AD-A063 821 RAND CORP SANTA MONICA CALIF F/G 9/2

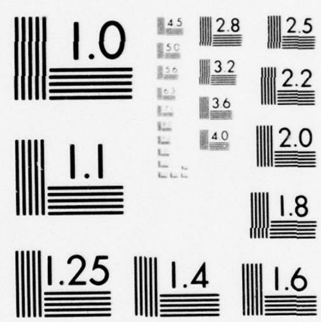
FORECASTING PROCESSING REQUIREMENTS OF STANDARD USAF BASE LEVEL--ETC(U)

JUL 78 E C POGGIO, N AINSLIE F49620-77-C-0023

UNCLASSIFIED RAND/R-2334-AF NL

1 OF 2

AD A063 821



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

RAND/R-2334-AF FORECASTING PROCESSING REQUIREMENTS OF STANDARD USAF BASE LEVEL COMPUTER SYSTEMS Poggio & Ainslie

LEVEL II

R-2334-AF
July 1978

Forecasting Processing Requirements of Standard USAF Base Level Computer Systems

Eugene C. Poggio, Naomi Ainslie

A Project AIR FORCE report
prepared for the
United States Air Force

Rand
SANTA MONICA, CA. 90406

The research reported here was sponsored by the Directorate of Operational Requirements, Deputy Chief of Staff/Research and Development, Hq USAF under Contract F49620-77-C-0023. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

Library of Congress Cataloging in Publication Data

Poggio, Eugene C 1945-
Forecasting processing requirements of standard USAF base level computer systems.

([Report] - Rand Corporation ; R-2334-AF)
1. United States. Air Force--Data processing.
2. Burroughs B3500 (Computer) I. Ainslie, Naomi L., 1921- joint author. II. United States. Air Force. III. Title. IV. Series: Rand Corporation. Rand report ; R-2334-AF.
AS36.R3 R-2334 [UG633] 081s [358.4'16'02854 78-15605
ISBN 0-8330-0046-2

Published by The Rand Corporation

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER R-2334-AF	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Forecasting Processing Requirements of Standard USAF Base Level Computer System.		5. TYPE OF REPORT & PERIOD COVERED Interim 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Eugene C. Poggio and Naomi Ainslie		8. CONTRACT OR GRANT NUMBER(s) F49620-77-C-0023
9. PERFORMING ORGANIZATION NAME AND ADDRESS The Rand Corporation 1700 Main Street Santa Monica, CA. 90406		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Project AIR FORCE Office (AF/RDQA) Directorate of Operational Requirements Hq USAF, Washington, D. C. 20330		12. REPORT DATE July 1978 13. NUMBER OF PAGES 116
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release; Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NO RESTRICTIONS

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Data Processing	Forecasting
Computers	Planning
Information Systems	Air Force

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

SEE REVERSE SIDE

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

This system

~~d~~develops a methodology for forecasting computer requirements to support functional systems that are operational at the time of forecasting. The approach is to develop regression models that can relate past base characteristics to past computer requirements, so that one can then employ estimates of future base characteristics, as obtained from planning documents, to predict future workload. Although originally developed within the context of the Burroughs 3500, the methodology should be useful for estimating, for any computer system, the effects on the processing requirements of activity changes within the Air Force and of organizational and basing options that the Air Force may consider. It can also be used for predicting the processing requirements of such alternatives as a regional or central system, which, the report finds, may well yield large benefits. (WH)

(14) RAND / R-2334-AF

(11) Jul[redacted]978

(12) 13 pp.

(6)

Forecasting Processing Requirements of Standard USAF Base Level Computer Systems.

(9) Interim rept.,

(10) Eugene C./Poggio Naomi/Ainslie

A Project AIR FORCE report prepared for the United States Air Force

(15) F49620-77-C-ØØ23

Rand
SANTA MONICA, CA. 90406

296 600

mt

PREFACE

In the past, Rand has worked with the Air Staff and the Air Force Data Systems Design Center (AFDSDC) in forecasting the processing requirements of USAF base-level computers. Numerous models have been developed for that purpose.

