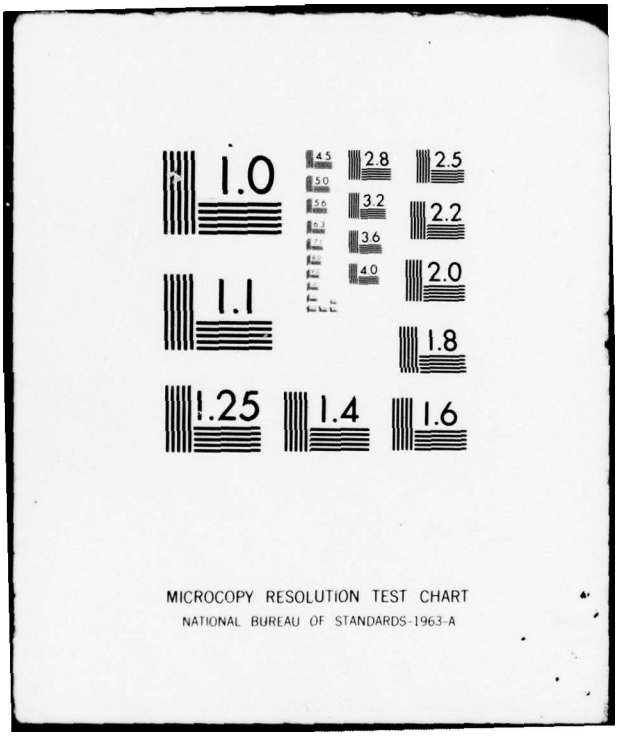
AD-A056 201 DUNLAY (WILLIAM J) JR BALA CYNWYD PA F/G 1/5
AIRPORT IMPROVEMENT TASK FORCE DELAY STUDY: DATA COLLECTION, RE--ETC(U)
NOV 77 W J DUNLAY W1-77-2412-1

UNCLASSIFIED FAA-EM-78-7 NL

| OF |
AD
A056201

END
DATE
FILMED
8 -78
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

LEVEL II

REPORT NO. FAA-EM-78-7

12

AIRPORT IMPROVEMENT TASK FORCE DELAY STUDY: DATA COLLECTION, REDUCTION AND ANALYSIS

NOVEMBER 1977

Prepared for

U.S. DEPARTMENT OF TRANSPORTATION
FEDERAL AVIATION ADMINISTRATION
Office of Systems Engineering Management
Washington, D.C. 20591

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FAA-EM-78-7		

4. Title and Subtitle	5. Report Date
AIRPORT IMPROVEMENT TASK FORCE DELAY STUDY: DATA COLLECTION, REDUCTION, AND ANALYSIS	22 November 1977
	6. Performing Organization Code
7. Author(s) William J. Dunlay, Jr.	8. Performing Organization Report No. 53 P.
9. Performing Organization Name and Address William J. Dunlay, Jr. 168 N. Latch's Lane Bala Cynwyd, PA 19104	10. Work Unit No. (TRAIS) 1
	11. Contract or Grant No. Order W1-77-2412-1
12. Sponsoring Agency Name and Address U. S. Department of Transportation Federal Aviation Administration Office of Systems Engineering Management Washington, DC 20591	13. Type of Report and Period Covered Final Report 18 August to 22 November 1977
	14. Sponsoring Agency Code FAA/AEM-100 (77-56)

15. Supplementary Notes

16. Abstract

A plan is presented for the collection, reduction, and analysis of data in support of the validation of an airside delay simulation model. Data collection forms are presented along with the forms on which data are reduced and finally merged from several sources into a single summary format by aircraft. Detailed suggestions are given for the statistical analysis of model estimates and for a time series analysis of these estimates vis-a-vis corresponding observed data. Included are guidelines for interpreting the results of the statistical tests.

17. Key Words	18. Distribution Statement
Airport Capacity and Delay Statistical Analysis Computer Simulation Model Validation	Document may be released to National Technical Information Service, Springfield, VA 22161 for sale to the public.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
UNCLASSIFIED	UNCLASSIFIED	46	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

ACKNOWLEDGEMENT

The writer gratefully acknowledges the significant contributions of the following individuals to the ideas of this plan:

(1) Dr. D. A. Hsu, University of Wisconsin, Milwaukee, who contributed ideas to and reviewed the statistical analysis section of this report.

(2) John R. VanderVeer, Anthony Bradley, Robert Holladay, LTC. John Hartnett, and Jacques Press of the National Aviation Facilities Experimental Center who provided data forms and details of the data collection and reduction.

(3) Ray H. Fowler, FAA/AEM-100, for his continuous support and guidance as Technical Officer of this contract.

(4) Drs. Andrew L. Haines of MITRE and Amedeo Odoni of MIT for their review of this plan.

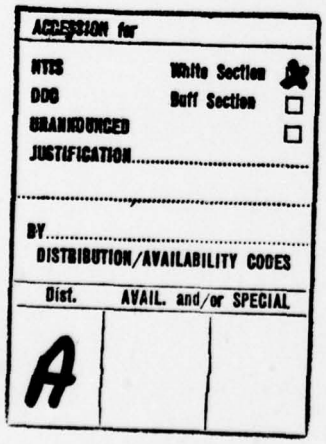


TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DATA COLLECTION, REDUCTION, AND ANALYSIS PLAN

by

William J. Dunlay, Jr.

I. INTRODUCTION

Background

This is the second report by the writer under the general title: Airport Improvement Task Force Delay Study. The first was the Delay Model Validation Plan, dated August 18, 1977. This second plan was prepared under supply contract No. W1-77-2412-1 with the Office of Systems Engineering Management of the U. S. Federal Aviation Administration.

The validation effort to which this plan is addressed is part of Phase I of contract No. DOT-FA77WA-3961 between the U. S. Federal Aviation Administration and Peat, Marwick, Mitchell and Company (PMM & Co.). The objective of the validation is to test whether the PMM & Co. delay simulation model is satisfactory (to the Technical Officer) for its intended application in Phase II of the contract, namely for delay estimation in support of six Airport Improvement Task Forces.

Purpose

The Delay Model Validation Plan presented an overall outline of the validation procedure. This second plan describes the approach taken to the collection of data for the validation (both for model input and comparisons with model output), the reduction of those data, and the statistical comparisons of model estimates and corresponding observed quantities.

Scope

This plan is process-oriented, that is to say, it does not present results of data collection, reduction or analysis; rather, it describes the methodology followed in these three processes.

The major emphasis of this plan is on the statistical analysis of the data and the hypothesis testing associated with comparing model estimates with collected data. The data collection and reduction steps are covered in less detail.

The remainder of this report is organized into three major sections: (1) data collection, (2) data reduction, and (3) statistical analysis of the data.

II. DATA COLLECTION

Description of Sample

Due to time and manpower constraints, only ten (10) days of data were collected. The first of these was for training the data takers. Thus, there were 9 days of useful data.

Data were collected on the following days:

(1) Monday, 20 June 1977 - training

(2) Tuesday, 21 June 1977 - Friday, 24 June 1977, inclusive

(3) Wednesday, 27 June 1977 - Friday, 1 July 1977, inclusive

On each day, data were collected during the periods 8:00 a.m. - 11:00 a.m. and 1:00 p.m. - 4:00 p.m., all local times, i.e., CDT. These periods correspond to 13:00 - 16:00 and 18:00 - 21:00 GMT, respectively.

Runway Configurations

Chicago's O'Hare International Airport has a total of twelve major runway ends (plus a couple of minor ones) as shown in Fig. 1. The six major runways are arranged in three sets of parallels: (9R/27L, 9L/25R); (4L/22R, 4R/22L); and (14L/32R, 14R/32L). Although there are many possible runway-use combina-

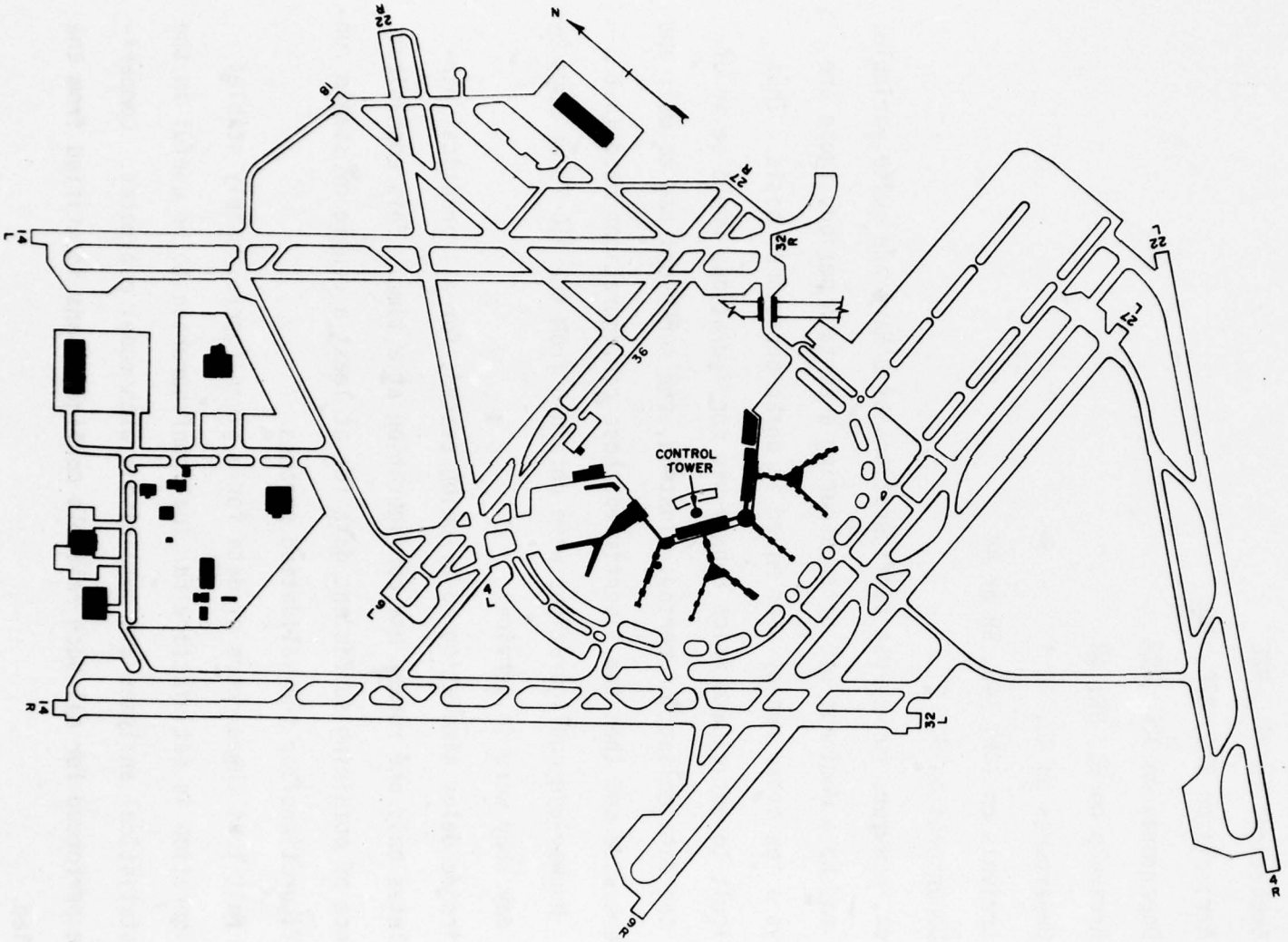


Fig. 1 - O'Hare Runway Configurations

tions, there are a number of preferred ones, including (but not limited to):

(1) Arrivals on 14R, 22R, 22L

Departures on 27L, 22L

(2) Arrivals on 27L, 27R, 32L

Departures on 32L, 32R

(3) Arrivals on 9L, 9R, 4R

Departures on 4L, 4R

(4) Arrivals on 14R, 14L, 9R or 4R

Departures on 9L, 27L

In fact, however, runway-use configurations at O'Hare are quite variable. While there may be a dominant use pattern during a certain period, there are nearly always a few other runway uses mixed in on an occasional basis. Thus it was difficult to anticipate which runway-use configurations would be in effect during the data collection period. Instead, the surveyors had to wait and see which were used and then they identified those configurations used most frequently. Runway-use configurations were obtained from the ATC Tower Logs by the time of day they were in service.

