CALIBRATING AGGREGATE MODELS OF FLIGHT DELAYS AND CANCELLATION PROBABILITIES AT INDIVIDUAL AIRPORTS

David J. Lovell and Andrew M. Churchill, Department of Civil and Environmental Engineering and Institute for Systems Research, University of Maryland, College Park, MD, lovell@eng.umd.edu, churchil@umd.edu

Amedeo R. Odoni, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, arodoni@mit.edu

Avijit Mukherjee, University Affiliated Research Center, University of California (Santa Cruz), Moffett Field, CA, avijit@isr.umd.edu

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

Abstract

This paper describes methods to calibrate aggregate models of internal delays and cancellations at a single airport. Internal delays are those specifically related to queuing effects at the arrival airport; i.e., they are induced because of a local imbalance between demand and capacity. Together with cancellations, these delays reflect important measures of system performance that are directly affected by carrier scheduling policies, as well as airport and FAA decisions about operating policies that affect capacity. We are concerned with models that might be able to produce outputs that are specific to carriers, but in any event cannot be calibrated with proprietary information, because the issues concern multiple carriers or are approached from the public perspective. Such models are appropriate, for example, when considering operational impacts of demand or capacity changes resulting from changes in infrastructure, resource allocation, or landing fees. Thus, the ground truth calibration data must come from publicly available aggregate data bases. The models themselves are unaware of some of the true drivers for cancellations and delays, and rely more on empirical correlations between these performance measures and public supply and demand information. The paper describes our experiences and recommendations about calibrating such models, including data filtering methods. Specific examples from our own efforts at model development are included, but the calibration methods described herein should be applicable to alternative model forms as well.

Introduction

There are a number of levels of detail and resolution with which it is useful to be able to predict delay levels and cancellation rates for flights arriving at a given airport. Any individual carrier has proprietary methods for estimating their own delays, and they know their own cancellation policies. They also probably each have predictive models for the delays and cancellations of their competitors. Outside the proprietary arena, however, for example from the perspective of the airport operator or of a higher level government agency, it is also important to be able to estimate these quantities, across all concerned carriers, given a reasonable projection of the scheduled demand.

There are many different definitions for delay. In this paper we are concerned with what we call "internal" delays – those delays imparted to scheduled arriving flights due to a demand/capacity imbalance at the arrival airport. Thus, for example, departure delays related to mechanical issues are not included. By focusing on internal delays at congested airports, we can study the impacts of scheduling decisions by the carriers, as well as policy decisions on the supply side that affect capacity.

A critical step in the development of models for internal delays, and one that we feel has not been given enough attention, is the calibration of these models. The scope of the models requires that they be calibrated against aggregate delay and cancellation performance metrics. Both types of statistics require care in indexing and aggregating properly. Delay statistics are more complicated than cancellation statistics because they encapsulate a large number of empirical effects simultaneously, not all of which are consistent with the notion of internal delay, while most delay models would only attempt to replicate a single effect. Thus, a significant amount of work is necessary to ensure that the calibration data and model outputs are compatible in some real sense. Finally, because these are multidimensional phenomena, care must be taken in choosing the measures for goodness-of-fit between model data and calibration data. In this

paper we provide examples of successful applications in these areas, for models that we have constructed for previous purposes. The models themselves are not as important as the calibration methods; we have framed the calibration arguments in such a way that alternative model forms with the same goals and output types could also be calibrated using similar procedures.

In the next section, we describe some of the related literature and highlight the context in which our methods reside. Following that are the two sections representing the bulk of the technical material in the paper on calibration processes, one for cancellation models and one for delay models, in that order. We close the paper with some discussions and our conclusions about applicability of the methods.

Background

The models for cancellation probabilities in the literature are predominantly constructed from the perspective of an individual carrier with proprietary information – see [1], [2], [3] for examples. An exception is a recent paper by these authors, in which we developed a broad model of cancellation activity at a single airport to support strategic simulation studies [4], based on a discrete network flow optimization platform. It is necessary, in cases such as this last paper, to understand that detailed information on carrier preferences and strategies is not available at the aggregate, non-proprietary level, and that therefore the prediction of cancellations has to be tackled more from an empirical standpoint, relating the probability of such events to other detectable system effects, such as delays.

The issue of temporal scope is also important; in this paper, as we are concerned with models whose outputs constitute a description of expected delays and cancellation probabilities over the course of an extended period of time, such as a month, quarter, or year. It is likely that over extended time periods, patterns in the daily time-varying airport capacities will appear. A number of different capacity scenarios should manifest themselves at a given airport, as a result of changes in weather and/or runway configuration. Thus, we are expecting that the models being calibrated, either for cancellation or delay purposes, have a random component that describes, through the use of probabilistic scenarios, the set of time-varying capacity profiles that are likely. The data we have used during the development of our own models result from scenario generation using a clustering algorithm [5].

