in the equal-weighted scenario as well – as the bottom row of Figure 5 shows. The top row shows no systematic patterns are discernible as the HHI changes for fixed consideration set sizes; however, like P1R2, the weighting scheme 4 has the smallest strategy-proneness measures.

4.6. P2R3: Convex preference structure, Differential weights with 25 players

In this scenario, the number of players is fixed at $N = 25$, and the consideration set size is one of $M = 5, 10, 15$. Players may have different weights; the five weighting schemes reported in Table 9 are simulated. This scenario closely resembles the likely final deployment phase of *COuNSEL*.

The key difference from P1R3 is that the players have a convex grading function of a special type – as specified for the P2R1 scenario. All the strategy-proneness measures are significantly lower as compared to the unrestricted domain case of P1R3. The top row shows no systematic patterns are discernible as the HHI changes for the smaller consideration set sizes of 5 and 10. However, a decline in the measures is apparent with increase in HHI for consideration set comprising of 15 candidates. The main observation continues from P1R2: compared to the equal-weighted scenario with convex grading

Number of candidates, M	Weighting Scheme				
	1	2	3	4	5
5	30.00	32.25	33.75	26.50	30.50
10	30.67	34.00	33.22	33.89	34.33
15	36.43	33.07	32.21	34.29	33.07

(a) P2R2: Mean φ^C (%)

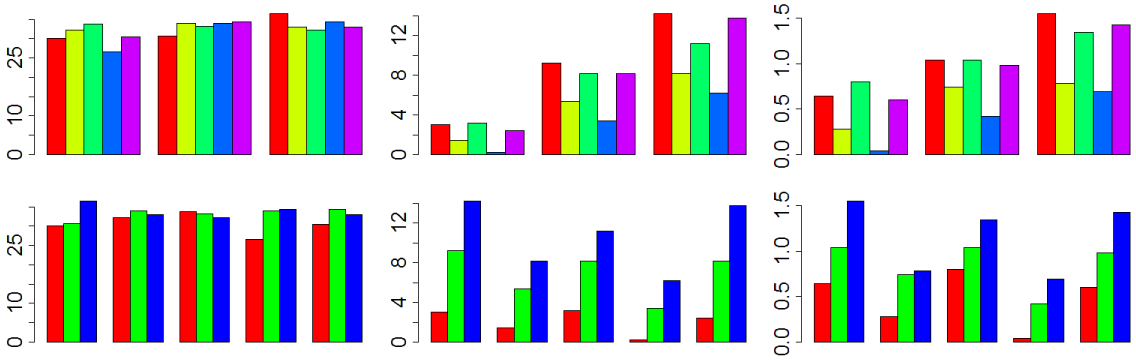
Number of candidates, M	Weighting Scheme				
	1	2	3	4	5
5	3.00	1.40	3.20	0.20	2.40
10	9.20	5.40	8.20	3.40	8.20
15	14.20	8.20	11.20	6.20	13.80

(b) P2R2: Mean φ^P (%)

Number of candidates, M	Weighting Scheme				
	1	2	3	4	5
5	0.64	0.28	0.80	0.04	0.60
10	1.04	0.74	1.04	0.42	0.98
15	1.55	0.79	1.35	0.69	1.43

(c) P2R2: Mean φ^S (%)

Table 14 Strategy-proneness Measures for P2R2



(a) Mean φ^C (b) Mean φ^P (c) Mean φ^S

Figure 5 Strategy-proneness Measures for P2R2

Measures for proneness to beneficial strategic opportunity in P2R2 scenario. Number of candidates N is fixed at 5. The players have same weight in the first weighting scheme, and different weights in the others. Within each of the three groups of bars in the top row, the consideration set size, M is fixed to one of 5, 10, 15. Across the five groups of bars in the bottom row, the weighting scheme is varied so that HHI increases from left to right. Players' grades are convex within the allowable range.

	Weighting Scheme				
Number of candidates, M	1	2	3	4	5
5	9.25	9.25	8.00	9.25	8.25
10	7.89	8.89	10.89	8.33	7.78
15	10.14	7.93	8.07	8.43	7.43

(a) P2R3: Mean φ^C (%)

	Weighting Scheme				
Number of candidates, M	1	2	3	4	5
5	1.16	2.00	1.36	0.80	1.56
10	4.12	5.24	4.64	2.48	2.72
15	8.40	7.28	5.44	5.88	4.24

(b) P2R3: Mean φ^P (%)

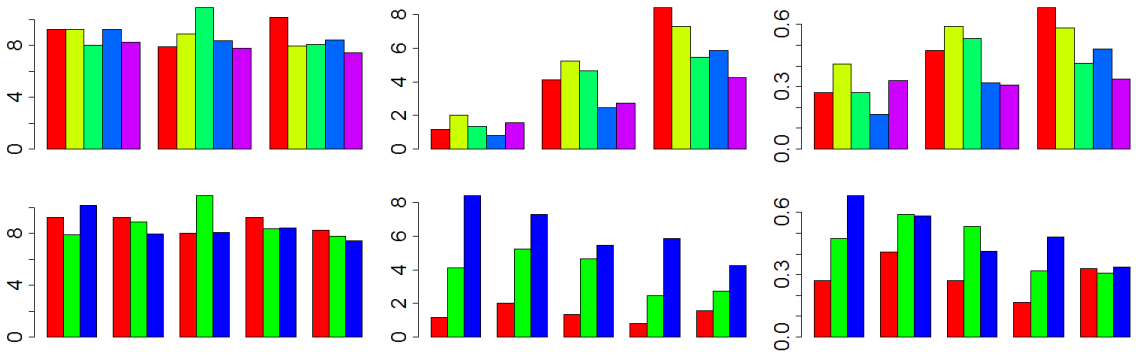
	Weighting Scheme				
Number of candidates, M	1	2	3	4	5
5	0.27	0.41	0.27	0.17	0.33
10	0.48	0.59	0.53	0.32	0.31
15	0.68	0.58	0.41	0.48	0.34

(c) P2R3: Mean φ^S (%)

Table 15 Strategy-proneness Measures for P2R3

functions, the measures do not change dramatically due to introduction of weights – especially for φ^C. Unlike P1R3, where φ^C decreased with increase in HHI, φ^C does not seem to have any pattern. The patterns for the other two remain similar to those in the P2R1 as well as P2R2 – as the bottom row of Figure 5 shows. That is, for each weighting scheme, φ^C and φ^S generally increase with increase in consideration set size.

These have implications on the implementation design parameters for the mechanism. As far as possible, consideration set sizes should be kept small, not only for increased cognitive load to the players, but also for strategy-proneness. Addition of weights not significantly impacting the strategy-proneness measures is a useful observation in itself. However, these should be investigated for different types of players – as it must be giving larger strategic opportunities to the larger players, while eliminating such opportunities for the smaller players.



(a) Mean φ^C (b) Mean φ^P (c) Mean φ^S

Figure 6 Strategy-proneness Measures for P2R3

Measures for proneness to beneficial strategic opportunity in P2R3 scenario. Number of candidates N is fixed at 25. The players have same weight in the first weighting scheme, and different weights in the others. Within each of the three groups of bars in the top row, the consideration set size, M is fixed to one of 5, 10, 15. Across the five groups of bars in the bottom row, the weighting scheme is varied so that HHI increases from left to right. Players' grades are convex within the allowable range.

5. Conclusion

Impossibility results due to Arrow, Gibbard and Satterthwaite, have ruled out existence of strategy-proof mechanisms in which no player has dictatorial powers – especially with unrestricted domain. Majority Judgment is a recent proposal that bypasses this result, and is claimed to be highly strategy resistant by its authors. In this paper, we characterized and quantified the proneness of Majority Judgment-based voting procedure to beneficial strategic opportunities by the players. We employed a framework similar to Nash equilibrium concept, which has been extensively used in mechanism design literature as a solution concept.

Specific to Majority Judgment in general, we developed the necessary and sufficient conditions for a player to benefit by reporting untruthful grades for one or more candidates. The conditions were then used as basis for quantifying three measures of strategy proneness.

Finally, we simulated several scenarios starting from basic Majority Judgment procedure, systematically varying assumptions and key parameters, leading up to scenarios that closely resemble initial and later deployment phases of *COuNSEL*. We found that the most obvious measure for strategy proneness, the one based on proportion of manipulable candidates, is deceptive – it consistently reports very high likelihood of manipulation, typically upwards of 50%. However, the likelihood of an individual player to find a beneficial strategic opportunity drops in the regions of 10% or less. Moreover, as the specific candidates via which the individual players may benefit are also brought into consideration, the likelihood drops to 1-2% levels.

A surprising, though useful, observation has been the rather insignificant impact of attaching weights to the players. Weights are a significant design element in *COuNSEL*, wherein unlike the democratic "one-person one-vote" scenario, it is essential to provide the airlines differential weight in the overall decision-making, for equity reasons.

Another key observation has been the drastic reduction in strategy proneness when the unrestricted domain of grades is replaced with a convex preference structure. Convexity, continuity, and monotonicity have been standard assumptions in the literature. These are also reasonable in our application area, whereby players would more likely have a possibly "single-peaked" preference structure over the feasible candidate space.

The results in themselves are quite encouraging. Even with complete knowledge of everyone's grades, and then being provided with an opportunity to benefit oneself, the likelihood of a particular player to find a beneficial opportunity via a candidate is in the region of 2% or below. In real-life, such opportunity would of course not exist. Moreover, untruthful reporting has a good possibility of hurting the player, as it may result with a new winner that is less preferred than the current winner.

These observations are based on simulations with simple preference structure, whereas *COuNSEL* design allows for a more nuanced structure over a multi-dimensional candidate space. Furthermore, it deals with feasibility constraints on the candidate space. Experiments incorporating these details, and with realistic application scenarios should be conducted before finalizing the design parameters of *COuNSEL*.

It should be mentioned here that the simulations assumed that a player had complete knowledge of other players' grades, and then had an opportunity to unilaterally deviate from truthful grading if it led to a more preferable candidate than the current winner. In practice, this will not be the case. There are three implications and possible directions for future research. One pertains to the information dissemination at the end of each round. The FAA could possibly release all the grading information, but that could incentivize airlines to collude among themselves – which would defeat the purpose of the entire mechanism. It could also lead to an information overload. On the other hand, the FAA need not release any information until the final round, but that may call into question the FAA's trustworthiness. A middle ground that encourages the airlines to productively contribute to the process without divulging unnecessary information needs to be found.

The second implication has to do with the possible strategic uses of the partial knowledge that does get disseminated at the end of each round. As the airlines gain experience, they may be able to anticipate

other airlines' behavior probabilistically, and use the information to update their beliefs. Instead of Nash equilibrium, a Bayesian Nash equilibrium may then serve as a more appropriate solution concept. The modeling details would depend on the type of information released.

Finally, with the probabilistic knowledge of other airlines' grade functions replacing the full knowledge as in this paper, it would be imperative to quantify the expected loss due to strategic grading. We have identified the best case scenarios for an airline to benefit from strategic grading; this investigation would form the worst case for an airline.

Acknowledgment

This work was supported by the Federal Aviation Administration through the NEXTOR-II Consortium.

Appendix A: Convex Preference Structure

The procedure for drawing grades so that they follow a convex structure is detailed in this section.

Suppose the candidates are drawn randomly from a fixed range: $x \sim [0 \dots 1]$. For a given candidate x, a special quadratic function maps these values into the grade for each player i: $y_i = a_i x^2 + b_i$, where a_i and b_i are player i-specific coefficients. The coefficients for each player are constrained such that: (a) the grade function is convex in the allowable grading range of $[0 \dots 1]$, (b) the grade function is non-negative in the allowable range, (c) the grade function has its global maxima within the allowable range, and (d) the grade function has its maxima as the largest allowable grade of 1.

(a) and (c) are inter-related for quadratic functions. For it to have a global maximum, following necessary and sufficient conditions must be met (dropping subscript i for ease of notation):

$$\frac{\mathrm{d}y}{\mathrm{d}x} = 0 \Rightarrow 2ax^* + b = 0 \Rightarrow x^* = \frac{-b}{2a},$$
$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} < 0 \Rightarrow 2ax^* < 0 \Rightarrow a < 0.$$

For the maxima to be within the given range as required in (c), we want:

$$0 \le x^* \le 1 \Rightarrow 0 \le \frac{-b}{2a} \le 1 \Rightarrow 0 \le b \le -2a.$$

For the last inequality, recall $-2a > 0$ as required in the previous statement.

Further, recall that the specified function has $y = 0$ at $x = 0$. To satisfy (b), we need to ensure that $y \ge 0$ at the largest allowable value of x – which is 1 in this case. Thus, we get another bounding constraint for b:

$$0 \le y|_{x=1} \le 1 \Rightarrow 0 \le a + b \le 1 \Rightarrow -a \le b \le 1 - a.$$

Putting the two bounding constraints for b, we get:

$$0 \leq -a \leq b \leq -2a \leq 1-a.$$

The tighter of the bounds require that:

$$-a \leq b \leq -2a.$$

For (d), we evaluate y at the maxima, and set it to the largest allowable grade, that is, 1:

$$y|_{x=x^*} = 1 \Rightarrow \frac{-b}{2a}\left[a(\frac{-b}{2a}) + b\right] = 1 \Rightarrow b = 2\sqrt{-a}.$$

Thus the bounds derived above imply:

$$-a \leq 2\sqrt{-a} \leq -2a \Rightarrow -1 \leq a \leq -4.$$

Some sample grade functions are shown in Figure 7.

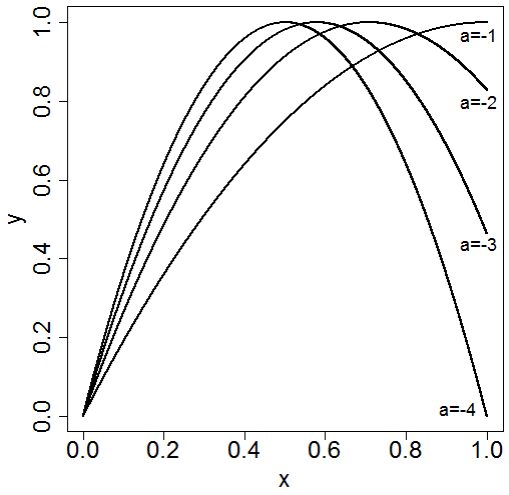


Figure 7 Sample convex grade functions

The procedure for generating the coefficients for each player is summarized as follows. For each player i, draw a coefficient: $a_i \sim [-4 \cdots -1]$, and compute $b_i = 2\sqrt{-a}$.

Generating the grades for a given consideration set of M candidates is straightforward. Player i's grade for a candidate x is computed as: $y_i = a_i x^2 + b_i x$. For all the M candidates the same coefficients are used for a given player. The procedure is repeated for all N players.

It should be mentioned here that *COuNSEL* employs a more nuanced grading function, and allows for arbitrary number of dimensions for the candidates. It also has the additional complexity of the feasible candidate space. Nonetheless, this simple model allows us to study the strategy proneness with structured preferences, and contrast the results with no structure.

References

Balinski, M., R. Laraki. 2011. *Majority Judgment: Measuring, Ranking, and Electing*. The MIT Press.

Maskin, Eric. 1985. The theory of implementation in nash equilibrium: A survey. *Social goals and social organization: Essays in memory of Elisha Pazner* 173–204.

Maskin, Eric. 1999. Nash equilibrium and welfare optimality. *The Review of Economic Studies* **66**(1) 23–38.

Moore, John, Rafael Repullo. 1990. Nash implementation: a full characterization. *Econometrica: Journal of the Econometric Society* 1083–1099.

APPENDIX V

Airline-Driven Performance-Based Air Traffic Management: Game Theoretic Models and Multi-Criteria Evaluation

Antony Evans[a], Vikrant Vaze[b] and Cynthia Barnhart[b]

[a]School of Engineering and Mathematical Sciences, City University London
[b]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology

Abstract

Defining Air Traffic Management as the tools, procedures and systems employed to ensure safe and efficient operation of air transportation systems, an important objective of future air traffic management systems is to support airline business objectives, subject to ensuring safety and security. Under the current model for designing air traffic management initiatives, the *central authority* overseeing and regulating air traffic management in a region makes trade-offs between specified performance criteria. The research presented in this paper aims instead to allow the *airline community* to set performance goals and thus make trade-offs between different performance criteria directly, before specific air traffic management strategies are determined. We propose several approaches for collecting inputs from airlines in a systematic way and for combining these airline inputs into implementable air traffic management initiatives. These include variants of averaging, voting and ranking mechanisms. We also propose multiple criteria for evaluating the effectiveness of each approach, including Pareto optimality, airline profitability, system optimality, equity, and truthfulness of airline inputs. We apply a game-theoretic approach to examine the potential for strategic (gaming) behavior by airlines. We offer a broad evaluation of each approach by simulating each of the approaches for a generic system using Monte-Carlo methods, sampling values for input parameters from a wide range. We also provide an indication of how the approaches might perform in a real system by simulating ground delay programs at two airports in the New York City area. We first apply a simplified model that simulates the process of selecting only planned end times of a ground delay program, using Monte-Carlo methods. Next, we apply a more detailed model that simulates the process of selecting planned end times and reduced airport arrival rates. Finally, we characterize the effectiveness of each of the considered approaches on the proposed criteria and identify the most desirable approaches. We conclude that voting schemes, which score highly on all criteria (including airline profitability, system optimality and equity), represent the most promising approaches (among those considered) to elicit airline preferences, thereby allowing the central authority to design air traffic management initiatives that optimize system performance while respecting the objectives of airlines.

Keywords

Performance-based Air Traffic Management, Airline preferences, Game theory, Nash equilibrium.

1. Introduction

Air Navigation Service Providers (ANSPs), such as the Federal Aviation Administration (FAA) in the United States and EUROCONTROL in Europe, are the central authorities responsible for safe and efficient operation of our air transportation systems. In order to ensure these goals, ANSPs employ various tools, procedures and systems, which together are termed Air Traffic Management (ATM). ATM systems in the U.S. and Europe are currently poised for a major overhaul, under projects titled Next Generation Air Transportation System (NextGen) in the U.S., and Single European Sky ATM Research (SESAR) in Europe. An important objective of future ATM systems as envisioned by the FAA is supporting the business objectives of airlines, subject to ensuring safety and security (JPDO, 2007; FAA, 2011). In addition to safety and security, airlines value many different operational aspects of the air transportation system, such as capacity, efficiency, flexibility, predictability etc. Better availability of sufficient *capacity* in the various components of the system reduces or eliminates congestion related delays. Greater *efficiency* in resource utilization translates into reduced operating costs. Greater *flexibility* in scheduling operations enables airlines to make appropriate changes closer to departure times, as their needs evolve with time. And better *predictability*, which refers to the reliability of the system to deliver on planned performance, leads to more certainty about future operations, which in turn helps airlines plan better. Different airlines might value these different performance criteria differently.

An ANSP may support airlines' business objectives by designing Traffic Management Initiatives (TMIs) in such a way as to maximize a single performance goal or some pre-defined combined measure based on multiple performance criteria, subject to ensuring safety and security. Here, performance goal refers to the quantified value, based on some defined metric, of a performance criterion. However, an ANSP cannot typically maximize all performance goals simultaneously, and must identify an appropriate trade-off between them. For example, consider a Ground Delay Program (GDP), a common TMI implemented by the FAA to control the flow of aircraft into an airport by delaying flights destined for that airport at their respective origin airports. A GDP is typically implemented for a period of time when increased aircraft spacing is considered necessary between landing aircraft, to ensure safety, and is often associated with adverse weather. However, weather forecasts are uncertain, so the point when conditions improve and additional spacing is no longer necessary is usually difficult to predict. Setting a GDP end time to be optimistically early maximizes the airport capacity and therefore throughput, because inbound flights are not delayed at their departure gates any longer than necessary. Therefore, no matter when conditions improve, and the airport capacity can be returned to the normal level, there are aircraft positioned to land. However, last minute extensions may have to be made to the GDP if the adverse weather continues longer than was forecast, potentially requiring airborne holding. So an early GDP end time would be at the expense of predictability (in addition to safety concerns and fuel costs associated with airborne holding). To maximize predictability, the GDP end time should be set conservatively late, allowing airlines to be confident that the GDP would not be extended. But this would be at the expense of throughput, as capacity might be underutilized if conditions were to improve earlier than the set GDP end time. There is therefore a trade-off between throughput and predictability.

Different airlines might have different preferences for prioritizing throughput over predictability. For example, for an airline operating a frequent shuttle service with low load factors, which allows easy

rebooking of passengers and easy reassignment of aircraft, some throughput reduction is not as detrimental as operating an unpredictable schedule. On the other hand, for an airline with lower frequency and higher load factors, for which delay recovery is difficult, high throughput may be preferred to predictability so that the airline does not have to cancel flights. The primary motivation for our research is to investigate various approaches for ANSPs to determine the trade-off between performance criteria, based on inputs (i.e., preferences) from airlines. We apply our research specifically to the case of GDPs.

In the existing literature, supporting airline preferences has typically been studied at the level of individual flight trajectories, through trajectory-based initiatives. In such initiatives, airlines are, for example, given authority to modify their own flight trajectories in a far- and mid-term time horizon to avoid an identified constraint, such as a region of airspace with high traffic or a region impacted by weather (e.g., Garcia-Chico *et al.*, 2008). Alternatively, airlines are given the opportunity to provide the ANSP with multiple prioritized flight trajectories, individual flight priorities, or route priorities (e.g., Sheth and Gutierrez-Nolasco, 2008). In this paper, we consider accommodating airline preferences at the more aggregate level, shifting the focus from flight trajectories to overall system performance. To the best of the authors' knowledge, ours is the first study that addresses the challenge of supporting airline preferences at a system level. Such a performance-based ATM system would be capable of making trade-offs between different performance criteria, such as capacity, efficiency, flexibility, predictability etc., at the system level, and could account for the system-level performance preferences of different airlines. The chosen system performance objectives would then serve as the basis for deciding on specific parameters, such as the length, scope and magnitude of TMIs, that include GDPs, Ground Stops, Miles-in-Trail (MIT) restrictions, traffic re-routes, etc.

The performance of the U.S. National Airspace System (NAS) is typically measured by the number of delayed flights and by the length, scope and magnitude of TMIs. The length of a GDP, Ground Stop, MIT restriction or re-route typically refers to the planned duration of the initiative. In the case of a GDP, Ground Stop, MIT restriction or re-route, scope refers to the subset of flights impacted by the initiative. For example, in the cases of a GDP or Ground Stop, scope refers to the set of flights delayed on departure as a consequence of the initiative. This set is comprised of all flights whose destinations are the constrained airport and whose origins are within a specified maximum distance from the constrained airport. The magnitude of a GDP refers to the specified airport arrival rate (AAR) at the destination airport, which in the case of a Ground Stop is zero. The magnitude of a MIT restriction is the actual in-trail spacing required of the traffic, while the magnitude of a re-route can be considered to be how far from the planned flight trajectory the reroute takes the traffic. These values, however, do not represent the *performance of the system* per se, but are rather *indicators of aspects of the system performance* (FAA, 2011). They are also inadequate to describe differences in system-level preferences and requirements of different airlines. It is therefore important to identify what the performance criteria of airlines are, and to describe them in quantifiable terms. The International Civil Aviation Organization (ICAO, 2005) lists performance criteria, or the "expectations of the ATM community", as follows: access and equity, capacity, cost-effectiveness, efficiency, environment, flexibility, global interoperability, participation by the ATM Community, predictability, safety, and security. In this paper, we focus directly on a subset of these performance criteria rather than dealing with indicators of aspects of performance, as has been done traditionally. The reader is referred to Liu and Hansen (2012) for an example of how performance goal vectors can be expressed as a function of these

indicators of aspects of performance. A performance goal vector refers to a vector with individual components that are the values or goals for specific performance criteria, such as capacity, predictability etc. In this paper, we make use of the expressions of capacity and predictability developed by Liu and Hansen (2012) when analyzing specific TMIs in Section 4.