The airside delay simulation model being tested, for all practical purposes, simulates only one runway-use configuration at a time. This emphasizes the importance of obtaining sufficient data for at least a couple of stable runway-use configurations for the validation process.

It is felt that three hours of data for a given (approximately stable) runway configuration is satisfactory for that configurarion to be useful in the subsequent statistical analyses and comparisons with model estimates. Comparisons will be performed for all such reliable configurations identified from the data collected.

Data Sources and Collection Techniques

There were ten distinct sources of data utilized in the data collection

effort:

(1) field observation

(2) airline data

(3) ATC voice tapes

(4) ATC departure strips

(5) ARTS-III tapes

(6) ATC tower personnel

(7) existing PMM & Co. data

(8) Official Airline Guide

(9) ASDE films

(10) ADR data

Each of the first eight sources was a primary source for at least one major piece of data; the last two served as back-up data for cross checking and for matching together ambiguous data items.

Field Observation. The following variables were observed directly in the field:

(1) aircraft identifications (arrivals & departures)

(2) lift-off times

(3) roll times

(4) intersection times (arrivals & departures)

(5) runway exit used and exit times (arrivals)

(6) departure-queue length

Several data collection forms were developed at NAFEC. Figure 2 shows a two-part form designed for recording flight details for both arrivals and departures. The upper half of the form was used to record the aircraft identification, runway used, intersection time (if any), exit time, and exit used by each arrival. The bottom half of the form is for recording data for each departure: identification, runway used, roll time, intersection time (if any),

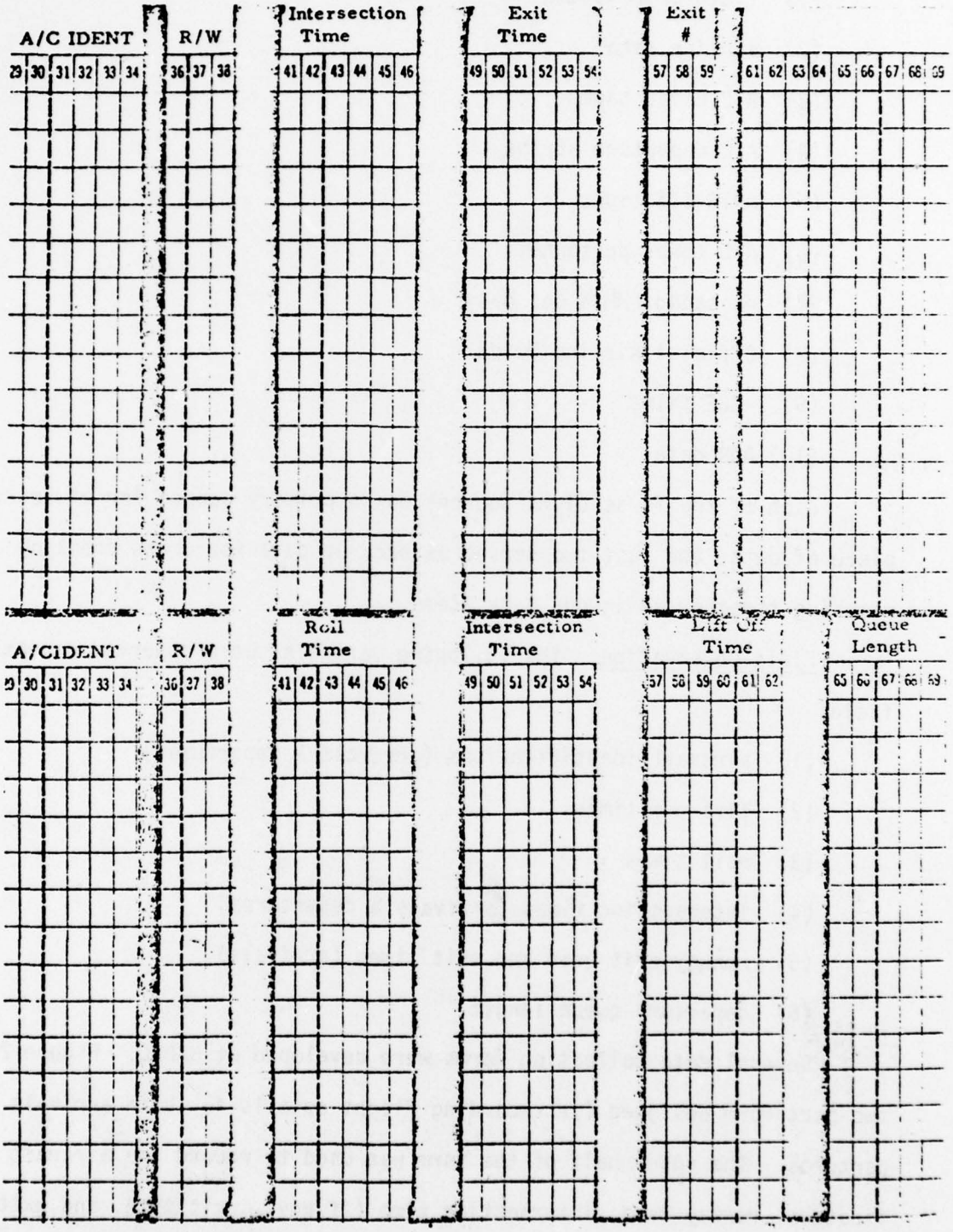


Fig. 2 - Data Form for Arrivals and Departures.

and lift-off time.

Figure 3 is designed for recording data on departure queue lengths. On this form the times that departures enter and leave each queue were recorded. A plus (+) sign beside the aircraft identification indicates an aircraft entering the departure queue; a minus (-) sign indicates one leaving, i.e., taking off. There is also space to record the runway number and queue length. The forms of Fig. 3 were filled out in the FAA air traffic control tower.

The third field collection form, Fig. 4, is for recording penalty-box delays and runway crossings. These were observed from the airline ramp control tower at O'Hare.

It was not possible to obtain a 100% sample of aircraft. Close to a complete record was attempted on the arrival/departure form of Fig. 2. The data on the second and third forms, however, were recorded for relatively small samples.

For field data collection purposes, the airport was divided roughly into two areas, either a so-called N-S division or an E-W division depending on the runway configuration in use. These correspond roughly to the areas under each local controller in the O'Hare tower. Two persons were assigned to each area, usually one for arrivals and one for departures or, alternatively, one per runway depending on the traffic situation.

Equipment for field data collection include clip boards, pencils, and blank data forms. Aircraft identifications were overheard in the ATC tower. A digital clock was used as a source of time data.

Airline Data. The following data were obtained directly from the airlines:

arrivals: (1) taxi-in times
(2) gate-arrival times

departures: (1) gate-push-back times

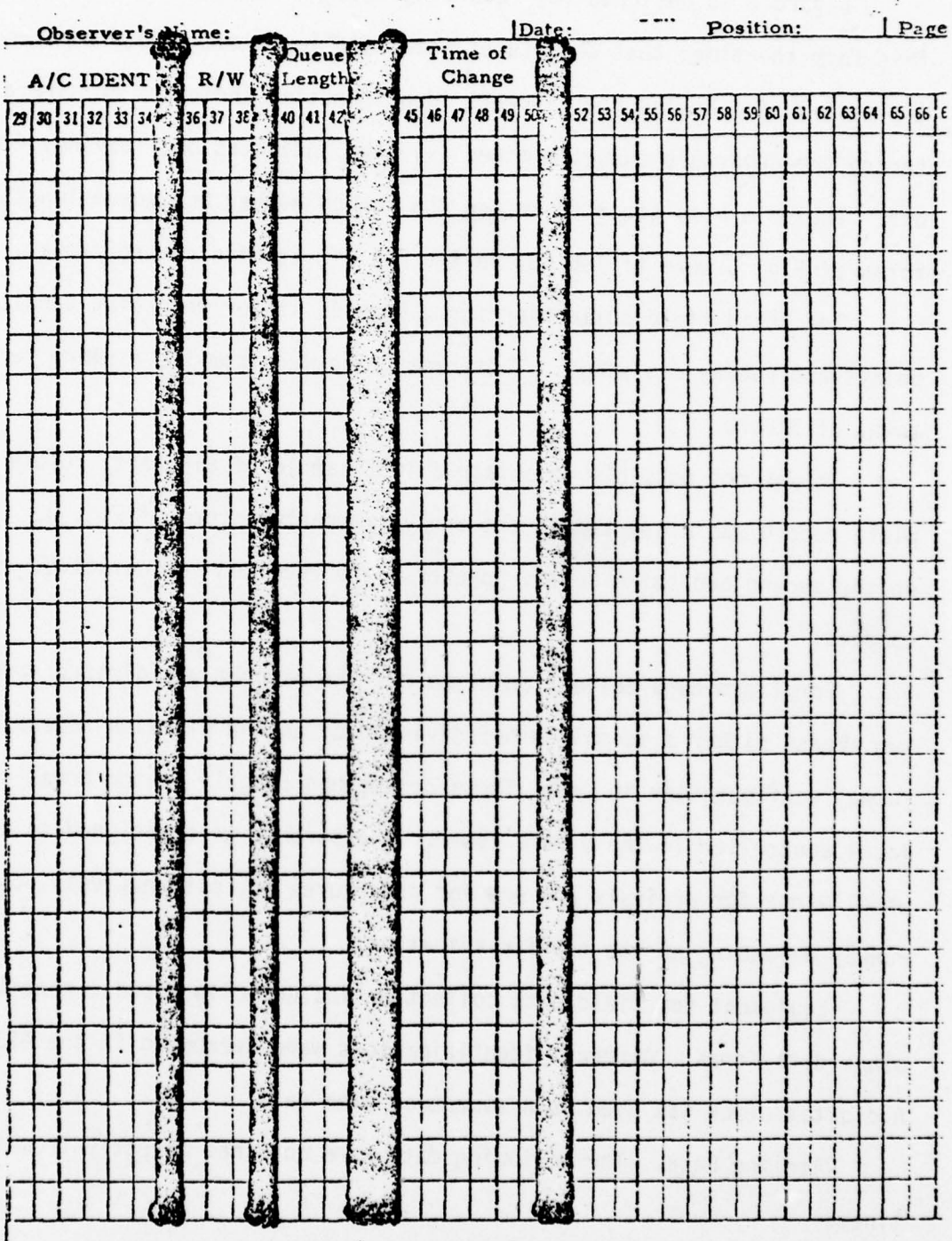


Fig. 3 - Departure Queue Length Data Form.