At a single airport, the primary mechanism to evaluate internal arrival delays is to build a queuing system. Importantly, it is generally not appropriate to use steady-state queuing models, since the classical steady-state results of queuing theory usually cannot be applied to runway queues. During the course of a typical day demand and service rates may vary significantly over time and the use of steady-state expressions often yields very poor approximations [6]. Moreover, demand rates often exceed service rates for periods of time that may last for as long as a few hours at some major airports. Steady-state results do not, of course, apply when the demand rate exceeds the service rate.

In our own work, we have been using a non-stationary analytical queuing model called DELAYS© developed at MIT. The theoretical underpinnings to the model are described in [7, 8], while some of our own experiences with the model are described in [4]. The model is stochastic and dynamic and treats the arrival process to an airport as an $M(t)/E_k(t)/n$ queuing system.

Of course, this is only an example; many other stochastic processes might be adopted, and a great range of model specifications are possible. What is important for any of these models is to determine how well they approximate delay values and dynamic behavior observed in the highly complex air traffic environment.

Model outputs

This subsection describes the outputs we expect from any model of cancellations and/or internal delays that is to be calibrated using the methods and data manipulations described herein. If we let q represent the scenario index, then p_q is the probability that scenario q takes place. We index time by t and air carriers by a. The scheduled demand is captured by the variables D_{ta}^- and D_{ta}^+, which are the inbound and outbound demand (number of scheduled flights), respectively, during time period t for airline a. We expect the cancellation model being calibrated to produce, for capacity scenario q, an estimate r_q of the overall cancellation rate for the scenario, and furthermore, a set of cancellation predictions for each of the time buckets, $\{r_{tq}\}$. Taking expectation over the scenario distribution, then, produces the average overall cancellation rate for airline a, determined as:

$$\rho_a = \sum_q p_q \frac{\sum_t r_{tq} D_{ta}^+}{D_{ta}^+} \qquad (1)$$

The expected outputs from a delay model to be calibrated are similar. Assume the model

produces d_{tq}^{-} and d_{tq}^{+}, which are the average inbound and outbound delay, respectively, during time period t under scenario q. These can be aggregated to d_{q}^{-} and d_{q}^{+}, which are the overall average inbound and outbound delays under scenario q. The ultimate outputs of a delay model of the type we are concerned with are Γ_{a}^{+} and Γ_{a}^{-}, the average outbound and inbound delay, respectively, for airline a. These statistics are computed from the above inputs by:

$$\Gamma_{a}^{+} = \sum_{q} p_{q} \frac{\sum_{t} D_{ta}^{+} d_{tq}^{+}}{\sum_{t} D_{ta}^{-}} \tag{2}$$

$$\Gamma_{a}^{-} = \sum_{q} p_{q} \frac{\sum_{t} D_{ta}^{-} d_{tq}^{-}}{\sum_{t} D_{ta}^{-}} \tag{3}$$

Calibrating cancellation models

In this section we describe the data needs and methods proposed for calibrating cancellation models. The following section does the same for delay models. That process is arguably more complicated, since the source data for delays can be confounded with a much greater set of other factors, and thus a significant amount of data processing and shaping is required to make comparisons appropriate for calibration purposes. Historical data for cancellations are more straightforward, so this section is more benign. That does not imply that the metrics of fit for cancellation models are necessarily better than those for delay models; it is still the case that the *real* constituent processes driving cancellations are unknown and not captured in an aggregate model, so even moderately good measures of fit should be interpreted as success in this setting.

Data concerning flight cancellations are available from many sources. One question that should be considered up front is whether specificity by carrier is required. If so, one data source is the Federal Aviation Administration (FAA) Aviation System Performance Metrics (ASPM) database, and particularly the "Cancelled Flights" database of that system. This data set provides information about each commercial flight operating in the National Airspace System (NAS) that was cancelled. For each flight, the hour of scheduled arrival should be extracted from the scheduled time of arrival field, ARRSCHTM. Then, the data should be aggregated by the scheduled arrival date field ARRSCHDT, carrier field FAACARRIER, and by the derived field which shows the hour of arrival. By this method, a time series of hourly cancellations for each carrier can be created.

If individual carriers do not need to be distinguished, then this aggregation has already been done and can be found elsewhere in the ASPM database. In the quarter-hour airport analysis database, the CANCARR field specifies the number of flights scheduled in the Official Airline Guide to arrive during that quarter hour to a given airport that were cancelled for any reason. The examples in this paper show models operating at a time resolution of one hour, so the quarter-hour data were aggregated to create an hourly time-series of cancellations.

