The Static Stochastic Ground Holding Problem with Aggregate Demands

Michael O. Ball, Robert Hoffman
Robert H Smith School of Business and Institute for Systems Research
University of Maryland
College Park, Maryland

Amedeo Odoni, Ryan Rifkin
Operations Research Center
MIT
Cambridge, MA

January 4, 1999

ABSTRACT

The ground delay program is a mechanism used to decrease the rate of incoming flights into an airport, when it is projected that the rate of flow into the airport will exceed capacity. In this paper we present an integer programming model for planning ground delay programs. The model considers a stochastic capacity profile, which is represented by a set of airport capacity scenarios and their probabilities. The demand on the airport as well as the model's output is represented at an aggregate level in terms of numbers of flights per unit time. This allows for the model to be used in conjunction with arbitrarily complex processes for allocating individual flights to slots. It was specifically designed to be used in the Collaborative Decision Making setting where individual flight assignments result from an iterative process involving both the airlines and traffic flow managers. We show that the linear programming dual of the model can be transformed into a network flow problem. This in turn implies that the integer program can be solved efficiently using linear programming or network flow methods.

1 Introduction

It is well-known that demand for runway operations (landings and take-offs) often exceeds available capacity for periods of several hours at major airports throughout the world. This is primarily true under unfavorable weather conditions but it happens occasionally under good weather conditions as well, especially in the United States, where some airports are "overscheduled" at high demand periods.

When the expected demand for landings is predicted to exceed the arrival capacity of a major airport in the United States for a significant period of time, the Federal Aviation Administration (FAA) uses various air traffic flow management (ATFM) measures to smooth out the arrival flow and bring arrival demand in line with capacity. Ground holding is the most important of these methods. The idea is simple: it is preferable to have a flight wait on the ground at its point of origin rather than to have it circle the airport at its destination, unable to land. Therefore, if it is known with certainty, or at least with high probability, that a flight will be unable to land due to lack of capacity, it may be advantageous to hold the flight on the ground at its point of origin. Ground holding saves fuel costs and preserves safety margins by relieving airborne congestion.

The FAA adpoted ground holding as a commonly-employed strategy in the early 1980's. For each possibly capacitated airport, the FAA generates an estimate, or forecast, of capacity for the day. Treating this forecast as deterministic, the FAA assigns ground delays to incoming aircraft so that the arrival flow will match the forecasted capacity. If the forecast is accurate, this ensures that air holds (planes forced to wait in the air at their destination due to lack of landing capacity) will be kept to a minimum. This policy will be referred to as the *deterministic* ground holding policy.

An airport acceptance rate (AAR) is the number of flights that can be landed at a given airport in a given unit of time (strictly speaking, this is a capacity not a rate, nonetheless, it is the established terminology in ATFM). The short-term forecast of capacity on which ground holding policies are based is known as the forecast of airport acceptance rates (AAR) and is usually given for each hour over several hours. For instance, it might be predicted that over a six-hour period, the AAR of an airport will be 36 flights for the first hour, then 30 fights for each hour thereafter, perhaps reflecting worsening weather conditions.

Forecasted weather conditions for an airport are converted to runway configurations and subsequent AARs with reasonable accuracy. The main problem with the deterministic ground holding policy is that it ignores the highly stochastic nature of the weather conditions that ultimately determine the AARs. For instance, if the forecasted AARs turn out to be lower than the AARs that actually materialize, then (in retrospect) too much ground holding has been applied and valuable airport capacity goes unutilized. Similarly, if the forecasted AARs are higher than the actual AARs, then demand will exceed capacity and there will be airborne holding that (again, in retrospect) could have been replaced with ground holding.

A large body of work exists on deterministic versions of the ground-holding problem. See e.g., Terrab [?], Vranas [?], Vranas [?], Bertsimas and Vranas [?], Vranas, Bertsimas and Odoni [?], and Bertsimas and Stock [?] (this list is by no means exhaustive).

