Validating an Operations Centre Model

Problem presented by

Mark Harrington

GE Aviation Systems

Executive Summary

An airline operations centre determines the imminent and future operations of an airline by assessing the current situation, considering current constraints and then issuing a high level plan, describing which flights are operating, which are cancelled, which are delayed or rerouted, impacting aircraft, air crew, flight cabin crew, passengers and ground staff. Other departments of the airline transform the high level plan into detailed execution instructions. In the event, the system consists of a mixture of humans and machines both of which often provide and receive incomplete information, on which they make a decision, striving to execute as close as possible the 'Ideal Plan', which is the published timetable. The Airline Operations Decision Making Process is the 'planning' portion determining the next course of action, and the Airline Operations is the execution portion, putting the plans into action. In this context, the Study Group was asked to address the following problem: How do you validate a model of an airline's decision processes to ensure a faithful representation of reality?

By analysing model outputs and actual data, we found that the model does not satisfy any of our proposed "validation" criteria and have therefore suggested a series of possible ways forward for the model development team. We have provided visualisation tools in both Matlab and R, which will be of use in analysing subsequent model versions. We have also provided a series of statistical indicators of model quality. Lastly, and perhaps most importantly, we have described in detail what can be done if additional information were to be made available, for instance further runs of the model or extra data from the operations centre.

Version 1.0
May 22, 2013
iii+14 pages

Report editors

Lorcan MacManus and Melvin Brown (KTN for Industrial Mathematics)

Authors / Contributors

Claudia Hecht (University of Regensburg)
Dan Hewitt (University of Bristol)
Karina Piwarska (Polish Academy of Sciences)
Leonard Smith (London School of Economics)
Erica Thompson (London School of Economics)
Eddie Wilson (University of Bristol)

ESGI91 was jointly organised by
University of Bristol
Knowledge Transfer Network for Industrial Mathematics

and was supported by
Oxford Centre for Collaborative Applied Mathematics
Warwick Complexity Centre

Contents

1 Introduction

1.1 Overview

(1.1) An airline operations centre determines the imminent and future operations of an airline by assessing the current situation, considering current constraints and then issuing a high level plan, describing which flights are operating, which are cancelled, which are delayed or rerouted, impacting aircraft, air crew, flight cabin crew, passengers and ground staff. Other departments such as customer service, cargo, catering, maintenance and fuelling add further detail to the plan. For example the customer service departments determine which passengers are allocated to which flights in the event of a cancellation or missed connections following the delay of an incoming flight, similarly the cargo department will plan revised routes for cargo. The maintenance department will organise to complete the required repairs to aircraft once the future locations of aircraft are known. The operations centre issues the high level plan to customer service, maintenance, catering, flight operations, etc. departments of the airline which transform the high level plan into detailed execution instructions, be they implicit (what normally happens) or explicit (modified to take account of abnormal circumstances).

(1.2) Expressed in a control theory framework, the operations centre is the outer loop controller with departments such as maintenance, customer servicing, fuelling, cargo, being the controller in the inner and middle level control loops. The action items, such as aircraft flying, loading, unloading, fuelling are part of the plant model. Events such as weather, mechanical failure, diversions, airport congestion, and employee industrial action (and many other factors) can be thought of as disturbances to the closed loop control system. The plant is the airline executing flights, transporting passengers and cargo. The control system analogy is flawed because of the ability of an inner controller to 'negotiate' with the high level controller is missing. A representation of this view is shown in Figure 1.

(1.3) The system consists of a mixture of humans and machines both of which often provide and receive incomplete information, on which they make a decision, striving to execute as close as possible the 'Ideal Plan', which is the published timetable. The Airline Operations Decision Making Process is the 'planning' portion determining the next course of action, and the Airline Operations is the execution portion, putting the plans into action.

1.2 Problem statement

(1.4) The Study Group addressed the following problem: How do you validate a model of an airline's decision processes to ensure a faithful representation of reality?