Figure 1 shows a plot used to compare predicted with actual cancellations for a 24-hour period. The predicted numbers were generated using a version of the cancellation model in [4]. The domain for this example is the Atlanta Hartsfield-Jackson International Airport (ATL), for September 15, 2004, for all carriers. The plot includes a vertical translation in the predicted curve, which is intended to account for ambient cancellation purposes not captured in the model. This detail is described in a later paragraph.

Our experience with many such comparisons has led to a number of qualitative conclusions that should help guide the calibration process. First, any cancellation model that uses predicted congestion as a proxy for the causes of cancellations (as ours did in [4]), necessarily ignores some of the other known sources of cancellations, including operational factors such as mechanical and crew problems. Thus, one might imagine that there is an underlying ambient rate of cancellation from these other causes that is stochastic and probably low compared to congestion effects, yet nevertheless might manifest itself as a systematic difference between the plots of actual and predicted cancellations. This might be particularly true in early morning and late evening periods when low demands would not imply enough delay effects to induce cancellations for congestion reasons alone.

Because congestion-related concerns tend to be the primary driver in cancellations, aggregate cancellation models based on congestion factors should (if they are good) do well in identifying the times of day when cancellations are most likely, and with predicting the shape of the cancellation profile over the course of the day. However, because of complications like the background cancellation levels mentioned above, there can be a systematic difference between actual and predicted profiles on these plots.

The calibration effort should not be penalized for this difference, and one way to deal with this is to use calibration metrics that are related more to the shape of curves than to their explicit values. An example of such a process would be to shift the

predicted curve vertically until it matches best with the actual cancellation curve. The metric for goodness of fit would then be the root-mean-square (RMS) difference between the two curves. Thus, the predicted curve shown in Figure 1 was shifted vertically from its original position by the amount 0.13542 to provide the best match with the actual curve. Importantly, however, this has no effect on what we believe to be the most salient feature of the chart, i.e., the close match between the times and relative magnitudes of most of the cancellation activity.

As mentioned above, we expect that for this type of cancellation model, one will observe some actual cancellations early and late in the day that are not predicted by the congestion-related model. Presumably, these are due to reasons other than congestion; so again, it is not unreasonable that the model does not match well at these times. If an absolute measure of cancellation activity were the goal, it would be important to account for this background cancellation level in some other way. However, the premise of this paper is that the models being calibrated were developed to compare, for example, the congestion impacts of several different scheduling policies at an airport, so their outputs should only be sensitive to the important inputs, which are related to the scheduled demand and the capacity of the airport. If one had reason to believe that the unaccounted-for cancellations had a markedly non-uniform distribution over the course of the day, then the RMS metric described above could be weighted to ensure that, for example, very early and very late cancellation patterns did not have an excessive impact on the calibration process.

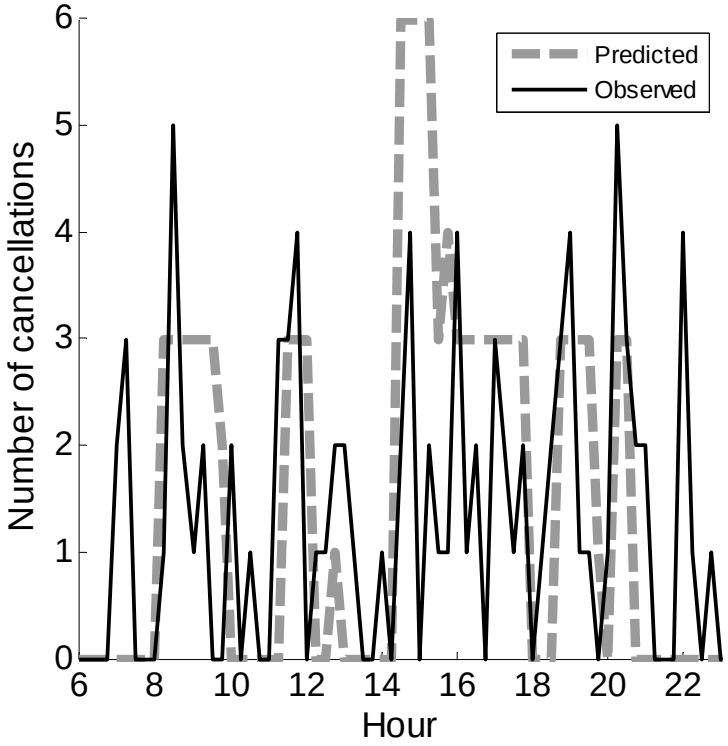

Figure 1 - Hourly cancellation profiles, ATL, September 15, 2004

At the aggregate level, partitioning predicted cancellations by airline can be difficult, since the data driving individual carrier decisions are not available. Two modeling approaches that might be adopted are to assume that carriers' cancellation propensities will continue as they have historically, despite any change in the schedule being evaluated by the model. This does not work, of course, when the scenario under study includes new entrants or significant realignment in carrier occupancy of the airport. An alternative is to assume that cancellations will be meted out to carriers commensurate with their fraction of the scheduled demand. This is more a question of the models used than the calibration process itself, and we do not endorse in this paper any particular modeling strategy over another. For the present purposes, it most important to note that the calibration process

described here for an airport could be applied for individual carriers as well.