One set of Rand models developed, documented, and used several years ago produced forecasts that allowed for changes in the pattern of base-level activity. These models could also be used in forecasting the requirements of a regional computer system. That work, although made available to AFDSDC staff members at the time, was not published. The methodology was developed within the context of the Burroughs 3500 used at base level and with data that are no longer current. Nevertheless, the methodology should still be useful for estimating the effects on computer processing requirements of activity changes occurring within the Air Force, and of organizational and basing options that the Air Force is currently considering. For this reason, the report is being published at this time.

The report establishes a methodology of very general applicability. As the activities and composition of a base change, so do its processing requirements. An increase in the authorized flying hours on base, for example, will usually increase the processing requirements to support the maintenance data system. The methodology developed here allows one to predict the computer processing needed to support functional systems for which past operational data are available. (The requirements to support any new functional system must be handled by other means.) The technique employs multiple regression to relate computer processing requirements to base characteristics. The models developed use past processing data together with data on base characteristics taken from planning documents. By using planned authorization figures for the future, the models can forecast the corresponding future workload.

The description of how these models were developed should be useful to anyone tasked with developing similar models. For example, the approach of determining good predictors for each major functional system as a means of obtaining candidate predictors for the entire system, and the use of separate models for different commands, should continue to be fruitful. Many of the variables found to be good predictors are likely to remain so. The discussion of a model with an autoregressive structure should be useful to anyone fortunate enough to have data that are longitudinal as well as cross-sectional.

Finally, the methodology developed here may prove to be the best instrument for predicting the processing requirements resulting from the regionalization (or centralization) of USAF base-level computer systems. To make such a prediction, one has to assume that the processing required to support several bases within a region will be the same as the processing required to support a single hypothetical base of the same size and composition as the several bases combined. If this assumption is not valid, the prediction based on it will have to be adjusted. The possible benefits to be gained from regionalization, as indicated in this report, are sufficiently large to warrant reexamination with current data and within the context of the reorganization and basing options currently being investigated by the Air Force.

Because of a Congressional restriction on Rand's logistics research for fiscal year 1978, all work primarily concerned with improving the efficiency of various functional areas in logistics was discontinued as of October 1, 1977. The present report, as noted above, documents unpublished Rand research from FY 1977 and prior years. It is being produced under the Project AIR FORCE project "Documentation of FY 77 Logistics Research."

SUMMARY

To design future computer systems, one must forecast the workload they will need to support. This report develops a methodology for forecasting the processing requirements of USAF base level computers to support those functional systems operational at the time of forecasting. The approach to forecasting these requirements is to develop models that can relate past base characteristics to past computer requirements, so that one can then employ estimates of future base characteristics, as obtained from planning documents, to predict future workload. The report deals specifically with only one of the USAF's standard base level computer systems: the Burroughs 3500. It presents a set of models that can now be employed to make such forecasts with high precision for A level installations of the Burroughs 3500.

Multiple linear regression analysis is used in constructing the mathematical model employed and estimating its parameters. Two measures of system requirements are modeled: total direct time (the primary measure) and total number of inputs and outputs. These are the dependent variables of the regression analysis. The base characteristics by which they are to be modeled, the independent variables, are selected on the basis of expected correlation with the dependent variables and the availability of estimated or planned figures for the future. The latter constraint confines our choices of characteristics to the manpower and aircraft authorizations of several years ago from three sources: the Manpower Authorization File (HAF-PRM(AR)7102); the *USAF Program: Bases, Units and Priorities* (known as the PD); and the *USAF Program: Aerospace Vehicles and Flying Hours* (known as the PA).