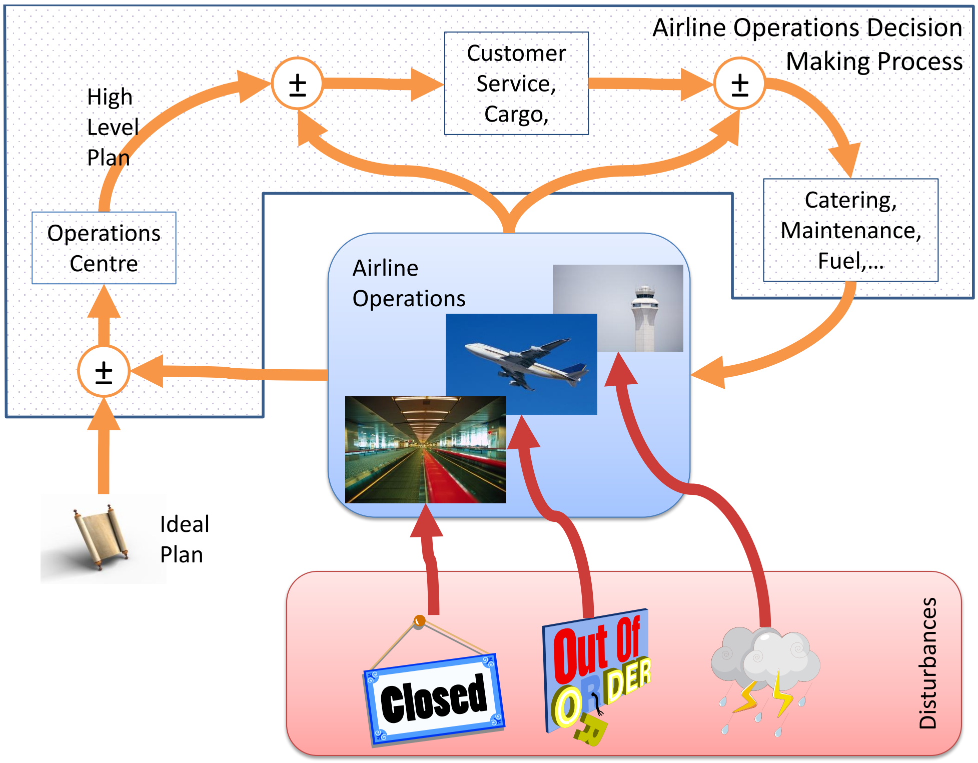

Figure 1: Airline viewed as a control system.

(1.5) GE needs the ability to demonstrate that a model of an Airline Operations Decision Making Process is an adequately fair and faithful representation of actual airline decisions. A pictorial representation of the problem posed to the study group is represented in Figure 2.

(1.6) This raises two questions: first how you measure 'adequate' and second, how do we demonstrate sufficient adherence to be able to claim that the model reflects reality?

(1.7) The myriad of potential measures is a complicating factor. Measures include things such as the number of cancellations, number of passenger delay minutes, misplaced passengers, employees working overtime, lost baggage items, late cargo, fuel used and serviceable aircraft availability, but the ultimate measure, driven by these factors and others, is the financial performance of the airline. Some measures are available in the public domain, for example Bureau of Transportation Statistics in the USA, others as part of a subscription service, for example [1], or others by private arrangement with an airline. The appropriate selection of output measures is a further issue, which partially reflect an airline's business goals, so for example one airline may be focused on minimal number of cancellations at the expense of delays, another may be focused on ensuring passengers reach destination with

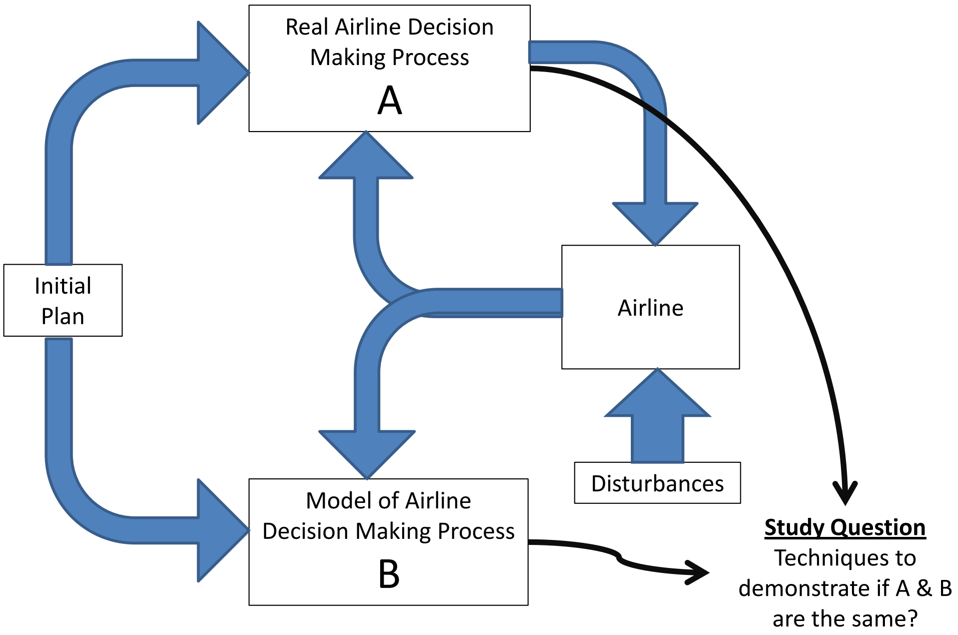
Real Airline Decision Making Process
A
Initial Plan
Airline
Disturbances
Model of Airline Decision Making Process
B
Study Question
Techniques to demonstrate if A & B are the same?