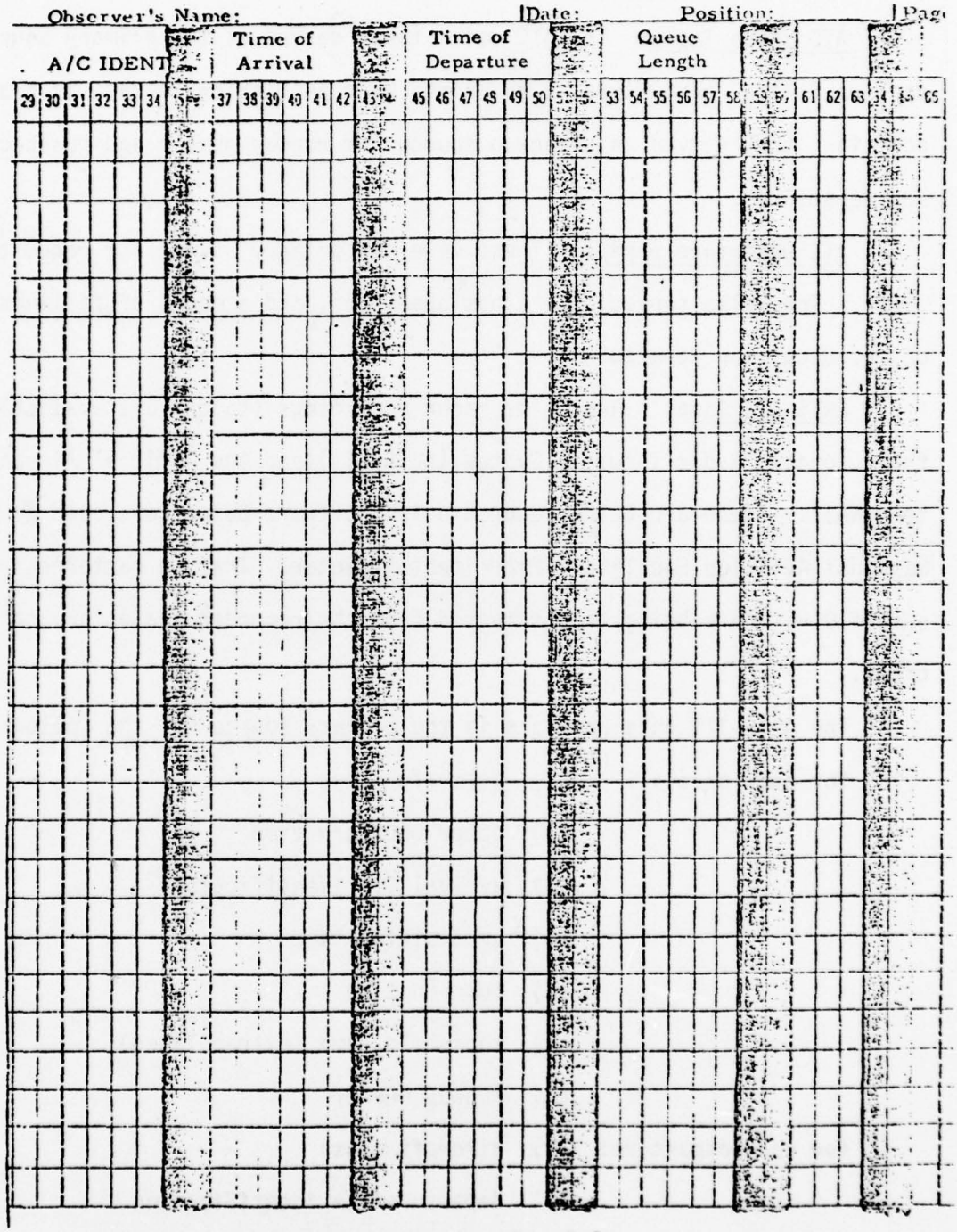


Fig. 4 - Data Form for Penalty-Box Delay.

(2) taxi-out times

Coded forms were provided by the airlines, and the airline data effort was coordinated through the Air Transport Association.

ATC Voice Tapes. The ATC voice tapes served as the primary source of data on initial taxi times, i.e., times when permission to taxi is granted. In addition, they served as a backup source for runway number and pushback times for departures.

ATC Departure Strips. These were the primary source for request-for-taxi times. They also served as a cross check for, and a means of fillings gaps in, observed data on departures.

ARTS-III Tapes. Real-time radar recordings (tapes) are avaliable from the Automated Radar Terminal System (ARTS-III) and the National Airspace System (NAS). These are being reduced using software being developed at the National Aviation Facilities Experimental Center. Traffic patterns into and out of O'Hare are being reconstructed from the recorded trajectory of each aircraft.

The ARTS-III data analysis is the primary source for the following:

for each arrival:
(1) aircraft class
(2) outer boundary time
(3) arrival fix identification
(4) arrival-fix time
(5) turn-on time
(6) threshold time (extrapolated)
(7) runway number

for each departure:
(1) aircraft class
(2) departure fix identification
(3) departure-fix time

In addition, the ARTS-III data serves as the basis for calculating sched-

uled threshold times for arrivals using nominal, undelayed flying times (from arrival fix to threshold) deduced from ARTS-III trajectories in low-activity periods. The ARTS-III data is also used as the basis for deducing various model inputs such as aircraft approach-speed distributions and distributions of minimum separations.

ATC Tower Personnel. From discussions with ATC tower personnel, and observations made in the tower, the following data items were defined or checked:

(1) lengths of common approach path for each runway, each aircraft class, and different levels of activity

(2) locations of holding stacks

(3) gate-hold locations and procedures (if any)

(4) from ATC tower logs, data on runway-use configurations and hourly traffic counts (by arrival vs. departure and type of flight)

(5) locations of runway-crossing problems

(6) taxiway routes and two-way paths

(7) departure runway reassignment procedures.

Existing PMM & Co. Data. The following quantities were defined from existing PMM & Co. data for O'Hare:

(1) runway exit utilization - distributions

(2) standard taxiway speeds by location

(3) link data

(4) runway-exit distances

The above data were checked against corresponding data from other sources.

Official Airline Guide. The Official Airline Guide (OAG) was the primary source of scheduled departure times.

ASDE Films. O'Hare International Airport has airport surface detection equipment (ASDE) radar. Films were taken of the ASDE scope during the data-collection periods. These were used in conjunction with the voice tapes to in-

vestigate departure queue behavior, as a primary source for penalty-box delays and locations, and to fill in gaps in the observed data. Thus, ASDE served as a secondary data source for cross-checking and matching ambiguous observed data.

ADR Data. Data from the Aircraft Delay Report (ADR) were obtained for the data-collection period. These served as additional backup for cross-checking data from other sources and for filling gaps.

III. DATA REDUCTION

Most of the data reduction activities are taking place at the National Aviation Facilities Experimental Center (NAFEC). The objective of the data reduction is to obtain reduced data in a format that facilitates the computation of required model inputs and level-of-service measures that are comparable with model outputs.

Approximately ten computer programs are being or have been developed by personnel at NAFEC.* These programs are described below:

(1) COMP - compares arrival and departure records and reduces them to a format similar to the PMM & Co. model output. Outputs from this program comprise the arrival queue data set.

(2) CONVERT - converts data on arrivals into a queue data set, i.e., the time joining queue and time leaving queue in selected time intervals.

(3) QSUM - computes average queue length, maximum and minimum queue lengths, and time in queue for both arrivals and departures, and by runway, using data broken down by 5-minute intervals from either the model or the field data collection.

(4) HISTO - constructs and prints histograms of taxi-in and taxi-out

*These programs were developed by Anthony Bradley, Robert P. Holladay, and Jacques Press of ANA-220.

times for the airline data set.

(5) TAXI - using the taxiway route structure as input, this program computes expected values of taxi-in and taxi-out times from estimated probabilities of runway-exit and gate utilization.

(6) ARTS III - a program made up of four steps which, when run successively, reduce data found on the Automated Radar Terminal System (ARTS III) extractor tapes collected at the TRACON.

(a) The first step converts the original seven-track tapes to nine-track format.

(b) The second step filters out and collects radar track messages for arrival and departure tracks for the airport being measured. Specific portions of the airspace may be specified as desired.

(c) The third step constructs a radar-track history, i.e., the trajectories flown by the aircraft across selected portions (areas) of the airspace.

(d) The fourth step analyzes the time histories created in step 3 and determines crossing times at key points in the airport/airspace system including: (1) the outer ring (about 45 miles out), (2) the arrival fix, (3) the common-approach point, and (4) the runway threshold. Travel times are also computed. The output is a record summary for each aircraft.

