

STATISTICAL ANALYSIS OF AN AIR TRAFFIC CONTROL PROBLEM

A R BRETNALL*, R C H CHENG*, A DREW**

**University of Southampton ,UK, **Eurocontrol, Bretigny, France*

Abstract: A discrete-event simulation has been built to evaluate effects of scheduling strategies on Air Traffic Control. This paper considers some practical issues in the application of standard statistical methods in validation and experimentation analysis on this type of simulation model. Using the model as an example we demonstrate how careful application of a variety of statistical methods including bootstrapping, hypothesis testing, goodness-of-fit techniques, Analysis of Variance, non-parametric methods and regression are needed to gain a solid understanding of the simulation models behaviour.

Keywords: Air Traffic Control, Simulation, Experimentation, Validation, Statistical Analysis

1 INTRODUCTION

This paper continues the story begun in [Brentnall et al, 2003] where a discrete-event simulation model built to evaluate effects of scheduling strategies on Air Traffic Control (ATC) was described. We concentrate on some statistical techniques that may be applied to models of this kind, demonstrating them on the ATC simulation.

Section 2 provides some background to the problem, motivating its importance. The discrete-event simulation model is described in Section 3. Section 4 shows how some standard statistical techniques may be used in the context of validation Question - Hypothesis - Test - Result procedures. The analysis of an incomplete 2-way experimental design raises a number of statistical issues that require care. These are outlined in section 5.

We hope that by outlining the application of some statistical procedures to our simulation the reader will agree that appropriately selected statistical methods are necessary to understand simulation models of this sort.

2 BACKGROUND TO PROBLEM

The primary purpose of Air Traffic Control (ATC) is to ensure that aircraft fly to their destination in a safe, orderly and expeditious manner. Demand on airspace has increased over the years and ATC has had to adapt in order to maintain a safe and efficient service. The Eurocontrol Experimental Centre is investigating the use of Arrival Manager (AMAN) decision-support tools to aid controllers with increasing traffic. These computer-driven tools advise on landing sequences and the control actions necessary to implement them for aircraft up to a specified distance away from an airport.

Traditionally Controllers have sequenced arrivals First-Come-First-Serve (FCFS). However, sequencing aircraft in a different order may help minimize delay or maximize use of runway. The problem of efficiently sequencing aircraft landings is not new, several authors have proposed techniques (a survey is included in [Beasley et al, 2000]) and some real-time systems already exist [Eurocontrol, 2000]. There is a contrast between the two - the majority of real systems reduce controller workload by generating advisories based on FCFS such that minimum separations between aircraft are respected, whereas published work proposes more complex sequencing techniques to work towards objective functions. Consequently there is a need to assess what effect using smart algorithms may have on the ATC system. When an aircraft landing sequence is changed from FCFS some aircraft will need to be delayed. The Control actions necessary to 'share' this delay are important because they may have knock-on effects elsewhere. The goal of this work was to develop an analysis tool to investigate the use of different scheduling and delay-sharing strategies when landing aircraft. Specifically, to assess how the ATC system reacts to changes in sequencing algorithms, optimization criteria and delay-sharing strategy.

3 THE SIMULATION MODEL

3.1 Model of airspace

A general model of airspace was developed for the ATC Sectors in a zone where sequencing may be applied (see Figure 1). The model ends at points all aircraft pass known as Intermediate approach fixes (IAFs). The simulation may be run using an actual day's flight plans or a random sample of all flight plans in the database. The process to generate traffic Sample follows 5 steps.