Because of several complications with B level installations (configured with 150K bytes of core), the models are built only for A level installations (configured with 100K bytes). To select a set of candidate independent variables by which to model the total system load at the A level installations, we first identify the major functional systems supported on the Burroughs 3500. We ascertain the function of each and then select base characteristics thought to be correlated with

the generated load. Using these characteristics as candidate independent variables, we then build intermediate models for the direct time charged to individual functional systems, the aim being to isolate the best predictive variables for each.

General models of total monthly requirements are then built for the A level bases. This is done by using stepwise regression to select those requirements that are the best predictors of the overall load. The total direct time model is based on the manpower authorizations for travel (a subfunction of accounting and finance), civil engineering, and mission equipment maintenance. This model achieves an R^2 of .72; that is, it explains 72 percent of the observed variance in monthly total direct time. The standard error of the estimate is 24 hours, the mean monthly direct hours being 248. The general model for the total monthly inputs and outputs (I/Os) is based on accounts control (another accounting and finance subfunction), civil engineering, medical material, and airmen; it achieves an R^2 of .80 and a standard error equal to only 12 percent of the mean number of I/Os.

Command-specific models are then developed in an attempt to improve upon the already good fit obtained with the general models. The 72 A level installations on which the general models are built are partitioned into those belonging to SAC, TAC, and Other Commands. For each "command," we obtain the best single predictors of direct time charged to the eleven major functional systems. Models of both direct time and total number of I/Os are built for each, again using stepwise regression to select the best predictive variables for the major systems. The direct time models for SAC and TAC have very high R^2s of .81 and .89, respectively. The standard errors are 17 and 13 hours, only 7 and 5 percent of the corresponding means. The direct time model for the Other Commands has an R^2 equal to .72 and a standard error equal to 12 percent of the corresponding mean. For the SAC and TAC I/O models, we obtain remarkable R^2s of .84 and .95, with standard errors equal to only 8 and 5 percent of the respective means. The Other Commands I/O model achieves an R^2 of .79 and a standard error of 15 percent of the corresponding mean.

The precision of estimation obtainable with both the general and command models is shown by the presentation of 90-percent confidence intervals computed under various assumptions of shift in the independent variables. The approximate half-width of intervals for the general direct time model is 17 percent of mean monthly direct time. The SAC, TAC, and Other Commands direct time models have half-widths equal to only 12, 10, and 19 percent of this mean, respectively. The SAC and TAC I/O models similarly improve on the precision obtainable with the general I/O model, while the Other Commands I/O model does only slightly less well.

Consequently, the command models substantially improve on the precision of estimation obtainable with the general models, an improvement sufficiently large as to recommend their use over the general models. Moreover, the level of precision obtainable with the command models is judged to be excellent.

The report discusses forecasts based on a model with an autoregressive structure, taking into account any autocorrelation between observations at a single installation. Since the data used for this study were entirely cross-sectional, there was no need to be concerned with autocorrelation in building the models. In forecasting, however, incorporation of autocorrelation into the model theoretically allows us to use the observed residuals to increase the precision of predictions. But since we cannot estimate the autocorrelation without longitudinal data, we simply formulate a model and recommend forecasting based on "bounding" assumptions concerning the value of the autocorrelation.

The report explains how the models can be used to predict the processing requirements for a regional (or central) computer system. Predictions made with the models indicate the possibility of very substantial savings with regionalization.

It is recommended that the command models be verified on an independent data base and then maintained by periodic verification and, if necessary, updating. It is further recommended that the models so maintained be used annually to forecast the processing requirements at each installation for the five subsequent years.

The techniques of this report should be used in evaluating *any* alternative computer system the Air Force may be considering.

It is thought that efforts to improve the models would most fruitfully be spent decomposing the Other Commands models into several command-specific models and estimating the autocorrelation to be used in forecasting with a model possessing autoregressive structure. Only small improvements are likely to be gained through using alternative variables or making additional observations.

The most profitable area for future work probably lies in extending these models to the few A level installations omitted from this analysis, to the B level installations, and to the Univac 1050. As the Air Force deems necessary, extensions could also be made to encompass only those currently operational systems that will continue to be operational in the future, and to systems not yet operational.