Figure 2: A pictorial representation of the study group question.

minimal delay, a third may be focused on minimizing flight crew disruptions, a fourth on lowest cost, a fifth on minimal flight arrival delay. Thus any validation techniques must account for differing objectives of different airlines.

(1.8) One approach considered is matching statistical indicators, a crude form of statistical inference, comparing actual performance to a modelled performance and checking to see if the resultant statistical distributions are sufficiently similar. An observation may be considered as a day of operations with multiple measures being recorded and a sample set could be a month's worth of data.

(1.9) Subordinate questions are: does anything similar exist already; what is the correct statistical inference test or tests; do alternates other than statistical inference exist; can the myriad of measures be catered for; can the different, and complex, goal sets of an airline be accounted for; should a validation techniques deal with component parts of the model or should a more holistic approach be taken?

(1.10) An engineering approach suggests making sure the component parts of the model are accurate as a precursor to integrating the whole. Is this the best approach?

(1.11) The validation must also consider the time element in a number of different ways. Does the model provide outputs at a similar time to reality? Are the time domain responses similar? How do the model's outputs vary with elapsed time compared to reality?

(1.12) In these types of models what level of accuracy can be achieved, correct to an order of magnitude, a small multiplier, or a few per cent? The elapsed time to complete the validation is of secondary importance.

(1.13) An area of concern is the multiplicity of non-linear and domino effects within the model. The easiest way to think about this is as a sensitivity analysis, so if a variable input to the model is adjusted slightly does it cause a large change when promulgated through the system? How is this catered for in a validation technique?

(1.14) Another way of expressing the problem is by analogy with government treasury economic or financial models, but substitute airline operations. How are the government economic, financial models verified? A further differing view of the problem is the validation of operations research models. The problem holder suspects the answer is a blend of these two views.

1.3 Anticipated Outcome

(1.15) The required output is a toolbox of techniques to check the validity of an airline operations model compared to actual performance.

(1.16) We have some output of a model which simulates decisions made by an airline's operations control centre, and we want to know how good the model is at reproducing their decisions. We are NOT trying to optimise the output of the model with respect to any airline goals (efficiency, passengers reaching destinations, etc).

2 Data Analysis

2.1 Visualisations of the data

(2.1) We worked on various visualisations to examine some of the discrepancies in the data. The Matlab and R code used to produce all of the below diagrams is made available and documented separately.

(2.2) In Figure 3, we display the aircraft by aircraft travel times and route numbers for 20 aircraft. Each blue circle represents a **decision point** (defined in (2.7)), where a disturbance occurs and a decision must be made to fly, delay, or cancel the flight. Where the model deviated from what actually happened, we have terminated that aircraft's narrative (to include the ongoing divergent pathways would clutter the graph without being directly comparable with model data).

Here, and in the following, a model is considered to deviate from the real data if one of the following incidents occur:

- More than 15 minutes difference between the flying time of the actual and the mimic data, where the flying time is calculated as the difference between gate arrival time and gate departure time;
- more than 15 minutes difference between the time at some airport of the actual and the mimic data, which is calculated as the difference between gate departure time of the current route segment and the gate arrival time of the previous route segment; or
- different routes of the actual and mimic data.

The continuous lines correspond to actual flying time. As soon as there is a divergence in the behaviour of the model and the actual data, the wrong decision of the model is additionally sketched in red colours, where the real behaviour stays black. It can be seen that some modelled aircraft remain on the same schedule as the real ones for a long period of time, whereas others diverge quickly, and some are even wrong before they have first taken off.

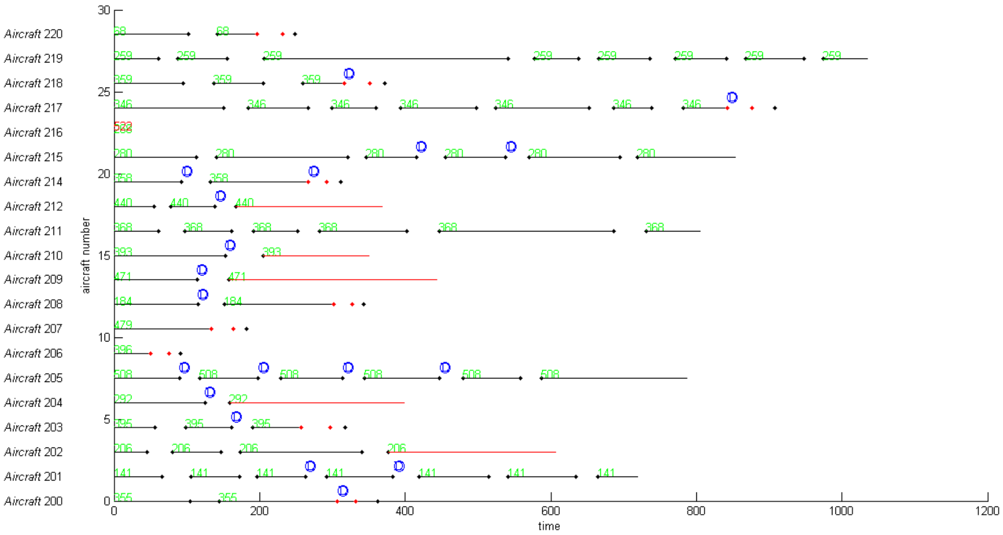

Figure 3: Here we display the aircraft by aircraft travel times and route numbers for 20 aircraft. Each blue circle represents a decision point. The model schedules are not shown after they have diverged from the actual, allowing us to see at a glance the average timescale for which the model is reasonably valid.