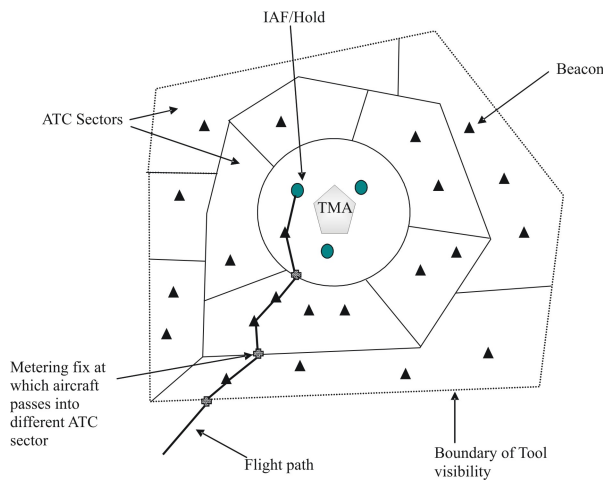


Figure 1: Schematic of the Arrivals Airspace

1. A non-stationary Poisson process generates an arrival sequence at the IAF's
2. Each of these times is assigned a IAF based on their probability (a multinomial model)
3. Each of the times and their IAF is assigned a plane type based on this probability (a multinomial model)
4. The database of flight plans is randomly sampled to get the right number of planes from each category
5. The flight plan for each of the sampled aircraft is changed so that the exit time is now the same as an exit time generated

Several algorithms and control mechanisms or delay-share strategies control the arrival process of aircraft during a simulation run.

4 VALIDATION ANALYSIS

4.1 Validation data

The model has been validated using three sources of data. The first is the flight plan data that aircraft file with authorities before they make their flight. A flight plan consists of fields such as wake-vortex category of aircraft, route and times of arrival. A second source of data comes from estimate messages sent over the network when aircraft are airborne. These messages contain estimated time over navigating beacons. The final source of data is radar track data. This data includes the 4-D position (longitude, latitude, altitude, time) of aircraft. The flight plan data essentially describes the underlying expected traffic situation, the estimate data the expected situation in around the next 30 minutes and the radar track data the actual situation. Sixteen radar track samples of varying lengths at different times of the day were gathered for arrivals to Stockholm Arlanda

in Autumn 2003. The basic validation idea is to feed the simulation model with input parameters fitted to the expected situation as observed in flight plan data and test if the model output tallies with performance indicators from estimate and radar track data. This may be done in a number of ways, some reported here.

4.2 Goodness-of-fit techniques

This section gives an example of hypothesis testing using goodness-of-fit techniques. When interested in validating the sampling procedure the basic question the hypothesis test tries to answer is:

Question Given the data sources above, is there enough evidence to reject the sampling procedure as a sufficiently accurate model of reality?

The first step in the process to generate a traffic sample is to generate a non-stationary Poisson process for the arrival sequence to the IAF's. To validate this we might ask:

Question Does a non-stationary Poisson process with rates that may change hourly accurately represent the arrival process of aircraft to the 4 IAF points (combined) at Stockholm Arlanda?

Hypothesis I In any time period of 1 hour during any day at Stockholm Arlanda arrivals to IAF points follow a Poisson process.

[Kelton and Law, 2000] point out that if a process is Poisson between times $[0, T]$ then the time of arrivals are distributed uniformly between $[0, T]$. So hypothesis I reduces to:

Hypothesis II In any time period of 1 hour during any day at Stockholm Arlanda the arrival times X are distributed Uniformly $[0, 1 \text{ hour}]$ i.e. $X \sim U[0, 1]$.

Test If there are K samples from the track data of arrivals at the 4 IAF points then we may wish to test simultaneously that all K samples are $U[0, 1]$. We can do this using K-Sample Empirical Distribution Function (EDF) Goodness-of-fit (GOF) statistics. Alternatively, since observations in the samples are independent we could pool them together into a single sample and test if this sample is distributed $U[0, 1]$. Again, GOF statistics may be used.

Result Table 1 shows the test scores for the K-Sample Cramer Von-Mises W_k^2 , K-Sample Anderson Darling A_k^2 test statistics, and for the pooled data the Anderson Darling (A^2), Kolmogorov-Smirnov (K) and Chi-Square (χ^2) tests (see [D'Agostino and Stephens, 1986]). There is not enough evidence to reject the null hypothesis at the 95% level using any of the tests.