In this article, we look at a *static stochastic* model for the single-airport ground holding problem: stochastic, in that it explicitly takes into account the stochastic nature of future capacity, and static, in that it requires all decisions over a given time horizon to be made in advance. In [?], Richetta and Odoni introduce and analyze the static stochastic ground holding problem. An integer program is developed, which represents uncertainty by assuming the existence of a probabilistic distribution of "scenarios", or possible realizations of capacity. By treating arrivals as flows rather than as individual flights and by making use of the fact that the linear programming relaxation always yields an integer solution, the model is solvable for reasonable problem sizes.

The present paper, based on Hoffman [?] and Rifkin [?], makes three main contributions. The first is to introduce a new model for the static stochastic ground holding problem. By taking advantage of pre- and post-processing our model can produce the same solutions as the Richetta and Odoni model under currently accepted practices, while reducing the number of decision variables by an order of magnitude. The second contribution is a proof that the integer program associated with the model is dual network, which implies that the model can be solved to optimality in polynomial time via network methods and that the LP relaxation yields an integer solution.

The third contribution of this paper is that the adopted perspective is entirely consistent with the recently-initiated (January 1998) Collaborative Decision-Making (CDM) approach to assigning ground holds to individual flights (see [?]). Prior work on ground delay problems focussed on assigning delays to individual flights in order to optimize a system-wide objective function. The CDM perspective is that the air traffic system consists of a large number of individual users (airlines) with diverse, often conflicting, objectives and it is inappropriate to apply a common objective function across all airlines.

Under CDM, the FAA and the airlines have jointly adopted the view that the FAA is responsible for forecasting airport capacity and, in the event that a GDP is warranted, partitioning arrival resources into "arrival slots" to be distributed amongst inbound flights in an equitable manner. For instance, if the AAR for the first hour of a GDP is 30 fights, then an arrival slot is created for every two minutes of that hour. In essence, fairness, as defined by CDM, is achieved by awarding the earliest slots to the flights with the earliest scheduled arrival times (as scheduled by the Official Airline Guide, created weeks earlier). Suppose that the capacity of an airport whose normal capacity is 50 flights per hour is forecasted to drop to 36 flights for the first hour and 30 flights for the next two hours. Then the first 36 scheduled flights would be assigned to the 36 slots in the first hour, the next 30 scheduled flights would be assigned to the 30 slots in the second hour, and so on. Each airline is then given an opportunity to redistribute its flights among the slots it has been awarded, subject to certain flight eligibilities. In practice, there are many practical considerations such as flights that cannot make their appointed slot time and flights exempted from FAA-assigned ground delay. See Hoffman, Ball, Hall, Odoni, Wambsganss [Rif: add ref] for

details.

The model presented in this paper outputs the *number* of arrival slots that should be made available. Thus, it can be viewed as a Phase I process, where in Phase II, the CDM procedures assign individual flights to slots.

In Section **??**, we formulate the problem more precisely, state our assumptions, and present our model. In Section **??**, we prove that the integer program defined by the model is dual network. In Section **??**, we present an example of the model's behavior. Finally, in Section **??**, we draw conclusions and point out future research directions.

2 Model Formulation

The static stochastic ground holding problem assumes that the only element of uncertainty is the arrival capacity at the airport in question; demand and travel times are deterministic and known in advance. We assume that the time interval of interest consists of $T \in \mathcal{Z}_+$ time periods. For each time period, there is a demand D_t, which is the number of flights predicted to arrive at the airport in time interval t, if there were no capacity restrictions. We also assume that probabilistic information about the uncertain capacity is available in the form of Q *scenarios*, M_q, for $1 \leq q \leq Q$, where $M_{q,t}$, $1 \leq t \leq T$, is the arrival capacity (AAR) of the airport during time t, if scenario q were realized. We assume that the probability of the q'th scenario occurring, p_q, is known. Let $c_g > 0$ be the cost of ground holding a single plane for one time period and let $c_a > 0$ be the cost of one period of airborne delay for a single plane. We assume that these costs are available, linear in the length of the delay, and that $c_a > c_g$ (if not, there is no need for ground holding). To ensure feasibility, we add a $T+1$'st time period during which the airport has an arbitrarily large capacity (A solution with delays extending

beyond into the $T+1$'st interval would indicate that the original time horizon (T) was not sufficient to ensure a recovery from lost capacity).