(2.3) In Figure 4 we display the distribution of error times, where 'error time' is defined as the time taken from the start of the schedule for an aircraft's modelled and actual flights to differ. We tried to fit a log-normal distribution but concluded that the error time data are not consistent with this distribution; we believe the reasons for this are:

- The maximum error time is limited to one day, and so there cannot be a tail as time goes to infinity.
- There is a spike at time zero, which corresponds to a large number of aircraft being rescheduled at the start of the day.

As seen in the distribution, one finds a mean value of about 300 minutes until the first error occurs; this coincides with the calculated mean value. See Section 2.2 for further analysis.

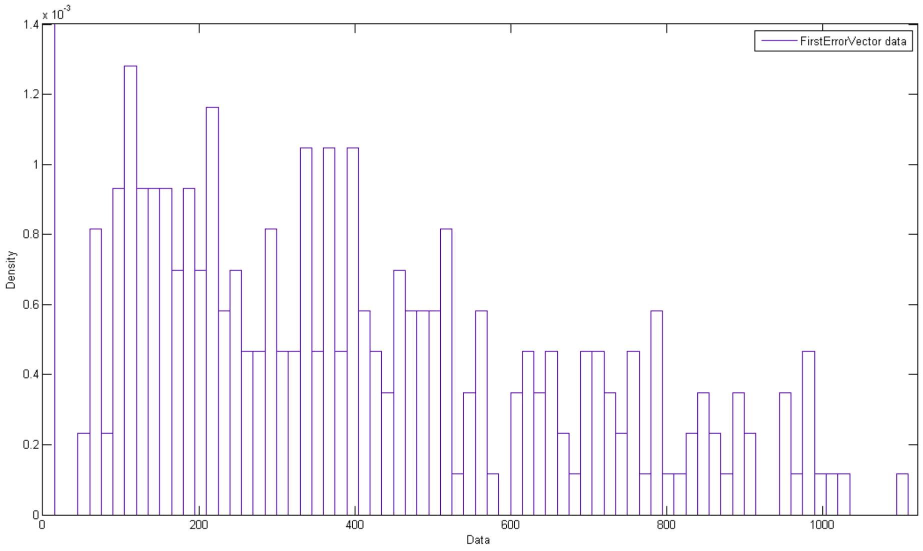

Figure 4: Error time distribution histogram. There are 71 of 500 aircraft without any error and 117 aircraft which differ at time zero.

(2.4) In Figure 5 we display the plan (blue), the actual data (green) and the model times (red) for a small subset of routes. We can see, in this diagram, that the modelled times and the plan agree extremely (perhaps unreasonably!) well, whereas the actual behaviour has a lot more delays.

(2.5) In Figure 6, we display the cumulative depart time delay of the actual data (in green) and the model (in red). It can be seen that the model does not produce anywhere near enough short delays, although it does occasionally produce a tail of longer delays. This is visible in the data sometimes as a delayed flight early in the day which is then made up at the end of the day