Given the differences in the valuation of these performance criteria by different airlines, an approach or a mechanism is needed to reconcile their competing preferences. In contrast, in the existing ATM system, the *ANSP* has sole responsibility for determining these trade-offs when designing TMIs. For example, the trade-off of throughput and predictability is determined in designing a GDP with the ANSP selecting the GDP end time, as described above, as well as scope and magnitude. The research presented in this paper aims instead to allow the *airline community* to influence TMI design by providing preferences in advance of the TMI design and implementation. In so doing, the ANSP can then design TMIs that capture airline preferences in the most effective way. A primary aim of this research is to design and assess various candidate mechanisms for this process, and demonstrate their applicability through general experiments and specific, real-world motivated case studies.

Consistent with the FAA objective of supporting airlines' business objectives, a large amount of research has been focused on formulating and solving the problem of system optimality in air transportation (e.g., Odoni and Bianco, 1987; Bertsimas and Stock-Patterson, 1998, 2000; Lulli and Odoni, 2007). In effect, these studies describe ways in which airlines and the ANSP might attempt to maximize total airline profits and minimize system cost by trading off performance goals, without compromising safety and security. The ICAO performance criteria that are most likely to be traded in such a scenario are capacity, efficiency, predictability and flexibility. Because each airline may prefer a different trade-off, it becomes important to ensure equity in how each airline's preferences are combined to set the final system-wide performance goals. It is noted that, as described by Bertsimas *et al.* (2011), the trade-off that minimizes system cost might not, in fact, be equitable, depending on how equity is defined. This is further complicated by the fact that when an airline requests certain performance goals, the request might not, in fact, represent the airline's true preferences. In other words, an airline might not be truthful about its preferences, and might behave strategically, in effect *gaming* the system, requesting performance goals different from its true preference in order to draw the finally selected system performance goals closer to its desired outcome. This can have significant consequences for equity, because, while it might appear that a solution is equitable based on the submitted preferences, it could be far from it. Furthermore, if airlines game severely (requesting solutions far from their true preferences), the ANSP is provided with an inaccurate picture of the airlines' preferences, and therefore of how well it is serving the airlines. Therefore, in this paper, in addition to airline profitability and system optimality, we also use equity and truthfulness to assess the effectiveness of any mechanism under consideration.

As mentioned earlier, due to safety and security concerns, not all of the relevant system-wide performance criteria, such as capacity, efficiency, predictability and flexibility, can be simultaneously maximized. This is especially true in cases of high traffic and/or adverse weather. Thus, there is a tradeoff between these different system-wide performance goals. An increase in one performance goal beyond a certain level necessarily requires a reduction in another. We define the *trade space* as the set of all combinations of

system-wide performance goals that are feasible, subject to ensuring safety and security. Note that the trade space as well as the performance goals are defined at the system level and all else being equal, each airline is assumed to prefer a higher value of a system performance criterion at least as much as a lower value of the same performance criterion. A subset of the boundary of the trade space is Pareto efficient, in that, for any point A in this Pareto efficient subset of boundary points, there does not exist another point B in the trade space such that the value of each performance criterion at point B is at least as high as the value of that performance criterion at point A and the value of at least one performance criterion at point B is strictly greater than the value of that performance criterion at point A. We define this subset of the boundary of the trade space as the *Pareto frontier*.

2. Contributions

In this paper, a number of contributions are made to performance-based ATM research:

- This is the first study that investigates various approaches to allow the airline community to set the system-level performance goals of the ATM system, and thus make trade-offs between performance criteria directly. We do this using a rigorous game-theoretic approach, which identifies the potential for gaming by airlines.
- We propose several approaches for collecting inputs from airlines in a systematic way and for combining these airline inputs into implementable TMIs. These include an approach which takes a weighted average of the airline-preferred performance goals; an approach that pushes the weighted average of the airline-preferred performance goals out to the Pareto frontier; an approach that makes a weighted random choice of airline-preferred performance goals; an approach that allows airlines to rank preferred performance goals; and an approach in which airlines vote on preferred performance goals.
- We propose multiple criteria for evaluating the effectiveness of each approach to performance-based ATM, including Pareto optimality, airline profitability, system optimality, equity, and truthfulness of airline preferences.
- By simulating each of the approaches for a generic system using Monte-Carlo methods (sampling values for input parameters from a wide range), we offer a broad evaluation of each approach to performance-based ATM.
- We also apply the approaches to more realistic cases in which we simulate GDPs at Newark Liberty International airport (EWR) and at LaGuardia airport (LGA), both in the New York City area. Two models are run: one simplified GDP model that simulates decisions regarding the planned GDP end time, and is applied using Monte-Carlo methods; and a second more detailed model that simulates decisions regarding both planned GDP end time and GDP magnitude, and is run for a single GDP case at each of EWR and LGA, respectively. The results of these simulations provide an indication of how the approaches would perform in a real system, and how the results differ from those for the generic experiments.

- Finally, we characterize the effectiveness of each of the considered approaches on the proposed criteria and identify the most desirable approach accordingly. Taking a weighted average of the user preferred performance goal vectors, making a weighted random choice of the user preferred performance goal vectors, and voting on ANSP provided candidate performance goal vectors were all found to be reasonable candidates for practical implementation. However, the voting scheme shows particular promise, scoring highly on all criteria.

3. Framework

Figure 1 provides a high level view of the process to be investigated.

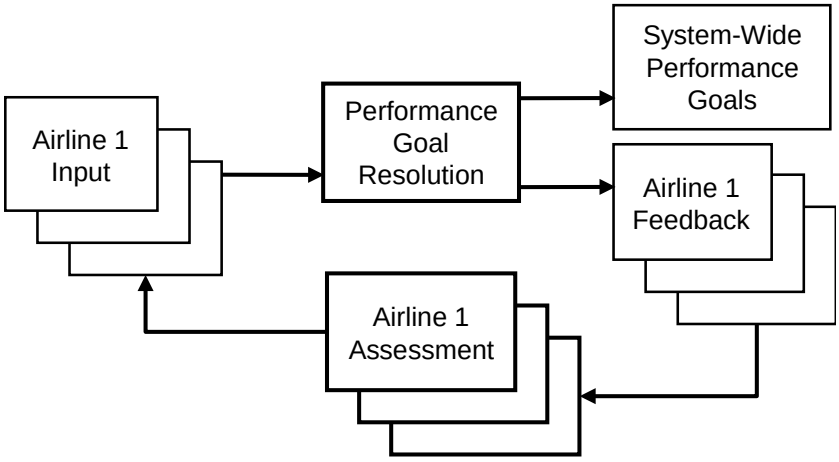


Figure 1. Architecture for process by which performance goals are set

The ultimate output of the process is a set of *system-wide performance goals* (upper right-hand box in Figure 1) that would be used by the ANSP to set specific TMIs. A possible form of these performance goals is described in Section 4. This set of performance goals could be for a single TMI within a single resource, such as the GDP at an airport described in the example in Section 1. Alternatively, the set of performance goals could be for multiple initiatives with multiple resources, or for the national airspace system as a whole. The process may start with an initial set of candidate performance goals suggested by the ANSP, or directly with *inputs from each airline* (set of boxes on upper left in Figure 1), which would be the performance goals preferred by each airline. A *performance goal resolution process* would then take the inputs from all airlines, confirm the feasibility of each input, and identify a set of system-wide performance goals by combining these different inputs in some way. The ANSP can then provide individual *airline feedback*, which would be a description of how the system-wide performance goals translate into changes in each specific airline's operational plan, e.g. delays to individual flights scheduled to arrive at an airport under a GDP. This allows each airline to assess the impact of the system-wide performance goals on its operational performance, such as propagated delays, passenger and crew schedule disruptions, additional fuel burn, etc., through an *airline assessment* process. In this process, each airline considers the feedback and determines what adjustments to make to its input in order to influence the system-wide performance goals in such a way as to improve its own operational performance. This feedback loop can be executed several times until an equilibrium is reached, where no airline can unilaterally adjust its own inputs to produce a "better" set of system-wide

performance goals (better being in terms of the performance objectives of that airline). This represents a pure strategy Nash equilibrium, a concept commonly used in game-theoretic literature to model situations with multiple, interacting, autonomous decision makers. We will use this concept to model the outcome of this iterative process.

The airline inputs and the performance goal resolution process may take a number of different forms. In this paper, different forms are studied in order to identify which has the best characteristics for setting system-wide performance goals. Each is described in detail in Sections 3.1 and 3.2 respectively. This is followed by a description of the metrics used to evaluate all approaches in Section 3.3.

3.1 Form of Airline Inputs

In this paper we analyze two forms of airline inputs:

1. *A preferred performance goal vector* – Each airline *k* simply specifies its preferred performance goal vector (I_k). This airline input is most applicable to a continuous trade space. In the example of a GDP described in Section 4.2, this input would take the form of each airline's preferred trade-off between capacity and predictability, using some pre-defined metrics. This trade-off could be input in the form of the parameters of the GDP (such as GDP end time and AAR), or it could be input in the form of capacity and predictability metrics that are calculated from these parameters. For example, a metric describing capacity could be the ratio of expected throughput, given known uncertainty in the GDP end time, to the maximum throughput that would be possible with perfect information. Similarly a metric for predictability could be the ratio of the expected flight delay assuming the GDP were to end as planned, to the expected delay given known uncertainty in the GDP end time. These metrics are described in more detail in Section 3.3.
2. *Votes or rankings on a set of candidate performance goal vectors* – Each airline *k* specifies its preferences in the form of votes or rankings I_{kp} for each of the *P* candidate performance goal vectors $G_p: p \in \{1, 2, \dots, P\}$. This airline input is most applicable for a discrete trade space, in which only a finite set of candidate performance goal vectors are valid or are under consideration. In the GDP example described in Section 4.2, this input would take the form of either a vote or a ranking, from each airline, for each of the candidate vectors.

3.2 Performance Goal Resolution Approaches

The ANSP determines the system-wide performance goal vector (G^*) by combining all airline inputs according to a defined resolution approach. Five different approaches are analyzed in this paper. For each approach described below and each airline *k*, the weights (w_k) are proportional to some non-decreasing function of the number of operations of *k* impacted by the initiative.

1. *Taking a weighted average of all airline-preferred performance goal vectors* – This is a simple and intuitive way of combining continuous-valued airline inputs. After each airline *k* has specified its preferred performance goal vector I_k, a weighted average performance goal vector is calculated. Mathematically, this can be represented as:

$$G^* = \Sigma_k w_k I_k \quad (1)$$

In the case of the GDP example, if specific parameters of the GDP representing capacity and predictability are traded off, such as GDP end time T and AAR C (specifying the duration and magnitude of the GDP), this equation is represented by (2), with the inputs from each airline k represented by T_k and C_k, and the system-wide solution represented by T^* and C^*.

$$[T^*, C^*] = [\Sigma_k w_k T_k, \Sigma_k w_k C_k] \quad (2)$$

The iterative framework described above is applied, allowing airlines to modify their preferred performance goal inputs according to inputs from other airlines. The process is continued for a set number of iterations, or until convergence to an equilibrium, as described in Section 4.

In order to gain further insight, the approach is simulated as described in Section 4. A sample result is illustrated for the case of two airlines and two performance criteria in Figure 2. This example uses an arc-shaped Pareto frontier (similar to that described by Equation 16) and a concave increasing quadratic payoff function (similar to that described in Equation 20). In this figure, sample truthful solutions for both airlines are shown (as red and blue circles), with lines of constant payoff overlaid (dashed lines in red and blue). A line of constant payoff represents the range of trade-offs between the simulated performance goals that would result in identical payoff to the airline. If the payoff functions are concave increasing functions, then these lines will always be convex. As we move towards the upper right of the figure, increasing the values of both performance goals, airline payoff increases. Therefore, the truthful solutions at which each airline maximizes its payoff fall where the lines of constant payoff are tangent to the Pareto frontier, as shown.

Figure 2 also shows sample strategic solutions for each airline (red × and blue ×), where each airline maximizes its payoff, given the use of a linear combination of all airline inputs to generate the system-wide performance goal vector. The approach by which these are identified is described in detail in Section 4.3. In the case of this sample instance, these points differ from the truthful solutions, because each airline is able to increase its payoff from the system-wide performance goal vector by gaming. As can be seen, each airline attempts to "pull" the system-wide solution towards its truthful solution. The final system-wide performance goal vector, calculated as a weighted average of the user preferred performance goal vectors, i.e., the strategic solutions submitted by each airline, is also shown on the figure (black ×). This is an interior point, and therefore not Pareto optimal.

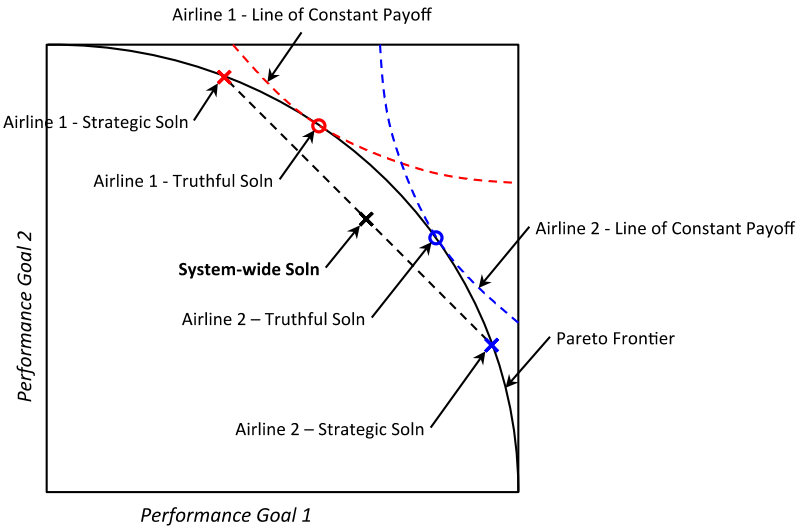


Figure 2. Sample result applying weighted average of all user preferred performance goal vectors.

2. *Taking a weighted average of all airline-preferred performance goal vectors (as in 1), and pushing the result out to the Pareto frontier* – This approach is similar to that presented above but avoids the issue of the combined output vector being in the interior of the trade space, and therefore not Pareto-optimal. In the case of a convex trade space where this Pareto frontier is non-linear, a weighted average of airline-preferred inputs might not fall on the Pareto frontier itself. After each airline k has specified its preferred performance goal vector I_k, a weighted average performance goal vector is calculated in the same way as in approach 1, above. This is then shifted out to the Pareto frontier. This shift can be done in many different ways. One reasonable approach is to do this in such a way as to maintain the same ratios of the values of each performance goal, generating a new system-wide performance goal vector that is on the Pareto frontier. This approach is also most applicable for a continuous trade space, for which the airline inputs are preferred performance goal vectors. Mathematically, this can be represented as:

$$G^* \in ParetoFrontier \qquad (3)$$
$$\text{such that:} \quad G_m^* / G_n^* = G_m' / G_n' \quad \text{for all } m, n$$
$$\text{where} \quad G' = \Sigma_k \, w_k \, I_k$$

Note that this point, G^*, on the Pareto frontier will always be unique because if there are two such points with the same ratio of the values of each performance goal then one of the two points will have a lower value of each performance goal compared to that for the other point and hence the former will not lie on a Pareto frontier.

In the case of the GDP example described above, this equation would be represented as follows:

$$T^* = f(C^*) \qquad (4)$$
$$\text{such that:} \quad T^* / C^* = T' / C'$$
$$\text{where:} \quad [T', C'] = [\Sigma_k \, w_k \, T_k, \Sigma_k \, w_k \, C_k]$$

As in approach 1, the iterative framework described above is applied, allowing airlines to modify their preferred performance goal inputs according to inputs from other airlines. The process is continued for a set number of iterations, or until convergence to an equilibrium, as described in Section 4.

Two sample results are illustrated for the case of two airlines and two performance criteria in Figure 3. Comparing Figure 3 to Figure 2, it is immediately clear that, by pushing the system-wide solution to the Pareto frontier, there is more gaming from both airlines. In Figure 3a, a case is shown in which the strategic solutions (red and blue ×s) for both airlines fall at corner points. In Figure 3b, a case is shown in which only one of the airlines requests a corner point. The other airline does not request a corner point because it is able to "pull" the system-wide solution to coincide with its truthful solution without moving to the corner point.

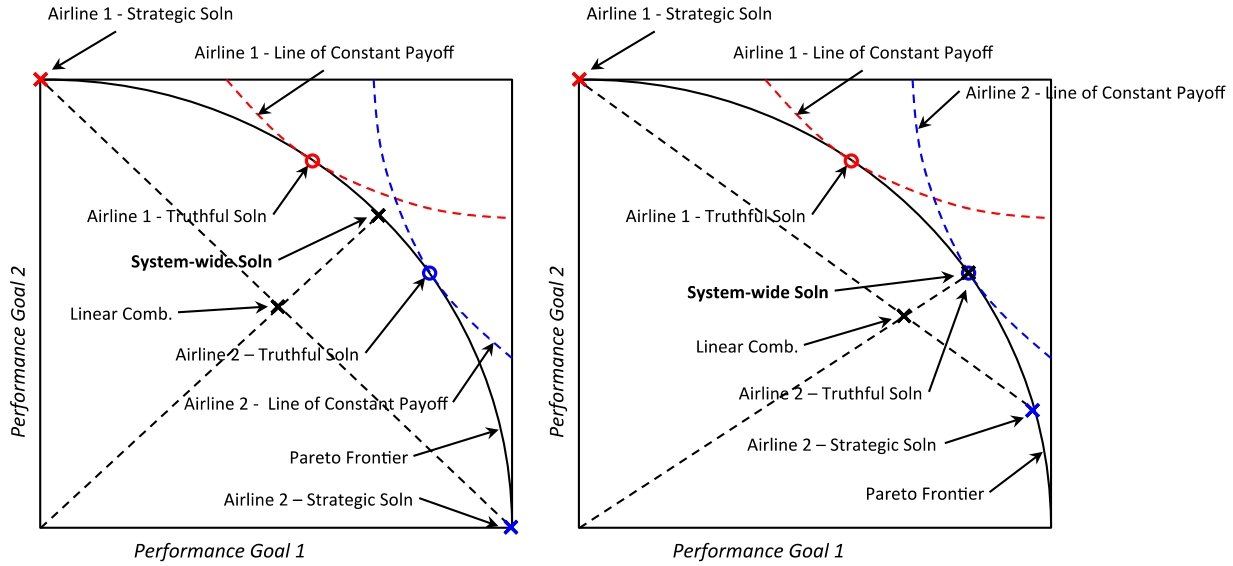


Figure 3. Sample result applying weighted average of all user preferred performance goal vectors, pushed to Pareto frontier: (a) both airlines request corner point solutions, (b) only one airline requests a corner point solution.

3. *Making a weighted random selection of the airline-preferred performance goal vectors* – The main motivation for considering this approach is that it eliminates strategic gaming behavior by the airlines, as we will see later in Section 5 and in Appendix E. After each airline k has specified its preferred performance goal vector I_k, one of the airline-preferred performance goal vectors I_k is randomly selected for G^*. The probability of each airline-preferred performance goal vector being selected is proportional to its weight w_k. This approach is applicable for both continuous and discrete trade spaces, for which the airline inputs are preferred performance goal vectors. Mathematically, this can be represented as:

$$G^* = I_j \qquad (5)$$

$$\text{where} \quad \Pr(j = k) = w_k \quad \text{for all } k$$

In the case of the GDP example, this equation would be represented as follows:

$$[T^*, C^*] = [T_j, C_j] \quad \text{where} \quad \Pr(j = k) = w_k \quad \text{for all } k \tag{6}$$

A sample result is illustrated for the case of two airlines and two performance criteria in Figure 4. As shown, the strategic and truthful solutions coincide. The reason for this is that airlines are not incentivized to game in any way because, if their solution is not chosen, their input does not affect the chosen solution in any way. Thus, they are incentivized to submit truthful solutions irrespective of how the probabilities are defined to randomly choose one of the airline inputs. This also means that the iterative framework described above is not necessary, as there is no incentive for any airline to change its preferred performance goal inputs based on other airline inputs. The system-wide goal vector (G^*) is also always Pareto optimal because each user input is Pareto optimal.

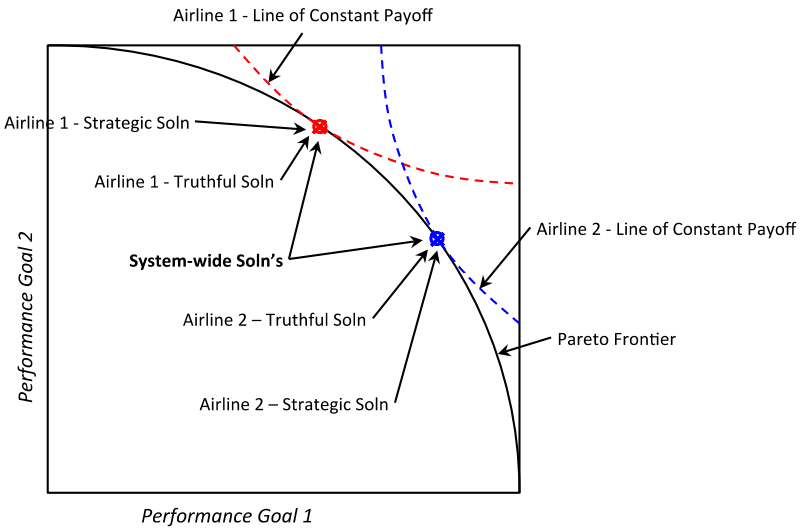


Figure 4. Sample result applying weighted random selection of the user preferred performance goal vectors.

One disadvantage of this approach is that it does not account for the fact that the payoff of any chosen solution may vary significantly across airlines. The chosen solution may therefore have highly disproportionate impacts on each airline. A solution may exist that has the lowest overall impact on all airlines, but is not the most preferred solution for any of the airlines.

4. *Ranking the candidate performance goal vectors based on airline preferences* – After each airline k has specified its preferences for each of the P performance goal vectors G_p in the form of descending ranks I_{kp} from P to 1 (P being the rank of that airline's most preferred performance goal vector, and 1 the rank of its least preferred performance goal vector), the combined rank for each vector is calculated as a weighted sum of individual ranks assigned by different airlines to that vector. The performance goal vector with the greatest combined rank is assigned to be the system-wide performance goal vector G^*. This approach is most applicable for a discrete trade space, in which only a candidate set of performance goal vectors is valid. Mathematically, this can be represented as:

$$G^* = G_q \quad (7)$$

where q is such that $R_q = \max (R_1 , R_2 , \ldots, R_P)$

and $R_p = \Sigma_k w_k I_{kp}$ for all $p \in \{1, 2, \ldots, P\}$

In the case of the GDP example described above, this equation would be represented as follows:

$$[T^*, C^*] = [T_q, C_q] \quad (8)$$

where q is such that $R_q = \max (R_1 , R_2 , \ldots, R_P)$

and $R_p = \Sigma_k w_k I_{kp}$ for all $p \in \{1, 2, \ldots, P\}$

As in approach 1 and 2, the iterative framework described above is applied, allowing airlines to modify their rankings according to rankings from other airlines. The process is continued for a set number of iterations, or until convergence to an equilibrium, as described in Section 4.