ACKNOWLEDGMENTS

Several of the author's colleagues at Rand have made valuable contributions to this study. Stephen Drezner and Richard Hillestad constructively reviewed an earlier draft of this report. Will Harriss did a careful and thoughtful editing of the manuscript. Finally, for having formulated the problem and for his encouragement, a special debt of gratitude is owed to Irving Cohen.

CONTENTS

TABLES

I. INTRODUCTION

The problem addressed in this study is the forecasting of processing requirements of Air Force standard base level computer systems. Such forecasting is critical in assessing the necessity for and benefits to be gained from alternatives to today's data processing systems. To design tomorrow's systems, one must "size" the workload those systems will need to support. It may be that tomorrow's requirements will necessitate only the addition of a few peripheral devices at several bases; on the other hand, an entirely new system may be needed worldwide. Possibly, a regional rather than a base level system, or, perhaps, several dedicated functional systems would better fill tomorrow's needs. To determine the best alternative system, one must be able to forecast the processing requirements of that system.

Those processing requirements consist of two components: (1) workload from functional systems operational at the time of forecasting and (2) workload from functional systems not yet operational.* The first is the primary concern of the present study. This workload is not likely to remain unchanged in the future; rather, it is likely to vary as a function of the amount of activity in the functional area supported. For example, more computer time would be required to support the military personnel system if the base military population increased. Analogously, a decrease in flying activity would likely result in a decrease in computer requirements to support the maintenance data system.

Forecasting the workload of a functional system not yet operational requires an analysis of a type not touched upon in this report. The technique of this report does have a potential application, however, as a complement to this other analysis. This will be discussed briefly in Sec. VIII.

---

*It is assumed throughout this report that the software supporting each operational functional system remains unchanged; any major change in software would require the system to be treated as one not yet operational.

The primary objective of this study is the development of a general methodology by which to forecast the workload from the functional systems operational at the time of forecasting. A secondary objective is the development of specific estimating equations that can now be used to forecast this workload. To accomplish the first objective, this report first describes a general mathematical model and then employs it to develop estimating equations. The high precision of forecasts with the equations obtained testifies to the power of the methodology. The specific equations developed in this process satisfy the second objective.

Our basic approach to forecasting these processing requirements is illustrated in Fig. 1. We first develop a mathematical model relating past base characteristics to the corresponding processing requirements. For example, we relate the number of airmen on a base to the processing time charged at that base. Then taking the planned authorization figures as estimates of future base characteristics, we can use the model to predict tomorrow's computer system requirements. If, for example, we built the model suggested relating the airmen population to processing time, we would then simply use the planned authorization figures for airmen at a base to estimate the future workload requirements at that base.

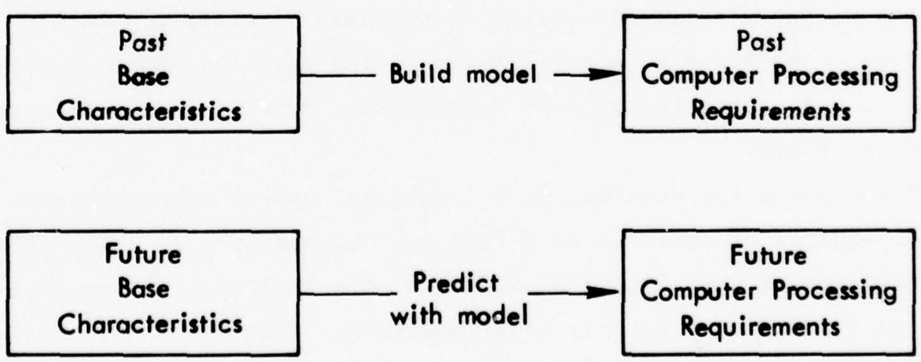


Fig. 1 — Basic approach of the study

Currently, the USAF has two standard base level computer systems. The primary one is a third-generation multiprogrammer, the Burroughs 3500. It supports a wide variety of functions, including military personnel, civil engineering, accounting and finance, transportation, and maintenance. The other is a second-generation Univac 1050-II; it is a "dedicated" computer supporting only supply.