Our decision variables are A_t, $1 \leq t \leq T+1$, the number of planes that should land during time interval t in the absence of airborne delays. We also introduce auxiliary variables G_t, and $W_{q,t}$, for $1 \leq t \leq T+1$, $1 \leq q \leq Q$, where G_t is the number of flights whose arrival time is adjusted from time interval t to time interval $t+1$ using a ground delay at their point of origin, and $W_{q,t}$ is the number of flights held in the air from time period t to $t+1$ by an airborne delay under scenario q.

The A_t values can be viewed as planned airport acceptance rates (PAARs) in the sense that they represent the number of aircraft that should land in each time interval based on the planned departure times. Of course, depending on

which AAR scenario is realized it may or may not be possible to land the planned number of aircraft. The G_t variables represent arrival time adjustments

made based on planned ground delay and the $W_{q,t}$ variables represent arrival time adjustments made based on unplanned (stochastic) airborne delays. The

model can be viewed as assigning ground delays in order to mitigate uncertain airborne delays.

The model provides information on adjustments to arrival times. The actual

control variables in the physical system are aircraft departure times. The revised departure times can be easily recovered by subtracting the travel time from the scheduled arrival time. Since the model does not assign delay to individual aircraft, another mechanism must do this. As indicated in the previous section, the model achieves compatibility with CDM procedures by dealing with planes in the aggregate.

The objective is to minimize the expected value of the sum of the air and ground delay costs. This gives rise to the following integer programming problem:

$$\min \sum_{t=1}^{T} c_g G_t + \sum_{q=1}^{Q} \sum_{t=1}^{T} c_a p_q W_{q,t} \tag{1}$$

$$A_t - G_{t-1} + G_t = D_t \qquad t = 1, \ldots, T+1 \quad (G_0 = G_{T+1} = 0) \tag{2}$$

$$-W_{q,t-1} + W_{q,t} - A_t \geq -M_{q,t} \qquad t = 1, \ldots, T+1, \quad q = 1, \ldots, Q$$

Figure 1: Illustration of Model for $Q = 1$ and $Q = 2$

$$(W_{q,0} = W_{q,T+1} = 0) \tag{3}$$

$$A_t \in \mathcal{Z}_+, W_{q,t} \in \mathcal{Z}_+, G_t \in \mathcal{Z}_+ \tag{4}$$

The objective function (1) is the sum of the (fixed) ground delay costs and the (expected) air delay costs. Constraint set (2) defines conservation of flow for planes departing from their destination airport: all planes that were delayed from the previous time period (G_{t-1}), plus new demand for this time period (D_t), must either be delayed until the next time period (G_t) or allowed to arrive (A_t). Constraint set (3) performs a similar function for the air delay queue at the capacitated airport. The "$\geq$" rather than an "$=$" is necessary because there may not be enough planes available to fill up the airport's capacity.

We note that this model is "almost" a network flow problem. Constraints (2) and (3) have the structure of flow conservation constraints, however, the presence of a common in-flow variable, A_t, in each of the Q constraints in (3) destroy the network structure except in the case of $Q = 1$. Figure **??** illustrates the model for the cases of $Q = 1$, where it is a

network flow problem and $Q = 2$ where it is not.