Ranking is intuitive to understand and use, and unlike the first three approaches discussed above allows the airline to input its relative preferences for all the candidate performance goal vectors, instead of just specifying its single most preferred performance goal vector. Furthermore, an airline's truthful input to the ranking mechanism does not require exact knowledge of the payoffs of each performance goal vector. Instead it only requires airlines to have a good idea of their comparative preference of one vector over the others. However, ranking is not devoid of drawbacks; the most significant is that convergence is not guaranteed, nor is a solution necessarily unique. Convergence is not guaranteed because airline rankings can alternate between two different rankings from iteration to iteration. In the cases run in this paper, this is a significant problem, with as little as 8% of runs converging (in the simplified GDP case at LGA). Convergence is highly dependent on the input parameters, and therefore different cases perform very differently (in contrast to the LGA case, 73% of runs converged in the simplified GDP case at EWR).

5. *Voting on the candidate performance goal vectors based on airline preferences* – After each airline k has specified its preferences for each of the P performance goal vectors G_p in the form of I_{kp} votes, the weighted sum of votes is calculated. This approach differs from ranking in that airlines can apply varying numbers of votes to each performance goal vector, according to their preferences, instead of just rank order. The total number of votes that can be assigned by an airline across different performance goal vectors is the same for each airline. The performance goal vector with the highest weighted sum of votes is assigned to be the system-wide performance goal vector G^*. This approach is most applicable for a discrete trade space, in which only a candidate set of performance goal vectors is valid. Mathematically, the approach can be represented as:

$$G^* = G_q \quad (9)$$

where q is such that $V_q = \max(V_1 , V_2 , \ldots, V_P)$

and $V_p = \Sigma_k w_k I_{kp}$ for all $p \in \{1, 2, \ldots, P\}$

In the case of the GDP example described above, this equation would be represented as follows:

$$[T^*, C^*] = [T_q, C_q] \quad (10)$$

where q is such that $V_q = \max(V_1 , V_2 , \ldots, V_P)$

and $V_p = \Sigma_k w_k I_{kp}$ for all $p \in \{1, 2, \ldots, P\}$

A very general voting framework is considered, in which each airline has a fixed maximum number of votes that it can distribute across available options (i.e., a form of range voting). We set this fixed number to 100. The airline may assign all its votes to its highest preference, or may distribute the votes across multiple options. The value of each airline's vote in determining the system-wide performance goal vector is proportional to that airline's weight. As in approaches 1, 2 and 4, the iterative framework described above is applied, allowing airlines to modify their votes according to what other airlines have voted. The process is continued for a set number of iterations, or until convergence to an equilibrium, as described in Section 4.

In the voting framework simulated, airlines are not required to allocate all their votes at any time, and can increase their votes from iteration to iteration. Only integer votes are considered. In order to ensure convergence, an airline is not permitted to reduce its vote for any candidate vector between iterations. An airline can only increase its vote, or maintain it at the same level.

As with ranking, voting allows airlines to input their relative preferences for all the candidate performance goal vectors, instead of just specifying their single most preferred performance goal vector. Unlike ranking, however, voting allows airlines to apply different values to different preferred performance goal vectors, beyond simply providing the rank order.

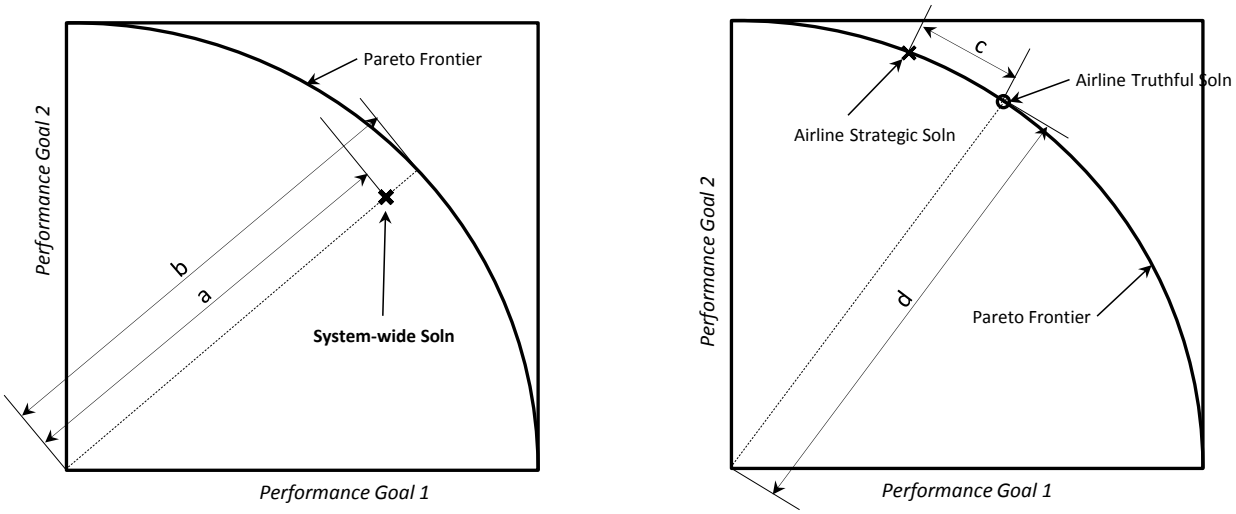
3.3 Characterization of Mechanism Performance Metrics

In order to evaluate different approaches to combine airline preferences to set system-wide performance goals, it is important to compare how each approach performs relative to the goals of each airline. A number of metrics are defined for this purpose: Pareto optimality, airline profitability, system optimality, equity and truthfulness. These are described and defined below. For the experiments modeling the generic initiative and the simplified GDP, each of these metrics is calculated for each run of a Monte-Carlo simulation, described in Section 4. A simple average is taken across different runs to estimate the expected value of each metric. In all cases, the metrics are designed to vary from 0 to 1, with larger values being better.

1. *Pareto Optimality* – Defined as how close the final system-wide solution is, on average, to the Pareto frontier. This metric provides a general indication of whether or not the selected system performance goals make maximum use of the available resources. It is defined as follows:

$$ParetoOpt = \frac{a}{b} \ . \tag{11}$$

where a and b are as defined in **Error! Reference source not found.**a. a is the distance of the system-wide solution from the origin and b is the length of the vector from the origin to the Pareto-frontier, which passes through the system-wide solution. The metric provides an indication of how far from the origin the system-wide solution is compared to how far it would be if on the Pareto frontier with the same ratio of the values of individual performance goals. The metric equals 1 when the system-wide solution is in fact Pareto optimal for every Monte-Carlo run.



(a) (b)

Figure 5. Parameters for defining metrics for (a) Pareto optimality, and (b) truthfulness.

2. *Airline Profitability* – Defined as the normalized difference between each airline's maximum payoff and its payoff applying the system-wide solution. Each airline's maximum payoff is the payoff obtained if we maximize that airline's payoff function over the trade space. The metric is averaged over all airlines, and provides an indication of how close each airline's profit is, at the system-wide solution, to its maximum possible profit over the trade space. It is defined as follows.

$$AirlineOpt = \frac{\sum_{k=1}^{K}\left(1-\frac{{P^M}_k - {P^*}_k}{max(|{P^M}_k|,|{P^*}_k|)}\right)}{K}. \quad (12)$$

where P^*_k represents the payoff for airline k applying the system-wide solution G^*, ${P^M}_k$ the maximum payoff for airline k, and K is the number of airlines. Note that the payoff can be negative, i.e., a cost. In order to ensure that the metric is meaningful (i.e. varying from 0 to 1) in this case, we define the denominator in equation (12) to be the larger of the absolute value of the maximum payoff and the absolute value of the payoff at the system-wide solution.

3. *System Optimality* – Defined as the difference between the total payoff across all airlines for the system optimal solution and the total payoff across all airlines applying the system-wide solution, normalized by the system optimal total payoff. The system optimal total payoff is calculated by maximizing the sum of airline payoff functions over the trade space. This metric provides an indication of how closely the ANSP goal of maximum system "effectiveness" is achieved. It is defined as follows.

$$SysOpt = 1 - \frac{\sum_{k=1}^{K} {P^{SysOpt}}_k - \sum_{k=1}^{K} {P^*}_k}{max(|\sum_{k=1}^{K} {P^{SysOpt}}_k|,|\sum_{k=1}^{K} {P^*}_k|)}. \quad (13)$$

where P^*_k represents the payoff for airline k applying the system-wide solution G^*, and ${P^{SysOpt}}_k$ the payoff for airline k at the point of system optimality (maximum total payoff across all airlines). Again, we define the denominator to ensure that the metric has meaningful value (i.e. varying from 0 to 1) even when payoffs are negative (costs).

4. *Equity* – Equity or fairness in resource allocation problems, in which some scarce resources must be allocated among multiple players by a central decision maker, has been extensively studied in social sciences, welfare economics and engineering. However, because of the multiple interpretations of concepts of fairness, and the different characteristics of different problems, no single criterion is universally accepted. For the purposes of this paper, we use one of the most prominent concepts in the literature: the *max-min* concept of fairness. This is one of the two concepts that Bertsimas *et al.* (2011) consider most applicable to air transportation (the other is the *proportional* concept of fairness). The max-min concept of fairness is a generalization of Rawlsian justice (Rawls, 1971) and the Kalai-Smorodinksy solution to the two-player game (Kalai & Smorodinksy, 1975). It maximizes the minimum (normalized) utility level that all players derive. In the context of this work, we denote this as P^{Fair}, the point of maximum minimum-payoff, or the "Fair" solution, which we calculate by solving a separate optimization problem in which the (non-positive) minimum normalized change in payoff relative to each airline's maximum payoff is maximized over the trade space. If we define e_k as the normalized change in payoff for airline k at the system-wide solution, and f_k as the normalized change in payoff for airline k at the "Fair" solution, our equity metric is defined as the ratio of the minimum value of $(1 - e_k)$ across all airlines to the minimum value of $(1 - f_k)$ across all airlines, as shown in equation (14). Subtracting the normalized change in payoff from 1 ensures that: (a) the metric is higher for a more equitable (as defined by the max-min fairness concept) strategic solution than for a less equitable strategic solution; and (b) the maximum possible value of the metric is 1, which is consistent with the definitions all our other performance metrics.

$$Equity = \frac{min_k(1-e_k)}{min_k(1-f_k)}, \quad \text{where } e_k = \frac{{P^M}_k - {P^*}_k}{max(|{P^M}_k|,|{P^*}_k|)} \text{ and } f_k = \frac{{P^M}_k - {P^{Fair}}_k}{max(|{P^M}_k|,|{P^{Fair}}_k|)}. \tag{14}$$

P^*_k represents the payoff for airline k applying the system-wide solution G^*, and ${P^{Fair}}_k$ the payoff for airline k at the point of maximum minimum-payoff, that is, at the "Fair" solution. Again, our definition of the denominator ensures that the metric always takes values between 0 and 1.

5. *Truthfulness* – Defined as how close a strategic solution (the airline's requested solution) is, on average, to the true preference of the airline (the airline's "truthful" solution). This "truthful" solution is the "maximum-payoff" solution referred to in metric 2 above. This provides an indication of the degree to which the airline is *gaming* the system. Truthfulness is not of value in itself, unlike the other metrics, but does provide an indication of whether the airline inputs are close to their true preferences. This is important because the larger the extent of gaming, the less likely it is that a fair mechanism can be implemented. The metric is defined as follows.

$$Truth = \frac{\sum_{k=1}^{K} max\left(1-\frac{c_k}{d_k},\ 0\right)}{K}. \tag{15}$$

where c_k and d_k are as defined in **Error! Reference source not found.**b for airline k, and K is the number of airlines. d_k is the distance of an airline's truthful solution from the origin. c_k is the distance from the truthful solution to the strategic solution of an airline. Subtracting the ratio c_k/d_k from 1 ensures that: (a) our truthfulness metric is higher for a strategic solution closer to the true solution than for a strategic solution farther from the true solution; and (b) the maximum possible

value of the metric is 1, which is consistent with the definitions of all the other performance metrics. It is noted that because c_k can be greater than d_k, we take a maximum of the numerator with 0 to ensure that the metric remains in the range from 0 to 1. (Note that in almost 100% of the cases in our experiments, the ratio c_k/d_k is less than or equal to 1.)

4. Computational Experimental Setup

In order to gain an understanding of how effective the approaches described in Section 3 are in setting system-wide performance goals, each approach is simulated, first for a generic TMI, and then for a specific TMI, a Ground Delay Program (GDP), at each of EWR and LGA airports. This allows us to derive general results, which cover various types of initiatives, and specific results, which provide an indication of likely results for a real-world case. In order to simplify the analysis, trade-offs are simulated between only two performance criteria, for a small number of airlines (between 3 and 4). This small number of airlines is reasonable given that the number of airlines with greater than 5% of operations at EWR and LGA, two of the busiest airports in the U.S., is 3 and 4, respectively (FAA, 2012). The forms of the trade space, Pareto frontier and airline payoff functions are described for each initiative in the following sections, followed by a description of the simulation methodology.

4.1 Generic Traffic Management Initiative

Trade Space and Pareto Frontier

For a generic TMI, we define a convex trade space, with the Pareto frontier defined by one of three alternative functions: an arc, a parabola, and a piecewise-linear function. The functional form of each of these is shown below as a function of G_1 and G_2, performance goals for two performance criteria (e.g., capacity and predictability). The alternative forms of the Pareto frontier, illustrated in Figure 6, are:

1. An arc: $G_1^2 + G_2^2 = 1$ (16)
2. A parabola: $G_2 = aG_1^2 + bG_1 + c$, (17)
   where $a = -1 / (G_{1TP} - 1)^2$; $b = -2aG_{1TP}$; and $c = -a - b$.

G_{1TP} represents the value of performance goal 1 at the turning point of the parabola, sampled from a uniform distribution between 0 and 1. Defining the parabola in this way results in a Pareto frontier that more closely resembles the Pareto frontiers in the more realistic GDP scenarios described below. Note that only the right half of the parabola forms the Pareto frontier. The upper portion from the turning point to the G_2 axis is formed by a horizontal straight line, as shown in Figure 6b.

3. A piecewise-linear function: $G_2 = -m_1G_1 + 1$, and (18)
   $G_2 = -m_2G_1 + m_2$. (19)

m_1 is sampled from a uniform distribution between 0 and 1, while m_2 is sampled from a uniform distribution between 1 and ∞. In our experiments, the Pareto frontier is described by only two lines, as shown in Figure 6c. In theory, any number of lines can be used.

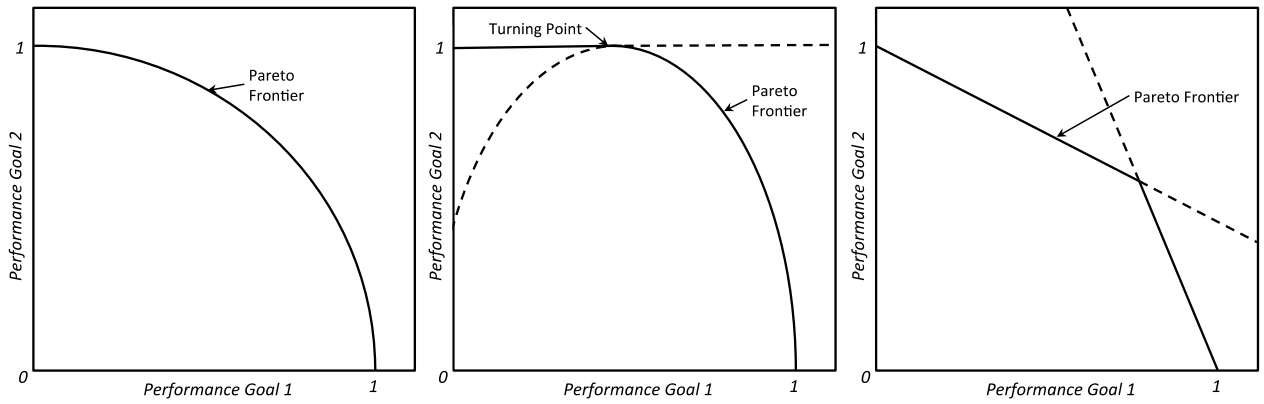


Figure 6. Alternative Pareto frontiers describing the trade space for the generic TMI: a) an arc, b) a parabola, and c) a piecewise-linear function.

Airline Payoff Functions

The airline payoff functions are defined by concave increasing functions in these performance goals, rather than linear functions, because of the small buffers and redundancies built into airline schedules to reduce the impact of performance goal reductions of smaller magnitudes. Beyond a certain threshold, decreases in, e.g., capacity, lead to faster than linear increases in passenger re-accommodation costs, crew delay and reserve crew costs, airline recovery costs, etc. For the generic TMI, the payoff function for airline *k* is defined as the sum of quadratic functions of each of the two performance goals G_1 and G_2, as follows.

$$P_k = \sum_g (a_{g,k}\left(G_g^*\right)^2 + b_{g,k}\left(G_g^*\right) + c_{g,k}), \qquad \forall k \qquad (20)$$
$$\text{with } a_{g,k} \leq 0 \text{ and } b_{g,k} > -2a_{g,k}.$$

The linear case is a special case of this function setting all $a_{g,k}$ values to 0. Note that the additive constant $c_{g,k}$ is later dropped from this expression without any loss of generality because it is inconsequential to any of the analysis. The parameters defining each airline's payoff function, $a_{g,k}$ and $b_{g,k}$, are sampled from uniform distributions from -1 to 0 (in the case of $a_{g,k}$) and from 0 to 2 (in the case of $b_{g,k}$, ensuring that $b_{g,k} > -2a_{g,k}$ so that the payoff is non-decreasing in each performance goal). It is noted that payoff functions do not in fact need to be additive functions of G^*_g, and may include coupling between different performance goals. The simpler additive function is, however, retained for this paper.

4.2 Ground Delay Program

Trade Space and Pareto Frontier

For the specific TMI, we consider a GDP under capacity uncertainty, in which we trade-off capacity and predictability. As described earlier, a GDP typically has three decision variables: duration, scope and magnitude. We apply two different models of a GDP. In the first, a simplified, computationally efficient model is applied in which we consider only duration, while in the second, we apply a more detailed model and consider both GDP duration and magnitude. In all cases we ignore the impact of GDP modification in

response to updated information. The first model is run using Monte-Carlo methods, making use of the expressions for the expected values of capacity and predictability metrics developed by Liu and Hansen (2012). The second model is run for two specific GDP scenarios, one at each of EWR and LGA respectively, making use of expressions for the capacity and predictability metrics described in Appendix A. In the absence of closed form expressions for their expected values, we resort to numerical integration. A detailed Monte-Carlo simulation of the second model is not within the scope of this paper, but is considered a useful next step in this research. The second model should be treated as an example of how our modeling framework is easily extendable to more complex forms of ATM initiatives and its results further validate our main conclusions, as shown later in Section 5.

For the simplified model of the GDP, we utilize metrics for capacity and predictability derived for a single airport by Liu and Hansen (2012), assuming a constant scheduled arrival demand rate, λ. When the GDP is initiated, the AAR is reduced from a known constant high level, C_H, which is assumed to be greater than λ, to a known constant low level, C_L, which is lower than λ. The planned duration of the GDP is T, at which time the AAR is expected to return to C_H. However, due to errors in prediction, the AAR may return to C_H at a different time, τ. When the GDP is initiated, T is set but τ is unknown, and assumed to be uniformly distributed between t_{min} and t_{max}. Conceptually, if T is set close to t_{min}, τ is likely to be larger than T, and the GDP ends late. In this case, capacity will be more fully utilized but there will be less predictability. Alternatively, if T is set close to t_{max}, then τ is likely to be smaller than T, and the GDP ends early. In this case, capacity will be underutilized and unnecessary delay will result. However, the delay is predictable. Thus, for this specific TMI, the only input from the airline is T, the planned duration of the GDP. Mathematical representations for capacity utilization and predictability, as developed by Liu and Hansen (2012), are presented below.

Capacity Utilization, α_c, is defined as the ratio of realized throughput, from the beginning of the GDP until the time when there is no more delay, to the maximum possible throughput with perfect information, were the airlines able to take advantage of the increase in AAR at time τ. α_c varies from 0 to 1, and is shown by Liu and Hansen (2012) to have the expected value shown below, which we use to define our metric for the performance goal of capacity G_c:

$$G_c = E[\alpha_c] = \frac{t_{max}-T}{t_{max}-t_{min}} + \frac{a/c}{t_{max}-t_{min}} \cdot log\left(\frac{B+cT}{b+ct_{min}}\right), \quad (21)$$

$$\text{where } a = \lambda\frac{C_H-C_L}{C_H-\lambda}T\ ,\ b = C_H\frac{C_H-C_L}{C_H-\lambda}T, \text{ and } c = C_L - C_H.$$

Predictability, α_p, is defined as the ratio of expected flight delay, assuming the GDP ends at the planned time T, to the total realized delay, i.e., the delay actually incurred given the early or late increase in AAR at τ. Again, α_p varies from 0 to 1 (given that we ignore GDP modifications in response to updated information), and is shown by Liu and Hansen (2012) to have the expected value shown below, which we use to define our metric for the performance goal of predictability G_p:

$$G_p = E[\alpha_p] = \frac{1}{t_{max}-t_{min}} \cdot \left(-\frac{T^2}{t_{max}} + 2T - t_{min}\right). \quad (22)$$

For the more detailed model of a GDP, we calculate expected values of the metrics for capacity and predictability through numerical integration for a single airport using a similar approach to that used by Liu

and Hansen (2012), but allowing for a varying scheduled arrival demand rate $\lambda(t)$, and planned and actual values of GDP AAR equal to C_L and C_l, respectively. When the GDP is initiated, a planned AAR, C_L, is specified, which is lower than λ. This represents the expected AAR for the duration of the GDP. As in the simplified model, the planned duration of the GDP is T, at which time the airport capacity is expected to return to the known C_H, which is greater than λ. Due to errors in prediction, the actual AAR (C_l) for the GDP may be different than that planned. Similarly, the AAR returns to C_H at a time τ, which may be different than that planned. When the GDP is initiated, C_L and T are set but C_l and τ are unknown. They are assumed to be uniformly distributed between $C_{L\,min}$ and $C_{L\,max}$, and t_{min} and t_{max}, respectively. Conceptually, if C_L is set close to $C_{L\,min}$, C_l is likely to be larger than C_L, and the rate at which aircraft arrive at the airport will be lower than the available capacity. In this case, capacity will be underutilized and unnecessary delay will result. However, the delay is predictable. Alternatively, if C_L is set close to $C_{L\,max}$, then C_l is likely to be smaller than C_L, and arriving aircraft will be required to hold because they will arrive at a faster rate than the airport can accommodate. In this case, the available capacity will be more fully utilized but there will be less predictability. As in the simplified model, if T is set close to t_{min}, τ is likely to be larger than T, and the GDP ends late. In this case, capacity will be more fully utilized but there will be less predictability. Alternatively, if T is set close to t_{max}, then τ is likely to be smaller than T, and the GDP ends early. In this case, capacity will be underutilized but predictability will be high.