It might be the case that inbound and outbound cancellation rates are estimated differently. Differences between those rates, however, obviously violate the principle of conservation of flow, and therefore might only be trustworthy if a good explanation were provided for this. In our work with specific cancellation models, such as [4], we estimate only cancellations among arriving flights and assume that the same rates apply to departing flights. Notice that this is not necessarily a simplifying assumption – in a very detailed model, one would have to think carefully about *which* departing flights were being cancelled, which depends strongly on the network structure and operating practices of the carriers involved. Some of these data are proprietary, of course, and outside the scope of the type of modeling we have envisioned for the models in this paper.

Calibrating delay models

The models envisioned for the methods in this paper are those that produce estimates of internal delays at an airport; e.g., a queuing model. We are concerned with decisions made by carriers and/or airport operators that affect either demand, capacity, or both, and delay performance measures associated with those decisions should not be confounded with external causes that are not a direct result of the decisions. This classification can be difficult – for example en route delay is not automatically excluded. Delays en route caused by congestion in an upstream sector or re-routing around a weather area would not be included, but en route delays incurred because of a miles-in-trail restriction imposed to meter flows into the specific arrival airport in question would be appropriate to include. When looking at data sources for delay, frequently only the difference between scheduled and actual arrival time is available for a particular flight, and that difference (which is a more general form of delay) contains elements of delay we are interested in but perhaps some that we are not. Beginning with methods to clean up the delay data, following are the steps that will be described in the recommended calibration methodology:

1. Ground truth data preparation – remove the impacts of propagated delay, unscheduled traffic, and external delays.
2. Construct time-dependent profiles of predicted internal delay and the ground truth delays from step 1.
3. Rectify those profiles using a pattern-matching method and metric.
4. Quantify the calibration quality with the optimal profile-matching metric from step 3.

The rest of this section is devoted to describing these steps in more detail, using a particular example from previous work by these authors [4]. The data for that example are from the ASPM database, for a full year (2004) at ATL airport. The first step in the process is to clean up the data. This step is driven partly by how the data in ASPM are represented and, as such, other data sources might require less, or different, degrees of manipulation to make them appropriate for calibration purposes. Whether ASPM data are used or not is less important than singling out the detailed distinctions that can exist between sets of numbers that all purport to represent "delays" at an airport. This section closes with explanation and demonstration of calibration measures of goodness-of-fit.

Propagated delay

The principal difficulty in performing a model validation with the ASPM database stems from the fact that the delay estimates that are obtainable directly through the ASPM database are severely contaminated in at least three important respects. First, the recorded delay metrics used most commonly measure the difference between the actual arrival time of a flight and its scheduled arrival time. For example, for most flights other than those arriving early in the morning, some of the recorded delay at ATL (relative to the scheduled arrival time) may be due to delays associated with flight legs flown by the same aircraft prior to the leg arriving at ATL. Of course, a queuing model applied for operations only at ATL would estimate only the delays resulting from the relationship between the schedule of demand (arriving flights in our case) and available runway capacity on any given day. Clearly, to make a fair comparison between field data and the results of a queuing model applied only at ATL, this "propagated delay" must be filtered out. Fortunately, the use of individual flight records makes this possible, as they track aircraft by tail number as they travel through the NAS. The scheduled and actual arrival and departure times for every leg of each aircraft's itinerary are available. One readily available source for this data is the "Individual Flights" database in ASPM, although others also exist. This method makes it possible to re-compute the recorded delays by subtracting propagated delay from the total delay incurred by each scheduled flight arriving at ATL for all days in 2004. The re-computed delay statistics thus contain only the delay incurred for operations arriving at ATL in 2004. Note,

however, that this step alone is not sufficient (see "Internal and External Delays" below).