This report deals specifically only with the Burroughs 3500. The wide variety of functional systems supported on the 3500 allows us to assess the generality of the methodology of the report to different functional areas without needing to analyze the 1050. In fact, the success of our efforts with the 3500 strongly suggest that an application of the methodology to the 1050 would produce excellent results. The Univac computer was not examined, primarily because hardware utilization data are lacking for this machine. Some limited work has been done, however, in attempting to predict such surrogates for utilization as number of inputs and number of transactions; this will be briefly discussed in Sec. VIII.

Section II of this report formulates the mathematical model and describes its components. In Sec. III, we begin the development of the models. Those base characteristics that should best predict total workload are isolated by building intermediate models for each of the major functional systems supported on the Burroughs 3500. We employ these variables in Sec. IV to develop general models of total processing requirements. In Sec. V, command-specific models of these requirements are built. Section VI discusses forecasting with the models. In Sec. VII, the use of these models in predicting the processing requirements of a regional computer system is discussed. Section VIII closes with recommendations on verification, maintenance, use, improvement, and extensions of the models.

II. THE MATHEMATICAL MODEL AND ITS COMPONENTS

BASIC MODEL

The mathematical model we employ is that of multiple linear regression analysis.* The theoretical relationship is assumed to be of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon ,$$

where Y is the dependent variable, X_1, X_2, ..., X_p are the independent variables, and ε is a normally distributed random error term with mean 0 and variance σ^2. For our application, Y is a measure of processing requirements, and the X_i are the base characteristics to which we attempt to relate Y. These are, respectively, the right- and left-hand sides of Fig. 1.

Frequently, the linearity assumption is made simply to attempt to approximate a function thought to be much more complex. For the relationships we wish to model, it seems reasonable that the actual functional relationships may be of this form. For example, the processing requirements to support a pay system may reasonably be expected to be a linear function of the number of people the system supports. Since the total requirements would simply be the sum of such functions, it too would be linear.

The assumption of the regression analysis model allows us to draw upon the techniques of that analysis both in building the model and in making predictions with it. Using observations of past data for the variables, we obtain least squares estimates of the β_i; replacing the β_i in the equation by these estimates, we have an estimate of the mathematical relationship between the Y and the X_i. This can then be used to predict future values of Y. A normality assumption for the probability distribution of the random error term, coupled with an

*N. R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley and Sons, New York, 1966.

assumption of independent observations, then allows us to test the statistical significance of the regression and of the individual coefficients, and to obtain confidence bounds for the predicted values. Having thus defined the model, we need to examine its components: the dependent variables, the independent variables, and the unit of analysis.

Dependent Variables: Direct Time and Total Number of I/Os

The choice of the dependent variables is simply the choice of that which we wish to predict, constrained only by the availability of data. As discussed in the Introduction, our interest in "sizing" future workload is to be able to assess the need for, and benefits of, alternative systems. Hence, our dependent variable should be the measure of "size" best suited to making such assessments, among those measures now available to us.

For the Burroughs 3500, there is fortunately an ample choice of such measures because of a rich data source: the Workload Analysis Model of the Air Force Data Systems Design Center.* Each of the utilization measures in Table 1 is obtainable for each installation. Those measures most appropriate to sizing are number of runs, direct time, total time on good runs, and total time.† Number of runs has the disadvantage that it measures size only indirectly; one must then "size" an average run in units of time to understand the capacity required. If, for example, it were predicted that future requirements would increase to 30,000 runs per month, one could not determine whether current hardware would handle the increase without knowing the average processing time of each run. Consequently, we prefer to restrict ourselves to the remaining measures, each expressed directly in units of time.