As in the simplified model, Capacity Utilization, α_c, is defined as the ratio of realized throughput, from the beginning of the GDP until the time when there is no more delay, to the maximum possible throughput with perfect information, were the airlines able to take advantage of the actual GDP AAR C_l, and the increase in AAR at time τ. The value of α_c varies from 0 to 1. Predictability, α_p, is defined as the ratio of expected flight delay, assuming the planned GDP AAR C_L and the planned GDP end time T, to the total realized delay, i.e., the delay actually incurred given the actual GDP AAR C_l and the early or late increase in AAR at τ. Again, the value of α_p varies from 0 to 1 (given that we ignore GDP modifications in response to updated information). Expressions for α_c and α_p are presented in Appendix A. Based on these expressions, the expected value of the performance goals for capacity G_c and predictability G_p, can be calculated through numerical integration assuming uniform distributions of C_l between $C_{L\,\mathrm{min}}$ and $C_{L\,max}$, and τ between t_{min} and t_{max}. For numerical integration, we divide the range of possible values of C_l into 60 discrete points (steps of 0.1 aircraft/hour, across 6 aircraft/hour) and the range of possible values of τ into 60 discrete points (steps of 2 minutes, across 2 hours).

Given the specification of performance goals for capacity and predictability, the Pareto frontier is identified by calculating G_c and G_p across a range of planned C_L values from $C_{L\,min}$ to $C_{L\,max}$, and T values from t_{min} and t_{max}. The relationship between the capacity and predictability metrics was found to be very close to parabolic. A parabolic function is therefore fitted to the resulting maximum values of G_c for each given value of G_p, of the form:

$$G_c = aG_p^{\,2} + bG_p + c \tag{23}$$

We used 100 data points to estimate 3 parameter values. The estimated parameters and fit performance at EWR and LGA are shown in Table 1. The parabolic nature of the relationship is confirmed by the consistently high R^2 values shown.

Airport	*a*	*bB*	*c*	R^2
EWR	-0.485	0.663	0.776	0.91
LGA	-1.086	1.626	0.395	0.93

Table 1. Estimated Pareto frontier parameters for the detailed GDP model

Airline Payoff Functions

For the simplified GDP model, airline payoff functions are defined based on the cost of the flight delays incurred as a result of the GDP (payoff is therefore negative), similar to the way in which Liu and Hansen (2012) define a metric for efficiency. We assume that all the flights depart their respective origin airports under the assumption that the GDP ends at exactly the planned GDP end time *T*. If the actual GDP end time, *τ,* is before or equal to *T*, flights incur a delay on the ground equal to the difference between the actual departure time based on the planned GDP end time and the originally scheduled departure time. We assume that this delay is incurred at a specific ground delay cost of $Cost_D$ dollars/minute. If, however, the GDP is extended, and the GDP end time is after *T*, airborne delay is incurred in addition to ground delay (assuming the flight has already departed by the time the GDP is extended). The cost of incurring delay in airborne holding is generally significantly higher than that of holding on the ground. We define the ratio of the airborne delay cost per minute to the ground delay cost per minute as *k*, which is typically greater than 1. Using Liu and Hansen's (2012) simplified model of a GDP, we derive the airline payoff as a function of τ as follows:

$$Payoff(\tau) = -\frac{1}{2} Cost_D \frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \begin{cases} T^2 + k\tau^2 - kT^2, if\ \tau > T \\ T^2, if\ \tau \le T \end{cases}. \quad (24)$$

Given our assumed uniform distribution of τ between t_{min} and t_{max}, we derive the expected value of the payoff as follows:

$$P_k = E[Payoff] = -\frac{1}{2} Cost_D \frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \frac{\left(T^2(t_{max} - t_{min} - kt_{max}) + \frac{k}{3}{t_{max}}^3 + \frac{2}{3}kT^3\right)}{t_{max} - t_{min}} \quad (25)$$

The derivations of equations (24) and (25) are shown in Appendix B. Yi and Hansen (2012) apply a special case of these equations with *k*=2 in the derivation of their efficiency metric.

For the more detailed GDP model, airline payoff functions are also defined based on the cost of the flight delays incurred as a result of the GDP, as in the simplified model, but this flight delay is calculated as a function of both GDP end time and AAR. We assume that all the flights take off from their respective origin airports assuming that the GDP AAR is exactly as planned, C_L, and that the GDP will end at exactly the planned GDP end time *T*. If the actual GDP AAR, C_l, is greater than or equal to C_L, flights are assumed to be delayed on the ground in order to meet the planned GDP AAR C_L. Similarly, if the actual GDP end time, *τ,* is before or equal to *T*, flights also incur a delay on the ground. As in the simplified model, we assume that ground delay is incurred at a specific ground delay cost of $Cost_D$ dollars/minute. If, however, the actual GDP AAR, C_l, is less than C_L, airborne delay is incurred (in addition to ground delay), because flights arrive at the airport at a rate higher than the actual AAR. Similarly, if the GDP is extended, and the GDP end time is after *T*, airborne delay is incurred (in addition to ground delay). As in the simplified model, we define the ratio of

the airborne delay cost per minute to the ground delay cost per minute as *k*, which is typically greater than 1. Expressions describing airline payoff as a function of C_L and T are presented in Appendix A. Based on these expressions, the expected value of airline payoff can be calculated numerically assuming uniform distributions of C_l between $C_{L\,\min}$ and $C_{L\,max}$, and τ between t_{min} and t_{max}.

Given the specification of airline payoff as well as performance goals for capacity and predictability, airline payoff functions are identified by calculating airline payoff, G_c and G_p across a range of planned C_L values from $C_{L\,min}$ to $C_{L\,max}$, and T values from t_{min} and t_{max}. The relationship between the performance goals and the airline payoffs was found to be very close to linear. As a result, a linear function was fitted to the resulting airline payoff as a function of G_c and G_p, of the form:

$$P_k = b_{c,k} * G_c + b_{p,k} * G_p + Constant_k \qquad \forall k \tag{26}$$

The estimated parameters and fit performance for each airline at EWR and LGA are shown in Table 2. The values of $Constant_k$ are not shown because they are inconsequential to the analysis. The linear nature of the relationship is confirmed by the consistently high R^2 values.

Airport	*Airline*	*Capacity* $b_{c,k}$	*Predictability* $b_{p,k}$	R^2
EWR	Delta	1,209	355	0.96
	US Airways	1,550	287	0.90
	United	1,528	343	0.93
LGA	Delta	1,154	621	0.97
	US Airways	1,436	495	0.93
	United	1,430	596	0.95
	American	1,554	608	0.95

Table 2. Estimated payoff-function parameters for the detailed GDP model

4.3 Simulation Methodology

The methodology for simulating the generic initiative and the simplified GDP model is to run a Monte-Carlo simulation, sampling values for key parameters for each simulation run from appropriate distributions, and solving for the system optimal, most equitable, truthful, and strategic airline-preferred performance goal vectors in each case. The mechanism performance metrics described in Section 3.3 are then calculated in each case for the performance goal vectors and the corresponding airline payoffs. The metrics are then averaged over all simulation runs for comparison. Other aggregate statistics, such as standard deviation, minimum and maximum are also calculated. This is done for each of the performance goal resolution approaches described in Section 3.2, allowing each averaged metric to be compared across the different approaches. In contrast, the detailed GDP model is run for only one GDP at each of EWR and LGA, simulating in greater detail how airlines would behave under those specific conditions. Because of issues with computational efficiency associated with the numerical integration process in calculating expected values, it is not possible to run the detailed GDP model using Monte-Carlo methods. Again, we solve for the system optimal, most equitable, truthful, and strategic airline-preferred performance goal vectors in each

case, before calculating the mechanism performance metrics described in Section 3.3 for each of the performance goal resolution approaches described in Section 3.2.

As described in Section 3.3, the system optimal solution is found by identifying the performance goal vector that, if applied, maximizes the total payoff across all airlines combined. The most equitable solution is found by identifying the performance goal vector that, if applied, maximizes the minimum (normalized) payoff of any airline. Each airline's truthful solution is found by identifying the performance goal vector that, if applied directly as the chosen system-wide performance goal vector, maximizes the individual payoff of that airline. Finally, each airline's strategic solution is found by simulating a myopic best-response game between airlines. This best-response game is solved iteratively, and is myopic in the following sense: In each iteration, each airline's input is determined by identifying the performance goal vector that, if combined with the performance goal vectors input by all other airlines in the previous iteration using the given performance goal resolution approach, maximizes the payoff for that airline. This applies to both a continuous trade space and a discrete trade space. In the latter case, the discrete performance goal vectors are ranked or voted upon by each airline in such a way that the winning performance goal vector maximizes the payoff for that airline, given the rankings or votes of all other airlines in the previous iteration. The performance goal vectors of all airlines are combined based on the different approaches described in Section 3.2. Each airline's input is re-optimized at every iteration, based on the corresponding inputs of all other airlines. The game is simulated until convergence, but stopped after 100 iterations if the stopping criterion is not met by then. Convergence is said to be achieved when the difference in each of the airline preferred performance goal vectors in successive iterations is less than 1×10^{-6} (for the case of a continuous trade space), or when each of the airline preferred performance goal vectors in successive iterations are the same (for the discrete case). Note that each performance goal has a value between 0 and 1. If convergence is not reached in 100 iterations, the results of that simulated game are not included in the calculation of the metrics described in Section 3.3. For each approach, the percentage of runs that converged to equilibrium before 100 iterations is provided in Section 5.7. The mathematical formulations for each of the optimizations described above are presented in Appendix C.

Applying Monte-Carlo methods for the generic initiative and the simplified GDP model, the simulations are run 1,000 times, randomly sampling for the key parameters in each case. In the generic initiative, the parameters sampled include those defining the Pareto frontier and each airline's payoff function, described in Section 4.1, and the airline weights (w_k). This last parameter, which represents the relative numbers of impacted flights of different airlines, is sampled from a uniform distribution from 0 to 1 for each airline, and then divided by the sum of sampled values across all airlines, to ensure that the sum of airline weights is 1.0.

In the case of ranking and voting, 5 candidate performance goal vectors are specified in each run, which the airlines must rank or vote on. The capacity metric value (G_c) for each candidate vector is assumed to vary uniformly from 0 to 1, and fall on the Pareto frontier. This automatically and uniquely determines the predictability metric value (G_p). The results were also run with the 5 discrete performance goal vectors distributed evenly along the Pareto frontier, with the results changing by less than 2% of the values shown in Section 5 in all cases.

For both the simplified and detailed GDP models, we simulate operations at both EWR airport and at LGA airport. This allows us to compare the performance of the different approaches applied to real airports, and also allows us to compare how each approach performs for different types of airports. EWR was chosen as an example of an airport that is dominated by a single airline: United Airlines. LGA was chosen as an example of an airport with no single dominant carrier. Both these airports are among the busiest and the most congested airports in the U.S. (Barnhart *et al.*, 2012). In each scenario, we simulate all airlines that operate at least 5% of the operations at that airport. For EWR, this includes only United Airlines, Delta Air Lines and US Airways; and for LGA, it includes United Airlines, Delta Air Lines, US Airways, and American Airlines (FAA, 2012).

For the simplified GDP model, we analyze historical data for 2011 (specifically, FAA Aviation Systems Performance Metrics (ASPM) data (FAA, 2012) and Form 41 data (DOT, 2012)) to define distributions from which to sample values for each parameter in the Monte-Carlo simulation. The parameters, and the type of distributions from which they are sampled, are described in Table 3. In each case, average values from the Operational Evolution Partnership (OEP) 35 airports (FAA, 2012) are also shown for comparison.

Parameter	***Distribution Assumed***	***Airline***	***EWR***	***LGA***	***OEP35***
Ratio of airline operations to total - w_k	None (Fixed)	American	-	0.207	0 to 1
		Delta	0.067	0.272	
		United	0.886	0.125	
		US Airways	0.112	0.282	
Avg. scheduled arrival demand rate - λ [ac/hr]	Poisson	All	33.7	35.1	40.0
Avg. AAR during GDP - C_L [ac/hr] (Std. dev.)	Normal	All	36.2 (5.32)	33.3 (4.44)	45.0 (20.9)
Avg. AAR not under GDP - C_H [ac/hr] (Std. dev.)	Normal	All	39.8 (6.40)	34.6 (9.9)	63.2 (27.2)
Avg. GDP duration [hrs] (Std. dev.)	Normal	All	6.95 (2.90)	8.01 (3.87)	6.14 (3.57)
Avg. difference between max and min GDP durations [hrs] (Std dev.)	Normal	All	1.59 (1.35)	1.62 (1.21)	2.03 (1.77)
Avg. airline flight operating cost during airport holding - $k \times Cost_D$ [US$/hr] (Std. dev.)	Normal	American	$6,592 ($3,770)		$7,968 ($5,997)
		Delta	$5,365 ($1,435)		
		United	$5,777 ($4,115)		
		US Airways	$5,211 ($2,798)		
Avg. airline flight operating cost on ground - $Cost_D$ [US$/hr] (Std. dev.)	Normal	American	$2,819 ($3,845)		$4,061 ($4,184)
		Delta	$2,074 ($922)		
		United	$2,592 ($2,195)		
		US Airways	$2,607 ($1,343)		
Ratio of airline cost during holding and on the ground - k	Normal	American	2.27 (0.28)		2.32 (1.40)
		Delta	2.79 (1.75)		
		United	2.36 (2.42)		
		US Airways	2.11 (0.53)		

Table 3. GDP parameter distributions

Based on equations (27) and (28) below, the maximum and minimum duration of the GDP (t_{max} and t_{min}), are calculated by sampling the average GDP duration t_{avg}, and the difference between the maximum and minimum durations t_{diff}. The distributions of these two factors (t_{avg} and t_{diff}) are described in Table 3. These distributions are based on historical GDP durations and historically observed differences between the initial GDP duration and the actual GDP duration, respectively. This data was made available by Metron Aviation®.

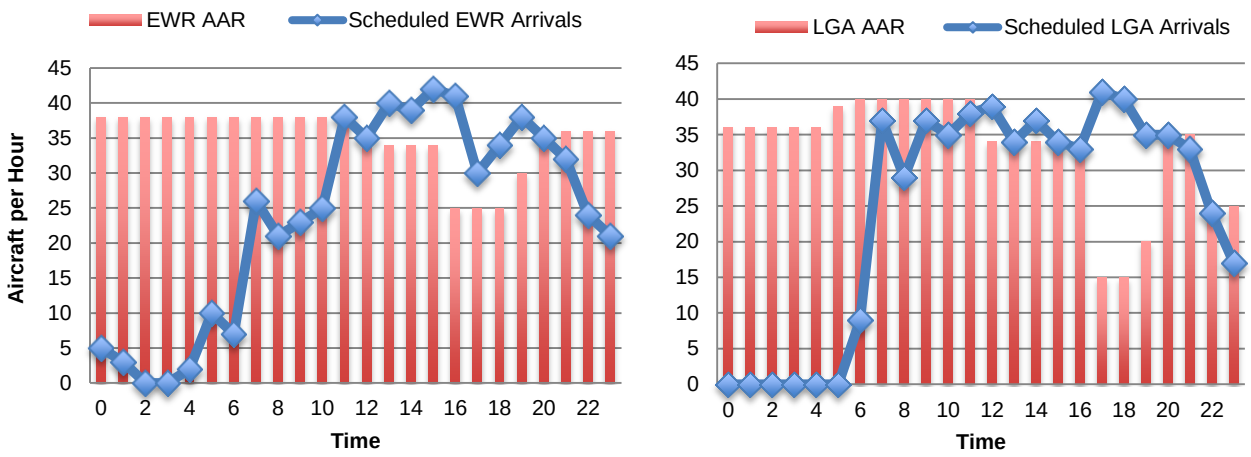
$$t_{max} = t_{avg} + \frac{t_{diff}}{2} \tag{27}$$

$$t_{min} = t_{avg} - \frac{t_{diff}}{2} \tag{28}$$

In the case of ranking and voting, 5 candidate performance goal vectors are specified in each run, which the airlines must rank or vote on. T values are sampled for each candidate vector from a uniform distribution between t_{min} and t_{max}. The capacity and predictability metric values (G_c and G_p) for each candidate vector are then calculated by substituting each sampled value of T in equations (21) and (22). Across all approaches and all runs of the simplified model, the minimum values for G_c and G_p at EWR are found to be 0.71 and 0.03, respectively, while at LGA, they are found to be 0.75 and 0.10, respectively.

We use historical data for the two simulated days to define each parameter, as follows.

- The ratio of operations of each airline to the total (w_k) and the airline delay costs ($k \times Cost_D$ and $Cost_D$) are as shown in Table 3.
- The scheduled arrival demand rate (λ), which varies with time, is specified per hour according to Figure 7.
- The AAR after the GDP (C_H) is assumed to be 36 aircraft/hour and 35 aircraft/hour at EWR and LGA, respectively, based on the specified AARs in Figure 7.
- The minimum AAR of the GDP ($C_{L\ min}$) is assumed to be 22 aircraft/hour and 12 aircraft/hour at EWR and LGA, respectively, while the maximum AAR of the GDP ($C_{L\ max}$) is assumed to be 28 aircraft/hour and 18 aircraft/hour, respectively. These compare to the actual GDP AARs in Figure 7 of 25 aircraft/hour and 15 aircraft/hour, respectively.
- The minimum duration of the GDP (t_{min}) is assumed to be 3 hours and 2 hours at EWR and LGA, respectively, while the maximum duration of the GDP (t_{max}) is assumed to be 5 hours and 4 hours, respectively. These compare to the actual durations in Figure 7 of 4 hours and 3 hours, respectively.
- In the case of ranking and voting, 5 candidate performance goal vectors are specified in each run, which the airlines must rank or vote on. The capacity metric value (G_c) for each candidate vector is assumed to be uniformly distributed between the minimum and maximum values of G_c. These are calculated based on the ranges of planned C_L values from $C_{L\ min}$ to $C_{L\ max}$, and T values from t_{min} and t_{max} described above. The corresponding value of the predictability metric is then uniquely determined by solving equation (23) that characterizes the Pareto frontier. At EWR, the minimum values for G_c and G_p are found to be 0.91 and 0.55, respectively, while at LGA, they are found to be 0.92 and 0.74, respectively.



(a) (b)

Figure 7. Scheduled arrival rates and AARs at EWR on July 25, 2011 (a) and LGA July 8, 2011 (b).

5. Analysis and Results

In this section, we describe the results of our analysis. We evaluate the five approaches for performance goal resolution (as described in Section 3.2) based on five mechanism performance metrics, namely Pareto optimality, system optimality, airline profitability, equity, and truthfulness (as described in Section 3.3). First, we note that none of these five approaches, except for the Weighted Random Choice approach, is completely immune to manipulation by players. Voting and ranking approaches have a long and notorious history of results about their potential manipulability, starting with Arrow (1951) who famously stated that "when voters have three of more distinct alternatives, no voting system can convert the ranked preferences of individuals into a community-wide (complete and transitive) ranking while also meeting a certain set of criteria, namely: unrestricted domain, non-dictatorship, Pareto efficiency, and independence of irrelevant alternatives." This result and the subsequent body of research (notably including Gibbard 1973 and Satterthwaite 1975) shows that our voting and ranking approaches are not completely immune to manipulation, and we cannot guarantee their system optimality and truthfulness, in general.

For the Weighted Average approach, Appendix D provides some guarantees of pure strategy equilibrium existence, uniqueness, truthfulness, and the convergence of the best response dynamic. Most of these results are only applicable for the case of linear payoff functions, which is a special case of the concave payoff functions that we have assumed in this paper. Appendix E states, and proves, three propositions related to the necessary and sufficient conditions for the truthfulness of the Weighted Random Choice approach and the Weighted Average approach. Propositions 2 and 3 in Appendix E prove that the truthfulness of the Weighted Average approach under the arc-shaped and parabolic Pareto frontiers can be disproved unless some very restrictive conditions are met. A similarly restrictive result can be proved for the piecewise linear case. It involves many more cases than the first two, and as a result is both lengthy and relatively less informative. Hence we decided to exclude this proof from this paper. Instead, we motivate the issues with the truthfulness of the Weighted Average approach for the piecewise linear Pareto frontier using the following simple

example. Consider a case of two equally weighted airline players. If their respective truthful solutions lie at two different points on the same line segment of the Pareto frontier, then each will be incentivized to move its strategic solution away from each other's solution in order to *'pull'* the resultant system-wide solution closer to the respective truthful solutions. The Weighted Average Pushed to Pareto Frontier approach (approach 2) is typically even more prone to manipulation than the Weighted Average approach, as we will see later in this section. The exact conditions for truthfulness are more difficult to prove in this case. Figure 3 provides some intuition towards this end. Finally, Proposition 1 in Appendix E proves that the Weighted Random Choice approach is always guaranteed to yield a truthful solution. However, it is easy to see that unless each player's truthful solution is identical to the system optimal solution, the system-wide solution will not be system optimal.

Note that all of these aforementioned results focus purely on the truthfulness and system optimality properties in an absolute sense. These results do not eliminate the possibility of these approaches yielding a system-wide solution that is relatively close to the Pareto frontier, the system-optimal solution, the most profitable solution, the most equitable solution, and/or the truthful solution. Through Monte-Carlo simulation, we evaluate and compare the five approaches in terms of their relative closeness to these idealized solution points. Here, the closeness is as defined by the five metrics in Section 3.3.

Each of the approaches described in Section 3.2 is simulated for the generic initiative and for the simplified and detailed models of a GDP at EWR and LGA, for a small number of competing airlines (three for the generic initiative and GDP at EWR, four for the GDP at LGA), trading off two performance criteria. Monte-Carlo methods are applied in the cases of the generic initiative and the simplified model of the GDP to simulate varying conditions by sampling values from distributions for each of the parameters, as described in Section 4.3. The more detailed model of the GDP is run for two sample GDP cases: July 25 2011 at EWR, and July 8 2011 at LGA. Results are presented for each model in the sections below.

5.1 Generic Traffic Management Initiative

The results for the generic TMI are shown in Table 4 below. Each of the metrics described in Section 3.3 is calculated for each of the performance goal resolution approaches described in Section 3.2. In each case, the simulations are run with three different types of Pareto frontiers, as described in Section 4.1: an arc, a parabola, and a piecewise-linear function. The airline payoff functions are as described in Section 4.1. The average metric values across all 1000 Monte-Carlo runs are presented, along with the standard deviation (in parentheses), and maximum and minimum values (in square brackets).

For each of the three Pareto frontier shapes, all the simulation experiments converged for all the approaches except for approach 4, i.e., ranking. For ranking, only between 21% and 25% of the simulation runs converged depending on the Pareto frontier shape. Given that the ranking approach converges in only a small fraction of the simulated cases, it seems unlikely to be useful for practical implementation in the form simulated here because a stable solution can rarely be achieved. However, convergence can be improved to some extent by adjusting the values assigned to the different ranks, although this has not been investigated in detail in this paper.