Statistic	Value	p-value
W_k^2	2.46059	0.716*
A_k^2	15.0984	0.723*
A^2 (Pooled)	2.19645	0.0755*
χ^2 (Pooled)	17.7567	0.6034 (df=22)
K (Pooled)	0.0571	0.1111

*2000 Bootstraps for reference distribution

Table 1: Hypothesis test results for U[0,1]

Percentile	2.5%	5%	95%	97.5%
Param	-0.3178	-0.2716	0.2081	0.262

Table 2: Bootstrapped Empirical Percentiles of differences in mean delay

4.3 Bootstrap confidence intervals of differences

The bootstrap is a very powerful and versatile statistical technique. We have already used it in this paper to form reference distributions for test statistics above. The bootstrap often lends itself to simulation output that does not satisfy normality assumptions of parametric methods. It is used here in this context to find a confidence interval of difference in means between model and reality.

Hypothesis There is no difference between mean positive delay in the model set with maximum likelihood estimates of inputs and mean positive IAF delay from track data (IAF delay = difference between first EST IAF time and track data IAF time).

Test The hypothesis that model mean delay follows a normal distribution is rejected at the 95% level by the Shapiro-Wilks test for normality [D'Agostino and Stephens, 1986]. A nonparametric test for the difference that may be used is the Mann-Whitney test statistic Z [Hollander and Wolfe, 1973]. Using this procedure we obtain a p-value of 0.5175, so we cannot reject the null hypothesis of no difference. This information is limited and it would be more useful to have a confidence interval on the size of difference. Since the normality test failed we cannot build a t confidence interval. A relatively straightforward method to obtain a confidence interval without making any assumptions about the form of the two distributions is to bootstrap the difference in distribution means [Efron and Tibshirani, 1993]. Table 2 shows empirical percentiles of the test. Both the 90% and 95% cover 0. Again we do not reject the null hypothesis but now have 95% level of confidence that the true difference lies between [-0.3178, 0.262].

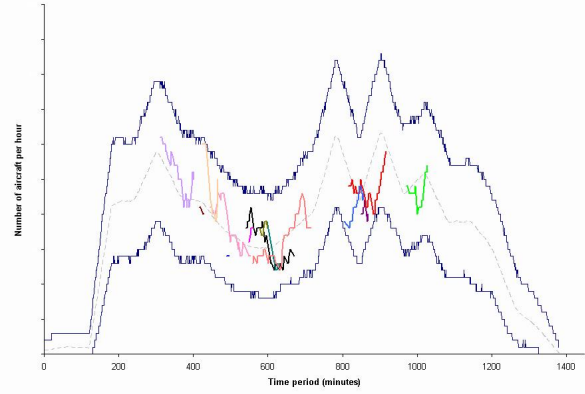


Figure 2: Simulation land rate mean and 95% range Vs track data samples

4.4 Confidence in performance indicators that change over time

Since landing rate is a function of time it will have a different distribution at each moment in time. This makes it more difficult to check if the model landing rate distribution matches track data. This section shows how graphical techniques may be used to get an idea of how the range of model behaviour behaves relative to actual data and how to set-up a hypothesis test for a difference.

Hypothesis There is no difference in landing rate distribution between the model and landing rate contained in the track data samples.

Test 1 Define landing rate $l(t)$ at time in minutes t to be the number of aircraft that landed since $t - 60$. Then compare a plot of landing rates from track data with a 95% range of landing rates from 500 runs of the model. This gives an idea of the range of landing rates we can expect from the model, compared to real data.

Result Visual inspection of Figure 2 suggests that the simulation model data contains and follows the behaviour of the track data land rate quite well.

Test 2 At each time point in Figure 2 the model forms a reference distribution of landing rate from its independent runs. If the distribution of landing rate was not time dependent then we could pool together independent landing rates and carry out GOF tests to compare this distribution with the model distribution. In this case let Y represent the set of observed landing rates and X the set of model landing rates. One way to test if the two distributions are the same would be to test if the set $\{Pr(y \leq X) : \forall y \in Y\}$ is distributed Uniformly on [0,1]. The same idea may be applied to test whether there is no difference between the time dependent model distributions and the actual situation. Changing the distribution of X based on the time period

Statistic	Value	p-value
K	0.1287	0.8147
A^2	0.25635	0.9665*
χ^2	3.4545 (df=6)	0.75

*2000 Bootstraps for reference distribution

Table 3: Test results for landing rate hypothesis test

t at which the observation $y \in Y$ was made and test if $\{Pr(y_t \leq X_t) : \forall y_t \in Y_t, t \in T\} \sim U[0, 1]$. $U[0, 1]$ may be tested using for instance EDF or χ^2 statistics.