(7) MERGE - reduces and synthesizes data from many sources (e.g., field observation, airlines, ARTS III, OAG, ADR, PMM & Co., ATC tower) and merges them together into a single format for subsequent processing - see Fig. 5. Note in Fig. 5 that the data are merged to an 80-column format, and that each data field is identified by an arrow through the appropriate columns. Listed below each arrow is the

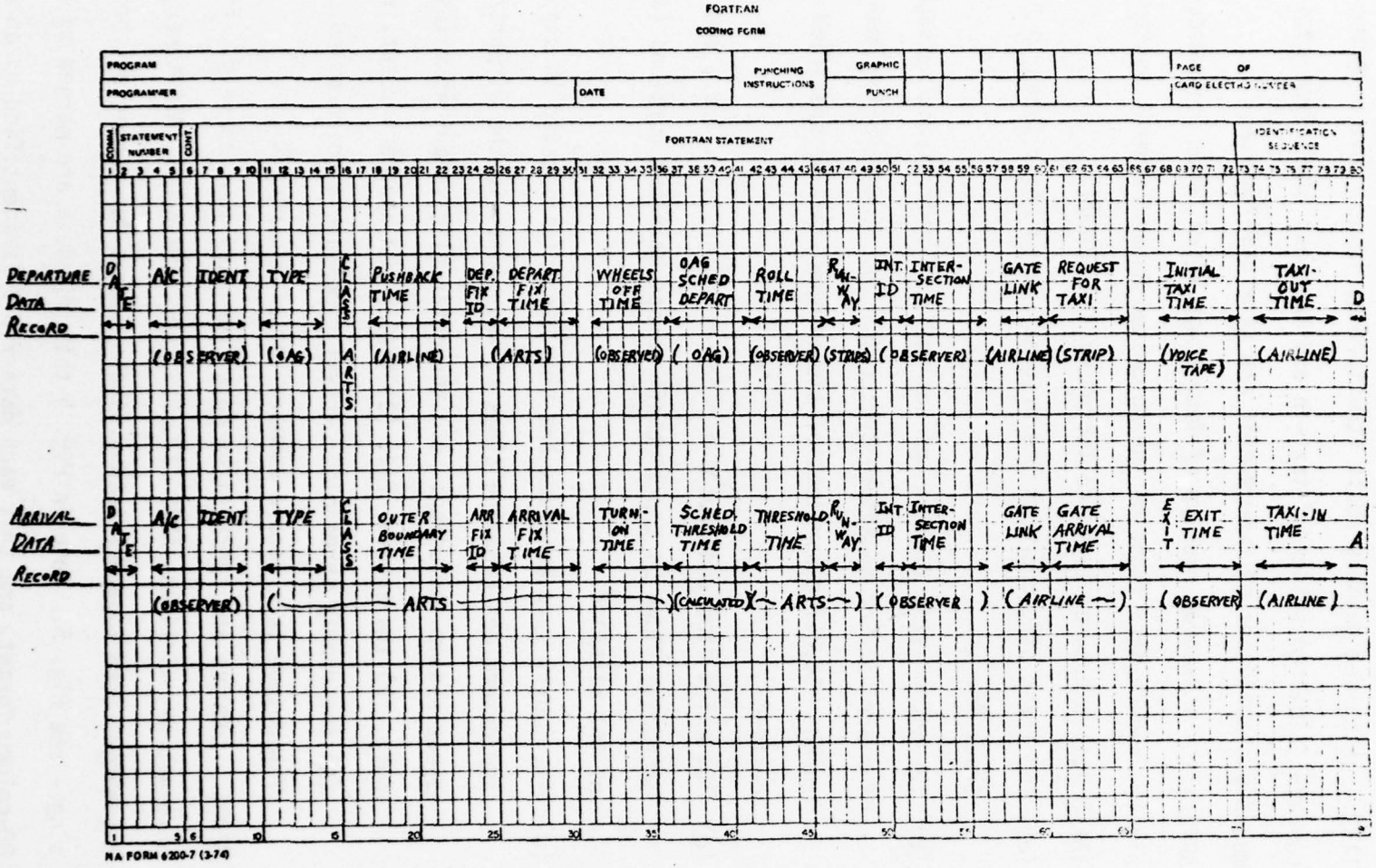


Fig. 5 - Final Format for the Merged Data Set.

source of that particular data item.

(8) ACSCHED - takes the merged data and generates an aircraft arrival and departure schedule by matching arrivals and departures according to one of several criteria (in descending, hierarchical order):
   (a) by aircraft identity, i.e., flight number or call sign
   (b) by assigned gate number, i.e., same gate assigned to two aircraft of the same class within a short time interval
   (c) by operation times of aircraft of same class and airline
   (d) by operation times.

(9) TRMSEP - using the outputs of MERGE, this program constructs statistical distributions for aircraft separations, in seconds between successive passings of the runway threshold. Separations are classified by (a) runway, (b) by dependent runway pair, (c) by aircraft-class pair, and (d) by type of operation, i.e., arrival or departure.

(10) TRVTIME - calculates means, standard deviations, and statistical distributions for aircraft travel times between the arrival fix and runway threshold. Results are presented by "arrival-fix/runway-used" pair and by aircraft class.

The data reduction process, as described in the foregoing, will provide the necessary "real-world" data for input to the simulation model and response variables for comparison with model outputs.

IV. STATISTICAL ANALYSIS OF THE DATA

Purpose and Scope

The purpose of the statistical tests of model output versus observed data is to obtain quantitative evidence of the model's ability to simulate airfield operations. These tests should, therefore, be viewed as "aids to de-

cision making" rather than numerical criteria that the model must satisfy. The significance of the statistical comparisons described below can only be judged in the context of the anticipated application of the model and the types of decisions that the model might support.

The description that follows is intended to guide the process of final data reduction and calculations of "real-world" comparison quantities, e.g., delays, queue sizes, travel times, and flow rates. It is also intended to guide the specification of the output of corresponding model estimates of those quantities. Finally, this section describes the actual steps in comparing model estimates with measured quantities and the interpretation of those comparisons.

Model Convergence

The internal convergence of the model estimates will be considered as a problem separate from (but not unrelated to) the validation comparisons with observed data. A simulation model might, for example, produce average values with extremely high precision, as measured by the standard error of the mean, but this says nothing about the accuracy of those average values in an absolute sense, i.e., relative to corresponding real-world values.

The question to be addressed in this section is "How many independent replications of the model (each with a different random number seed) are required to obtain a desired degree of precision in model estimates of mean values of delays, flow rates, etc.?" The degree of precision will be assumed to be expressed as a confidence interval. A commonly used measure is the 95% confidence interval, which has a certain intuitive appeal.

Suppose that we run the model n times, each with the same input data but with a different random number seed. From each run suppose we obtain an estimate of some response variable for the simulated period: L_1,

$L_2, \ldots, L_n$.[1] The point estimate of the overall average for the n replications is

$$\hat{L} = \frac{1}{n}\sum_{i=1}^{n} L_i \tag{1}$$

and the sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (L_i - \hat{L})^2. \tag{2}$$

If n is sufficiently large, say at least twelve, then the sample average, $\hat{L}$, may be assumed to have a normal distribution with mean $\frac{1}{n}\sum_{i=1}^{n} L_i$ and variance

$$s_{\hat{L}}^2 = \frac{1}{n}\,\frac{\sum_{i=1}^{n} (L_i - \hat{L})^2}{n - 1} \tag{3}$$

Note that the assumption of normality is supported by the central limit theorem even if the L_i are not normally distributed. Note also that the variance is not known, a priori, but must be approximated by the sample variance.

The above assumption allows us to obtain a confidence interval estimate for $\hat{L}$ as

$$\hat{L} \pm z_\alpha\, s_{\hat{L}}$$

if n is 30 or larger or

$$\hat{L} \pm t_\alpha(n-1)\, s_{\hat{L}}$$

---

[1] The L_i can be total or average values without loss of generality. Furthermore, the fact that the components of each L_i might be autocorrelated can be ignored for the sake of this discussion. The requirement that the L_i, i=1, ..., n, be independent cannot be ignored.

for n less than 30, where Z_α is from the standard normal tables, $1 - \alpha$ is the confidence level, and $t_\alpha(n-1)$ is from a table of the Student's t distribution with n-1 degrees of freedom.[2]

The 95% confidence interval, $\hat{L} \pm Z_{.05}\, s_{\hat{L}}$ or $\hat{L} \pm t_{.05}(n-1)\, s_{\hat{L}}$, can be specified in advance as either an absolute value, say $\hat{L} \pm A$ minutes of delay, or as a fraction B of the mean value, e.g., $\hat{L} \pm B\hat{L}$. In this latter case, which is a common way to express convergence, it must be realized that $\hat{L}$ is a random variable, and so the specified confidence interval size, $B\hat{L}$, is a random variable and not a fixed range. Thus the usual equations for such intervals are only approximate since they ignore this fact. This applies to Eqs. (6), (7), and (9) below. In either case one can solve for the required number of replications to achieve the specified precision as follows:

for absolute diff. - A

$$n < 30: \quad n^2 - n - \frac{t_\alpha^{\,2}(n-1)\sum_{i=1}^{n}(L_i - \hat{L})^2}{A^2} = 0 \tag{4}$$

$$n > 30: \quad n^2 - n - \frac{Z_\alpha^{\,2}\sum_{i=1}^{n}(L_i - \hat{L})^2}{A^2} = 0 \tag{5}$$

for diff. as a fraction - B

$$n < 30: \quad n^2 - n - \frac{t_\alpha^{\,2}(n-1)\sum_{i=1}^{n}(L_i - \hat{L})^2}{B^2\hat{L}^2} = 0 \tag{6}$$

$$n > 30: \quad n^2 - n - \frac{Z_\alpha^{\,2}\sum_{i=1}^{n}(L_i - \hat{L})^2}{B^2\hat{L}^2} = 0 \tag{7}$$

which can be solved for n using the quadratic formula.

In the cases where n > 30, the Z_α value is usually assumed to be based on a known fixed population variance $\sigma^2_{L_i}$, even though we estimate it with s^2.

---

[2] Use of the t-statistic implies the additional assumption that the L_i are normally distributed.

Therefore, Eqs. (5) and (7) are usually written:

$$n \geq \left(\frac{z_\alpha s}{A}\right)^2 \tag{8}$$

or

$$n \geq \left(\frac{z_\alpha s}{B \hat{L}}\right)^2 \tag{9}$$

Similarly, even when $n < 30$, it is usually assumed that the estimator, s, is not a function of n. Therefore, Eqs. (4) and (6) are usually written:

$$n \geq \left(\frac{t_\alpha(n-1)\ s}{A}\right)^2 \tag{10}$$

and

$$n \geq \left(\frac{t_\alpha(n-1)\ s}{B \hat{L}}\right)^2 \tag{11}$$

Note, however, that in Eqs. (10) and (11), $t_\alpha(n-1)$ is itself a function of n. Therefore, n must be estimated by trial, i.e., assume a value, say n*, plug in $t_\alpha(n^*-1)$ and solve for n and check to see if n* is sufficiently close to the n computed. If not, repeat until n* and n are sufficiently close.

How rapidly the model converges for any given n depends on how many aircraft are processed in each replication which, in turn, affects the total number of variates randomly drawn (for a given number of variates per aircraft) in each replication. Thus, it is not possible, and in fact might be wasteful, to make any blanket statements about how many replications are necessary to achieve convergence. Instead, a relationship should be developed between the number of replications, the number of aircraft per replication, and the desired confidence interval. Such a relation could be depicted graphically as in Fig.

6 (which is schematic only), or in tabular form as in Table 1, for a given response variable expressed in units of minutes.

The convergence of the model probably depends on the variance of the particular quantity being considered. For example, a greater number of replications might be needed for average arrival delay to converge to its criterion than for average flow rate to converge to its criterion. This should be investigated. In such cases, one might choose the maximum of the various derived numbers of replications associated with the different comparison quantities.

There is an obvious tradeoff between run length, i.e., the length of time period being simulated, and the number of replications. Clearly, more replications will be required, for any desired degree of convergence, if one hour is simulated than for three hours, other things being equal, e.g., the level of activity. This is why it is desirable to express the number of required replications as a function of the level of activity and the length of time interval being simulated.