Approach	*Pareto Frontier*	*Pareto Optimality*	*Airline Profitability*	*System Optimality*	*Equity*	*Truthfulness*
1. Weighted Avg.	Arc	0.95 (0.04) [1.00;0.76]	0.91 (0.07) [1.00;0.56]	0.96 (0.04) [1.00;0.72]	0.88 (0.12) [1.00;0.15]	0.91 (0.05) [1.00;0.72]
	Parabola	0.98 (0.02) [1.00;0.84]	0.93 (0.06) [1.00;0.62]	0.95 (0.05) [1.00;0.70]	0.87 (0.13) [1.00;0.25]	0.96 (0.05) [1.00;0.66]
	Piecewise-Linear	0.98 (0.04) [1.00;0.66]	0.90 (0.07) [1.00;0.60]	0.95 (0.05) [1.00;0.60]	0.87 (0.14) [1.00;0.17]	0.86 (0.14) [1.00;0.24]
2. Weighted Avg. Pushed to Pareto	Arc	1.00 (0.00) [1.00;1.00]	0.96 (0.05) [1.00;0.66]	0.98 (0.03) [1.00;0.82]	0.94 (0.11) [1.00;0.36]	0.43 (0.12) [0.86;0.04]
	Parabola	1.00 (0.00) [1.00;1.00]	0.94 (0.06) [1.00;0.67]	0.96 (0.05) [1.00;0.67]	0.88 (0.13) [1.00;0.23]	0.84 (0.18) [1.00;0.20]
	Piecewise-Linear	1.00 (0.00) [1.00;1.00]	0.92 (0.07) [1.00;0.60]	0.97 (0.04) [1.00;0.68]	0.90 (0.14) [1.00;0.31]	0.54 (0.15) [1.00;0.14]
3. Weighted Random Choice	Arc	1.00 (0.00) [1.00;1.00]	0.92 (0.08) [1.00;0.44]	0.96 (0.07) [1.00;0.52]	0.85 (0.16) [1.00;0.11]	1.00 (0.00) [1.00;1.00]
	Parabola	1.00 (0.00) [1.00;1.00]	0.96 (0.08) [1.00;0.41]	0.98 (0.06) [1.00;0.21]	0.93 (0.14) [1.00;0.09]	1.00 (0.00) [1.00;1.00]
	Piecewise-Linear	1.00 (0.00) [1.00;1.00]	0.89 (0.13) [1.00;0.36]	0.95 (0.11) [1.00;0.24]	0.79 (0.25) [1.00;0.00]	1.00 (0.00) [1.00;1.00]
4. Ranking	Arc[1]	1.00 (0.00) [1.00;1.00]	0.88 (0.13) [1.00;0.44]	0.90 (0.14) [1.00;0.37]	0.83 (0.20) [1.00;0.09]	0.71 (0.21) [1.00;0.00]
	Parabola[2]	1.00 (0.00) [1.00;1.00]	0.88 (0.15) [1.00;0.24]	0.88 (0.16) [1.00;0.24]	0.82 (0.22) [1.00;0.12]	0.79 (0.17) [1.00;0.18]
	Piecewise-Linear[3]	1.00 (0.00) [1.00;1.00]	0.86 (0.14) [1.00;0.28]	0.89 (0.15) [1.00;0.28]	0.80 (0.23) [1.00;0.11]	0.69 (0.20) [1.00;0.12]
5. Voting	Arc	1.00 (0.00) [1.00;1.00]	0.94 (0.07) [1.00;0.49]	0.98 (0.05) [1.00;0.57]	0.91 (0.15) [1.00;0.13]	1.00 (0.03) [1.00;0.68]
	Parabola	1.00 (0.00) [1.00;1.00]	0.98 (0.06) [1.00;0.56]	0.99 (0.03) [1.00;0.70]	0.97 (0.10) [1.00;0.30]	1.00 (0.02) [1.00;0.66]
	Piecewise-Linear	1.00 (0.00) [1.00;1.00]	0.92 (0.09) [1.00;0.42]	0.98 (0.05) [1.00;0.45]	0.88 (0.18) [1.00;0.04]	0.99 (0.03) [1.00;0.62]

Table 4. Generic TMI: Comparison of average metrics (with standard deviation in parentheses) [and maximum and minimum values in square brackets] for all approaches.

[1] These results apply only to the 23% of runs that converged using this approach.
[2] These results apply only to the 25% of runs that converged using this approach.
[3] These results apply only to the 21% of runs that converged using this approach.

The first metric, describing how close the system-wide solution is to the Pareto frontier, shows that Pareto optimality is achieved by all approaches with the exception of approach 1, which takes a weighted average of the user preferred performance goal vectors. This is expected, as this is the only approach that is not specifically designed to achieve Pareto optimality. Even in this approach, however, the metric is consistently high – average values are between 0.95 and 0.98 (with standard deviations between 0.02 and 0.04), while the lowest value achieved in the Monte-Carlo simulation is 0.66 (piecewise-linear). This is because the preferred solutions input by each airline, from which the weighted average is calculated, are always Pareto optimal.

The second, third and fourth metrics, describing how close the system-wide solution is to maximizing each airline's payoff; how close the system-wide solution is to system optimality; and how equitable the system-wide solution is, all show similar results. The metrics are generally lowest for approach 4, which applies ranking. This approach also shows the highest variability across different Monte-Carlo runs, and the lowest minimum metric values (0.24, 0.24 and 0.09 for metrics two, three and four, respectively). Each of these three metrics is generally highest for approaches 2 and 5, which push the weighted average of the user preferred performance goal vectors to the Pareto frontier, and apply voting, respectively. Approach 2 also shows the lowest metric variability and highest minimum values (0.60, 0.67 and 0.23 for metrics two, three and four, respectively) of all the approaches. Approach 1 and approach 3 (a weighted random choice of the user preferred performance goal vectors) show results that are slightly lower than those for approaches 2 and 5, and with slightly greater variability. The results are also generally similar across the different Pareto frontiers, with the piecewise-linear results typically slightly lower than for the arc or parabola. In many cases the piecewise-linear Pareto frontier leads to the system-wide solution falling at the intersection of the lines (or as close to it as possible in the case of discrete options), which is generally further from the airline-optimal, system-optimal, and most-equitable solutions than for the arc or parabola.

The fifth metric describing the truthfulness of the airline inputs is particularly low (between 0.43 and 0.84) for approach 2. This is a recognized concern about approach 2, as discussed in Section 3.2. In contrast, the metric is highest (1.00) for approach 3. This is expected as this approach is specifically designed to prevent gaming. Also notable, however, is that the metric is also high for approach 5 (between 0.99 and 1.00), while it is lower for approach 4 (between 0.69 and 0.79), and to a lesser extent for approach 1 (between 0.86 and 0.91). The variability of the truthfulness metric is greatest for approach 4 (between 0.17 and 0.21), and lowest for approach 3 (0) and approach 5 (between 0.02 and 0.03). Similarly, the minimum value of truthfulness is lowest for approach 4 (0.00), highest for approach 3 (1.00), and second highest for approach 5 (0.62). The metric also shows some variation across the different Pareto frontiers, with the piecewise-linear results typically slightly lower than for the arc or parabola. This is because gaming leads each airline to request solutions at the vertices of the Pareto frontier, as illustrated in Figure 3 (or as close to the vertices as possible in the case of discrete options), which are generally further from their truthful solutions than for the arc and parabola cases. Furthermore, the truthfulness metric has a higher value for the parabolic frontier than for the arc-shaped frontier, which is consistent with our theoretical results (Propositions 2 and 3 in Appendix E) that show that the necessary and sufficient conditions for truthfulness are less restrictive for the parabolic frontier than for the arc-shaped frontier.

In summary, based on the results from this generic initiative, we can derive a number of practical insights. First, we note that amongst the five approaches for performance goal resolution, all but ranking converge to equilibrium in all the simulation runs. Ranking, on the other hand, frequently runs into convergence issues, and is thus unsuitable for practical implementation, at least in this form. It is worth noting here that voting (approach 5) could potentially have run into convergence issues had it not been for the constraint that prohibits airlines from reducing their votes for any of the candidate solutions in subsequent iterations. Of the four approaches 1, 2, 3, and 5, approach 2 performs substantially worse than all the other approaches in terms of truthfulness. Therefore, even though approach 2 shows reasonable levels of performance on some of the other metrics, the submitted preferences by the airlines are unlikely to have much meaning. Approaches 1, 3 and 5, therefore, seem to be reasonable candidates for practical implementation. There is still, however, quite a large variation in performance across these approaches, as displayed in Table 4. Overall, voting out-perform approaches 1 and 3, with none of the average values of the metrics being below 0.88 across all the three shapes of the Pareto frontier. These conclusions based on the generic initiative set the stage for further evaluation of these approaches on GDP initiatives at EWR and LGA.

5.2 Ground Delay Program

Simplified Model of a GDP

The results applying the simplified model of the GDP are shown in Table 5. Each of the metrics described in Section 3.3 is calculated in each case, for each of the performance goal resolution approaches described in Section 3.2. The Pareto frontiers and airline payoff functions are simulated as described in Section 4.2. The average metric values across all 1000 Monte-Carlo runs are shown, along with the standard deviation (in parentheses), and maximum and minimum values (in square brackets).

It is noted that no results are presented in Table 5 for approach 2, which pushes the weighted average of the user preferred performance goal vectors out to the Pareto frontier. This is because, for this model, approach 2 is no different from approach 1, which just takes the weighted average of the user preferred performance goal vectors. This is because the simplified GDP scenario is defined for a single ANSP decision variable, that is, the planned GDP end time, T. Performance metrics for capacity and predictability are calculated as a function of this single decision variable, as described in Section 4.2. Therefore, for each planned GDP end time, there are unique values for both the capacity and predictability performance goals. The Pareto frontier represents the corresponding values for these two performance goals at all possible planned GDP end times. The entire trade space therefore lies on the Pareto frontier, and no interior points are feasible. For this reason, a more detailed GDP model is also run, modeling multiple decision variables, as described above, although only for a single GDP case at each of EWR and LGA.

For the GDP at LGA, like for the generic initiative, we find that all simulation runs, except under ranking, converge to an equilibrium within 100 iterations. Under ranking, only 8% of the simulation runs converge. This result reinforces our conclusion that ranking is unlikely to yield stable outcomes for practical implementation, at least in the form simulated here. For the GDP at EWR (Table 5), 73% of simulation runs converged under ranking which is significantly better, but still not 100% reliable. The reason for this improved performance is the highly skewed distribution of operations across airlines at EWR, with United

operating 89% of flights. However, for an initiative involving a more even distribution of operations across airlines, ranking seems to be unsuitable for practical implementation in its current form.

Approach	*Airport*	*Pareto Optimality*	*Airline Profitability*	*System Optimality*	*Equity*	*Truthfulness*
1. Weighted Avg.	EWR	1.00 (0.00) [1.00;1.00]	0.95 (0.05) [1.00;0.76]	0.98 (0.03) [1.00;0.74]	0.90 (0.12) [1.00;0.49]	0.95 (0.03) [0.99;0.72]
	LGA	1.00 (0.00) [1.00;1.00]	0.83 (0.11) [1.00;0.41]	0.82 (0.13) [1.00;0.35]	0.67 (0.16) [1.00;0.22]	0.96 (0.05) [1.00;0.59]
2. Weighted Avg. Pushed to Pareto Frontier	EWR	-	-	-	-	-
	LGA	-	-	-	-	-
3. Weighted Random Choice	EWR	1.00 (0.00) [1.00;1.00]	0.97 (0.04) [1.00;0.67]	0.99 (0.03) [1.00;0.51]	0.96 (0.06) [1.00;0.52]	1.00 (0.00) [1.00;1.00]
	LGA	1.00 (0.00) [1.00;1.00]	0.98 (0.03) [1.00;0.67]	0.99 (0.02) [1.00;0.75]	0.96 (0.06) [1.00;0.45]	1.00 (0.00) [1.00;1.00]
4. Ranking	EWR[4]	1.00 (0.00) [1.00;1.00]	0.95 (0.06) [1.00;0.62]	0.95 (0.08) [1.00;0.49]	0.93 (0.10) [1.00;0.46]	0.95 (0.05) [1.00;0.55]
	LGA[5]	1.00 (0.00) [1.00;1.00]	0.98 (0.04) [1.00;0.70]	0.99 (0.04) [1.00;0.70]	0.98 (0.06) [1.00;0.63]	0.98 (0.03) [1.00;0.80]
5. Voting	EWR	1.00 (0.00) [1.00;1.00]	0.98 (0.04) [1.00;0.78]	1.00 (0.01) [1.00;0.87]	0.98 (0.07) [1.00;0.64]	1.00 (0.00) [1.00;1.00]
	LGA	1.00 (0.00) [1.00;1.00]	0.99 (0.02) [1.00;0.83]	1.00 (0.01) [1.00;0.87]	0.98 (0.05) [1.00;0.56]	1.00 (0.00) [1.00;0.94]

Table 5. Simplified GDP model: Comparison of average metrics (with standard deviation in parentheses) [and maximum and minimum values in square brackets] for all approaches.

The first metric shows that all approaches simulated are Pareto optimal at EWR and LGA, as was the case for the generic initiative in Table 4. All other metrics, with the exception of those under approach 1 at LGA, are high (between 0.90 and 1.00), have relatively low variability (between 0 and 0.12), and have minimum values above 0.49. They are also very similar to the results presented for the generic initiative in Table 4, with the exception of somewhat lower values of metrics 2, 3 and 4 at LGA for approach 1 (between 0.67 and 0.83, in comparison to values between 0.87 and 0.96 in Table 4), and somewhat higher values of metrics 2, 3, 4 and 5 in approach 4 (between 0.93 to 0.99, in comparison to values between 0.69 and 0.90 in Table 4). The similarities between the results presented here and in Table 4 suggest that the conclusions drawn for the generic initiative are indeed applicable to more realistic cases, as simulated here. Particularly, the best

[4] These results apply only to the 73% of runs that converged using this approach.

[5] These results apply only to the 8% of runs that converged using this approach.

performing approach across all metrics at both EWR and LGA is approach 5, (with the average of all metrics for both airports consistently at least 0.98), with approach 3 not far behind (with the average of all metrics for both airports consistently at least 0.96).

The results for EWR and LGA differ relatively little. In approach 1, airline profitability, system optimality and equity are all somewhat higher at EWR, where a single airline is dominant, than at LGA, where no airline is dominant. For all other approaches, the results differ very little across the two airports. This suggests that only approach 1 is likely to be significantly affected by the dominance of an airline at the airport. (Note that approach 2, not considered here, is also likely to be affected by airline dominance.)

The overall results from the simplified GDP scenarios at EWR and LGA therefore reinforce our conclusions from the generic initiative. Once again, ranking in its current form can be eliminated from consideration because of convergence issues for the GDP scenario at LGA. Approaches 3 and 5 (and to a lesser extent approach 1) seem to be reasonable candidates for practical implementation. However, based on these results, voting (approach 5) performs the best, with none of the metrics values being below 0.98 for either the EWR or the LGA scenario.

More Detailed Model of a GDP

The results applying the more detailed model of the GDP are shown in Table 6. Each of the metrics described in Section 3.3 is calculated in each case, for each of the performance goal resolution approaches described in Section 3.2. The Pareto frontiers and airline payoff functions simulated are as estimated in Section 4.2. The results are presented with a higher number of significant figures than for the other models in order to distinguish differences in the metric values. Note that we present only a single value per metric, rather than average, standard deviation, minimum, and maximum as for the other models, because we only do this evaluation for a single GDP case rather than using a Monte-Carlo simulation approach.

Comparing the results in Table 6 to the results in Table 5, it is immediately clear that the values of almost all the metrics are closer to unity than modeled using the simplified GDP model, hence the requirement for increased significant figures in the table. This is because of the size of the feasible region in the detailed GDP model. As described in Section 4.3, while the metrics G_c and G_p can each nominally vary from 0 to 1, because of the variation in GDP end time T from t_{min} to t_{max} and GDP magnitude C_L from $C_{L\,min}$ to $C_{L\,max}$, G_c and G_p only vary from 0.91 to 1 and 0.55 to 1, respectively, at EWR, and from 0.92 to 1 and 0.74 to 1, respectively, at LGA. In the generic initiative, G_c and G_p both vary from 0 to 1, while in the simplified GDP, G_c and G_p vary from 0.71 to 1 and 0.03 to 1, respectively at EWR, and from 0.75 to 1 and 0.10 to 1, respectively at LGA. Hence the range over which the different solutions can be found in the detailed GDP is reduced relative to the previous results. All the solutions also lie on or close to the Pareto frontier, combining to make most of the metric values close to unity. Also, note that this is a single computational experiment and not a set of 1000 Monte-Carlo simulation runs, as in the case of the generic and the simplified GDP initiatives. Therefore, the low variation in the metric values is not surprising.

Approach	*Airport*	*Pareto Optimality*	*Airline Profitability*	*System Optimality*	*Equity*	*Truthfulness*
1. Weighted Avg.	EWR	0.9999	0.9989	0.9998	0.9986	1.0000
	LGA	0.9992	0.9976	0.9985	0.9965	1.0000
2. Weighted Avg. Pushed to Pareto	EWR	1.0000	0.9995	0.9999	0.9997	0.3784
	LGA	1.0000	0.9892	0.9895	0.9701	0.4203
3. Weighted Random Choice	EWR	1.0000	0.9991	1.0000	0.9993	1.0000
	LGA	1.0000	0.9992	1.0000	0.9992	1.0000
4. Ranking	EWR[6]	-	-	-	-	-
	LGA[6]	-	-	-	-	-
5. Voting	EWR	1.0000	0.9997	1.0000	1.0000	1.0000
	LGA	1.0000	0.9999	1.0000	1.0000	1.0000

Table 6. Detailed model of GDP at EWR (July 25 2011) and LGA (July 8 2011): Comparison of metrics for all approaches.

As in the previous results, we find that all simulation runs, except ranking, converge to an equilibrium within 100 iterations. At both EWR and LGA, ranking does not converge for the sample GDP run, reinforcing our conclusion that ranking is unlikely to yield stable outcomes for practical implementation, at least in the form simulated here. As in Table 4, all approaches are Pareto optimal except approach 1, as expected. However, because of where the airline preferred solutions lie on the Pareto frontier in this case, even approach 1 is almost Pareto optimal. All other metrics, with the exception of truthfulness in approach 2, are very high. The very low values of truthfulness in approach 2 at both EWR (0.38) and LGA (0.42) are consistent with the low values of truthfulness in approach 2 for the generic initiative in Table 4. This confirms that truthfulness may be a problem for this approach in real initiatives as well. Similar to the results for the generic initiative and the simplified GDP, the best performing approach is 5 (voting), with all metric values consistently above 0.9997, and approach 3 is close behind with all metric values consistently above 0.9991. Similar to the results for the simplified GDP, the results for EWR and LGA differ very little.

The overall results from the detailed GDP scenarios reinforce our conclusions from the generic initiative and the simplified GDP scenarios. Ranking (approach 4) in its current form can be eliminated because of convergence issues, and approach 2 performs relatively poorly on at least one metric compared to the rest of the approaches. Approaches 1, 3 and 5 seem to be reasonable candidates for practical implementation, but voting, approach 5, consistently performs the best.

[6] This approach failed to converge at this airport.

6. Conclusions

In this paper a number of approaches are considered within the context of an air traffic management system for setting system-wide performance goals based on preferred trade-offs as expressed by airlines. This is the first study to investigate how system-wide performance goals can be set based on airline inputs. We investigate this by evaluating differing approaches for doing so using a rigorous game-theoretic method, which identifies the potential for gaming in each approach. Each approach is evaluated based on a number of proposed criteria representing stakeholder objectives, including Pareto optimality, airline profitability, system optimality, equity and truthfulness of airline preferences. By simulating each of the approaches using Monte-Carlo methods, sampling values for input parameters from representative distributions, we offer a broad evaluation of each approach to performance-based ATM. A generic TMI is simulated, as well as a simplified model of a ground delay program at Newark Liberty International airport and LaGuardia airport. Finally, a more detailed model of a ground delay program is also run, for a single case at each of Newark Liberty International airport and LaGuardia airport, confirming the validity of the previous Monte-Carlo results.

The results presented in Section 5 suggest that taking a weighted average of the user preferred performance goal vectors (approach 1), making a weighted random choice of the user preferred performance goal vectors (approach 3), or voting on ANSP provided candidate performance goal vectors (approach 5) may all be reasonable candidates for practical implementation. However, approach 1 has somewhat inferior performance for the simplified GDP case, while approach 3 might suffer from lack of stakeholder buy-in because the process of combining airline preferences involves a non-deterministic (randomized) step. Overall, voting shows the highest promise. Our results based on multi-criteria evaluation suggest that this approach has the most potential to satisfy the objectives of all stakeholders.

While the results presented here suggest that voting is likely to be the best way to satisfy the objectives of all stakeholders in setting performance goals, different voting schemes have not been examined in detail. It is therefore recommended that future research include the detailed analysis of specific voting schemes, including instant run-off voting (Lewyn, 2012; Robb, 2012) and majority judgment voting (Balinski and Laraki, 2007; 2011), with the goal of developing implementable airline-driven approaches for performance-based air traffic management that will benefit both airlines and ANSPs.

Acknowledgements

This work was funded through the National Center of Excellence for Aviation Operations Research (NEXTOR), a consortium of 8 universities contracted by the FAA to provide research support for a wide variety of aviation issues. Their support is gratefully acknowledged. The authors would also like to thank colleagues on the Service Expectations project (Distributed Mechanisms for Determining NAS-Wide Service Level Expectations), including Rich Jehlen of the FAA, Mike Ball and Prem Swaroop of the University of Maryland, and Mark Hansen and Yi Liu of the University of California, Berkeley for helpful discussions, and information on the GDP models.

References

- Arrow, K. 1951. Individual values and social choice. New York: Wiley.
- Balinski, M., and R. Laraki. 2007. A theory of measuring, electing and ranking. Proceedings of the National Academy of Sciences of the United States of America 104(21): 8720-8725.
- Balinski, M., and R. Laraki. 2011. Election by majority judgment: experimental evidence. B. Dolez, B. Grofman, A. Laurent, eds. In Situ and Laboratory Experiments on Electoral Law Reform: French Presidential Elections. Springer, New York.
- Barnhart, C., Fearing D., V. Vaze. 2012. Modeling passenger travel and delays in the National Air Transportation System. Working paper.
- Bertsimas, D., S. Stock-Patterson. 1998. The air traffic flow management problem with enroute capacities. Operations Research 46(3): 406–422.
- Bertsimas, D., S. Stock-Patterson. 2000. The traffic flow management rerouting problem in air traffic control: A dynamic network flow approach. Transportation Science 34(3): 239–255.
- Debreu, G. 1952. A social equilibrium existence theorem. Proceedings of the National Academy of Sciences 38: 886-893.
- DOT (Department of Transportation), 2012. *Air Carrier Financial Reports (Form 41 Financial Data), Schedule P52*, U.S. Department of Transport, Research and Innovative Technology Administration, Bureau of Transportation Statistics, Washington DC, USA.
- FAA (Federal Aviation Administration), 2012. *Operations and Performance Data, Aviation System Performance Metrics (ASPM)*, http://www.apo.data.faa.gov/aspm/ASPMframe.asp [Cited April 13, 2012].
- FAA (Federal Aviation Administration), 2011. *Performance (Service) Based NAS ATM*, White Paper, Version 0.2, System Operations Services, Planning & Performance, AJR-5, Washington DC, June, 2011.
- Garcia-Chico, J.L., H. Idris, J. Krozel, K.S. Sheth, 2008. *Task Analysis for Feasibility Assessment of a Collaborative Traffic Flow Management Concept*, 8th AIAA Aviation Technology, Integration and Operations Conference, Anchorage, Alaska, 14-19 September 2008.
- Fan, K. 1952. Fixed point and minimax theorems in locally convex topological linear spaces. Proceedings of the National Academy of Sciences 38: 121-126.
- Gibbard, A. 1973. Manipulation of voting schemes: A general result. Econometrica 41: 587-601.
- Glicksberg, I.L. 1952. A further generalization of the Kakutani fixed point theorem with application to Nash equilibrium points. Proceedings of the American Mathematical Society 3: 170-174.
- ICAO (International Civil Aviation Organization), 2005. *Global Air Traffic Management Operational Concept*, Appendix D, Doc. 9854.
- Kalai, E., M. Smorodinsky. 1975. Other solutions to Nash's bargaining problem. Econometrica 43(3): 513–518.
- Lewyn, M.E. 2012. Two Cheers for Instant Runoff Voting, The Selected Works of Michael E Lewyn, Working Paper, Available at: http://works.bepress.com/lewyn/74.
- Liu, Y., M. Hansen. 2012. *Performance trades and cost optimization in ground delay programs: single airport case*, 5th International Conference on Research in Air Transportation, Berkeley, CA, 22-25 May 2012.