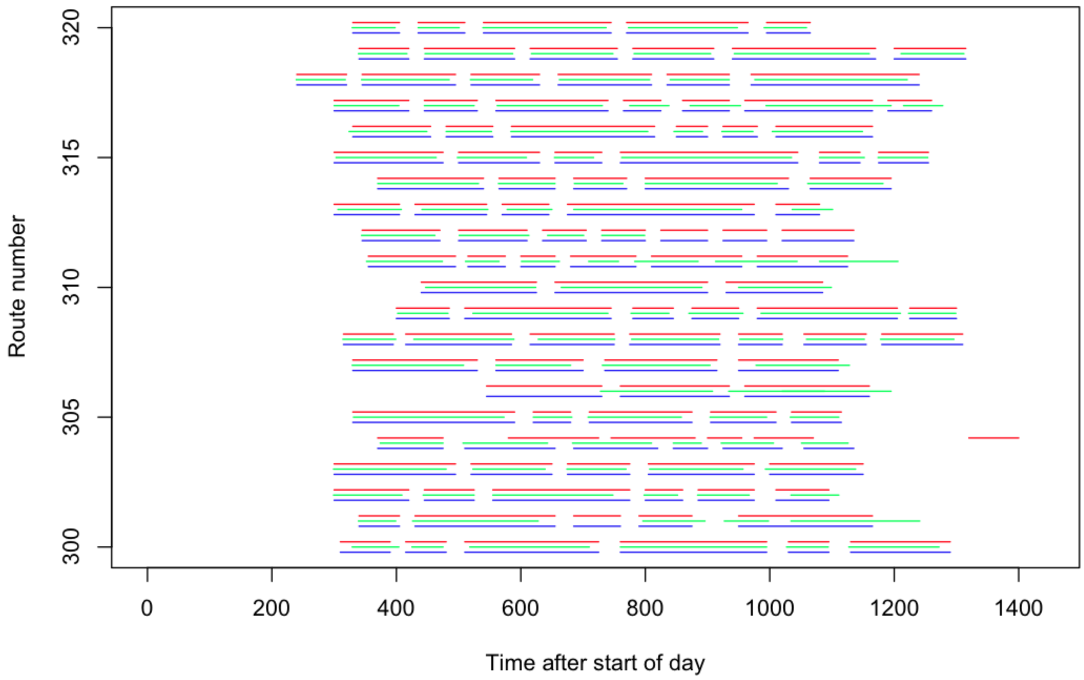

Figure 5: Planned time (blue), Actual time (green) and modelled time (red).

with an additional flight. This again shows that almost of all the model values are zero (flights depart exactly on time).

2.2 Statistical indicators

(2.6) We calculated the average time until the model and the real data diverge. For the four original data sheets (datasets 1-4) provided to the Study Group, each consisting of about 520 routes and 700 different aircrafts, the average mean time to the first divergence of the mimic and actual data sets is about 300 minutes.

(2.7) In all the available data we found that about 68% of all the first divergence points happen on decision points, and only 32% of the errors are made at non–decision point. A decision point is defined on the basis of the actual data and is said to be reached if:

- there is a swap,
- the scheduled arrival airport differs from the actual arrival airport,
- the scheduled and gate departure time differ from each other more than 15 minutes, or
- the scheduled and gate arrival time differ from each other more than 15 minutes.

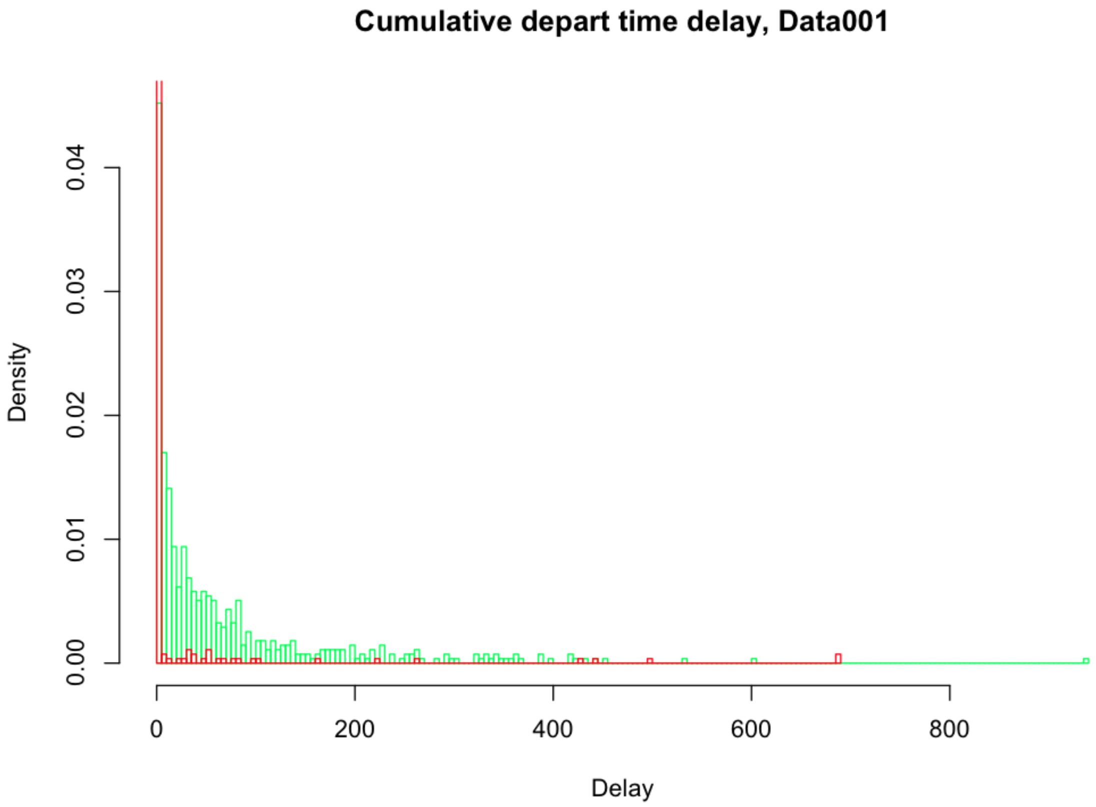

Figure 6: Showing the cumulative departure delay (Dataset 1). The actual data is green and the model is in red. Note that the first bar of the red (model) histogram, representing the flights with zero delay, is actually much higher (well off the scale of these axes).