This model is closely related to Richetta and Odoni's model (see [**?**]), but is much simpler. In Richetta and Odoni's model, the decision variables are of the form $X_{i,j}$, $1 \leq i \leq j \leq t$, the number of planes that were originally scheduled to arrive at time i that were rescheduled to arrive at time j. Additionally, slightly superlinear ground holding costs are used, in order to avoid solutions where the delay distribution is perceived to be unfair. Assuming the superlinear cost structure, with minimal pre- and post-processing, our model produces solutions identical to Richetta and Odoni's model but is substantially faster to solve. Additionally, the model admits a formal proof that the LP relaxation of the IP is guaranteed to yield integral solutions (see Section **??**).

We now turn to the cost parameters c_a and c_g. An inspection of the model indicates that the optimal solution depends only on the *ratio* of these parameters $r = \frac{c_a}{c_g}$. A possible objection to this model is the necessity to specify this ratio. We emphasize that this ratio need not represent actual relative dollar costs of ground and air delay — indeed, the relative costs of air and ground delay may vary across airlines and even across flights by the same airline. Instead, the cost ratio r need only be a quantification of the FAA's willingness to trade ground delay for air delay, taking into account the tradeoff between its desire to serve the industry efficiently and its operational and safety concerns.

3 Theoretical Results

In this section, we show that the static stochastic ground holding problem can be solved in polynomial time. In particular, we show that the integer program defined in the previous section is a dual network flow problem. As immediate corollaries, the constraint matrix associated with this IP is totally unimodular, and the LP relaxation yields integral solutions.

We say that an m row, n column $(0, 1, -1)$ matrix M is a *network matrix* if there exists a directed tree R on $m + 1$ nodes and a one-to-one mapping of the rows of our matrix onto the edges of R with the property that each column of the matrix corresponds to the characteristic vector of a path in R. This definition was introduced in [**?**] and is explored in [**?**]. The more familiar *node-arc incidence matrices*, with a single 1 and -1 in each column, are a special case of this construction. Network matrices are desirable because they are totally unimodular and give rise to integral polyhedra.

Theorem 3.1 *Let A^T be the transpose of the constraint matrix associated with the static stochastic ground holding problem. Then A^T is a network matrix.*

We will illustrate the construction of the tree R for the case $(T = 2, Q = 2)$ (two time periods and two scenarios) from which it will be clear how to extend the construction to an arbitrary number of time periods and scenarios. The general construction is explicitly formulated as a matrix transformation in [**?**], thus providing a more formal proof.

For a two-time period, two-scenario problem with slack variables, the primal problem is:

Where M is any sufficiently large constant (e.g., $M = \sum_{t=1}^{T} D_t$). We write the matrix A associated with this problem as below;

Note: actually, A^T is a $(0, 1, -1)$ matrix but, for ease of association of each column with its respective variable, we have substituted the corresponding variable name for each 0, 1, -1 entry in its column. To show that A^T is a network matrix, we must find a directed tree R on 16 nodes such that every column of A^T (equivalently, every *row* of A) corresponds to the characteristic vector of a path in R. Figure **??** shows the required tree. We have labeled each arc of R with the corresponding primal variable. It can be verified that every row of A corresponds to the characteristic vector of a path in R. For example, the fifth row of A corresponds to the path formed by traversing arc S_{12} backward, then arc W_{12} forward, then arcs A_2 and W_{11} backward. Figure **??** shows the graph G spanned by the tree R. We label each arc with the associated dual

Figure 2: The dual tree R

Figure 3: The dual tree R, with columns of A^T

cost coefficient (in the primal, the associated RHS element). This construction extends to an arbitrary of number of scenarios and time periods.

Corollary 3.2 *The constraint matrix associated with the static stochastic ground holding problem is totally unimodular, its LP relaxation yields an integral solution, and the stochastic ground holding problem can be solved in polynomial time.*

The dual problem can be recast in a more familiar form as a min-cost flow problem by negating the objective function and multiplying the dual LP in equality form by N, where N is the node-arc incidence matrix associated with R with a single row deleted. Figure **??** shows the resulting flow problem for the previous example ($T = 2, Q = 2$). The labeled arcs can accommodate either positive or negative flow, as they are associated with equality constraints in the original primal problem; the flow must be nonnegative on the unlabeled arcs, which are associated with dual slack variables. The unlabeled arcs are zero-cost arcs and the unlabeled nodes are neither sources nor sinks.