Unscheduled flights

The second challenge in comparing empirical data from ASPM with the results of any queuing model lies in properly accounting for unscheduled flights. Such "pop-up" flights occur without prior notice in the form of general aviation and other types of traffic, and do consume part of the available capacity of the runway system (the "airport acceptance rate" – AAR). Moreover, no meaningful delay statistics are recorded for unscheduled flights. On the other hand, the demand profiles that can be provided to the queuing model do not include unscheduled flights as, by definition, it is not possible to assign a scheduled demand time to them. We recommend that this problem be dealt with, in an approximate way, by taking advantage of the fact that it is possible to extract a count of the actual number of unscheduled arrivals taking place in each hour from a part of the ASPM database. In our analysis, this number was subtracted from the AAR provided by ASPM, for each time period, to arrive at the capacity which is truly available to the scheduled traffic. In other words, if during a time period *t*, the number of scheduled flights is *s* and of unscheduled flights *u*, the queuing model is operated with demand *s* and capacity (AAR-*u*) whereas, in truth, the demand is (*s*+*u*) and the capacity is AAR. This approximation is justified by the fact that the number of unscheduled flights at such major commercial airports as ATL, ORD and LGA is quite small compared to the number of scheduled flights.

Internal and External Delays

The third and most challenging problem with the data is that even after subtracting delay propagated from upstream flight legs, the arrival delay experienced by a flight is not necessarily due to a shortage of capacity at the arrival airport. Numerous other factors, such as en-route congestion and weather, congestion related to the take-off phase of the incoming flight, late crew arrival, long boarding times and any number of mechanical problems may significantly lengthen the previous turn-around, as well as the en-route segment of the flight. We refer to the delays due to the relationship between capacity and scheduled demand at the arrival airport as "internal" delay – the delay that the queuing model estimates – and the delay due to all other causes, other than propagated delay, as "external" delay.

To investigate this further, a linear regression model was constructed using the observed delay (minus propagated delay) as the dependent variable and two independent variables, the predicted delay from the DELAYS queuing model (using scheduled demand minus cancellations as the input and the revised AARs as the capacities, as described previously) and a measure reflecting external factors causing arrival delays at ATL as described below. The objective was to decompose arrival delays at ATL into internal and external components, and investigate whether the queuing model performs adequately at estimating the former.

En route congestion and weather are key factors that cause delays to air traffic. On any given day, for example, flights bound to ATL may suffer additional delay if there is convective weather activity in the greater vicinity of ATL. However, this delay will also be felt by many other flights that arrive at airports in the vicinity of ATL. In the event of convective weather, therefore, there will be strong correlations amongst delays of flights in the same area at similar times. Using the ASPM data, we computed the average arrival delay of flights arriving at airports within 400 miles of ATL (measured in Great Circle distance) and called this statistic the "regional delay." This regional delay was treated as the second independent variable in the regression model, i.e., as a proxy for the external component of delay. Flights that departed from ATL to other airports within this region were not considered, because any departure delays suffered by those flights may be linked with terminal area weather conditions at ATL.

The results from the regression analysis are summarized in Table 1 below. The squared multiple correlation coefficient R^2 is 0.52; i.e., the model is able to account for 52% of the variation in the data. The parameter estimates in Table 1 are all significant at the 95% confidence level. It is particularly noteworthy that the coefficient related to the estimate of delay from the queuing model is statistically indistinguishable from 1. Another important fact is that this parameter's estimate is not biased, suggesting that the delay due to Atlanta's capacity reduction is related primarily to weather in the terminal airspace of ATL, and hence not correlated with unobserved external factors – i.e. the error term. Note also that, according to the model, each extra minute of delay experienced by the regional flights, which in turn may be caused by convective weather within 400 miles from ATL, causes 0.56 minutes of delay to flights inbound to ATL. As was already noted, there are many factors, in addition to capacity shortfall at ATL and en route convective weather, which induce arrival delay to flights and cannot be included in the simple regression model for lack of adequate, publicly-available information.

Table 1: Internal vs. External delays

Explanatory Variable	Parameter Estimate	Standard Error	t-statistic	p-value
Intercept	-1.73	0.88	-1.95	0.05
Predicted average delay from queueing model	0.98	0.06	16.25	0.00
Regional delay	0.56	0.06	8.88	0.00

R^2 : 0.52; Adjusted R^2 : 0.52

The small negative intercept may suggest an interesting point. On a hypothetical day when the DELAYS model predicted no internal delay and there was no delay for flights arriving within 400 miles of ATL, the model would predict that the arrival delay at ATL would be -1.73 minutes, on average. Although such days are extremely rare, the negative intercept may be a reflection of the practice of airline "padding" of scheduled arrival times, which is used to take into consideration expected delays and thus improve on-time-arrival performance. One may also argue that the negative intercept is immaterial as a few minutes of delay are always caused by the variety of other factors mentioned above, which are not accounted for in the model using this level of detail. Finally, it is possible that the negative intercept is simply correcting for some amount of collinearity between the two independent variables, within reasonable ranges.