The last, total time, has the advantage of allowing an immediate assessment of saturation; a predicted total time exceeding an average

---

*See Capt. J. W. Kurina and First Lt. Joel Kizer, *Workload Analysis Model, Report #1*, OR Project A09-72, AFDSDC/SYO (AFDAA), Gunter AFB, Alabama.

†See note to Table 1 for definition of direct time and total time.

Table 1

MACHINE UTILIZATION DATA AVAILABLE
FOR THE BURROUGHS 3500

Number of runs (executions)

Number of bad runs (executions that did not end in normal-end-of-job)

Direct time (in hours)

Prorated time (in hours)

Total (chargeable) time on good runs (in hours)

Total (chargeable) time on bad runs (in hours)

Total (chargeable) time (in hours)

Overlay count total time (in hours)

Number of cards read

Number of cards punched

Number of lines printed

Number of logical tape records processed

Number of physical tape records processed

Number of logical disk records processed

Number of physical disk records processed

NOTE: Direct time is defined as the sum of the central processing unit time spent actually performing the instructions of an application program and the operating system (Master Control Program) processing time generated as a result of an application program's requests (e.g., I/Os and overlay calls). Prorated time is the time when the operating system is in a "nothing-to-do" loop while all programs in the mix are waiting for I/Os. Total or chargeable time equals the sum of direct and prorated time.

of 24 hours per day would obviously indicate the need for more computing capacity (as would a somewhat smaller number, in view of the reality of downtime, both planned and unplanned). To make similar assessments, each of the other two measures requires a conversion to total time by the estimation and addition of another quantity: to a predicted direct time an estimate of prorated time must be added, and likewise, to a predicted total time on good runs must be added an estimate of the time required on bad runs. In each case, the estimated addition would likely be simply a fixed percentage of the first predicted quantity. For example, to a predicted direct time, we might simply add 72 percent* as the estimate of the corresponding prorated time, in order to predict the total chargeable time.

However, because prorated time depends even upon operating and scheduling procedures,† we prefer to eliminate it from the dependent variable. Thus, since direct time is precisely the difference between total and prorated time, we choose it as our primary dependent variable.

In addition, we select total number of inputs and outputs (I/Os) of any type‡ as a secondary dependent variable. Though it is not a particularly useful measure of load on the major system components, it provides a single, albeit gross, measure of workload on peripheral equipment.

The data for these two variables, on which models are developed, are for the period January through June 1972. Each variable is measured in terms of mean monthly utilization, where the mean for each installation is taken across all months in this period for which data

---

*According to the WAM Report, this is the ratio of prorated time to direct time for A level bases in each of the months of January, February, and March 1972.

†"The nature and interdependence of prorated time and direct time are being investigated and correlated with the various B3500 computer configurations in the Air Force. A determination is being made of the extent to which operating and scheduling procedures of a DPI [Data Processing Installation] can reduce prorated time." See Kurina and Kizer, p. 44.

‡This is defined to be the sum of number of cards read, number of cards punched, number of lines printed, number of physical tape records processed, and number of physical disk records processed.

were available. Table 2 presents the means and standard deviations of these variables across all installations for which we are to build our models.*

Table 2

DEPENDENT VARIABLES

Total Direct Time (hours per month)		Total Number of I/Os (millions per month)	
Mean	Standard Deviation	Mean	Standard Deviation
248	45.1	22.7	5.95

NOTE: These means and standard deviations are obtained by first averaging, for each installation, the monthly utilizations as measured by these two variables for each of the months for which data were available in the last half of fiscal year 1972, and then averaging these across the 72 A level installations listed in Table 4.

Independent Variables: Manpower and Aircraft

The choice of independent variables should be based on two criteria: (1) expected correlation with the dependent variable and (2) availability of estimates for the future. If no correlation exists, the candidate variable will be of no help in building the model. If a correlation exists but no estimates are available, a model can be built relating it to the dependent variable, but the model cannot be used for prediction. It would be like building a model to find that your car gets fifteen miles to the gallon and then trying to estimate the gallons required for a trip without knowing how far you are going.