(2.8) We have observed that in 36% of cases the model makes the same decision as is made in reality when it comes to a decision point, and the remaining 64% of cases it decides differently.

2.3 Parameter perturbations

(2.9) GE provided model runs (datasets 5-8) for the same day (same validation data) using versions of the model with different parameter values, which are meant to change the preference of the model for alternative strategies (such as weighting the preference for delay rather than cancellation, for instance). Analysis of this output (not shown here) demonstrated that the model is remarkably insensitive, even to large (order of magnitude) parameter variations. This suggests that the main component determining the performance of the model is not the variable parameters, but the structure of the model itself.

(2.10) However, between datasets 5-8 and datasets 1-4, we have found some differences in the time to first error (as in Figure 4). In particular, we notice that, in datasets 5-8, the mean time to first error is now about 170 minutes instead of the 300 minutes measured in datasets 1-4, see Section 2.2. This is also indicated in the distribution of the error time; compare Figure 7. Compared to the histogram of the original datasets 1-4, Figure 4, one observes that the data for the parameter perturbed datasets 5-8 is shifted to the left, which means that the error time tends to be earlier.

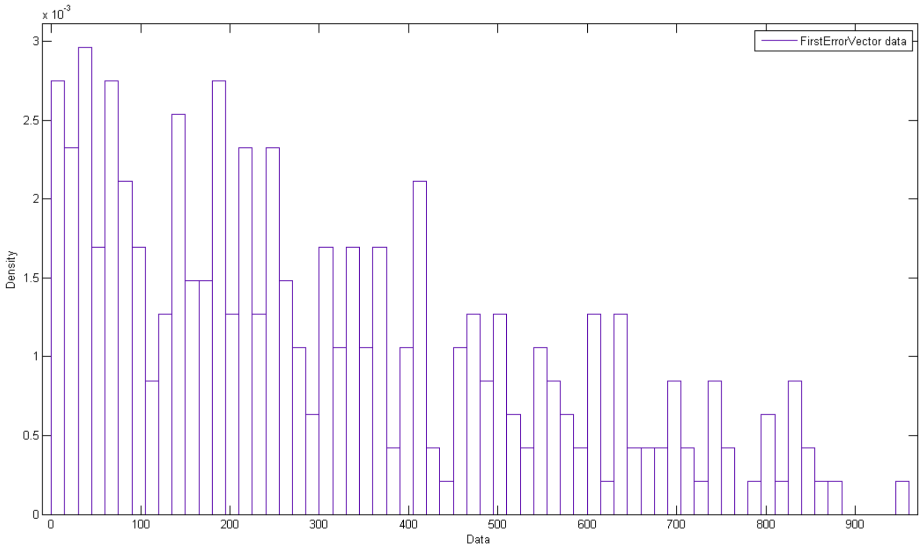

Figure 7: Error time distribution histogram for the perturbed model with extreme parameter values. The mean time to first error is now 170 minutes as opposed to 300 minutes for the unperturbed model version.

2.4 Some comments on the data

(2.11) Some "phantom flights" have been discovered, for instance aircraft 240 flying route 224 (in the first set of data).

(2.12) It has been suggested that the reason for the unexpectedly good performance of the model in terms of flight scheduling (almost always corresponding to the plan) may be due to the model taking account of future information, before it would have been available to the airline controllers. This was an attempt to incorporate the possibility of some kind of prognostics (weather forecasts, or routine maintenance issues) but appears to have been too optimistic. It would be useful to see the performance of the model when only real-time data is available, before this addition is considered. (It will then also be possible to compare the performance with and without prognostics.)

(2.13) The real data shows a large number of slightly early departures (0-10 min ahead of schedule). These could be included, perhaps in a stochastic way.

(2.14) Swaps of the aircrafts are much more common in reality than in the model:

001		Model swaps	
		Y	N
Real	Y	3	173
swaps	N	23	3237

(2.15) In reality, an incoming delay is often followed by another delayed flight for the same flight number; the series of delays can have 7 flights in a row, whereas for the model the longest series has 3 flights (data set 1). This is probably for the same reasons as the general underestimation of delays and cancellations, but it would be a useful indicator for the structure of the model error, in particular to distinguish between the policies of different airlines.

(2.16) We have some suspicions that the correction for time zone may not be quite right; it seems to look problematic in some data but not others. We would recommend a quick check.

(2.17) We would recommend using raw data as far as possible when evaluating the performance of the model. It was suggested that the data have been considerably tidied, partly for anonymisation and partly to remove areas where the model is not expected to do well. It would be more statistically rigorous to evaluate the full data set and then to consider separately the areas where the model is and is not expected to do well.