- JPDO (Joint Planning and Development Office), 2007. *Concept of Operations for the Next Generation Air Transportation System*, Version 2.0, Joint Planning and Development Office, Washington DC, June 13, 2007
- Lulli, G., A. R. Odoni. 2007. The European air traffic flow management problem. Transportation Sci. 41(4) 431–443.
- Odoni, A. R., L. Bianco. 1987. The flow management problem in air traffic control. A. R. Odoni, L. Bianco, G. Szego, eds. Flow Control of Congested Networks. Springer-Verlag, Berlin.
- Rawls, J. 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.
- Robb, D.M. 2012. The Effect of Instant Runoff Voting on Democracy, PhD Thesis, Department of Political Science, University of California, Irvine, CA.
- Rosen, J.B. 1965. Existence and Uniqueness of Equilibrium Points for Concave N-Person Games. Econometrica 33(3): 520-534.
- Satterthwaite, M.A. 1975. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. Journal of Economic Theory 10: 187-217.
- Sheth, K.S., Gutierrez-Nolasco, S., 2008. *Incorporating User Preferences in Collaborative Traffic Flow Management*, AIAA Guidance, Navigation and Control Conference, Honolulu, Hawaii, 18 - 21 August 2008.

Appendix A – Calculation of Metrics and Airline Payoff in Detailed GDP Model

Four scenarios exist:

1. The actual GDP end time τ is smaller than or equal to the planned GDP end time T, *and* the actual GDP AAR C_l is smaller than the planned GDP AAR C_L, i.e., $\tau \leq T$ and $C_l < C_L$
2. The actual GDP end time τ is greater than the planned GDP end time T, *and* the actual GDP AAR C_l is smaller than the planned GDP AAR C_L, i.e., $\tau > T$ and $C_l < C_L$
3. The actual GDP end time τ is smaller than or equal to the planned GDP end time T, *and* the actual GDP AAR C_l is greater than or equal to the planned GDP AAR C_L, i.e., $\tau \leq T$ and $C_l \geq C_L$
4. The actual GDP end time τ is greater than the planned GDP end time T, *and* the actual GDP AAR C_l is greater than or equal to the planned GDP AAR C_L, i.e., $\tau > T$ and $C_l \geq C_L$

Each of these scenarios is illustrated in Figure A-1 below.

As defined by Liu and Hansen (2012), Capacity Utilization, α_c, is defined as the ratio of realized throughput (or expected throughput given known uncertainties in actual AAR C_l and GDP end time τ), from the beginning of the GDP until the time when there is no more delay, defined here as N_R, to the maximum throughput that would have been possible with perfect information, were the airlines able to take advantage of the actual AAR C_l, and the increase in AAR at time τ, defined here as N_I. α_c varies from 0 to 1. Predictability, α_p, is defined as the ratio of expected flight delay, assuming the planned AAR C_L and the GDP end time at the planned time T, defined here as D_P, to the total realized delay, i.e., the delay actually incurred given the actual AAR C_l and the early or late increase in AAR at τ (or the expected delay given known uncertainties in C_l and τ), defined here as D_R. Again, α_p varies from 0 to 1 (given that we ignore GDP modifications in response to updated information). Therefore:

$$\alpha_c = N_R / N_I \quad \text{(A-1)}$$

$$\alpha_p = D_P / D_R \quad \text{(A-2)}$$

The payoff for airline a, defined as $Payoff_a$, is a function of ground delay D_G, airborne delay D_A, ground delay cost $Cost_{Da}$, and the ratio of airborne delay cost to ground delay cost k_a, as follows:

$$Payoff_a = Cost_{Da} \times (D_G + k_a \times D_A) \quad \text{(A-3)}$$

Scenario 1: $\tau \leq T$ and $C_l < C_L$

Figure A-1a and b show cumulative diagrams of passenger demand and throughput for the detailed GDP case, under the scenario in which $\tau \leq T$ and $C_l < C_L$. In Figure A-1a the difference between τ and T is sufficiently large that, once the AAR returns to C_H at τ, the holding stack is cleared before T, so no more airborne delay is incurred by the later incoming flights. In contract, in Figure A-1b, the difference between τ and T is sufficiently small that, once the AAR returns to C_H at τ, the holding stack in not cleared before the later incoming flights arrive. These incoming flights do therefore incur some airborne delay. The figures allow calculation of α_c, α_p and $Payoff_a$ based on equations A-1, A-2 and A-3. This requires calculation of D_G, D_A, D_P, D_R, N_R, and N_I, as follows:

$$D_G = \int_{t=0}^{t_G} \lambda(t)dt - \frac{1}{2}T^2 C_L - \frac{1}{2}(t_G - T)^2 C_H - (t_G - T)C_L T \quad \text{(A-4)}$$

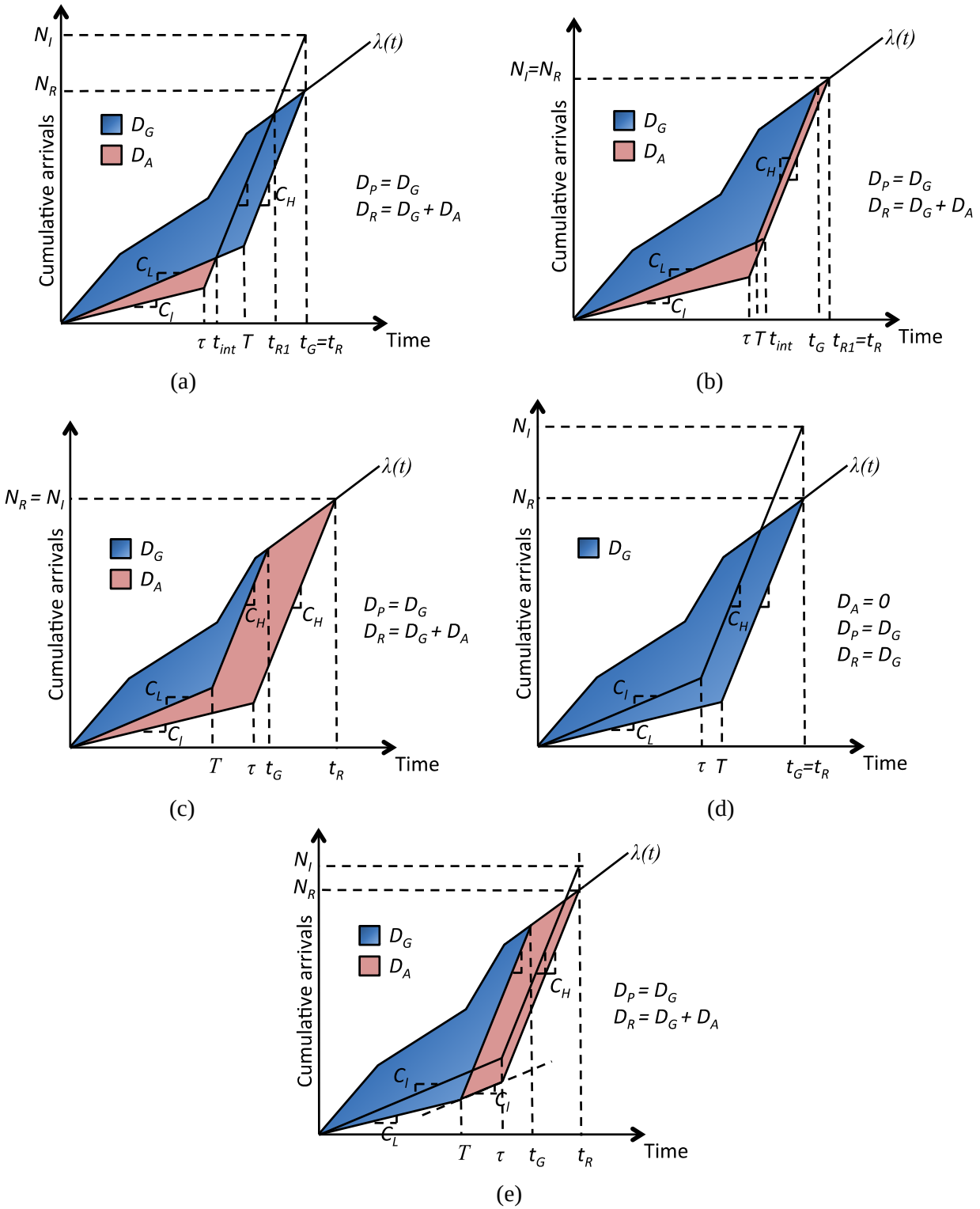


Figure A-1. Cumulative diagrams of passenger demand and throughput in detailed GDP cases, (a) Scenario 1a: $\tau \le T$, $C_l < C_L$ and $t_{int} \le T$, (b) Scenario 1b: $\tau \le T$, $C_l < C_L$ and $t_{int} > T$, (c) Scenario 2: $\tau > T$ and $C_l < C_L$, (d) Scenario 3: $\tau \le T$ and $C_l \ge C_L$, and (e) Scenario 4: $\tau > T$ and $C_l \ge C_L$.

$$D_A = \begin{cases} {t_{int}}^2 C_L - \frac{1}{2}\tau^2 C_l - \frac{1}{2}(t_{int} - \tau)^2 C_H - (t_{int} - \tau) C_l \tau & \text{if } t_{int} \le T \text{ (i. e., Fig. A1a)} \\ \int_{t=0}^{t_R} \lambda(t)dt - \frac{1}{2}\tau^2 C_l - \frac{1}{2}(t_R - \tau)^2 C_H - (t_R - \tau) C_l \tau - D_G & \text{if } t_{int} > T \text{ (i. e., Fig. A1b)} \end{cases} \quad \text{(A-5)}$$

where $t_{int} = \tau \frac{C_H - C_l}{C_H - C_L}$ (A-6)

$$D_P = D_G \quad \text{(A-7)}$$

$$D_R = D_G + D_A \quad \text{(A-8)}$$

$$N_R = \lambda(t_R) \quad \text{(A-9)}$$

$$N_I = C_l \tau + C_H (t_R - \tau) \quad \text{(A-10)}$$

t_G can be identified by solving the following equality for t_G:

$$\lambda(t_G) = C_H t_G + C_L T - C_H T \quad \text{(A-11)}$$

t_R can be identified by solving the following equality for t_{R1}:

$$\lambda(t_{R1}) = C_H t_{R1} + C_l \tau - C_H \tau \quad \text{(A-12)}$$

$$t_R = \max(t_{R1}, t_G) \quad \text{(A-13)}$$

Scenario 2: $\tau > T$ and $C_l < C_L$

Figure A-1c shows this scenario. Given that $N_R = N_I$, in this case, $\alpha_c = 1$. D_G, D_P and D_R are calculated according to equations A-4, A-7 and A-8 while D_A, is calculated as follows.

$$D_A = \int_{t=0}^{t_R} \lambda(t)dt - \frac{1}{2}\tau^2 C_l - \frac{1}{2}(t_R - \tau)^2 C_H - (t_R - \tau) C_l \tau - D_G \quad \text{(A-14)}$$

In this case, t_R can be identified by solving the following equality for t_R:

$$\lambda(t_R) = C_H t_R + C_l \tau - C_H \tau \quad \text{(A-15)}$$

t_G can be identified by solving equation A-11 for t_G.

Scenario 3: $\tau \le T$ and $C_l \ge C_L$

Figure A-1d shows this scenario. Given that $D_A = 0$, and therefore $D_P = D_R$, in this case, $\alpha_p = 1$. D_G is calculated according to equation A-4, while N_R, and N_I are calculated as in equations A-9 and A-10. t_R, which equals t_G can be identified by solving equation A-11 for t_G.

Scenario 4: $\tau > T$ and $C_l \ge C_L$

Figure A-1e shows this scenario. N_R and N_I are calculated according to equations A-9 and A-10. D_G, D_P and D_R are calculated according to equations A-4, A-7 and A-8, while D_A is calculated as follows.

$$D_A = \int_{t=0}^{t_R} \lambda(t)dt - \frac{1}{2}T^2 C_L - \frac{1}{2}(t_R - \tau)^2 C_H - \frac{1}{2}(\tau - T)^2 C_l - (t_R - T) C_L T - (t_R - \tau) C_l (\tau - T) - D_G \quad \text{(A-16)}$$

In this case, t_R can be identified by solving the following equality for t_R:

$$\lambda(t_R) = C_H t_R + (C_l - C_H)\tau + T(C_L - C_l) \quad \text{(A-17)}$$

t_G can be identified by solving equation A-11 for t_G.

Appendix B – Derivation of Expected Value of Airline Payoff in a Simplified GDP

We consider a simplified model of a GDP with uncertain duration, as described by Liu and Hansen (2012). We consider a simple queuing scenario with a constant scheduled arrival demand rate, λ. When the GDP is initiated, the airport capacity is reduced from a constant high level, C_H, which is assumed to be greater than λ, to a constant low level, C_L, which is lower than λ. The planned duration of the GDP is T, at which time the airport capacity is expected to return to C_H. However, due to errors in prediction, the capacity returns to C_H at a time, τ, which might be different from T. When the GDP is initiated, T is set but τ is unknown, assumed to be uniformly distributed between t_{min} and t_{max}. Conceptually, if T is set close to t_{max}, then τ is likely to be smaller than T, and the GDP ends early. In this case, capacity will be underutilized and unnecessary delay will result. However, the delay is predictable, and all the delay is incurred on the ground at the origin airport, at an assumed ground delay cost of $Cost_D$ \$/min. This scenario is illustrated in Figure B-1a below. Alternatively, if T is set close to t_{min}, τ is likely to be larger than T, and the GDP ends late. This means the GDP has to be extended. Capacity is fully utilized, but the delay is less predictable, and a portion of the delay is incurred in the air, at higher cost, assumed to be k time greater than the ground delay cost of $Cost_D$ \$/min. This scenario is illustrated in Figure B-1b below.

The total ground delay incurred in the two scenarios illustrated in Figure B-1 is represented by the area of the blue triangles in the two respective figures. The airborne delay is represented by the area of the red quadrilateral in Figure B-1b. Based on the parameters introduced above, and the geometry shown in the figures, the ground delays can be shown to be as follows:

$$Delay_{Ground}(\tau) = \frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot T^2, \qquad \text{for all } \tau \tag{B-1}$$

Airborne delay can be shown to be as follows:

$$Delay_{Airborne}(\tau) = \frac{1}{2}\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot (\tau^2 - T^2), \qquad \text{for all } \tau > T \tag{B-2}$$

Given the cost of ground delay ($Cost_D$ \$/min) and the ratio of airborne delay cost to ground delay cost (k), the airline payoff (i.e., −Cost) can therefore be shown to be:

$$Payoff(\tau) = -\frac{1}{2}Cost_D\frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \begin{cases} T^2 + k\tau^2 - kT^2, if\ \tau > T \\ T^2, if\ \tau \le T \end{cases} \tag{B-3}$$

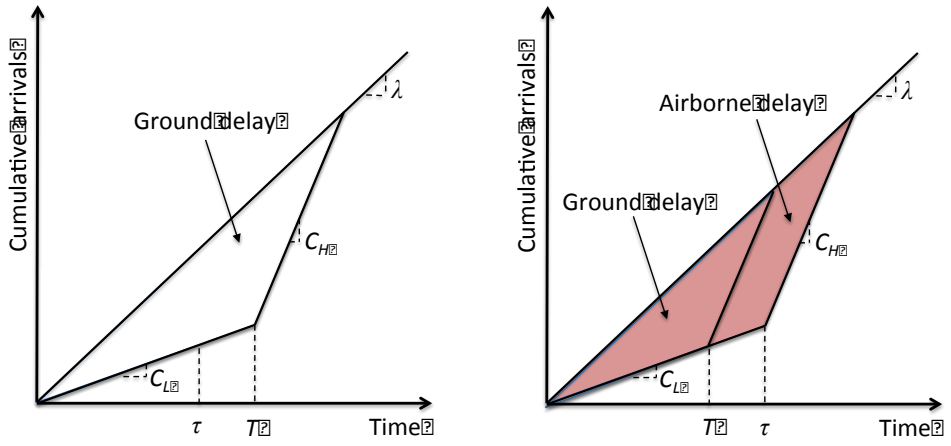


Figure B-1. Flight delay in simplified model of a GDP (modified from Liu and Hansen, 2012), a) when GDP planned end time is late, and b) when GDP planned end time is early

τ is unknown, but because we assume that it is uniformly distribution between t_{min} and t_{max}, we can calculate the expected value of the payoff ($g(\tau)$), based on the equation:

$$E[g(\tau)] = \int_{-\infty}^{\infty} g(\tau) \cdot f(\tau) d\tau \tag{B-4}$$

where $f(\tau)$ is the probability density function for τ. Given that we define τ uniformly distribution between t_{min} and t_{max}:

$$f(\tau) = 1/(t_{max} - t_{min}) \qquad \text{for } \tau \epsilon [t_{min}, t_{max}], \text{ 0 otherwise} \tag{B-5}$$

Therefore, substituting equation B-3 for $g(\tau)$ and equation B-5 for $f(\tau)$ in equation B-4,

$$E[Payoff] = -\frac{1}{2} Cost_D \frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \left(\int_{t_{min}}^{T} \frac{T^2}{t_{max} - t_{min}} d\tau + \int_{T}^{t_{max}} \frac{T^2 + k\tau^2 - kT^2}{t_{max} - t_{min}} d\tau \right)$$

$$E[Payoff] = -\frac{1}{2} Cost_D \frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \frac{\left(T^2 (t_{max} - t_{min} - k t_{max}) + \frac{k}{3} {t_{max}}^3 + \frac{2}{3} k T^3 \right)}{t_{max} - t_{min}} \tag{B-6}$$

Appendix C – Mathematical Formulations of Optimization Problems

In this appendix, we present the mathematical formulations to solve the optimization problems described in Section 4.

C.1 Generic Traffic Management Initiative

Each optimization problem is formulated as follows, based on Pareto frontiers and payoff functions defined in equations 16 to 20. Note that for the system optimal, most equitable, and truthful solutions, the constraints are specified for the continuous trade space. When the trade-space is discrete, an additional discreteness constraint is also added specifying the set of feasible solutions to choose from.

System Optimal Solution:

$$max\left(\sum_{k=1}^{K}\left(\sum_{g=1}^{2} a_{g,k}\left(G_g^{SO}\right)^2 + b_{g,k}\left(G_g^{SO}\right) + c_{g,k}\right)\right) \quad \text{(C-1)}$$

Decision variables: G_g^{SO} for $g \in \{1,2\}$

Subject to: $G_g^{SO} \geq 0$ for $g \in \{1,2\}$ (C-2)

If Pareto frontier is arc: $G_1^{SO^2} + G_2^{SO^2} \leq 1$ (C-3)

If Pareto frontier is parabolic: $G_2^{SO} \leq aG_1^{SO^2} + bG_1^{SO} + c$ (C-4)

If Pareto frontier is piecewise-linear: $G_2^{SO} \leq -m_1 G_1^{SO} + 1$ (C-5)

$G_2^{SO} \leq -m_2 G_1^{SO} + m_2$ (C-6)

Most Equitable Solution:

$$max\left(\min_{k=1,\ldots,K}\left(\sum_{g=1}^{2} a_{g,k}\left(G_g^{Eq}\right)^2 + b_{g,k}\left(G_g^{Eq}\right) + c_{g,k}\right)\right) \quad \text{(C-7)}$$

Decision variables: G_g^{Eq} for $g \in \{1,2\}$

Subject to: constraints defined by inequalities C-2 to C-6, replacing G_g^{SO} with G_g^{Eq}.

Truthful Solution

For each airline *k*: $$max\left(\sum_{g=1}^{2} a_{g,k}\left(G_g^{k}\right)^2 + b_{g,k}\left(G_g^{k}\right) + c_{g,k}\right) \quad \text{(C-8)}$$

Decision variables: G_g^{k} for $g \in \{1,2\}$

Subject to: constraints defined by inequalities C-2 to C-6, replacing G_g^{SO} with G_g^{k}.

Strategic Solution

For **Approach 1**: Linear combination

For each airline *k*: $$max\left(\sum_{g=1}^{2} a_{g,k}\left(\sum_{l=1}^{K} w_l G_g^{l}\right)^2 + b_{g,k}\left(\sum_{l=1}^{K} w_l G_g^{l}\right) + c_{g,k}\right) \quad \text{(C-9)}$$

where G_g^{l} for all airlines other than *k* are assumed given.

Decision variables: G_g^{k} for $g \in \{1,2\}$

Subject to: constraints defined by inequalities C-2 to C-6, replacing G_g^{SO} with G_g^k.

This optimization is solved iteratively, updating the values of G_g^l for each airline successively in each iteration, solving a myopic Nash best–response game.

For **Approach 2**: Linear combination pushed to Pareto Frontier

For each airline *k*: $max\left(\sum_{g=1}^{2} a_{g,k}\left(G_g^*\right)^2 + b_{g,k}\left(G_g^*\right) + c_{g,k}\right)$ (C-10)

Decision variables: G_g^k and G_g^* for $g \in \{1,2\}, C$

Subject to: constraints defined by inequalities C-2 to C-6 replacing G_g^{SO} with G_g^k, as well as the following:

$G_g^* = C * \left(\sum_{l=1}^{K} w_l G_g^l\right)$ where G_g^l for all airlines other than *k* is given. (C-11)

If Pareto frontier is arc: $G_1^{*2} + G_2^{*2} = 1$ (C-12)

If Pareto frontier is parabolic: $G_2^* = aG_1^{*2} + bG_1^* + c$ (C-13)

If Pareto frontier is piecewise-linear: $G_2^* \leq -m_1 G_1^* + 1$ (C-14)

$G_2^* \leq -m_2 G_1^* + m_2$ (C-15)

This optimization is solved iteratively, updating the values of G_g^l for each airline successively in each iteration, solving a myopic Nash best–response game.

For **Approach 3**: Weighted Random Choice

For each airline *k*: $max\left(\sum_{l=1}^{K} w_l \left(\sum_{g=1}^{2} a_{g,k} {G_g^l}^2 + b_{g,k} G_g^l + c_{g,k}\right)\right)$ (C-16)

Decision variables: G_g^k

Subject to: constraints defined by inequalities C-2 to C-6 replacing G_g^{SO} with G_g^k.

For **Approach 4**: Ranking

For each airline *k*: $max\left(\sum_{p=1}^{P} y_p^k \cdot \left(\sum_{g=1}^{2} a_{g,k} {G_g^p}^2 + b_{g,k} G_g^p + c_{g,k}\right)\right)$ (C-17)

Decision variables: $y_p^k, R_p^k, \Delta R_{pq}^+, \Delta R_{pq}^-$ for all *p* and *q* options made available by the ANSP

y_p^k represents a binary variable that equals 1 for the winning option p, and 0 otherwise.