Pseudo Random Number Generator

The contractor should document how they use the particular pseudo random number generator contained in the model. This documentation may take the form of: (1) existing descriptions of statistical testing of the generator as contained in library subroutine descriptions or the general literature; and (2) accepted methods of choosing different random number seeds when running the model. No new statistical testing is anticipated for this part of the validation.

The contractor has provided a copy of their random number generator routine and a series of random number streams to the model validation working subgroup. Preliminary tests of the provided streams indicate that they are satisfactory from the standpoint of serial independence and goodness-of-fit to a uniform distribution on the interval (0, 1).

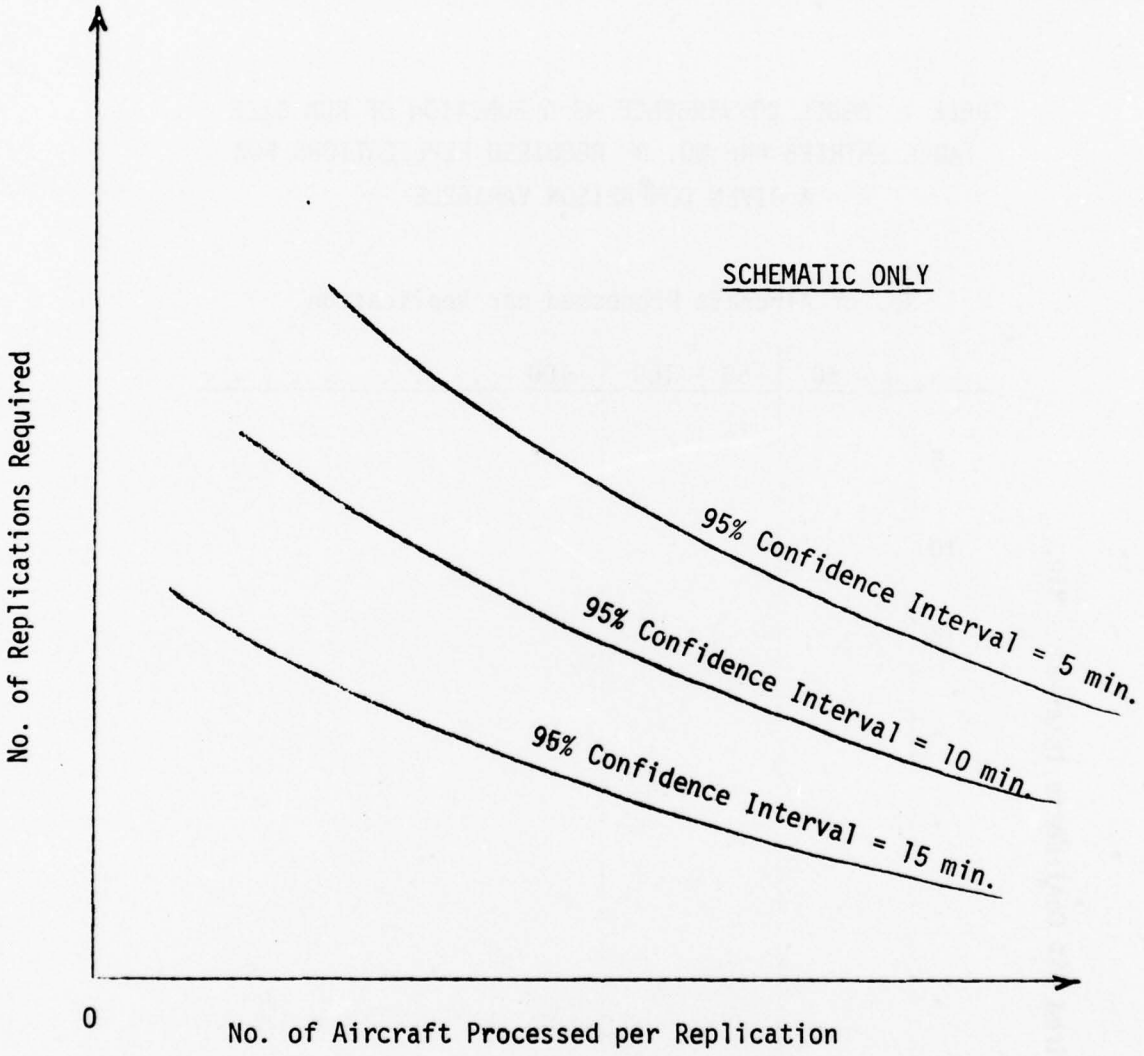


Fig. 6. Model Convergence as a Function of Run Size for a Given Comparison Variable.

TABLE 1 MODEL CONVERGENCE AS A FUNCTION OF RUN SIZE - TABLE ENTRIES ARE NO. OF REQUIRED REPLICATIONS FOR A GIVEN COMPARISON VARIABLE

No. of Aircraft Processed per Replication

Desired 95% Confidence Interval - Min.	0 - 50	50 - 100	100 - . . . . .
5			
10			
15			
.			
.			
.			
.			
.			

Model Output Responses to be Tested

As described in the validation plan, the following are the output responses to be compared with corresponding observed data (arranged in order of importance):

(1) arrival threshold flow rates (by time interval)

(2) departure threshold (roll point) flow rates (by time interval)

(3) arrival airspace delay (by time interval and aircraft)

(4) taxi-in time (by time interval and aircraft)

(5) taxi-out time (by time interval and aircraft)

(6) departure-queue size (by time interval)

(7) penalty-box delay (by aircraft)

These quantities are computed for individual five-minute intervals or individual aircraft, or both, as indicated above. Consider, for example, arrival delays. The model should output estimates of arrival delays by runway, by five-minute time interval, and by individual flight number.

In testing the model output against real-world data, we may use time intervals longer then five minutes and groups of flights larger than one. By obtaining five-minute and one-flight delay summaries, however, we will have sufficient flexibility in choosing interval length (multiples of 5 min.) and group size. We will thus be able to choose an optimal (in some sense) combination of interval size, which affects the reliability of individual observations, and the number of intervals, which constitutes the sample size for subsequent statistical tests. Similarly, we will be able to choose efficient combinations of flight-group size and the number of flight groups.

Table 2 summarizes the model outputs and the necessary detailed specifications for the model estimates.

Data Reduction of Output Responses

The reduction of the observed data should result in real-world output

TABLE 2. MODEL AND REAL-WORLD OUTPUT RESPONSES FOR STATISTICAL TESTING.

Response Variable in Order of Importance	Output Specification	Stratification	Comments
1. Aircraft Flow Rates (No. of Aircraft per Hour)	By five-minute interval.	By arrival and departures and by runway.	Tally threshold times and roll times by five minute interval for both model and real-world data.
2. Arrival Airspace Delay (minutes)	By five-minute interval and by individual flight.	By arrival runway.	Should be tallied in the time interval when the delay is finally computed.
3. Ground Travel Times (minutes)	By five-minute interval and by individual flight.	By link.	(same as above)
4. Departure-Queue Size (No. of Aircraft)	By five-minute time interval.	By departure runway.	Check to see that model output is comparable in concept to reduced data on departure queue size.
5. Penalty-Box Delay (minutes)	By aircraft.	By box or holding position.	

responses identical in definition to the model output specifications of Table 2, hence the title "Model and Real-World Output Response."

From the comments column of Table 2 it is clear that it is important to tally delays by time interval (more specifically, the time interval in which they are finally calculated) in the same way in the model as in the observed data reduction and delay calculations. This could present certain problems that can be avoided by testing delays tallied by individual flight number. In the latter case, delays for groups of individual flights can be averaged together. Hence, the two series to be compared could be denoted X_i, i=1, ..., m and Y_i, i=1, ..., m, where i represents the individual flight group, X_i are model estimates, and Y_i are calculated from the collected data. The above two series are then treated as time series.

It is expected that arrival delays, ground travel times, and penalty-box delays will be treated on an individual-flight basis in addition to the time-interval basis. Which is the best approach will have to be judged after attempting the tests and seeing what problems are encountered in each one.

Nature of the Statistical Hypothesis Tests

The Hsu-Hunter method of time-series analysis will be used in the hypothesis testing of the model estimates against observed data.[3] Computer programs will be available to aid in conducting the statistical comparisons. To facilitate the process the required comparison data of Table 2 should be punched on cards in the reduction of the model's detailed (individual aircraft) output tape.

Unfortunately, it is not feasible to obtain data on a large number of days that constitute independent and identical conditions. Because conditions and runway configurations are so variable at O'Hare, we are probably constrained to compare observed data for individual days to model outputs for those days.

---

[3] Hsu, D. A. and J. S. Hunter, "Analysis of Simulation-Generated Responses using Autoregressive Models," Management Science, Vol. 24, No. 2, October 1977, pp. 181-190.

Model outputs are averaged over, say, n different replications to obtain the desired model convergence as discussed earlier. The sample size for the model estimates is, therefore, n-fold larger than the sample of observed data. The two samples are depicted in Table 3.

The first step of the Hsu-Hunter method is to obtain an autoregressive time series model of order p for the observed data as follows:

$$y_t = \phi_1^* y_{t-1} + \phi_2^* y_{t-2} + \cdots + \phi_p^* y_{t-p} + \beta_t \qquad (12)$$

where $y_t = Y_t - \bar{Y}$; $\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$, so that $E\{y_t\} = 0$ for all t,

ϕ_i^*, i=1, ..., p = autoregressive coefficients for the observed data,

β_t = normally distributed random error term with mean 0 and variance σ_2^2.

Before fitting the autoregressive model, one can gain insight into a time series by estimating its autocorrelation function. Consider a time series Y_t that has the properties of a covariance-stationary process: namely that neither the covariance nor the expected value of the time series is a function of time. The autocovariance function of Y_t is defined as

$$\text{Cov}\{Y_t, Y_{t+s}\} = E\{[Y_t - E(Y_t)]\,[Y_{t+s} - E(Y_{t+s})]\}$$

The assumption of stationarity implies that Cov $\{Y_t, Y_{t+s}\}$ depends only on s and not on t. The autocovariance function can be estimated using the following estimator for the "sample covariance of lag s":

$$c_s = \frac{1}{n-s}\sum_{t=1}^{n-s} (Y_t - \bar{Y})\,(Y_{t+s} - \bar{Y})$$

Note that c_0 is the sample variance of Y_t based on a divisor n instead of the usual divisor n-1. The autocorrelation function, ρ_s, obtained from

TABLE 3 - SAMPLES OF MODEL ESTIMATES AND OBSERVED DATA FOR INDIVIDUAL DAYS

Interval No. or Flight Group No. - i

Replication No. - j	1	2	i	m
1	$X_1^{(1)}$	$X_2^{(1)}$	$X_i^{(1)}$	$X_m^{(1)}$
2	$X_1^{(2)}$	$X_2^{(2)}$	$X_i^{(2)}$	$X_m^{(2)}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
j	$X_1^{(j)}$	.	$X_i^{(j)}$	$X_m^{(j)}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	$X_1^{(n)}$	.	$X_i^{(n)}$	$X_m^{(n)}$

(a) Model Estimates

Interval No. or Flight Group No. - i

1	2	i	m
Y_1	Y_2	Y_i	Y_m

(b) Observed Data

$$\rho_s = \frac{c_s}{c_0} \quad ,$$

is a measure of the linear dependence of Y_t on its past history. The function ρ_s is equal to unity for s = 0 and lies in the interval (-1, 1) for all other values of s.