In choosing base characteristics as independent variables, then,

---

*As discussed below, the models are developed for the 72 A level installations listed in Table 4. Of these, data were available for all six of the months at 28 installations, for five at 24 installations, for four at 12, for three at 4, for two at 3, and for only one at 1.

the second criterion restricts us to characteristics for which estimated, or planned, future figures are available. Such figures are generally available for only two classes of base characteristics: manpower and weapon systems. The former are obtainable in aggregated fashion in *USAF Program: Manpower and Organization* (known as the PM) and in disaggregation in the Manpower Authorization File (HAF-PRM(AR) 7102). The latter are available in detail in *USAF Program: Bases, Units and Priorities* (known as the PD) and in *USAF Program: Aerospace Vehicles and Flying Hours* (known as the PA). In this report, we use the Manpower Authorization File, the PD, and the PA as our sources of independent variables.

The former provides the number of authorized personnel at each base for each quarter from the present one to that five years hence. The authorizations are given for each unique combination of functional account code, personnel identity code, Air Force specialty code, rated position indicator, military grade, civilian-employment-category-group code, major command, organization kind, and organization type.*

The PD gives the authorized number of aircraft by type and the number of each type of missile at each base. The PM provides information which, in conjunction with the information provided by the PD, allows the computation of the authorized quarterly flying hours for each aircraft type at each base.** All of these figures are also available for the present and for each quarter for the next five years.

Choosing variables from these sources is, by the first criterion, a question of choosing those thought to be correlated with the dependent variable. We first determine the major functional systems

---

*For definitions of these terms, see U.S. Department of the Air Force, *Data Automation, Data Elements and Codes*, Vol. XII, General Purpose, Washington, D.C., June 1971.

**In developing our forecast models, we actually used authorized quarterly flying hour figures provided directly to us by USAF. Subsequently, we were unable to precisely reproduce these figures from the PM. We have, however, found a method that reproduces them very closely which can and should be used in making predictions with those of the models developed herein that require flying hour figures in order to compute a base maintenance cost variable. Appendix A describes this method in conjunction with a description of the method of computation of this maintenance cost variable.

supported on the Burroughs 3500, ascertain the specific functions performed by each, and then, on the basis of the functions performed, choose for each system those base characteristics thought to be correlated with the system workload. For example, the load from the civilian pay system (which computes pay and leave entitlements based on input time and attendance reports) is probably related to the number of civilians, or perhaps the number of personnel in the civilian pay subfunction of accounting and finance. Having selected these base characteristics, we then build intermediate models of the same general form discussed thus far. The direct time charged to the individual functional system is the dependent variable, and the base characteristics are candidate independent variables. Finally, those base characteristics found to be the best predictors of the direct time for the individual systems are used as candidate independent variables for modeling both total direct time* and total number of I/Os.

The actual selection of the candidate independent variables for each major functional system is discussed in Sec. III. Table 3 presents a composite list of these. The models are built based upon the values for the third quarter of fiscal year 1972, chosen to correspond with the period of our dependent variables, the last half of the fiscal year. Table 3 also presents the means and standard deviations of each, computed across the installations for which the models are built.

All but the last three variables in the list are authorized manpower figures derived from the Manpower Authorization File. Those followed by a parenthesized code are the authorizations in the functional account indicated by the code, with the Xs simply indicating aggregation across all digits in the corresponding position. For example, civilian pay is the authorized manpower in the functional account 1513; accounting and finance operation (151X) is the authorization in functional accounts 1510 through 1519. The authorizations for transport, fighter, bomber, and reconnaissance and trainer pilots are

---

*We will sometimes use the expression "total direct time" to distinguish the direct time as summed across all systems from the direct time of individual systems; it is not to be confused with total time. See pp. 5-7.