Subject to:

$\sum_{p=1}^{P} y_p^k = 1$ (C-18)

$\sum_{p=1}^{P} R_p^k = 1 + 2 + \cdots + P$ (C-19)

$y_p^k \;\; binary$ for all $p \in P$ (C-20)

$R_p^k \;\; integer^+$ for all $p \in P$ (C-21)

$R_p^k \neq R_q^k$ for all $p, q \in P; p \neq q$ (C-22)

$\Delta R_{pq}^+ \geq 0$ for all $p, q \in P; p \neq q$ (C-23)

$\Delta R_{pq}^- \leq 0$ for all $p, q \in P; p \neq q$ (C-24)

$\Delta R_{pq}^+ \geq \sum_{l=1}^{K} w_l R_p^l - \sum_{l=1}^{K} w_l R_q^l$ for all $p, q \in P; p \neq q$ (C-25)

$\Delta R_{pq}^- \leq \sum_{l=1}^{K} w_l R_p^l - \sum_{l=1}^{K} w_l R_q^l$ for all $p, q \in P; p \neq q$ (C-26)

$\Delta R_{pq}^+ \leq M\left(\left(y_p^k - y_q^k\right) + 1\right)$ for all $p, q \in P; p \neq q;$ and M is large (C-27)

$$-\Delta R_{pq}^{-} \le M\left(-\left(y_p^k - y_q^k\right) + 1\right) \quad \text{for all } p, q \in P; p \ne q; \text{ and } M \text{ is large} \tag{C-28}$$

This optimization is solved iteratively, updating the values of R_p^l for each airline successively in each iteration, solving a myopic Nash best–response game.

For **Approach 5**: Voting

For each airline *k*: $$max\left(\sum_{p=1}^{P} y_p^k \cdot \left(\sum_{g=1}^{2} a_{g,k} {G_g^p}^2 + b_{g,k} G_g^p + c_{g,k}\right)\right) \tag{C-29}$$

$V_{p,i}^k$ represents the vote by airline k in round i for candidate p;

y represents a binary variable that equals 1 for the winning option, and 0 otherwise;

Decision variables: $y_p^k, V_{p,i}^k, \Delta V_{pq}^+, \Delta V_{pq}^-$ for all p and q options made available by the ANSP

Subject to:

$$\sum_{p=1}^{P} y_p^k = 1 \tag{C-30}$$

$$\sum_{p=1}^{P} V_{p,i}^k \le 100 \tag{C-31}$$

$$y_p^k \ \ binary \qquad \text{for all } p \in P \tag{C-32}$$

$$V_{p,i}^k \ \ integer^+ \qquad \text{for all } p \in P \tag{C-33}$$

$$\Delta V_{pq}^+ \ge 0 \qquad \text{for all } p, q \in P; p \ne q \tag{C-34}$$

$$\Delta V_{pq}^- \le 0 \qquad \text{for all } p, q \in P; p \ne q \tag{C-35}$$

$$\Delta V_{pq}^+ \ge \sum_{l=1}^{K} w_l V_{p,i}^l - \sum_{l=1}^{K} w_l V_{q,i}^l \qquad \text{for all } p, q \in P; p \ne q \tag{C-36}$$

$$\Delta V_{pq}^- \le \sum_{l=1}^{K} w_l V_{p,i}^l - \sum_{l=1}^{K} w_l V_{q,i}^l \qquad \text{for all } p, q \in P; p \ne q \tag{C-37}$$

$$\Delta V_{pq}^+ \le M\left(\left(y_p^k - y_q^k\right) + 1\right) \quad \text{for all } p, q \in P; p \ne q; \text{ and } M \text{ is large} \tag{C-38}$$

$$-\Delta V_{pq}^- \le M\left(-\left(y_p^k - y_q^k\right) + 1\right) \quad \text{for all } p, q \in P; p \ne q; \text{ and } M \text{ is large} \tag{C-39}$$

$$V_{p,i}^k \ge V_{p,i-1}^k \qquad \text{for all } p \in P \tag{C-40}$$

This optimization is solved iteratively, updating the values of V_p^l for each airline successively in each iteration, solving a myopic Nash best–response game.

C.2 Simplified Ground Delay Program

Each optimization problem is formulated as follows, based on payoff functions defined in equation 25.

System Optimal Solution:

$$max\left(\sum_{k=1}^{K}\left(-\frac{1}{2} Cost_{D,k} \frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \frac{\left({T^{SO}}^2 (t_{max} - t_{min} - k_k t_{max}) + \frac{k_k}{3} {t_{max}}^3 + \frac{2}{3} k_k {T^{SO}}^3\right)}{t_{max} - t_{min}}\right)\right) \tag{C-41}$$

Decision variables: T^{SO}

Subject to:

$$t_{min} \le T^{SO} \le t_{max} \tag{C-42}$$

Most Equitable Solution:

$$max\left(\min_{k=1,\dots,K}\left(-\frac{1}{2}Cost_{D,k}\frac{(C_H-C_L)(\lambda-C_L)}{C_H-\lambda}\cdot\frac{\left(T^{Eq^2}(t_{max}-t_{min}-k_k t_{max})+\frac{k_k}{3}t_{max}{}^3+\frac{2}{3}k_k T^{Eq^3}\right)}{t_{max}-t_{min}}\right)\right) \quad \text{(C-43)}$$

Decision variables: T^{Eq}

Subject to: constraints defined by inequality C-42, but replacing T^{SO} with T^{Eq}.

Truthful Solution

For each airline *k*: $$max\left(-\frac{1}{2}Cost_{D,k}\frac{(C_H-C_L)(\lambda-C_L)}{C_H-\lambda}\cdot\frac{\left(T^{k^2}(t_{max}-t_{min}-k_k t_{max})+\frac{k_k}{3}t_{max}{}^3+\frac{2}{3}k_k T^{k^3}\right)}{t_{max}-t_{min}}\right) \quad \text{(C-44)}$$

Decision variables: T^k

Subject to: constraints defined by inequality C-42, replacing T^{SO} with T^k.

Strategic Solution

For **Approach 1**: Linear combination

For each airline *k*:

$$max\left(-\frac{1}{2}Cost_{D,k}\frac{(C_H-C_L)(\lambda-C_L)}{C_H-\lambda}\cdot\frac{\left(\left(\sum_{l=1}^{K}w_l T^l\right)^2(t_{max}-t_{min}-k_k t_{max})+\frac{k_k}{3}t_{max}{}^3+\frac{2}{3}k_k\left(\sum_{l=1}^{K}w_l T^l\right)^3\right)}{t_{max}-t_{min}}\right) \quad \text{(C-45)}$$

where T^l for all airlines other than *k* are assumed given.

Decision variables: T^k

Subject to: constraints defined by inequality C-42, replacing T^{SO} with T^k.

This optimization is solved iteratively, updating the values of T^l for each airline successively in each iteration, solving a myopic Nash best–response game.

For **Approach 2**: Linear combination pushed to Pareto Frontier

Same formulation as in Approach 1.

For **Approach 3**: Weighted Random Choice

For each airline *k*: $$\max\left(\sum_{l=1}^{K}w_l\left(-\frac{1}{2}Cost_{D,k}\frac{(C_H-C_L)(\lambda-C_L)}{C_H-\lambda}\cdot\frac{\left(T^{l^2}(t_{max}-t_{min}-k_k t_{max})+\frac{k_k}{3}t_{max}{}^3+\frac{2}{3}k_k T^{l^3}\right)}{t_{max}-t_{min}}\right)\right) \quad \text{(C-46)}$$

Decision variables: T^k

Subject to: constraints defined by inequality C-42 replacing T^{SO} with T^k

For **Approach 4**: Ranking

For each airline *k*:

$$max\left(\sum_{p=1}^{P} y_p^k \cdot \left(-\frac{1}{2} Cost_{D,k} \frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \frac{\left(T^{p^2}(t_{max} - t_{min} - k_k t_{max}) + \frac{k_k}{3} t_{max}^{\ 3} + \frac{2}{3} k_k T^{p^3}\right)}{t_{max} - t_{min}}\right)\right) \quad \text{(C-47)}$$

Decision variables: $y_p^k, R_p^k, \Delta R_{pq}^+, \Delta R_{pq}^-$ for all p and q options made available by the ANSP

y_p^k represents a binary variable that equals 1 for the winning option, and 0 otherwise.

Subject to: constraints defined by constraints C-18 to C-28.

This optimization is solved iteratively, updating the values of R_p^l for each airline successively in each iteration, solving a myopic Nash best–response game.

For **Approach 5**: Voting

For each airline *k*, iteration *i*:

$$max\left(\sum_{p=1}^{P} y_p^k \cdot \left(-\frac{1}{2} Cost_{D,k} \frac{(C_H - C_L)(\lambda - C_L)}{C_H - \lambda} \cdot \frac{\left(T^{p^2}(t_{max} - t_{min} - k_k t_{max}) + \frac{k_k}{3} t_{max}^{\ 3} + \frac{2}{3} k_k T^{p^3}\right)}{t_{max} - t_{min}}\right)\right) \quad \text{(C-48)}$$

where $V_{p,i}^k$ represents the vote by airline k in round i for candidate p;

y_p^k represents a binary variable that equals 1 for the winning option, and 0 otherwise.

Decision variables: $y_p^k, V_{p,i}^k, \Delta V_{pq}^+, \Delta V_{pq}^-$ for all p and q options made available by the ANSP

Subject to: constraints defined by constraints C-30 to C-40.

This optimization is solved iteratively, updating the values of V_p^l for each airline successively in each iteration, solving a myopic Nash best–response game.

C.3 Detailed Ground Delay Program

Each optimization problem is formulated as follows, based on Pareto frontiers and payoff functions defined in equations 23 and 26, respectively.

System Optimal Solution:

Objective function defined by expression C-1.

Decision variables: G_g^{SO} for $g \in \{1,2\}$

Subject to: constraint defined by equation C-2, and the following:

$$G_c^{SO} \leq aG_p^{SO^2} + bG_p^{SO} + c \quad \text{(C-49)}$$

Most Equitable Solution:

Objective function defined by expression C-7.

Decision variables: G_g^{Eq} for $g \in \{1,2\}$

Subject to: constraints defined by inequalities C-2 and C-49, replacing G_g^{SO} with G_g^{Eq}.

Truthful Solution

Objective function defined by expression C-8.
Decision variables: G_g^k for $g \in \{1,2\}$
Subject to: constraints defined by inequalities C-2 and C-49, replacing G_g^{SO} with G_g^k.

Strategic Solution

For **Approach 1**: Linear combination

Objective function defined by expression C-9.
decision variables: G_g^k for $g \in \{1,2\}$
subject to: constraints defined by inequalities C-2 and C-49, replacing G_g^{SO} with G_g^k.
This optimization is solved iteratively, updating the values of G_g^i for each airline successively in each iteration, solving a myopic Nash best–response game.

For **Approach 2**: Linear combination pushed to Pareto Frontier

Objective function defined by expression C-10.
Decision variables: G_g^k and G_g^* for $g \in \{1,2\}, C$
Subject to: constraints defined by inequalities C-2 and C-49 replacing G_g^{SO} with G_g^k, equation C-11, and the following:

$$G_c^* = aG_p^{*2} + bG_p^* + c \qquad \text{(C-50)}$$

This optimization is solved iteratively, updating the values of G_g^l for each airline successively in each iteration, solving a myopic Nash best–response game.

For **Approach 3**: Random Choice

Objective function defined by expression C-16.
Decision variables: G_g^k for $g \in \{1,2\}$
Subject to: constraints defined by inequalities C-2 and C-50 replacing G_g^{SO} withG_g^k.

For **Approach 4**: Ranking

Objective function defined by equation C-17.
Decision variables: $y_p^k, R_p^k, \Delta R_{pq}^+, \Delta R_{pq}^-$ for all p and q options made available by the ANSP
y_p^k represents a binary variable that equals 1 for the winning option, and 0 otherwise.
Subject to: constraints defined by equation C-18 to C-28.
This optimization is solved iteratively, updating the values of R_p^l for each airline successively in each iteration, solving a myopic Nash best–response game.

For **Approach 5**: Voting

Objective function defined by equation C-29.

Decision variables: $y_p^k, V_{p,i}^k, \Delta V_{pq}^+, \Delta V_{pq}^-$ for all p and q options made available by the ANSP

Subject to: Constraints defined by equation C-30 to C-40.

This optimization is solved iteratively, updating the values of V_p^l for each airline successively in each iteration, solving a myopic Nash best–response game.

Appendix D – Theoretical Results on Weighted Average Approach

Some conclusions can be drawn about the Weighted Average approach by examining the problem from the theoretical standpoint using game theory. As per Glicksberg's Theorem (Glicksberg, 1952), a mixed Nash equilibrium exists, because the trade space is a non-empty compact metric space, and the airline payoff functions are continuous. Furthermore, as per Debreu, Glicksberg and Fan's Theorem (Debreu, 1952; Glicksberg, 1952; Fan, 1952), a pure Nash equilibrium exists, because the payoff functions are also quasi-concave. For the special case of linear payoff functions, the objective function of each player is a linear function of its own preferred performance goal vector. Therefore, for linear payoff functions, the best response of any player will always be the same, irrespective of other airlines' inputs. We can assume the absence of multiple optimal points, because the parameters of the objective function are real numbers drawn from a continuous distribution. Thus, we can be assured that there exists a unique pure strategy Nash equilibrium for linear payoff functions. For the general convex quadratic payoff functions, as we have assumed here, in order to be assured of uniqueness, applying Rosen's Theorem (Rosen, 1965), the payoffs must be diagonally, strictly concave. This is difficult to prove theoretically. Therefore, we cannot be assured of a unique solution to the game in general.

In the case of linear payoff functions, the unique pure strategy Nash equilibrium will always be truthful, because the best response of each airline is independent of the inputs from other airlines and also independent of the weights of that airline's input. By deduction, if that airline's weight was 1.0 (and everyone else's 0), that airline's best response would not change. For the case where the weight is 1.0, the response will obviously correspond to the actual preference of that airline. Therefore, the pure strategy Nash equilibrium will be truthful. However, this need not be the case with general concave increasing payoff functions – it is dependent on the problem parameters. Finally, with linear payoff functions, any best response dynamic will converge to a truthful pure strategy Nash equilibrium in exactly 1 iteration (per player), because the best response of each player is unchanged by other players' inputs. Again, this is not necessarily the case for the general concave increasing payoff functions.

Appendix E – Derivation of Truthfulness Results

Proposition 1: Under the Weighted Random Choice approach, there is no incentive to game (i.e. deviate from truthful behavior).

Proof: We will prove this by contradiction. Let K denote the set of airlines. Let I_k denote the vector submitted by airline k ($\forall k \in K$) at equilibrium and let w_k be the weight of airline k ($\forall k \in K$). Also, let $P_k(I)$ be the payoff function for airline k ($\forall k \in K$). At the Nash equilibrium, the payoff of a particular airline k' under the Weighted Random Choice approach will equal $\sum_k w_k P_{k'}(I_k) = w_{k'} P_{k'}(I_{k'}) + \sum_{k,k\neq k'} w_k P_{k'}(I_k)$. Let I_k^* denote the truthful optimal solution for an airline k ($\forall k \in K$). Therefore, $P_{k'}(I_{k'}^*) \geq P_{k'}(I_{k'})$.

Let us assume the contrary, that is, there is an incentive to deviate from the truthful behavior at Nash equilibrium. Then, $w_{k'} P_{k'}(I_{k'}^*) + \sum_{k,k\neq k'} w_k P_{k'}(I_k) < w_{k'} P_{k'}(I_{k'}) + \sum_{k,k\neq k'} w_k P_{k'}(I_k)$. Therefore, $P_{k'}(I_{k'}^*) < P_{k'}(I_{k'})$. This leads to a contradiction. Thus we have proved that under the Weighted Random Choice approach, there is no incentive to deviate from truthful behavior.

Proposition 2: Under the Weighted Average approach with an arc-shaped Pareto frontier as a function of performance goal vectors G_1 and G_2, a airline's strategic solution (G_1, G_2) coincides with the truthful solution if and only if the weighted average of all other airlines' strategic solutions lies on a line segment passing through that airline's truthful solution and having a slope $\frac{a_1/a_2}{G_1/G_2}$.

Proof: The truthful solution for an airline can be calculated by solving the following non-linear optimization problem.

maximize $a_1 {G_1}^2 + b_1 G_1 + a_2 {G_2}^2 + b_2 G_2$

s.t. ${G_1}^2 + {G_2}^2 \leq 1$, $G_1 \geq 0$, $G_2 \geq 0$

Applying Karush-Kuhn-Tucker (KKT) conditions, we get,

$$-\begin{bmatrix} 2a_1 G_1 + b_1 \\ 2a_2 G_2 + b_2 \end{bmatrix} + \mu_0 \begin{bmatrix} 2G_1 \\ 2G_2 \end{bmatrix} + \mu_1 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + \mu_2 \begin{bmatrix} 0 \\ -1 \end{bmatrix} = 0,\ {G_1}^2 + {G_2}^2 \leq 1, G_1 \geq 0,\ G_2 \geq 0,\ \mu_0 \geq 0,\ \mu_1 \geq 0,$$
$$\mu_2 \geq 0, \mu_0\left({G_1}^2 + {G_2}^2 - 1\right) = 0, \mu_1 G_1 = 0, \mu_2 G_2 = 0. \quad \text{(E-1)}$$

We know that the truthful optimal solution lies on the Pareto frontier. Therefore, at the optimal solution, ${G_1}^2 + {G_2}^2 = 1$. We consider the following 3 possible cases. Case I: ($G_1 < 1$ and $G_2 < 1$); Case II: ($G_1 = 1$ and $G_2 < 1$); and Case III: ($G_2 = 1$ and $G_1 < 1$). Let's consider the cases one-by-one.

Case I: $G_1 < 1$ and $G_2 < 1$.

$\Rightarrow G_1 > 0, G_2 > 0, \mu_1 = \mu_2 = 0.$

Solving conditions (E-1) we get, $2(a_1 - a_2) = \frac{b_2}{G_2} - \frac{b_1}{G_1}$ (E-2)

This case is true if and only if: $\mu_0 \geq 0, 1 > G_1 > 0, 1 > G_2 > 0$.

Case II: $G_1 = 1$ and $G_2 < 1$

$\Rightarrow G_1 = 1, G_2 = 0, \mu_0 = a_1 + \frac{b_1}{2}, \mu_1 = 0, \mu_2 = -b_2$ (E-3)

This case is true if and only if $a_1 + \frac{b_1}{2} \geq 0, b_2 \leq 0$. But $b_2 > 0$. So this case is impossible.

Case III: $G_2 = 1$ and $G_1 < 1$. By symmetry, we can see that this case is also impossible.

Thus, at the truthful optimal solution, $2(a_1 - a_2) = \frac{b_2}{G_2} - \frac{b_1}{G_1}$.

Now consider the strategic solution. Let the weighted average of everyone else's inputs be: (G_1^{avg}, G_2^{avg}). Also, let the ratio of the total weight of all other airlines to the weight of a specific airline under consideration be $w:1$. The strategic solution can be calculated by solving the following non-linear optimization problem.

maximize $a_1(\frac{G_1 + wG_1^{avg}}{w+1})^2 + b_1(\frac{G_1 + wG_1^{avg}}{w+1}) + a_2\left(\frac{G_2 + wG_2^{avg}}{w+1}\right)^2 + b_2\left(\frac{G_2 + wG_2^{avg}}{w+1}\right) = \frac{a_1}{(w+1)^2}{G_1}^2 + \left(\frac{2a_1 wG_1^{avg}}{(w+1)^2} + \frac{b_1}{w+1}\right)G_1 + \frac{a_2}{(w+1)^2}{G_2}^2 + \left(\frac{2a_2 wG_2^{avg}}{(w+1)^2} + \frac{b_2}{w+1}\right)G_2 +$ constant terms

s.t. ${G_1}^2 + {G_2}^2 \leq 1, G_1 \geq 0, G_2 \geq 0$.

Reusing the previous results (E-2 and E-3), by analogy, we get

Case I: $2\left(\frac{a_1}{(w+1)^2} - \frac{a_2}{(w+1)^2}\right) = \frac{\left(\frac{2a_2 wG_2^{avg}}{(w+1)^2} + \frac{b_2}{w+1}\right)}{G_2} - \frac{\left(\frac{2a_1 wG_1^{avg}}{(w+1)^2} + \frac{b_1}{w+1}\right)}{G_1}$.

This case is true if and only if $\mu_0 \geq 0, 1 > G_1 > 0, 1 > G_2 > 0$.

Case II: $G_1 = 1, G_2 = 0, \mu_0 = \frac{a_1}{(w+1)^2} + \frac{\left(\frac{2a_1 wG_1^{avg}}{(w+1)^2} + \frac{b_1}{w+1}\right)}{2}, \mu_1 = 0, \mu_2 = -\left(\frac{2a_2 wG_2^{avg}}{(w+1)^2} + \frac{b_2}{w+1}\right)$.

This case is true if and only if $\frac{a_1}{(w+1)^2} + \frac{\left(\frac{2a_1 wG_1^{avg}}{(w+1)^2} + \frac{b_1}{w+1}\right)}{2} \geq 0, \left(\frac{2a_2 wG_2^{avg}}{(w+1)^2} + \frac{b_2}{w+1}\right) \leq 0$.

Case III: $G_1 = 0, G_2 = 1, \mu_0 = \frac{a_2}{(w+1)^2} + \frac{\left(\frac{2a_2 wG_2^{avg}}{(w+1)^2} + \frac{b_2}{w+1}\right)}{2}, \ \mu_1 = -\left(\frac{2a_1 wG_1^{avg}}{(w+1)^2} + \frac{b_1}{w+1}\right), \mu_2 = 0$.

This case is true if and only if $\frac{a_2}{(w+1)^2} + \frac{\left(\frac{2a_2 w G_2^{avg}}{(w+1)^2} + \frac{b_2}{w+1}\right)}{2} \geq 0, \left(\frac{2a_1 w G_1^{avg}}{(w+1)^2} + \frac{b_1}{w+1}\right) \leq 0.$

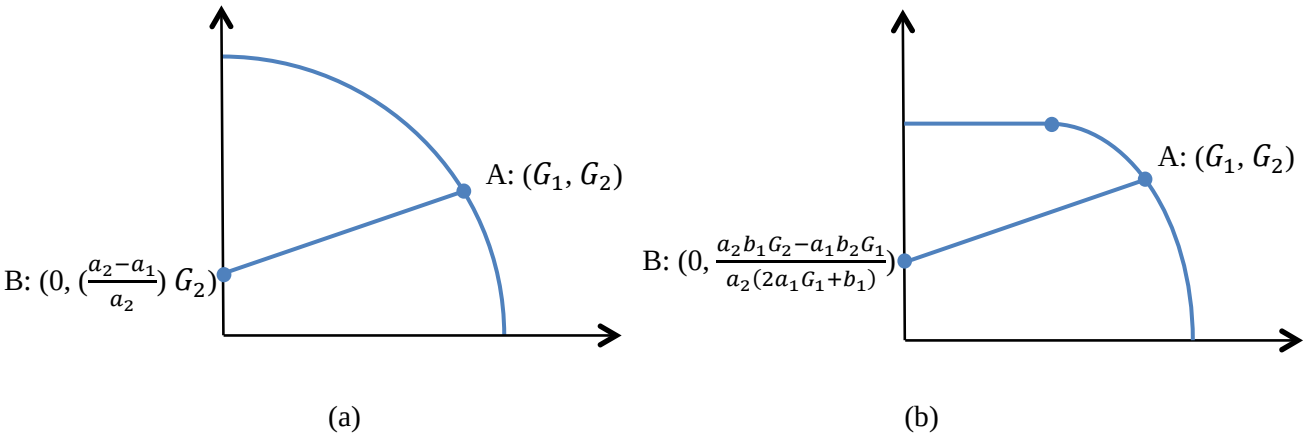
The only way the truthful and the strategic solutions can be identical is if both belong to case I.