Result It was possible to form 22 independent samples (length 1 hour) of landing rates from the track data. These were compared to the 22 different reference distributions at the corresponding time points. The test statistics of the test $\{Pr(y_t \leq X_t) : \forall y_t \in Y_t, t \in T\} \sim U[0, 1]$ are reproduced in Table 3. Based on these we do not reject the hypothesis that there is no difference in landing rate distribution between the model and landing rate contained in the track data samples.

4.5 Sensitivity analysis using regression

Sensitivity analysis is the study of what effect change to input parameters has on output performance indicators. Validation of sensitivity looks to see if change in input parameters produces similar changes in both the model output and real life. Presented here is use of linear regression to quantify the relationship between the two.

Question Does change in input arrival rate have the same effect on landing rate in the model as in real life?

Hypothesis There is a positive correlation between model and actual landing rates in different time periods.

Test Fit a regression line to points (v_i, w_i) paired by time where v_i is mean landing rate in the model and w_i landing rate sample i from track data.

Result The points used were the 22 independent sample landing rates paired with mean landing rate from model. A fit of the one-way linear model $w_i = u + v_i + \epsilon_i$ yields the following:

Coefficient	Value	$Pr(> t)$
u	4.4055	0.2688
v_i	0.7649	0.0006

A Shapiro-Wilks normality test of ϵ_i has $W = 0.9732$ with a p-value = 0.7836 so there is not enough evidence to reject the normality assumption.

	α_1	α_2	α_3	α_4	α_5	α_6
β_1	X				X	X
β_2	X	X	X	X	X	
β_3	X	X	X	X	X	
β_4	X	X	X	X	X	

Table 4: An incomplete two-way experimental design

The model mean and actual landing rate samples are positively correlated: the t-test tells us that we reject the hypothesis that the simulation mean land rate has no effect on the observed land rate at the 99% level. Also, the estimate of v_i is close to 1, so a unit change in mean landing rate in the model almost corresponds to a unit change in real-life data. This is an excellent result when you consider the variability observed in track data landing rate shown in Figure 2.

5 EXPERIMENTATION ANALYSIS

We present here some issues in applying Analysis of Variance and estimation procedures to an experimental design for a simulation model such as ours.

5.1 An incomplete two-way design

[Brentnall et al, 2003] reported some analysis on a complete two-way experimental design. They asked if there was any difference between algorithms at different levels of traffic sample intensity. Another two-way experimental design has also been run on the simulation model to examine the effect of changing algorithm and delay-share strategy when other input parameters are set to those used in the validation procedures. An incomplete two-way design was forced due to the nature of algorithm-delay share combinations. The design is shown in Table 4, where α_i is algorithm i and β_j delay-share strategy j and X marks a combination that may be run in the simulation.

5.2 Analysis: ANOVA and Estimation

5.2.1 Setting up the analysis

Given a performance indicator Y we are interested in experimenting with a form of output model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_k + \epsilon_{ijk} \quad (1)$$

where μ is an overall mean, α_i the effect of algorithm i , β_j the effect of delay-share strategy j , $(\alpha\beta)_{ij}$ their interaction, γ_k the effect of traffic sample k and ϵ_{ijk} a random error. Use of Analysis of variance (ANOVA) and the F-test is a standard way to test for significance of individual terms in the model. Nonparametric tests have also been developed to test for significance of a wide variety of models where normality of ϵ_{ijk} breaks down [Hollander and Wolfe, 1973]. If F-tests or their nonparametric cousins are significant one would really