The autocorrelation function is used in this validation to gain insight into the two time series being compared and to help decide on the order, p, of the autoregressive time series models described below. We will also compare the autocorrelation functions for the model estimates to the ones for observed data by plotting each one as a function of s.

For the estimates output by the simulation model we have the autoregressive time series model in Table 4 in which:

$$x_t^{(j)} = X_t^{(j)} - \bar{X}^{(j)} \text{ and } \bar{X}^{(j)} = \frac{1}{m}\sum_{i=1}^{m} X_i^{(j)}$$

TABLE 4 - AUTOREGRESSIVE TIME SERIES MODEL FOR THE MODEL ESTIMATES

Replication No. - j	
	$x_t^{(1)} = \phi_1 x_{t-1}^{(1)} + \phi_2 x_{t-2}^{(1)} + \cdots + \phi_p x_{t-p}^{(1)} + \alpha_t^{(1)}$
	$x_t^{(2)} = \phi_1 x_{t-1}^{(2)} + \phi_2 x_{t-2}^{(2)} + \cdots + \phi_p x_{t-p}^{(2)} + \alpha_t^{(2)}$
	. . . .
	$x_t^{(n)} = \phi_1 x_{t-1}^{(n)} + \phi_2 x_{t-2}^{(n)} + \cdots + \phi_p x_{t-p}^{(n)} + \alpha_t^{(n)}$
Pooled:	$\sum_{j=1}^{n} x_t^{(j)} = \phi_1 \sum_{j=1}^{n} x_{t-1}^{(j)} + \phi_2 \sum_{j=1}^{n} x_{t-2}^{(j)} + \cdots + \phi_p \sum_{j=1}^{n} x_{t-p}^{(j)} + \sum_{j=1}^{n} \alpha_t^{(j)}$

where $\alpha_t^{(j)}$ = normally distributed random error term with mean 0 and variance σ_1^2

and $\sum_{j=1}^{n} \alpha_t^{(j)}$ = normally distributed random error term with mean 0 and variance $n\,\sigma_1^2$.

We can pool together the results of the n replications, as shown on the previous page by the last line of Table 4, without loss of generality except that the variance of $\sum_{j=1}^{n} \alpha_t^{(j)}$ would not be comparable with the variance of β_t: if $\alpha_t^{(j)}$ has variance σ_1^2, then $\sum_{j=1}^{n} \alpha_t^{(j)}$ has variance $n\,\sigma_1^2$. Because we know that $\mathrm{Var}\left\{\frac{\underset{\sim}{Z}}{\sqrt{n}}\right\} = \frac{1}{n}\,\mathrm{Var}\,\{\underset{\sim}{Z}\}$, we can divide the last line through by $\sqrt{n}$ as follows:

$$\frac{\sum_{j=1}^{n} x_t^{(j)}}{\sqrt{n}} = \phi_1 \frac{\sum_{j=1}^{n} x_{t-1}^{(j)}}{\sqrt{n}} + \phi_2 \frac{\sum_{j=1}^{n} x_{t-2}^{(j)}}{\sqrt{n}} + \cdots + \phi_p \frac{\sum_{j=1}^{n} x_{t-p}^{(j)}}{\sqrt{n}} + \frac{\sum_{j=1}^{n} \alpha_t^{(j)}}{\sqrt{n}}$$

for which the varaince of the error term, $\frac{\sum_{j=1}^{n} \alpha_t^{(j)}}{\sqrt{n}}$, is σ_1^2, which is now comparable to the variance of the observed data error term, σ_2^2.

The foregoing arguments can be summed up by saying that the results of the n replications of the model should be added together and divided by the $\sqrt{n}$ before performing the autoregression. Thus, the two autoregressive time series models to be compared in step 1 of the Hsu-Hunter method are given by Eqs. (12) and (13).

The model series has parameters $\underset{\sim}{\phi}_1 = (\phi_1, \phi_2, \ldots, \phi_p)$ and σ_1^2 while the observed series has $\underset{\sim}{\phi}_2 = (\phi_1^*, \phi_2^*, \ldots, \phi_p^*)$ and σ_2^2. The parameters of the inferential statistic $G(\underset{\sim}{\psi}, \gamma)$ are $\underset{\sim}{\psi} = \underset{\sim}{\phi}_1 - \underset{\sim}{\phi}_2$ and $\gamma = \frac{\sigma_2^2}{\sigma_1^2}$. In the following discussion p is assumed to be strictly less than m (in previous applications p has usually been 2 or 3 depending on the nature of the autocorrelation function - more will be said about this later).

Hsu has shown that the joint probability density function of $\underset{\sim}{\psi}$ and γ, $G(\underset{\sim}{\psi}, \gamma)$, is asymptotically distributed as $\frac{1}{2}\chi^2(p+1)$.[4]

[4] Details of the derivation of $\frac{1}{2}\chi^2(p+1)$ asymptotic distribution are available in Hsu, D. A., Stochastic Instability in the Behavior of Stock Prices, unpublished Ph.D. dissertation, Department of Statistics, University of Wisconsin, Madison, May 1973.

The inferential statistic $G(\underset{\sim}{0}, 1)$ is used to test the hypothesis that $\underset{\sim}{\psi} = \underset{\sim}{0}$ and $\gamma = 1$, i.e., to test, simultaneously, the potential difference in both the autoregressive parameters and the residual variance between the two time series. Hsu and Hunter recommend, based on experience with Monte Carlo experiments, that the χ^2 approximation is satisfactory when the length of both time series, m, is no less than 60.[5]

The second step of the Hsu-Hunter method is to compare the means of the two time series under consideration. In this second step an inferential statistic, distributed as a Student's t random variable, is used assuming that the values of the autoregressive parameters (from step 1) are known quantities.

For comparison of means we have the two autoregressive time series models of order p:

$$\sum_{j=1}^{n} x_t^{(j)} = \phi_1 \sum_{j=1}^{n} x_{t-1}^{(j)} + \phi_2 \sum_{j=1}^{n} x_{t-2}^{(j)} + \cdots + \phi_p \sum_{j=1}^{n} x_{t-p}^{(j)} + \sum_{j=1}^{n} \alpha_t^{(j)}$$

and $$y_t = \phi_1^* y_{t-1} + \phi_2^* y_{t-2} + \cdots + \phi_p^* y_{t-p} + \alpha_t$$

The sample means of the two original time series are $\bar{X} = \sum_{j=1}^{n} \sum_{i=1}^{m} X_i^{(j)} / (n + m)$ and $\bar{Y} = \frac{1}{m} \sum_{i=1}^{m} Y_i$. Denote the two population means as μ_1 and μ_2, respectively. Then the two above series can be rewritten:

$$\sum_{j=1}^{n} X_t^{(j)} - \mu_1 = \phi_1 \left(\sum_{j=1}^{n} X_{t-1}^{(j)} - \mu_1 \right) + \phi_2 \left(\sum_{j=1}^{n} X_{t-2}^{(j)} - \mu_1 \right) + \cdots + \phi_p \left(\sum_{j=1}^{n} X_{t-p}^{(j)} - \mu_1 \right) + a_t ,$$

where $a_t = \sum_{j=1}^{n} \alpha_t^{(j)}$,

and

---

[5] Hsu and Hunter, p. 185.

$$Y_t - \mu_2 = \phi_1^*\left(Y_{t-1} - \mu_2\right) + \phi_2^*\left(Y_{t-2} - \mu_2\right) + \cdots + \phi_p^*\left(Y_{t-p} - \mu_2\right) + \beta_t$$

for $t = p + 1, \ldots, n$.

We will now obtain a transformed variable by shifting all terms of the above equations that involve μ_1 or μ_2 to the right side and the $\sum_{j=1}^{n} X^{(j)}$'s and Y's to the left as follows:

$$u_t \equiv \sum_{j=1}^{n} X_t^{(j)} - \phi_1 \sum_{j=1}^{n} X_{t-1}^{(j)} - \cdots - \phi_p \sum_{j=1}^{n} X_{t-p}^{(j)} = (1 - \phi_1 - \cdots - \phi_p)\,\mu_1 + a_t$$

(the symbol "≡" means "by the definition" or "is defined as")

and

$$u_t^* \equiv Y_t - \phi_1^* Y_{t-1} - \cdots - \phi_p^* Y_{t-p} = (1 - \phi_1^* - \cdots - \phi_p^*)\mu_2 + \beta_t \ .$$

Dividing both sides of the above equations by the coefficients of μ_1 and μ_2, i.e., by $(1 - \phi_1 - \cdots - \phi_p)$ and $(1 - \phi_1^* - \cdots - \phi_p^*)$, respectively, the desired transformed variables result:

$$w_t \equiv \frac{u_t}{1 - \phi_1 - \cdots - \phi_p} = \mu_1 + \frac{a_t}{1 - \phi_1 - \cdots - \phi_p} = \mu_1 + c_t \qquad (12)$$

and

$$w_t^* \equiv \frac{u_t^*}{1 - \phi_1^* - \cdots - \phi_p^*} = \mu_2 + \frac{\beta_t}{1 - \phi_1^* - \cdots - \phi_p^*} = \mu_2 + b_t \qquad (13)$$

where c_t is a normally distributed residual with mean 0 and variance $\sigma_1^2 / (1 - \phi_1 - \cdots - \phi_p)^2$ and b_t is normally distributed with mean 0 and variance $\sigma_2^2 / (1 - \phi_1^* - \cdots - \phi_p^*)^2$.

The transformed variables w_t and w_t^* are independent normal variables with mean μ_1 and μ_2 and variances $\sigma_1^2 / (1 - \phi_1 - \cdots - \phi_p)^2$ and $\sigma_2^2 / (1 - \phi_1^* - \cdots - \phi_p^*)^2$, respectively.