Hourly Delay Profiles

The most demanding and critical test of the queuing model is whether, in addition to overall average delay, it can predict the dynamic behavior of delay over the course of the day. To this end we compare the average predicted delays on an *hourly* basis obtained from the queuing model with the hourly profiles of average filtered observed delays for each month of 2004. Figure 2 shows, as an example, this comparison for February 2004. A very similar pattern was obtained for the other months of 2004. As suggested by Figure 2, these figures provide compelling evidence that the queuing model captures well the dynamic shape of the delay profile over the course of the day. The consistent under-prediction of the size of the delays is also to be expected, as the queuing model estimates only the "internal" delays at ATL, as discussed previously. This is very similar to the ambient cancellation levels described in the previous section of this paper.

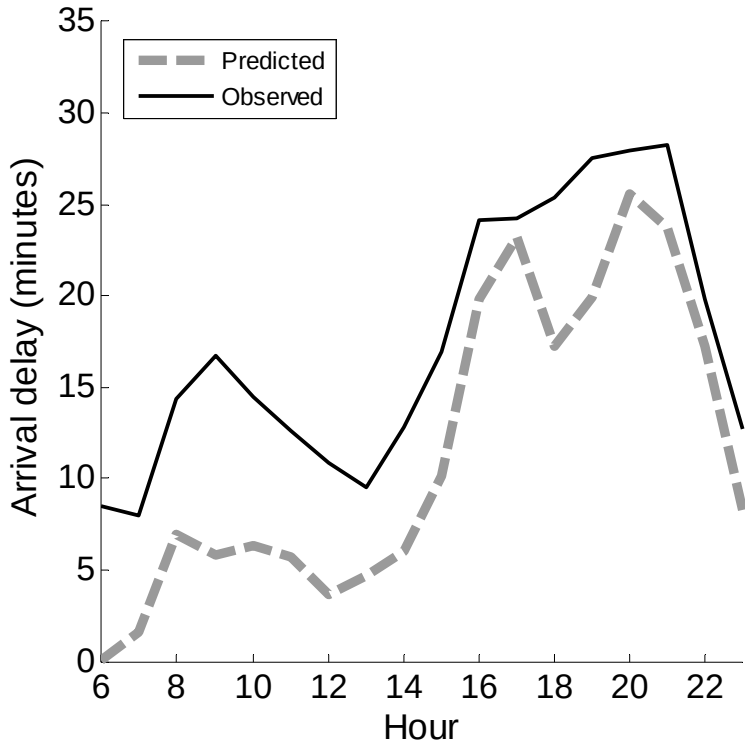

Figure 2 - Hourly profiles for February 2004

As was done previously with the cancellation profiles, the estimated delay profile is translated vertically in order to superimpose "best" upon the actual delay profile. The metric for "best" is the

same as before – the RMS difference between the vertices (i.e., the hourly values of average delay) of these two piecewise-linear profiles. The values of this metric for all 12 months of 2004 are shown in Table 2. The associated offset can also be thought of as a rough indicator of the average amount by which internal delay, as estimated by the DELAYS queuing model, under-predicts the average delay per flight during the course of an average day in that month. This process is illustrated for the February data in Figure 3. The optimal offset is 5.9 minutes, with a resulting sum of squared deviations equal to 100.4 min^2.

In conclusion, both in the aggregate and on a dynamic basis, one might conclude on the basis of this calibration exercise that the DELAYS model is in fact predicting the airport capacity-related delays (internal delays) at a level of accuracy which is satisfactory and entirely consistent with the strategic and approximate nature of the applications for which it was intended.

Table 2: Goodness-of-fit between hourly delay profiles

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Offset	2.9	5.9	2.5	2.6	6.4	10.2	5.6	3.8	7.3	1.6	5.0	7.0
Metric	187.9	100.4	50.3	53.5	95.4	546.8	129.1	78.6	220.8	150.9	25.3	50.5

As with the cancellation model, it should be noted that the example used in this paper does not explicitly show results that are distinguished by carrier. Partly this is true because delay effects as caused by queuing interactions are felt by all carriers. This is one of the primary difficulties in securing voluntary demand reductions at congested airports: carriers' delay experiences do not follow directly and proportionately from their own scheduling volume.

Nevertheless, there are applications where one might want to "assign" delays to carriers. For example, the authors of this paper (with a number of others) participated in a large strategic simulation to study collaborative scheduling practices with market mechanisms for assigning prices to airport resources [9]. In an iterative fashion, each carrier provided tentative schedule requests, and then aggregate and carrier-specific performance metrics related to delays and cancellations were provided as feedback. This enabled carriers to understand the implications of their own scheduling decisions, and in fact represented a novel fiduciary mechanism by which they might actually be expected to pay for their impacts.