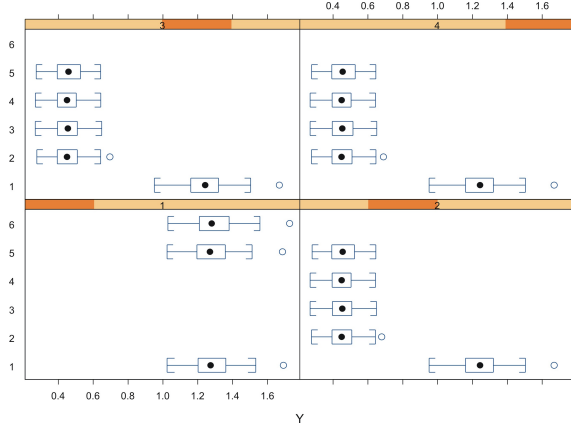


Figure 3: Mean delay output boxplots, split by algorithms 1 - 6 and delay-share strategies 1 - 4

like to know more about the differences. Estimation of the coefficients, or fitting a metamodel/response surface are standard ways to proceed.

The common problem in estimation of ANOVA type models is overparameterization: there are more parameters to estimate than there can be independent normal equations to solve. This means that there will be an infinite number of least squares solutions unless extra restrictions are added. Careful thought on the form of restriction is needed to make sure estimates are useful and understandable with respect to the objectives of experiment [Draper and Smith, 1998].

If we were to run a regression analysis using a standard statistical package then the usual restrictions put on the model would be to set $\sum \alpha_i = \sum \beta_j = \sum \gamma_k = 0$. This form of constraint is readily interpretable for balanced, complete designs. A general alternative to this is to set $\alpha_1 = \beta_1 = \gamma_1 = 0$. In this case each coefficient estimate compares with level 1 of α , β and γ . This makes good sense for interpretation of an experiment run to compare new treatment combinations with a base. This is true for our experiment where we are interested in comparing the base design point (α_1, β_1) used in validation procedures with different treatment combinations.

5.2.2 Building models

Once an approach to estimation has been decided upon the experimenter follows a well-trodden path in trying out various models to see if effects are significant, and checking assumptions underlying the model. Presented here is the process for building a model for mean delay from our simulation.

Figure 3 shows boxplots of mean delay split by algorithm and delay-share strategy. This visualization is quite useful in getting an idea of general trends. An

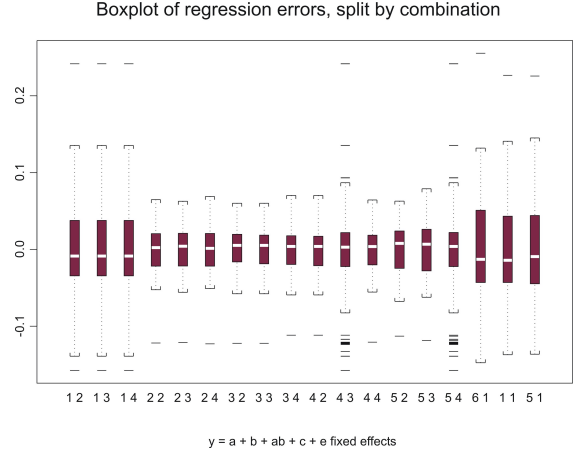


Figure 4: Residuals from fit of model (1), split by algorithm α_i and delay-share strategy β_j

ANOVA decomposition of the full model (1) found all terms were significant. However, to satisfy assumptions behind ANOVA and least-squares estimation, errors should be identically distributed in each of the α - β combinations. A box plot of the residuals from the least-squares fitted model is shown in Figure 4. One can see that they are not identically distributed. It turned out that removing the γ_k term from (1) produced the best distribution of model residuals. However, they still appeared non-normal.

A common method to deal with this situation is to fit a model to a transformation of the response such as one of the Box-Cox family [Draper and Smith, 1998]. In our example a 95% confidence interval of the maximum likelihood estimate of the Box-Cox parameter λ that specifies the transformation included the easy-to-understand square root. A QQ-Normal plot of the studentized residuals showed significant improvement and subjective acceptance of the normality assumption. The final fitted model was thus:

$$\sqrt{Y_{ijk}} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad (2)$$

Statistically significant estimated coefficients are shown in Table 5. Interpretation of these is carried out with one eye on the matrix of correlation between regressors. In particular interactions of $(\alpha\beta)_{5j}$ are correlated with α_5 and β_j . The sketch of mean delay in Figure 3 has been quantified by a routine that simultaneously takes all experimental design points into account. Based on the chosen approach to estimation, model term significance tests and size of coefficients the model is interpreted. The simulation experiment suggests that operating on traffic levels in Autumn 2003 at Stockholm Arlanda algorithms operating with delay share routines $\beta_2, \beta_3, \beta_4$ result in less mean delay than the current set up, less mean delay than combinations of the current algorithm with different delay-share procedures and less mean delay