Thus, we now have two transformed series $\underset{\sim}{w} = \{w_{p+1}, \ldots, w_m\}$ and

$\underset{\sim}{w}^* = \{w^*_{p+1}, \ldots, w^*_m\}$ which have means μ_1 and μ_2. We can express the distribution form of the inferential statistic for $(\mu_2 - \mu_1)$ as follows

$$t \equiv \frac{\bar{w}_2 - \bar{w}_1}{g_1\left[\frac{s_1^2}{m-p} + \frac{s_2^2}{m-p}\right]^{\frac{1}{2}}} \sim t\,(g_2) \qquad (14)$$

This statistic is distributed approximately as a Student's t random variable with g_2 degrees of freedom. The parameters s_1^2 and s_2^2 are the sample variances of $\underset{\sim}{w}$ and $\underset{\sim}{w}^*$, respectively, and g_1 and g_2 are both functions of s_1^2, s_2^2, and m.[6]

The $\bar{w}_1$ and $\bar{w}_2$ are the means of the transformed variables $\underset{\sim}{w}$ and $\underset{\sim}{w}^*$. Thus we can use the above t-statistic to test the hypothesis H_0: $\mu_1 = \mu_2$. By using the transformed variables w_t and w_t^*, both being serially independent, the condition that $\bar{w}_1$ and $\bar{w}_2$ be normally distributed is satisfied. Furthermore, we have taken account of and have removed the autocorrelation of the two series by dividing their variances σ_1^2 and σ_2^2 by $(1 - \phi_1 - \cdots - \phi_p)^2$ and $(1 - \phi_1^* - \cdots - \phi_p^*)^2$, respectively.

Even more so than most traditional methods of statistical analysis, it is essential that an experienced statistician be involved in the various steps of autoregressive time series analysis. This type of analysis involves, in addition to the usual estimation phase, a model identification phase. This phase provides considerable flexibility to the trained statistician in fitting an autoregressive model to a time series. The autocorrelation function, for example, provides a clue to choosing the order p of the autoregressive model. Furthermore, an analysis of the residuals, more precisely, the autocorrelation function of the residuals, of an autoregressive model can point to appropriate modifications of the model. For example, such an analysis of residuals may

---

[6] Details of the derivation and computation of the statistic t above are available in Box, G.E.P. and Tiao, G.C., Bayesian Influence in Statistical Analysis, Addison Wesley Publishing Co., Chapter 2, pp. 107.

point to an improved identification of the model.

Given the foregoing flexibility in the fitting of an autoregressive time series model, it is expected that there will be no major difficulties in fitting such a model to the observed or model-generated time series of this validation effort. For the unlikely event that model-fit problems do arise, however, a simplified alternative approach is presented below.

Alternative Methods of Statistical Analysis

In the unlikely event that difficulties are encountered in attempting to fit an autoregressive time series model, we will have to fall back on more traditional statistical hypothesis tests that assume that a true random sample can be achieved and, consequently, that the items of that random sample are mutually independent; this implies that they would be uncorrelated.

In such cases, for example, a standard Student's-t test can be used to test the difference in the means of two random samples that, say, correspond to the model outputs and observed data, respectively. Details of such a test are given here as an alternative to the Hsu-Hunter method.*

Described herein is a standard statistical test that can be used to test whether a set of delay estimates produced for a specific set of circumstances by a simulation model and a real world measurement taken under the same circumstances have the same mean value.

These delay estimates are considered to be averages taken over some time period. Care must be taken to insure that the time period used be sufficiently long that the model and the real world might reasonably be expected to show approximately the same behavior. The test described below can be used for average delays taken over any sufficiently long time period.

The assumptions required are that the estimates produced by the simulation

*This t-test description was provided by N. J. Kirkendall of The MITRE Corporation, memo to A. L. Haines, dated 5 January 1978.

model be independent and identically distributed $N(\mu, \sigma^2)$ where μ and σ^2 are fixed but unknown, and that the real world observation be independent of the model estimates and normally distributed with the same variance, σ^2.

Since the test described below is fairly robust to departures from normality,** the only really questionable assumption is the equality of the variances of the processes underlying the model estimates and the real world measurement. If the simulation model were perfect, of course, the variances as well as the means would be equal. In any case, a single observation from the real world will provide no information concerning the variance of the average delays observed in the real world.

Details. Let x_o represent the real world measurement for a given time period, and $x_1, \ldots, x_N$ represent the N estimates from the model for that time period. Assuming that the x_i are mutually independent and normally distributed with the same variance we would like to test the following hypothesis:

$$H_o: \quad x_o, \ldots, x_n \sim N(\mu, \sigma^2)$$

$$H_1: \quad x_o \sim N(\mu_1, \sigma^2), \quad x_1, \ldots, x_N \sim N(\mu, \sigma^2), \quad \mu \neq \mu_1$$

The test of this hypothesis is the standard likelihood ratio test for the equality of the means of two samples, as given, for example, on page 288 of Hogg & Craig, Introduction to Mathematical Statistics, 2nd ed., 1966.

The result is that the test with significance level α is to reject H_o when

$$t = \frac{\sqrt{\frac{N}{N+1}}\left|\bar{x} - x_o\right|}{\sqrt{\frac{NS_x^2}{N-1}}} > t_{N-1}(\alpha)$$

where

** Kendall & Stuart, The Advanced Theory of Statistics, Vol. II, 2nd ed., 1967, pp. 465-467.

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$S_x^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

and $t_{N-1}(\alpha)$ is chosen from tables of the Student's t distribution with N-1 degrees of freedom to have

$$P\left\{ |t| > t_{N-1}(\alpha) \right\} = \alpha$$

This implies that confidence limits for x_o are given by

$$P\left\{ \bar{x} - t_{N-1}(\alpha) \sqrt{\frac{N+1}{N-1}} S_x^2 \leq x_o \leq \bar{x} + t_{N-1}(\alpha) \sqrt{\frac{N+1}{N-1}} S_x \right\} = 1 - \alpha$$

Details of the derivation of this likelihood ratio test can be found in most textbooks on mathematical statistics.

It should be emphasized that this method considers a single model average for each replication for a given time period versus a single observed value rather than a time history, say by five minute interval, of model estimates for each replication versus a time history of observed data as described earlier. Thus, this alternative method does not provide as much information as the time series comparisons of the Hsu-Hunter Method. Nevertheless, it is presented here as a possible alternative in case there are difficulties encountered in attempting to apply the Hsu-Hunter method.

Decisions Based on Statistical Hypothesis Tests

It is recommended that the results of the tests not be judged on the basis of a priori significance levels, e.g., the 0.05 level. Instead, significance probabilities should be estimated for each test and the entire set of significance probabilities then judged by the Model Validation Group as described be-

low. Technical advisors on the validation working subgroup will provide assistance in interpreting the results of the tests.

Another important consideration in judging the statistical test results is that not all response variables are equally important. Some are more important than others in the types of decisions made by airport operators, airlines, and FAA. This will be taken into account in the evaluation of results by the Model Validation Group.

A significance probability (usually denoted as P_I) is the probability of obtaining a value of the inferential statistic, i.e., a χ^2-value or t-value, as large as the one computed in the test, given that the hypothesis tested is actually true.[7] If this probability is very small, one would tend to reject or at least suspect the hypothesis. If the significance probability is not small, then one would tend not to reject the hypothesis. The conclusion in the latter case might be that the differences obtained in the test are due to chance rather than to defects in the model. No precise universal definition of "small" can be offered - this is a matter of judgement and confidence in the statistical methods employed.

Further insight into the test results may be obtained by considering the results of the whole set of tests. Suppose for example that test results in the form of significance probabilities, P_{ij}, where i is the test number and j is the comparison variable, are tabulated as shown on the following page.

If the hypotheses tested were all really true, then at any arbitrary significance level, α, one would expect the number of tests that failed, i.e., the number of $P_{ij} < \alpha$, to not exceed (100 α) percent of the total number of tests, $k\ell$. Thus not more than about 5% of the P_{ij} should be less than 0.05, not more than about 10% should be less than 0.10, etc., by definition of significance

[7] Recall that the general nature of the hypotheses tested in this validation is that there is no difference between the model estimates and the observed data.

Test No. \ Comparison Variable	1	2	j	ℓ
1	P_{11}	. . .	. . .	$P_{1\ell}$
2	.		.	
.	.		.	
.	.		.	
.	.		.	
i		. . .	P_{ij}	
.	.		.	
.	.		.	
.	.		.	
k	P_{k1}	. . .	. . .	$P_{k\ell}$

probability.[8]

The foregoing percentages are not suggested as hard-and-fast criteria for acceptance of the model, because the corresponding tests are not all identical; besides, the number of tests will probably not be large. The above ideas do, however, provide one means of roughly interpreting the results of the whole set of tests to be performed in this validation.

V. SUMMARY DESCRIPTION OF STEPS IN STATISTICAL ANALYSIS OF DATA

Outlined below are the steps to be followed to accomplish the statistical analysis suggested in the foregoing sections:

I. Model Convergence

A. Prerequisites:

The model convergence as a function of the number of aircraft processed per replication and the number of replications can be done as part of the sensitivity analysis using, perhaps, the LaGuardia data set. One caveat, however, is that the results should be spot

[8] If k was very large, one could check this, more appropriately, within each column: $P_{1j}, \ldots, P_{kj}$; j=1, ..., ℓ. This will probably not be the case, however.

checked later to see if they hold for the O'Hare data set. Thus the only prerequisite is that the model be running for the LaGuardia input data. We must also know, however, the number of aircraft processed in each replication, which, in turn, affects the number of random variates drawn each time.

B. Procedure:

The model should be run for a large number of replications, say at least 30. Cumulative[9] averages, Eq. (1), should be computed for each response variable at every 5 replications. Also, compute the cumulative sample variance, Eq. (2), after every 5 replications. This will yield the results, illustrated in the table below, for each response variable and for each activity level, i.e., for each number of aircraft processed.

One or the other of the last two columns of the following table contains the 95% confidence interval for a given number of replications. These can then be compared to an a priori confidence interval, and an

n = No. of Replications	Sample Mean	Sample Variance	Standard Deviation of Mean	for $n < 30$ $t_\alpha(n-1)\ S_{\hat{L}}$	for $n \geq 30$ $z_\alpha\ S_{\hat{L}}$
5	$\hat{L}_5$	s_5^2		C_5	
10	$\hat{L}_{10}$	s_{10}^2		C_{10}	
15	$\hat{L}_{15}$	s_{15}^2		C_{15}	
.	.	.			
.	.	.			
30	$\hat{L}_{30}$	s_{30}^2			C_{30}

[9] The term "cumulative" here means for all previous replications at each stage and not just for the groups of 5.

appropriate number of replications can, thereby, be selected. To aid this process, a graph similar to Fig. 6 can be prepared.

Note that the curves of Fig. 6 are essentially contours of equal-confidence-interval values. These can be plotted as follows:

(1) Create a grid made up of horizontal lines at 5-replication intervals and vertical lines corresponding to the different activity levels, say A, B, C, etc. - see Fig. 7.

(2) The computed confidence intervals for each combination of number of replications and number of aircraft processed should be recorded at the corresponding intersections on the grid - see the C-values of Fig. 7.

(3) Interpolate contours on the grid for convenient, say rounded interger-valued confidence intervals - see, for example, the 5, 10 and 15-min. contours of Fig. 6.

This may seem a tedious process, but the result is valuable: namely, an approximate guide for choosing the number of runs for using the model at any airport given specifications on convergence, the approximate activity level under consideration, say in aircraft per hour, and the length of time period simulated in hours; note that the abscissa of Fig. 7 is the product of these latter two quantities.