The means to partition aggregate delay results to individual carriers might be as simple as apportioning according to their fraction of the schedule, as was offered as a possibility for the cancellation part of the paper. This is a potentially troublesome approach, however, as delays are incurred later than, and for longer durations than, the flights whose presence in the schedule might be thought of as responsible for those delays. Like the cancellation process mentioned above, if the general carrier representation at the airport was not imagined to change, then one could look at historical proportions of delays across carriers and use those to weight such an assignment. Alternatively, one could utilize the fact that during congested periods, the accrual of delay (i.e., over-scheduling) in early time periods has a much greater impact on total delay than later periods that are followed closely by under-scheduled recovery periods. Some mechanism might be used to account for these effects – see for example [10]. In any event, this process would take place *after* all of the important data shaping steps described in this paper had already taken place, and the final carrier-specific predictions could be compared using identical graphical forms and numerical metrics of goodness-of-fit.

Conclusions

We believe that there are a number of important applications of aggregate models of flight delays and cancellation probabilities at congested airports. Because of their scope and/or purpose, the models in mind cannot be supported by proprietary information from carriers that might give insight into the detailed causal processes which result in cancellations and delays. Nevertheless we assume that these models can be calibrated to produce sufficiently reliable results, and real examples exist to support this notion.

Furthermore, we assume that there is a real niche for analytical models (rather than, for example, simulations), particularly because of the strategic insights that they provide. Our own research into such models has been motivated by the need to support studies related to the allocation of airport capacity through administrative measures or through market-based mechanisms. However, such models have potential application in a wide range of settings where one wishes to estimate airport performance while varying certain operational characteristics.

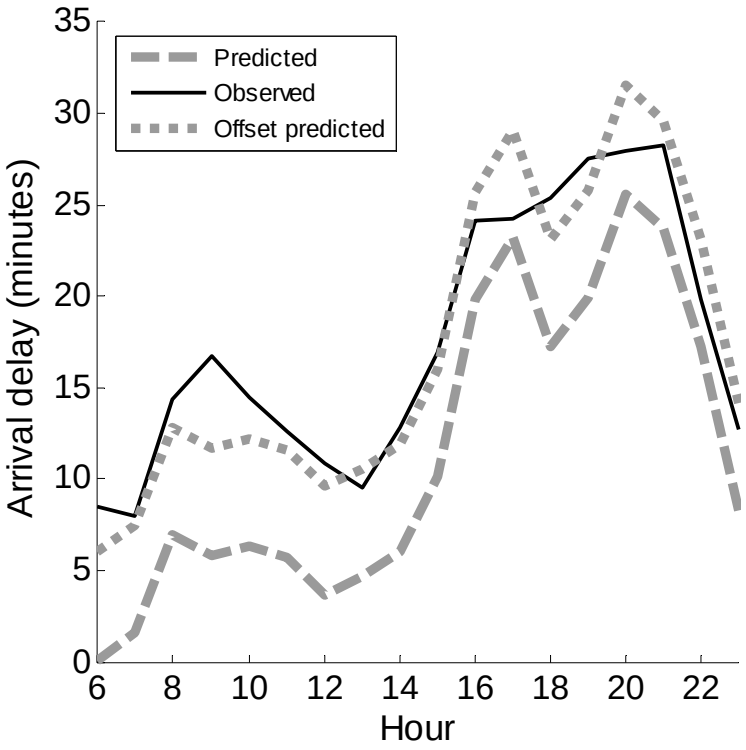

Figure 3 - Hourly profiles for February 2004 with vertical shift

The primary issue with such models is calibrating them to real data. Because their internal mechanisms do not closely reflect actual carrier practices, they rely heavily on patterns observed in historical data. These data are noisy and confounded with a multitude of effects, not all of which are desired in the calibration process. The goal of this paper, therefore, has been to describe specific experiences we have had in developing processes for data shaping and model calibration for this class of models. We do not endorse any particular models (although we offer examples from our own recent experience); rather, the purpose is to provide recipes and advice that can be useful for any number of model forms and applications, provided that they match in the basic description of the forms of output.

Our detailed recommendations were provided separately for calibration of models for cancellations and delays. Both applications have two important things in common, however. The first is that because not all causal factors related to cancellations and delays can be captured by aggregate models, there will always be a systematic difference between model outputs and real data for these effects. These differences should be accounted for, but not unfairly penalized, in the calibration process. Secondly, in applications where it is important to be able to partition predicted cancellation and delay statistics across the affected carriers, the means to do so fairly are complicated. When the nature of the airport operations is not expected to change drastically, historical data helps to provide weights by which this attribution to carriers might take place. In other cases, weighting by presence in the schedule might be appropriate. In any event, this is a difficult and qualitative step and should be approached with great care.