Coefficient	Value	Pr(> t)
μ	1.1333	0.0000
α_2	-0.4446	0.0000
α_3	-0.4429	0.0000
α_4	-0.4481	0.0000
$(\alpha\beta)_{52}$	-0.4397	0.0000
$(\alpha\beta)_{53}$	-0.4373	0.0000
$(\alpha\beta)_{54}$	-0.4379	0.0000

Table 5: Significant estimated mean delay coefficients

than different algorithms with the current delay-share procedure. For instance, if algorithm α_2 is used with delay-share strategy β_2 then the model mean delay $Y = (1.133 - 0.4446)^2 = 0.4739$ minutes. Also, \sqrt{Y} has standard error = 0.00864, so Y has an estimate 97.5% confidence interval of [0.2305 0.8039]. This compares to the base mean $Y = (1.133)^2 = 1.2837$ minutes, and 97.5% confidence interval of [0.8552 1.7988]. We can use the Bonferonni method to say that the simultaneous confidence interval is at the 95% level. This difference is due to effect of algorithm and delay-share strategy.

6 CONCLUSION

Computer simulation models often present statisticians with the chance to forget problems such as missing data and outliers by watching their data flow out of a tap. However, the practicality of conducting useful analysis still necessitates careful use of standard methods such as ANOVA and regression. This paper has presented a number of statistical methods applied to an ATC simulation model. By demonstrating the value careful use of these techniques adds to a simulation project we hope the reader will agree with the moral to the story: appropriately selected statistical methods are needed to understand simulation models of this sort.

REFERENCES

- Beasley J., Krishnamoorthy M., Sharaiha Y. M. and Abramson D. 2000 "Scheduling Aircraft Landings - The Static Case" In *Transportation Science* 34, Pp 180-197
- Brentnall A.R., Cheng R.C.H, Drew A. and Potts C.N. "The Air Traffic Control Arrival Management problem" In *Proc. UKSIM 2003* (Cambridge, UK, April 2003), Pp 121-127
- D'Agostino R.B. and Stephens M.A. 1986 "Goodness-of-fit tests" *Marcel Dekker* New York
- Draper N.R. and Smith H. 1998 "Applied Regression Analysis" *John Wiley & Sons* New York
- Efron B. and Tibshirani R.J. 1993 "An Introduction to the bootstrap" *Chapman & Hall* New York

Eurocontrol 2000 "AMAN Feasibility Study (part 1 and 2)", *Eurocontrol*, Paris, France

Hollander M. and Wolfe D.A. 1973 "Nonparametric statistical methods" *John Wiley & Sons* New York

Kelton W.D., Law A.M. 2000 "Simulation modeling and analysis" *McGraw-Hill* Singapore

AUTHOR BIOGRAPHIES



ADAM ROBERT BRENTNALL is a Research Student in Operations Research at the University of Southampton. His MMath Mathematics degree from The University of Sheffield included one year at The University of North Texas, USA. He obtained an Msc in Operations Research from The University of Southampton in 2002, and is a member of The Operational Research Society. His email address is <a.r.brentnall@maths.soton.ac.uk>.



RUSSELL C. H. CHENG is Professor, Head of Operational Research, and Deputy Dean of the Faculty of Mathematical Studies at the University of Southampton. He has an M.A. and the Diploma in Mathematical Statistics from Cambridge University, England. He obtained his Ph.D. from Bath University. He is a former Chairman of the U.K. Simulation Society, a Fellow of the Royal Statistical Society, and Member of the Operational Research Society. His research interests include: variance reduction methods and parametric estimation methods. He was a Joint Editor of the IMA Journal of Management Mathematics. His email and web addresses are <r.c.h.cheng@maths.soton.ac.uk> and <www.maths.soton.ac.uk/staff/Cheng>.