(2.18) Figure 8 shows an analysis of the possible modes of error in the model, including some possible ways of identifying and/or dealing with these errors.

3 Model development

3.1 Model development strategies

(3.1) Since the output of the "validation" is that the model is not a good representation of the decisions made by airlines' operation control centres, we then considered the opportunities for improving the fidelity of the model representation, taking into account limitations of both data and model.

(3.2) If no additional data are available, then the strategy should include

- Relaxing assumptions in the model structure;

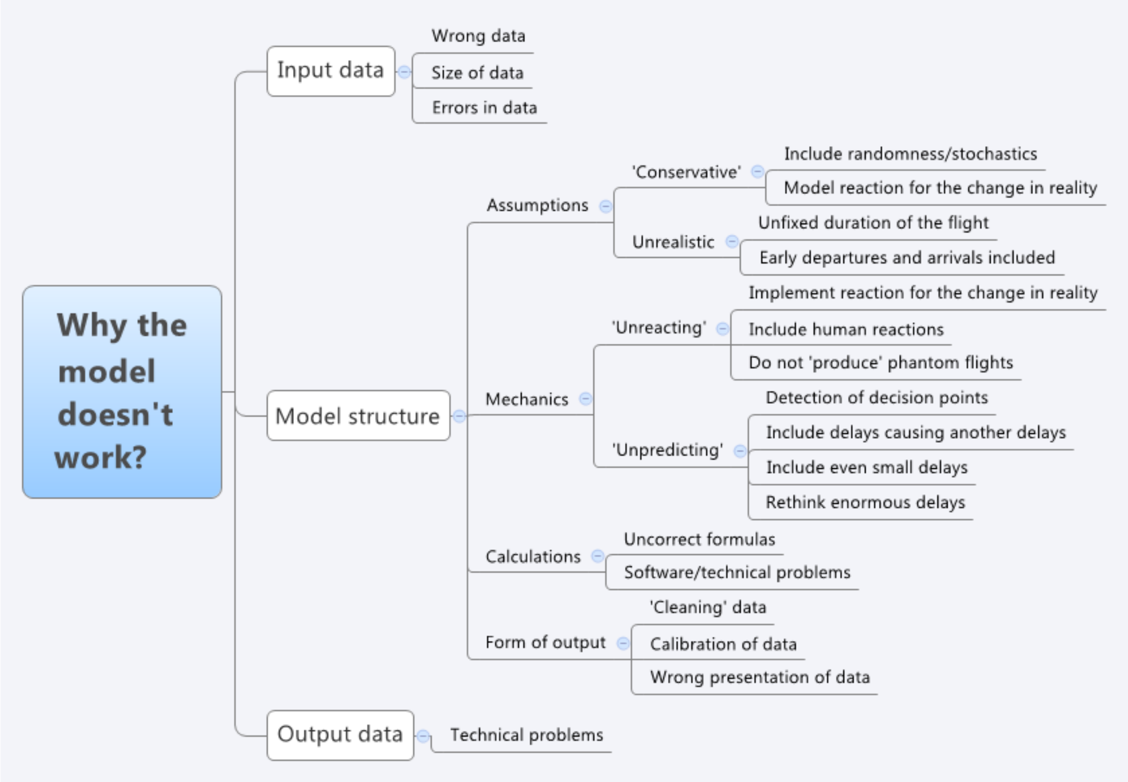
Why the model doesn't work?
Input data
Wrong data
Size of data
Errors in data
Model structure
Assumptions
'Conservative'
Include randomness/stochastics
Model reaction for the change in reality
Unrealistic
Unfixed duration of the flight
Early departures and arrivals included
Mechanics
'Unreacting'
Implement reaction for the change in reality
Include human reactions
Do not 'produce' phantom flights
'Unpredicting'
Detection of decision points
Include delays causing another delays
Include even small delays
Rethink enormous delays
Calculations
Uncorrect formulas
Software/technical problems
Form of output
'Cleaning' data
Calibration of data
Wrong presentation of data
Output data
Technical problems

Figure 8: Analysis tree showing possible reasons for error and ways to improve the model.

- Running the model more times with alternate parameters to try to fit the data better, using the visualisation tools described in Section 2.1 and statistical indicators (compare Section 2.2);
- Construction of a “benchmark” model based only on statistics of decisions (e.g., some fraction of decisions will be “cancellation”, and a stochastic function could be used to simulate the delay times. Then it could be conditioned on the time of occurrence or the existence of known weather, maintenance or other problems.)
- Re-designing the model to allow input of conditions and output of decisions part way through the day, so that individual decisions can be examined in greater detail.

(3.3) If there is the opportunity to collect additional data about the real system (human decisions in control centres), then further analysis would be possible:

- The most useful data would be a time-evolving flight plan, which records the updates and changes made throughout the day. If this were available then the decision-making capability of the model could be assessed by comparing like-with-like decisions made at the same time.