The convergence results, as expressed in Figs. 6 and 7, would be translated into a simplified set of guidelines for using the model in terms of the approximate number of replications desirable for different ranges of activity levels.

One question that needs to be addressed is whether conclusions about convergence at one airport are transferable to other airports where there may be a different pattern of operations and, hence,

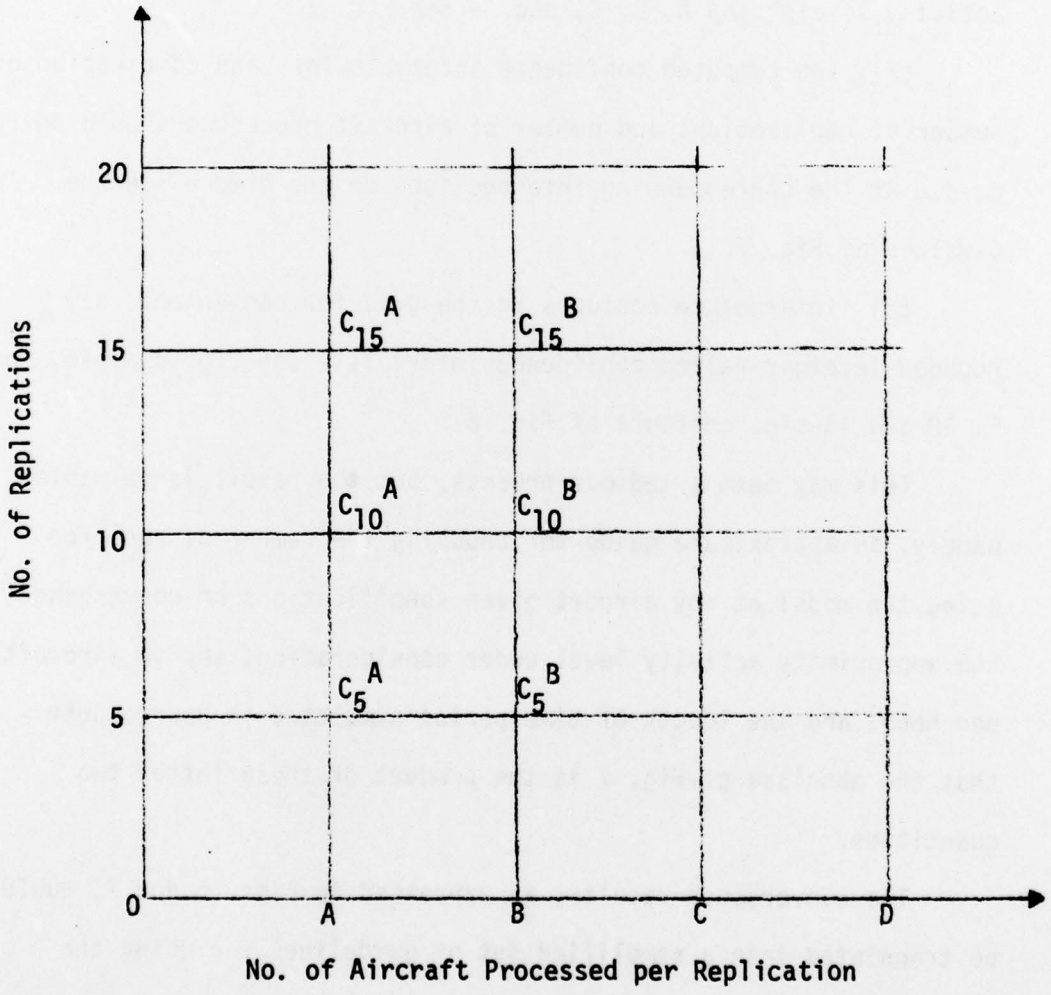


Fig. 7 Illustration of Plotting

Contours of Equal Confidence Interval

different types of aricraft interactions. The answer to this question is not, a priori, obvious. This should be investigated by comparing the LaGuardia convergence results with O'Hare results.

II. Statistical Analysis

A. Prerequisities:

1. Selection of final set of output responses to be validated - see Table 2.
2. Set up Hsu-Hunter computer programs so that they run on the NAFEC computer. Have Dr. Hsu instruct NAFEC personnel on using the programs. The computer program used to execute the Hsu-Hunter method produces computed results of: (1) the value of $G(\underset{\sim}{0}, 1)$ and its significance level determined based upon an appropriate χ^2 distribution; (2) the values of $G(\hat{\underset{\sim}{\psi}}, 1)$ and $G(\underset{\sim}{0}, \hat{\gamma})$, and their significance levels compared with appropriate χ^2 variables; (3) the t-value for testing the difference in means and its significance level; and (4) supplementary information including estimates of μ, ϕ and σ^2 for both observed and generated series. The preliminary stage of model identification using Box-Jenkins techniques, however, requires a separate subroutine package that is operational at Princeton University. An attempt will be made to get this preliminary subroutine running at NAFEC.
3. Prepare computer programs for reading required data from the contractor's simulation model, i.e., the individual aircraft output. The comparison variables of Table 2 are to be defined from that tape.

B. Procedure

1. Prepare graphical and tabular summaries of the model output and the corresponding real-world quantities. Perform visual comparisons of the two for each response variable. This will consist mainly of looking at the summary output of the model and corresponding summaries of observed data.
2. Choose appropriate size time intervals (or aircraft group sizes) for the time-series analysis. Keep in mind that the number of intervals (groups) is the sample size for subsequent analysis. This is desirably as large as possible except that another important consideration is the number of events (aircraft) in each interval (group). Thus there is a tradeoff here. It may not be reasonable to expect the model to accurately duplicate the real world on an aircraft-by-aircraft basis or minute-to-minute basis, given the way random variates are generated in the model. It is appropriate, however, to expect the model to provide reasonable estimates for larger time intervals, say five to ten minutes, or for larger aircraft group sizes, say five to ten aircraft. The problem here is to choose an optimal combination (or at least a good one) of sample size and the number of items in each sample element. Thus, a certain amount of data evaluation and experimental design is required before the actual time series analysis can proceed. Note that this second task must be performed for each output response (comparison) variable of Table 2, for both the model estimates and the observed data. Of course, any given comparison variable will have a common sample size

for model and observed values. Once the two time series are defined for each comparison variable, the remaining steps, described below, can proceed on each one.

3. Compute autocorrelation function, ρ_s, for the model series and the observed data sereis (see p. 28). Based on the form of the two autocorrelation functions, say as determined from a graphical plot of each one vs. s, choose the order, p, for the autoregressive time-series model.
4. Perform step one of the Hsu-Hunter method - see pp. 26-30.
5. Perform step two of the Hsu-Hunter method - see pp. 30-32.
6. Determine significance probabilities for each of the two above steps.
7. Repeat steps 3 through 6 for all response variables and all simulated periods. Note that there are essentially two tests for each output response variable and each simulated period; these correspond to step one and step two of the Hus-Hunter method. The results of all of these test-pairs should be summarized in tabular form as on page 37. In this table, the different comparison variables should be clearly identified and labeled, and so should the different simulated periods. This will facilitate later decision making; in this way the fact that some comparison variables are more important than others, and some time periods are more reliable than others, can be taken into account in making judgements about the collective outcomes of the test as described on pp. 35-37.
8. Apply standard t-test of the equality of means of two random samples in place of or in addition to Hsu-Hunter compari-

sons, particularly if difficulties are encountered in fitting an autoregressive model.

9. The foregoing is only one measure of the model's ability to simulate airfield and airspace operations. The results of these statistical comparisons will have to be weighed along with other evidence, e.g., graphical and tabular comparisons, results sensitivity analyses, and the evaluation of model logic, in making the final judgement as to the adequacy of the model for its intended applications.

C. Priority Ranking of Comparison Variables:

1. The variables of Table 2 may be priority ranked according to two main criteria:
   (a) their importance as figures of merit of the airfield, and
   (b) how accurately it is felt, a priori, the model should estimate the variables given the nature of the input data for validation.
2. The following priority ranking is selected:
   (a) arrival threshold flow rates
   (b) departure threshold (roll) flow rates
   (c) arrival airspace delay
   (d) taxi-in time
   (e) taxi-out time
   (f) departure queue size
   (g) penalty-box delay
3. It is further suggested that the foregoing variables be obtained for 5-minute intervals from both the model output and the observed data. This will enable us to use the

summary output format from the model, but for each 5-minutes instead of each hour which is the usual model output.

4. The 5-minute interval data should be punched directly on cards by both the model and the data-reduction programs. This will entail inserting an additional punch statement and corresponding format statement in the model and data reduction programs.

D. Analyzing Differences Between Model Estimates and Observed Data:

1. An important model application is to investigate differences among alternative improvements and runway-use configurations. It is, therefore, of interest in the validation to test the model's ability to estimate such differences. This will be attempted by testing differences in two time series as estimated by the model versus the corresponding differences in two time series from the observed data. In each case, a new time series, say Z_t, will be obtained as the difference between two time series, i.e.,

$$Z_t = X_t^{(1)} - X_t^{(2)}$$

where the superscripts, (1) and (2), refer to two different configurations or two alternative improvements. We will explore with D. A. Hsu the problems associated with analyzing such a time series.

VI. BIBLIOGRAPHY

1. de Neufville, R. and Stafford, J. H., Systems Analysis for Engineers and Managers, New York: McGraw-Hill Book Co., 1971.

2. Fishman, G. S., "Problems in the Statistical Analysis of Simulation Experiments: The Comparison of Means and the Length of Sample Records," Comm. ACM, Vol. 10, pp. 94-99, 1967.

3. Fishman, G. S., Concepts and Methods in Discrete Event Digital Simulation, New York: John Wiley & Sons, 1973.

4. Fishman, G. S. and Kiviat, P. J., "The Analysis of Simulation-Generated Time Series," Management Science, Vol. 13, No. 7, March 1967.

5. Gafarian, A. V. and Ancker, C. J., Jr., "Mean Value Estimation from Digital Computer Simulation," Operations Research, Vol. 6, pp. 25-44, 1966.

6. Gross, D. and Harris, C. M., Fundamentals of Queueing Theory, New York: John Wiley & Sons, 1974.

7. Hsu, D. A. and Hunter, J. S., "Analysis of Simulation-Generated Responses Using Autoregressive Models," Management Science, Vol. 24, No. 2, Oct. 1977, pp. 181-90.

8. Hunter, J. S. and Hsu, D. A., Simulation Model for New York Air Traffic Control Communications, Report N. FAA-RD-74-203, February 1975.

9. Naylor, T. H., Balintfy, J. L., Burdick, D. S. and Chu, Kong, Computer Simulation Techniques, New York: John Wiley & Sons, 1966.

10. Van Horn, R. L., "Validation of Simulation Results," Management Science, Vol. 17, No. 5, January 1971.