We note that the primal problem cannot be recast as a network flow problem. This can be verified by applying standard network recognition algorithms (see [**?**] or [**?**]). See [**?**] for further discussion of this issue. Additional investigations into these problem classes are given in Reference [**?**], which shows that scenario-stochastic min-cost flow problems in outerplanar graphs can always be recast as dual min-cost flow problems.

4 Experiments

In [**?**], Richetta and Odoni performed computational experiments in which they compared the quality of solutions generated by their stochastic model to those of the deterministic algorithm, and to the passive algorithm of no ground delays, under various cost and scenario assumptions. The primary conclusion was that in many cases, the stochastic algorithm found solutions with only slightly more

Figure 4: The transformed min-cost flow problem

Test Case	Numb flights	T = Numb periods	Min per period t	Time (sec)	Simplex Iteration	Node B&B
1a	624	12	60	0.07	51	0
1b	624	48	15	0.30	267	0
2a	624	12	60	0.07	59	0
2b	624	48	15	0.37	289	0
3a	624	12	60	0.05	55	0
3b	624	48	15	0.20	223	0

Table 1: SGH model performance

ground delay than the passive algorithm, with much lower total expected costs. Since our model finds identical solutions, their results apply.

To demonstrate the tractability of our model, we used CLPEX 4.0 with default settings on a Sparc Station 10 to solve three realistic instances of the stochastic ground holding problem, each comprised of a demand profile and three AAR scenarios with varying capacities (in the range of 30-60 flights per hour). Since each data set spanned a 12-hour time interval and 624 flights, these problem instances represent unusually long GDPs at a major airport. Each data set was solved in two ways: once with the time horizon divided into 60-minute time periods and once with 15-minute time periods, for a total of six test cases (1A, 1B, 2A, 2B, 3A, 3B). In each test case, we set $c_g = 2.0$ and $c_a = 5.0$.

The results in Table **??** show that the problem is highly tractable. The integer solution was obtained in zero nodes of the mixed integer program algorithm of CPLEX, empirically confirming the integer solution can be obtained directly from the linear program relaxation. The largest number of iterations of the simplex procedure was 289 and the longest run time was barely more than half a second. Note that the run time is almost linear in the coarseness (length of) the time periods.

Next, we explore the solution of an additional test case to gain a qualitative understanding of the solutions generated by the model. In our hypothetical example we know that some poor weather is approaching, but we do not know exactly when it will arrive, how long it will last, or how severe its impact will be. Figure **??** shows the expected demand at our airport over the next several hours late in a day (hence the decreasing demand), as well as three possible realizations of capacity. Each period is 30-minutes long. Capacity Scenario 1, with probability 0.4, corresponds to a severe, earliest, medium length impact, with an aftershock at period 7. Scenario 2 assumes a longer, more moderate impact, and Scenario 3 corresponds to the shortest, mildest impact; both these scenarios have probability 0.3. By comparing demand to the capacity scenarios, it is clear that some severe delays are bound to occur during periods 3-6, no matter what really happens on the capacity side.