- It would also be useful to have more records of the reasons for making decisions. In some cases it is unknown what the delay or cancellation is due to. If this were available, it would allow the model to be assessed under each condition, to see which it is better at and therefore improve the representation of each section individually.

(3.4) If no additional data about reality can be obtained (recognising the limitations of airline confidentiality, etc), then the model can be developed by

- Testing the sensitivity to different initial conditions
- Testing the sensitivity to parameter variations (we have begun this process this week, and found remarkably little sensitivity)
- Testing the sensitivity to the original assumptions and model structure. Given the above observations about the lack of sensitivity in the model, this seems to be a good place to start.

(3.5) If both of the above can be obtained, then a direct comparison can be made and a statistically robust comparison carried out. Again, this would give much greater opportunity to understand the strengths and weaknesses of the model representation, and to tune parameters to the observed human decision factors. The first statistical indicator that would be useful is a simple percentage of decision points where the model makes the correct decision (based on the same input conditions, i.e., with the model run again after every decision point, so that it has the same history up to that point).

3.2 Alternative model structures

(3.6) Statistical model: We have suggested, slightly tangentially, the construction of a statistical model with no mechanistic representations, see for instance the point about mimicking early flight departures in (2.13). This would form a “benchmark” for the evaluation of the mechanistic model, against which it can be assessed. At present it is clear that a well-calibrated statistical model, even a very crude one, would be able to beat the current model on almost any performance metric.

(3.7) Currently it is assumed that decisions are made by a single entity in possession of all relevant information. In reality, this scheduling problem is too great for a single agent, and operations centres are distributed (usually geographically, see Figure 9).

(3.8) Propagation of information model: A model describing the rate and cost of information exchange could predict scheduling closer to existing organisations (e.g. poor scheduling of long haul aircraft, sub-optimal use of distributed maintenance resources etc.). Design will prove difficult owing to

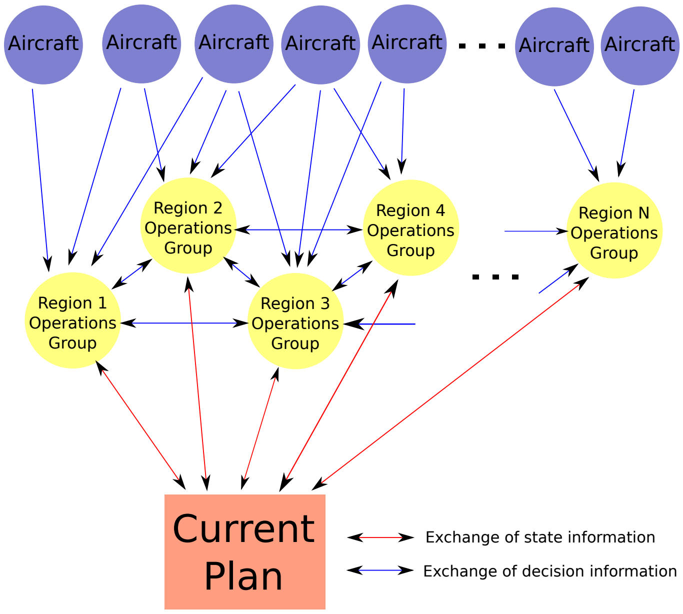
Aircraft
Aircraft
Aircraft
Aircraft
Aircraft
. . .
Aircraft
Aircraft
Region 2 Operations Group
Region 4 Operations Group
Region N Operations Group
Region 1 Operations Group
Region 3 Operations Group
. . .
Current Plan
Exchange of state information
Exchange of decision information

Figure 9: A basic model of the distribution of incomplete information

the complexity of the information, and would require further knowledge of the organisation's structure. An initial attempt could assign each agent an estimate of required maintenance time and urgency per aircraft and update it iteratively.

(3.9) Distributed decision making: Agents make decisions very quickly and in general these decisions are unlikely to reflect a full consideration of every other agent's optimal plan. This will likely lead to knock-on effects. Game theoretic models of collaborative decision making exist [2] but it may be of interest to examine ensembles of partially selfish agents.

4 Summary and recommendations

4.1 Overview

(4.1) In this report we have uncovered various inconsistencies and difficulties with the data sets provided.

(4.2) With the data we currently have we can:

- devise visualisation tools to see where model is performing well/badly;
- perform limited statistical descriptions of model performance;
- identify some inconsistencies in the data; and
- identify some inconsistencies in the model

(4.3) We have described above a wide range of possible strategies which would be possible with more data or greater access to the model itself (which GE's engineers of course have). This will allow GE to develop their model and to evaluate its progress.

Bibliography

[1] *www.masflight.com* Accessed 6 March 2013

[2] Yoav Shoham and Kevin Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations.* Cambridge University Press 2008.