$\Rightarrow 2(a_1 - a_2) = \frac{b_2}{G_2} - \frac{b_1}{G_1}, 2\left(\frac{a_1}{(w+1)^2} - \frac{a_2}{(w+1)^2}\right) = \frac{\left(\frac{2a_2 w G_2^{avg}}{(w+1)^2} + \frac{b_2}{w+1}\right)}{G_2} - \frac{\left(\frac{2a_1 w G_1^{avg}}{(w+1)^2} + \frac{b_1}{w+1}\right)}{G_1}.$

$\Rightarrow a_1\left(\frac{G_1^{avg} - G_1}{G_1}\right) = a_2\left(\frac{G_2^{avg} - G_2}{G_2}\right)$

So truthfulness requires that the weighted average of other players' chosen vectors at the equilibrium (G_1^{avg}, G_2^{avg}) should lie on the line segment A-B passing through that airline's truthful solution and having a slope $\frac{a_1/a_2}{G_1/G_2}$ as shown in Figure E-1(a). QED.

Note that the requirement that the weighted average of other airlines' strategic solutions should lie on a line segment, is extremely restrictive. So this condition is rarely satisfied exactly.



(a) (b)

Figure E-1. Necessary conditions for truthfulness: Line segment AB should contain the weighted average of other players' submitted vectors, for the truthful solution to be the same as the strategic solution when the Pareto frontier is a) arc-shaped, b) parabolic.

Proposition 3: Under the Weighted Average approach with a parabolic Pareto frontier as a function of performance goal vectors G_1 and G_2, an airline's strategic solution (G_1, G_2) coincides with the truthful solution if and only if one of the following two conditions is true: 1) The weighted average of all other airlines' strategic solutions lies on a line segment passing through that airline's truthful solution and having a slope $\frac{2a_2 G_2 + b_2}{2a_1 G_1 + b_1} * \frac{a_1}{a_2}$ and $\left(2a_1 + 2a_1 w G_1^{avg} + b_1(w+1)\right)(1 - G_{1TP}) < 2\left(2a_2 w G_2^{avg} + b_2(w+1)\right)$ holds or 2) the solution lies at (1,0) and $(2a_1 + b_1)(1 - G_{1TP}) \geq 2b_2$ holds.

Proof: The truthful solution for an airline can be calculated by solving the following non-linear optimization problem.

maximize $a_1{G_1}^2 + b_1G_1 + a_2{G_2}^2 + b_2G_2$

s.t. $G_2 \leq -\frac{1}{(G_{1TP}-1)^2}{G_1}^2 + \frac{2G_{1TP}}{(G_{1TP}-1)^2}G_1 + \frac{1}{(G_{1TP}-1)^2} - \frac{2G_{1TP}}{(G_{1TP}-1)^2}$ for $G_1 \geq G_{1TP}$, $G_2 \leq 1, G_1 \geq 0, G_2 \geq 0$.

Note that, because of $a_1, a_2 < 0, b_1, b_2 > 0, b_1 > -2a_1, b_2 > -2a_2$, an optimal value has to lie on $G_2 \leq -\frac{1}{(G_{1TP}-1)^2}{G_1}^2 + \frac{2G_{1TP}}{(G_{1TP}-1)^2}G_1 + \frac{1}{(G_{1TP}-1)^2} - \frac{2G_{1TP}}{(G_{1TP}-1)^2}$. So this problem is equivalent to,

maximize $a_1{G_1}^2 + b_1G_1 + a_2{G_2}^2 + b_2G_2$

s.t. $G_2 \leq -\frac{1}{(G_{1TP}-1)^2}{G_1}^2 + \frac{2G_{1TP}}{(G_{1TP}-1)^2}G_1 + \frac{1}{(G_{1TP}-1)^2} - \frac{2G_{1TP}}{(G_{1TP}-1)^2}, G_1 \geq 0, G_2 \geq 0$.

Applying Karush-Kuhn-Tucker (KKT) conditions, we get,

$$-\begin{bmatrix}2a_1G_1 + b_1\\ 2a_2G_2 + b_2\end{bmatrix} + \mu_0\begin{bmatrix}2G_1 - 2G_{1TP}\\ (G_{1TP}-1)^2\end{bmatrix} + \mu_1\begin{bmatrix}-1\\ 0\end{bmatrix} + \mu_2\begin{bmatrix}0\\ -1\end{bmatrix} = 0, G_2 \leq -\frac{1}{(G_{1TP}-1)^2}{G_1}^2 + \frac{2G_{1TP}}{(G_{1TP}-1)^2}G_1 + \frac{1}{(G_{1TP}-1)^2} - \frac{2G_{1TP}}{(G_{1TP}-1)^2}, G_1 \geq 0, G_2 \geq 0, \mu_0 \geq 0, \mu_1 \geq 0, \mu_2 \geq 0, \mu_0\left(G_2 + \frac{1}{(G_{1TP}-1)^2}{G_1}^2 - \frac{2G_{1TP}}{(G_{1TP}-1)^2}G_1 - \frac{1}{(G_{1TP}-1)^2} + \frac{2G_{1TP}}{(G_{1TP}-1)^2}\right) = 0, \mu_1G_1 = 0, \mu_2G_2 = 0. \quad \text{(E-4)}$$

We know that the truthful optimal solution lies on the Pareto frontier. Therefore, at optimality, $G_2 = -\frac{1}{(G_{1TP}-1)^2}{G_1}^2 + \frac{2G_{1TP}}{(G_{1TP}-1)^2}G_1 + \frac{1}{(G_{1TP}-1)^2} - \frac{2G_{1TP}}{(G_{1TP}-1)^2}$. We consider the following 2 possible cases. Case I: $(G_1 = 1)$; and Case II: $(G_1 < 1)$. Let's consider the cases one-by-one.

Case I: $G_1 = 1, G_2 = 0,\ \mu_1 = 0$

Solving conditions (E-4) we get, $\mu_0 = \frac{2a_1+b_1}{2(1-G_{1TP})}, \mu_2 = -b_2 + \left(a_1 + \frac{b_1}{2}\right)(1 - G_{1TP})$ (E-5)

This case is true if and only if: $(2a_1 + b_1)(1 - G_{1TP}) \geq 2b_2$. (E-6)

Case II: $G_1 < 1 \Rightarrow G_1 > 0, G_2 > 0$ (as long as $G_{1TP} < 1$), $\mu_1 = \mu_2 = 0$.

Solving conditions (E-4) we get, $\frac{2a_1G_1+b_1}{2a_2G_2+b_2} = \frac{2(G_1-G_{1TP})}{(G_{1TP}-1)^2}$. (E-7)

This case is true if and only if: $(2a_1 + b_1)(1 - G_{1TP}) < 2b_2$ (E-8)

Now consider the strategic solution. Let the weighted average of everyone else's inputs be: (G_1^{avg}, G_2^{avg}). Also, let the ratio of the total weight of all other airlines to the weight of a specific airline under

consideration be w: 1. The strategic solution can be calculated by solving the following non-linear optimization problem.

maximize $a_1(\frac{G_1+wG_1^{avg}}{w+1})^2 + b_1(\frac{G_1+wG_1^{avg}}{w+1}) + a_2\left(\frac{G_2+wG_2^{avg}}{w+1}\right)^2 + b_2\left(\frac{G_2+wG_2^{avg}}{w+1}\right) = \frac{a_1}{(w+1)^2}{G_1}^2 + \left(\frac{2a_1wG_1^{avg}}{(w+1)^2} + \frac{b_1}{w+1}\right)G_1 + \frac{a_2}{(w+1)^2}{G_2}^2 + \left(\frac{2a_2wG_2^{avg}}{(w+1)^2} + \frac{b_2}{w+1}\right)G_2 +$ constant terms

s.t. $G_2 \leq -\frac{1}{(G_{1TP}-1)^2}{G_1}^2 + \frac{2G_{1TP}}{(G_{1TP}-1)^2}G_1 + \frac{1}{(G_{1TP}-1)^2} - \frac{2G_{1TP}}{(G_{1TP}-1)^2}$, $G_1 \geq 0$, $G_2 \geq 0$.

Reusing the previous results (E-5 and E-7), by analogy, we get,

Case I: $G_1 = 1, G_2 = 0,\ \mu_1 = 0$

$$G_1 = 1, G_2 = 0, \mu_0 = \frac{2a_1+2a_1wG_1^{avg}+b_1(w+1)}{2(w+1)^2(1-G_{1TP})}, \mu_2 = \frac{-\left(2a_2wG_2^{avg}+b_2(w+1)\right)+\left(a_1+a_1wG_1^{avg}+\frac{(w+1)b_1}{2}\right)(1-G_{1TP})}{(w+1)^2} \quad \text{(E-9)}$$

This case is true if and only if:

$$\left(2a_1 + 2a_1wG_1^{avg} + b_1(w+1)\right)(1-G_{1TP}) \geq 2\left(2a_2wG_2^{avg} + b_2(w+1)\right). \quad \text{(E-10)}$$

Case II: $G_1 < 1 \Rightarrow G_1 > 0, G_2 > 0, \mu_1 = \mu_2 = 0.$

$$\frac{2a_1G_1+2a_1wG_1^{avg}+b_1(w+1)}{2a_2G_2+2a_2wG_2^{avg}+b_2(w+1)} = \frac{2(G_1-G_{1TP})}{(G_{1TP}-1)^2} \quad \text{(E-11)}$$

This case is true if and only if:

$$\left(2a_1 + 2a_1wG_1^{avg} + b_1(w+1)\right)(1-G_{1TP}) < 2\left(2a_2wG_2^{avg} + b_2(w+1)\right) \quad \text{(E-12)}$$

The only way the truthful and strategic solutions can be identical is if both belong to Case I or if both belong to Case II.

a. If both belong to Case I: i.e. $G_1 = 1, G_2 = 0$, then truthfulness requires that condition (E-6) is satisfied. (Note that condition (E-10) is satisfied automatically if condition (E-6) is satisfied.)

b. If both belong to Case II: Then truthfulness requires that conditions (E-7), (E-11) and (E-12) are satisfied. Note that condition (E-8) is automatically satisfied of condition (E-12) is satisfied. Thus, truthfulness requires the following two conditions:

   ⇨ $\frac{2a_1G_1+b_1}{2a_2G_2+b_2} = \frac{2a_1G_1^{avg}+b_1}{2a_2G_2^{avg}+b_2}$ and $\left(2a_1 + 2a_1wG_1^{avg} + b_1(w+1)\right)(1-G_{1TP}) < 2\left(2a_2wG_2^{avg} + b_2(w+1)\right)$ (E-13)

So truthfulness requires that either the weighted average of other players' chosen vectors at the equilibrium should be lying on the line segment A-B shown in Figure E-1(b) and $\left(2a_1 + 2a_1 w G_1^{avg} + b_1(w+1)\right)(1 - G_{1TP}) < 2\left(2a_2 w G_2^{avg} + b_2(w+1)\right)$ holds, or $G_1 = 1, G_2 = 0$ and $(2a_1 + b_1)(1 - G_{1TP}) \geq 2b_2$ holds. QED.

Note that the requirement that the weighted average of other airlines' strategic solutions should lie on a line segment is extremely restrictive. So condition 1 in proposition 3 is rarely satisfied exactly. Similarly, the condition $(2a_1 + b_1)(1 - G_{1TP}) \geq 2b_2$ is also not very likely to hold in general because, for general concave increasing payoff functions, we only need $(2a_1 + b_1) \geq 0$ to hold. Note that because $0 \leq (1 - G_{1TP}) \leq 1$ and $b_2 \geq 0$, condition 2 in proposition 3 is also quite restrictive.

APPENDIX VI

Table 1: Survey Questions and Responses, Delta Airlines, Managers

Questions	**Responses**
Part 1: Regarding the current TMI planning process	
What are the most common situations in which your airline is significantly impacted by TMIs?	Low ceilings at LGA/JFK/ATL Airspace Flow Programs Problems: • Planned and implementation • Inconsistent collaboration • Discussion often seems like a formality
Who at your airline participates in the strategic telecons	ATC sector managers • 65 employed • 7 on duty at any given time • 1-2 participating in the telecons
How much time and effort are required from the participants in the telecon?	30 minutes of preparation time • Review weather forecasts • Reviews demand/FSM • Determine appropriate rates • Allow for realistic debate 15 minutes of the telecon • Some multi-tasking • Usually 90% engaged
In general, what do you think about the current TMI decision making process?	Inconsistent collaboration Not a lot of science • Departure rates • AFP rates Lack of transparency in decision making process Revisions particularly problematic
Briefing presentation on COuNSEL	
Part 2: Regarding the proposed TMI planning concept: COuNSEL	
What do you think of the COuNSEL concept? • How would it affect the quality of the traffic management decisions made by FAA? • How would it affect the traffic management decision process?	Good concept, long overdue Current process has devolved to one based mainly on big-dog concept with interference from non-big dogs Will allow FAA to know what operators need Will be mutually beneficial More collaborative More based on data
Who at your airline would provide the COuNSEL input?	Sector managers The 20 or so that responsible for interacting with ATC
How do you think COuNSEL would affect the time and effort required for you airline to participate in TMI decision making?	Should not require more time and effort; will probably shorten the process
What do you think about the performance goals?	There are always tradeoffs between performance goals The performance goals identified are appropriate Predictability is an important goal

How much iteration is acceptable?	Confusing question Could not really answer
What concerns do you have for COuNSEL?	Need for training and communication Need for participants to see value early Cannot just “throw it over the transom” Cannot be perceived by other operators as special deal for Delta (in the context of Delta participating more actively in the development)
What do you think will be the benefits of COuNSEL compared to current practice?	Benefits already discussed
What do you think will be the cost or negative impacts of implementing COuNSEL?	Not much cost if folded into current training process

Table 2: Survey Questions and Feedback, Delta Airlines, Potential COuNSEL Users

Questions	**Responses**	**Comments**
Part 1: Regarding the current TMI planning process		
On the whole, Traffic Management Initiative decisions made by the FAA are usually: Poor(1)—Good(10)	6 7	
When FAA plans Traffic Management Initiatives, the logic behind the decisions is usually: Obscure(1)—Transparent(10)	5 6	Not sure how to distinguish obscure from transparent. Better options: Unclear(1)—Clear(10)
Faced with the same situation, the decisions made by different traffic managers are likely to be: Very different(1)—Very similar(10)	5 2	
How effective are the planning telecons in obtaining flight operator input in Traffic Management Initiative decisions? Not at all effective(1)—Very effective(10)	8 6	
How responsive are the FAA Traffic Management Initiative decisions to the differing needs of individual flight operators? Not responsive at all(1)—Very responsive(10)	6 6	
More vocal participants have greater influence on TMI decisions than less vocal participants? Strongly disagree—Strongly agree	Agree Agree	
TMI telecons require too much time and effort? Strongly disagree—Strongly agree	Disagree Strongly disagree	
The planned durations in the initial GDPs are usually: Too short(1)—Too long(10)	5	
The planned airport acceptance rates are usually: Too low(1)—Too high(10)	4	
The GDPs scope is usually: Too small(1)—Too large(10)	5	GDP scopes could be too small for some airports but too large for other airports. Better options: Unreasonable(1)—Reasonable(10)
Relative to program start time, the report time of GDPs is usually: Too far in advance(1)—Too soon before(10)	7	
The influence of the airlines with large numbers of affected operations on GDP decisions is: Too small(1)—Too large(10)	7	

Program revisions are made: Too infrequently(1)—Too often(10)	7 	
Importance of the following variables in assessing GDP performance: Not at all important(1)—Extremely important(7)		
• GDP lead time	6 6	Additional Factors: • GDP scope • Experiences of Staff on duty • Impact of other TMIs on GDP • GDP duration • Scope v.s. duration • Accuracy of forecast of GDP end time
• Average flight delay of non-exempted flights	3 3	
• Percentage of total delay that is taken in the air	3 3	
• Unrecoverable delay	6 6	
• Number of extensions	6 6	
• Maximum fight delay	6 6	
• Accuracy of initial delay estimates	6 6	
Briefing presentation on COuNSEL		
Part 2: Regarding the proposed TMI planning concept: COuNSEL		
How important are each of the following features of COuNSEL Not at all important(1)—Extremely important(7)		
• Flight operators provide structured input to the TMI planning process through a web portal	6 6	
• Input of all affected flight operators is considered	7 6	
• Flight operator votes are weighted based on the number of flights affected	6 6	
• TMI decisions are tied explicitly to performance goals	5 5	Performance goals may be easily taken as airlines' performance goals. Suggestion: replace performance goals with performance vectors.
• Less time and effort are required for telecons	5 4	
• Less opportunity for personal interaction in planning process	3 1	
• Effort required to learn a new system	4 1	
Please indicate your level of agreement of disagreement with each of the following statements about COuNSEL Strongly disagree(1)—Strongly agree(7)		

• Difficult to decide how to rate different performance vectors	4 6	
• Process is not transparent enough	3 4	
• The set of performance goals—efficiency, predictability, and throughput—is appropriate	6 5	

COuNSEL User Survey

Dear Participant:

Thank you for agreeing to take part in this survey.

There are two parts. In this part, we ask for your viewpoints concerning traffic management initiative (TMI) planning and decision making. In answering your questions, please focus on your experiences during the recently concluded convective weather season.

In the second part of the survey, we will ask for your opinions on a new concept for TMI planning, known as COuNSEL. You will complete this survey after we have briefed you on COuNSEL.

Sincerely,

COuNSEL Research Team
The National Center of Excellence for Aviation Operations Research

Section 1: Current practice

In this section, we would like to obtain your view on current Traffic Management Initiative (TMI) planning practice.

1. On the whole, Traffic Management Initiative decisions made by the FAA are usually:

	1	2	3	4	5	6	7	8	9	10	
Poor	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	Good

2. When FAA plans Traffic Management Initiatives, the logic behind the decisions is usually:

	1	2	3	4	5	6	7	8	9	10	
Unclear	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	Clear

3. Faced with the same situation, the decisions made by different traffic managers are likely to be:

	1	2	3	4	5	6	7	8	9	10	
Very different	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	Very similar

4. How effective are the planning telecons in obtaining flight operator input in Traffic Management Initiative decisions?

	1	2	3	4	5	6	7	8	9	10	
Not at all effective	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	Very effective

5. How responsive are the FAA Traffic Management Initiative decisions to the differing needs of individual flight operators?

	1	2	3	4	5	6	7	8	9	10	
Not responsive at all	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	Very responsive

How much do you agree with the following statements:

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
6. More vocal participants have greater influence on TMI decisions than less vocal participants	❍	❍	❍	❍	❍
7. TMI telecons require too much time and effort	❍	❍	❍	❍	❍

Section 2: Ground Delay Program (GDP)

GDP is one of the most common TMIs. In this section, we would like you to evaluate the performance of current GDP plans.

8. In your opinion, the planned durations in the initial GDPs are usually:

	1	2	3	4	5	6	7	8	9	10	
Too short	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	Too long

9. The planned airport acceptance rates are usually:

	1	2	3	4	5	6	7	8	9	10	
Too low	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	Too high

10. The GDPs scope is usually:

	1	2	3	4	5	6	7	8	9	10	
Unreasonable	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	Reasonable

11. Relative to program start time, the report time of GDPs is usually:

	1	2	3	4	5	6	7	8	9	10	
Too far in advance	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	Too soon before

12. The influence of the airlines with large numbers of affected operations on GDP decisions is:

	1	2	3	4	5	6	7	8	9	10	
Too small	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	Too large

13. Program revisions are made:

	1	2	3	4	5	6	7	8	9	10	
Too infrequently	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	Too Often

Section 3: GDP performance

In this section, we would like to understand what criteria flight operators use to assess GDP performance outcomes. Here, we list several variables that could be measured at the end of a GDP. Please rate the importance of these variables in assessing GDP performance.

14. Please rate the importance of each variable from 1 to 7:

	Not at all important	Very unimportant	Somewhat unimportant	Neither important nor unimportant	Somewhat important	Very important	Extremely important
GDP lead time (lead time= GDP start time - GDP report time)	❍	❍	❍	❍	❍	❍	❍
Average flight delay of non-exempted flights	❍	❍	❍	❍	❍	❍	❍
Percentage of total delay that is taken in the air	❍	❍	❍	❍	❍	❍	❍
Unrecoverable delay	❍	❍	❍	❍	❍	❍	❍
Number of extensions	❍	❍	❍	❍	❍	❍	❍
Maximum fight delay	❍	❍	❍	❍	❍	❍	❍
Accuracy of forecast on GDP end time	❍	❍	❍	❍	❍	❍	❍
Length of Program	❍	❍	❍	❍	❍	❍	❍
Scope of the Program	❍	❍	❍	❍	❍	❍	❍
Impact on GDP by other TMIs	❍	❍	❍	❍	❍	❍	❍
Accuracy of initial delay estimates	❍	❍	❍	❍	❍	❍	❍

15. If there are other variables that you think should be added to the list in the question above, please list them here.

	Not at all important	Very unimportant	Somewhat unimportant	Neither unimportant nor important	Somewhat important	Very important	Extremely important
Additional factor 1: _____	❍	❍	❍	❍	❍	❍	❍
Additional factor 2: _____	❍	❍	❍	❍	❍	❍	❍
Additional factor 3: _____	❍	❍	❍	❍	❍	❍	❍

COuNSEL User Survey Part 2

How important are each of the following features of COuNSEL?	Not at all important	Very unimportant	Somewhat unimportant	Neither important nor unimportant	Somewhat important	Very important	Extremely important
1. Flight operators provide structured input to the TMI planning process through a web portal	❍	❍	❍	❍	❍	❍	❍
2. Input of all affected flight operators is considered	❍	❍	❍	❍	❍	❍	❍
3. Flight operator votes are weighted based on the number of flights affected	❍	❍	❍	❍	❍	❍	❍
4. TMI decisions are tied explicitly to performance goals	❍	❍	❍	❍	❍	❍	❍
5. Less time and effort are required for telecons	❍	❍	❍	❍	❍	❍	❍
6. Less opportunity for personal interaction in planning process	❍	❍	❍	❍	❍	❍	❍
7. Effort required to learn a new system	❍	❍	❍	❍	❍	❍	❍

Please indicate your level of agreement of disagreement with each of the following statements about COuNSEL	Strongly disagree	Disagree	Somewhat disagree	Neither disagree nor agree	Somewhat agree	Agree	Strongly agree
8. Difficult to decide how to rate different performance vectors	❍	❍	❍	❍	❍	❍	❍
9. Process is not transparent enough	❍	❍	❍	❍	❍	❍	❍
10. The set of performance goals—efficiency, predictability, and throughput—is appropriate	❍	❍	❍	❍	❍	❍	❍