Figure 5: Demand and Scenarios

Figure 6: Optimal Arrival Schedules, with Delay Profiles

We explore several values of $R = c_a/c_g$, the ratio of air delay cost to ground delay cost. Figure **??** shows, for each value of R, the optimal schedule of PAARs and the resulting distribution of both ground and (expected) airborne delays. Note first that as R increases, the optimal schedule grows progressively more conservative, allowing fewer and fewer flights to arrive during the earlier, possibly congested periods. Assuming a First Scheduled, First Served discipline, the X-axis indicates the scheduled arrival time period and the Y-axis the number of arrivals. The color scheme indicates the amount of ground and airborne delay experienced, with delays ranging from 0 up to 4 time periods. The ground and air delay profiles further demonstrate our theme: as R increases, arriving planes have experienced progressively more ground delays, and expect to receive progressively fewer airborne delays. Note that at $R = 25$ no airborne delay can occur. $R = 25$ therefore corresponds to a maximally conservative schedule; increasing R still further will have no additional effect. If the FAA were to apply the deterministic policy to a forecast that consisted of the lowest capacity of any scenario at each time period, this is the schedule of PAARs that would be produced. Declaring airborne delays to be 25 times as costly as ground delays is almost certainly excessive in practice; the $R = 25$ schedule is shown to indicate how far R must be increased before no airborne delays can occur (at $R = 20$, there are still some expected airborne delays). The high value of R needed to eliminate all airborne delays indicates that applying the deterministic policy to pessimistic forecasts roughly corresponds to optimizing expected costs under an unrealistically high ratio of air delay cost to ground delay cost, producing excessive ground delays. Figure **??** shows, for each value of R, the distribution of ground and (expected) airborne delays with respect to the original arrival schedule.

Please note that R may be interpreted as a rough reflection of (perhaps subjective) preferences regarding ground vs. airborne delays, instead of a cost ratio in the strict sense. By varying R the FAA can observe quickly the consequences of alternative settings of the PAARs. For example, at $R = 6$ only some flights during periods 3, 4 and 5 have expected airborne delay in excess of 30 minutes (and of under one hour), a situation that may be acceptable to both the airlines and the FAA. Note that the PAAR for this case is set to 15 and 20 arrivals during periods 3 and 4 –as opposed to 10 for $R = 25$ – thus resulting in a large reduction in total ground delay assigned to flights that will arrive after period 3 and until the end of the day.

In conclusion, the example illustrates our expectation that the proposed model will be most useful as a "what if" tool: because of its speed and simplicity, it could assist the flow management specialists in the FAA to determine PAARs that, on the one hand, take into consideration the level of

uncertainty associated with short-term capacity forecasts and, on the other, strike a "comfortable" balance between airborne and ground delay.

Figure 7: The Original Arrival Schedule, with Delay Profiles

5 Conclusion

The stochastic ground holding model presented in this paper finds the optimal trade-off between airborne and ground holding in the formulation of a ground delay program. It represents a substantial simplification of an earlier model developed by Richetta and Odoni; although it finds the same solutions via post-processing, it uses far fewer decision variables. The fact that the model is dual network allows us to relax the integrality constraints, as we are guaranteed integer optimal solutions directly from the LP relaxation. The combination of these two facts results in a highly tractable model which requires only modest computing power. Additionally, the model is designed to be an integral component of the collaborative decision-making process, and can be easily integrated into existing tools.

Practical implementation of our model requires that the following issues be addressed:

- Exempt flights: In practice, not all flights bound for an airport during a GDP can be assigned a ground delay (e.g.,

  international flights). The effect of these exempt flights is to subtract from the capacity of the airport. Our model can fully accommodate these exemptions by pre-processing the AAR scenarios. See [?] for a formal algorithm.

- Our work here assumes the real-time generation of multiple weather scenarios but does not offer a rigorous method for quantifying the probability of each scenario occurring. However, we have reason to believe that this can be done through a combination of specialized forecasting techniques and historical statistics on weather conditions and their corresponding arrival capacities.

- Experimentation is required as to the behavior of the model with respect to changes in the cost ratio. An inroad to this has been made in (see [?]).

An ideal future model for ground delay would incorporate several effects which are not modeled here and are not currently included in ATFM practice, including non-linear delay costs, the interaction between arrival and departure capacity at an airport and the "network effects" (propagation to other airports) of ground holding policies.

Finally, it is worth noting that our model with $Q = 1$ is equivalent to a classic production-inventory model in which an item can be held in one of two states (see Section 4.5 of [?]). Thus, it would seem likely that there may be applications of the more general model outside the domain of air traffic

management.

$W_{2,2}$

$W_{2,2}$

g.

titles)

Inc.,