In summary, we believe that the insights into data sources, structure, and shaping possibilities are an important contribution to the body of literature in this area of operational modeling. The calibration steps proposed are tailored to the specifics of the data, models, and applications, and for this reason should be better than applying a generic calibration process. Certainly, a consistent process for calibrating models of this type would be helpful in comparing, on an equal footing, the performance of competing models. This represents another important contribution of this paper.

Acknowledgements

This work was supported by the National Center of Excellence for Aviation Operations Research (NEXTOR), under contracts from the Federal Aviation Administration (FAA). Opinions

expressed herein do not necessarily reflect those of the FAA.

References

[1]. Arguello, M.F., J.F. Bard, G. Yu, 1997, "Models and Methods for Managing Airline Irregular Operations," in *Operations Research in Airline Industry*, G. Yu (Ed.), Kluwer Academic Publishers, Boston, MA, pp. 1-45.

[2]. Rosenberger, J., E. Johnson, G. Nemhauser, 2003, "Rerouting Aircraft for Airline Recovery," *Transportation Science* 37, pp. 408-421.

[3]. Bratu, S., C. Barnhart, 2004, "Flight Operations Recovery: New Approaches Considering Passenger Recovery," Global Airline Industry Program Working Paper, Massachusetts Institute of Technology, Cambridge, MA.

[4] Mukherjee, A., D.J. Lovell, M.O. Ball, A.R. Odoni, G. Zerbib, 2005, "Modeling Delays and Cancellation Probabilities to Support Strategic Simulations," in *Proceedings of 6th USA/Europe Air Traffic Management R&D Seminar*, Baltimore, MD, USA.

[5]. Liu, B., M. Hansen, A. Mukherjee, 2006, "Scenario-based Air Traffic Flow Management: Developing and Using Capacity Scenario Trees," *Transportation Research Record*, 1951, pp. 113-131.

[6]. Odoni, A. R., E. Roth, 1983, "An Empirical Investigation of the Transient Behavior of Stationary Queuing Systems," *Operations Research,* 31, pp. 432-455.

[7]. Kivestu, P., 1976, "Alternative Methods of Investigating the Time-dependent M/G/K Queue," S.M. Thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA.

[8]. Malone, K., 1995, "Dynamic Queuing Systems: Behavior and Approximations for Individual Queues and Networks," Ph.D. Thesis, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.

[9]. Ball, M.O., K. Hoffman, G. Donohue, P. Railsback, D. Wang, L. Le, D. Lovell, A. Mukherjee, 2005, "Interim Report: The Passenger Bill of Rights Game," NEXTOR Working Paper NWP-2005-002, National Center of Excellence for Aviation Operations Research.

[10]. Lovell, D.J., M.O. Ball, A. Mukherjee, 2005, "Airport Congestion Prices based on Deterministic Queuing Effects," presented at the 2005 meeting of the Institute for Operations Research and the Management Sciences (INFORMS), San Francisco, CA.

Key Words

Delay estimation, cancellation probabilities, aggregate model calibration, traffic flow management

Biographies

David J. Lovell is an Associate Professor in the Department of Civil and Environmental Engineering at the University of Maryland. He holds a joint appointment with the Institute for Systems Research (ISR). Dr. Lovell received his B.A. degree in Mathematics from Portland State University, and his M.S. and Ph.D. degrees in Civil Engineering from the University of California, Berkeley.

Andrew M. Churchill is a Ph.D. student in the Department of Civil and Environment Engineering at the University of Maryland, where he serves as a Graduate Research Assistant for NEXTOR. He received his B.S. degree in Aerospace Engineering from the University of Maryland. He has previously worked in the airline industry.

Amedeo R. Odoni is T. Wilson Professor of Aeronautics and Astronautics and of Civil and Environmental Engineering and Co-Director of the Global Airline Industry Program at MIT. He was Co-Director of NEXTOR from 1996 to 2002. His recent books are *Urban Operations Research*, co-authored with Richard C. Larson (Dynamic Ideas, 2007) and *Airport Systems*, with Richard de Neufville (McGraw-Hill, 2003).

Avijit Mukherjee is an Associate Project Scientist at the University Affiliated Research Center, University of California (Santa Cruz), Moffett Field, CA. He is involved in the research on Air Traffic Management at the Aviation Systems Division of the NASA Ames Research Center. Dr. Mukherjee received his Ph.D. in Civil and Environmental Engineering from University of California, Berkeley, in 2004.

Michael O. Ball is the Orkand Corporation Professor of Management Science in the Robert H. Smith School of Business at the University of Maryland. He also holds a joint appointment within the Institute for Systems Research (ISR) in the Clark School of Engineering. Dr. Ball received his PhD in Operations Research in 1977 from Cornell University. He is co-Director of NEXTOR, the National Center of Excellence for Aviation Operations Research, and he leads the NEXTOR Collaborative Decision Making project.