Table 3

INDEPENDENT VARIABLES

Variable	Mean	Standard Deviation
Accounting and Finance Operation (151X)[a]	65.18	29.74
Accounts Control (1511)	7.17	2.99
Military Pay (1512)	17.82	18.10
Civilian Pay (1513)	4.42	2.38
Travel (1514)	7.67	3.73
Commercial Services (1515)	14.17	6.27
Paying and Collecting (1518)	5.24	1.74
Management Analysis (152X)	6.88	4.08
Budget (153X)	5.64	2.99
Data Automation/Operational (154X)	27.72	17.34
Audit Staff (155X)	5.14	2.53
Data Control/Consolidated Base Personnel Office (165X)	18.58	5.56
Base Civilian Personnel (1680)	14.24	9.16
Civil Engineering Staff (17XX)	2.22	8.12
Mission Equipment Maintenance (2XXX)[b]	1651.50	887.44
Chief of Maintenance (21XX)	129.97	63.82
Organization Maintenance (22XX)	365.83	206.09
Flight Line/Site Maintenance (2210)	317.35	183.62
Periodic/Mobile Maintenance (222X)	34.76	40.22
Field Maintenance (23XX)	543.19	312.35
Avionics Maintenance (24XX)	257.65	212.99
Munitions Management (25XX)	156.89	202.65
Ground Communications/Electronics Maintenance (26XX)	118.03	90.52
Ground-Launched Missile Equipment Maintenance (28XX)	78.22	205.59
Ground Support Equipment Maintenance (29XX)	1.71	7.59
Aircraft Crew (3110)	228.07	255.00
Vehicle Operations (4210)	59.07	24.52
Vehicle Maintenance Control (4240)	14.69	12.62
Vehicle Maintenance (4241)	46.96	25.69
Civil Engineering (44XX)	428.26	140.22
Pavements and Grounds (444X)	46.61	21.56
Structures (445X)	74.28	31.11
Mechanical-Civil Engineering (446X)	58.22	28.63
Electrical-Civil Engineering (447X)	29.04	12.00
Electrical Power Production (448X)	11.19	9.47
Sanitation (449X)	26.51	16.36
Medical (5XXX)	259.08	132.41
Medical Material (5110)	10.28	5.81
Hospital/Dispensary Services (52XX)	133.17	83.54
Physicians (5201)	13.46	9.12

Table 3--continued

Variable	Mean	Standard Deviation
Total Base Population	4,860.60	1,778.50
Total Military[c]	4,007.40	1,633.00
Total Civilian[d]	853.19	576.47
Airmen[e]	3,410.00	1,441.10
Officers[f]	597.46	274.29
Transport Pilots[g]	78.19	108.77
Fighter Pilots[h]	34.36	54.06
Bomber Pilots[i]	19.50	50.01
Reconnaissance and Trainer Pilots[j]	38.86	92.61
Rated Pilots[k]	164.92	131.57
Aircraft[l]	70.72	49.36
Flying Hours[m]	8,731.90	7,422.20
Base Maintenance Cost ($)[n]	2,385,400.00	1,672,700.00

NOTE: The footnotes to follow define the variables of the table. The terms and codes employed in these definitions are documented in AFM 300-4, Vol. 12. The means and standard deviations are for the A level installations listed in Table 4 for the third quarter of fiscal year 1972.

[a]This, as well as the thirty-seven subsequent variables, is the manpower authorized for the indicated function code. The Xs indicate aggregation across all digits in the corresponding position.

[b]Depot Maintenance (27XX) is excluded.

[c]Personnel identity code is "O" or "A."

[d]Personnel identity code is "G" or "P."

[e]Personnel identity code is "A."

[f]Personnel identity code is "O."

[g]Personnel identity code is "O" and AFSC is "10."

[h]Personnel identity code is "O" and AFSC is "11."

[i]Personnel identity code is "O" and AFSC is "12."

[j]Personnel identity code is "O" and AFSC is "13."

[k]Personnel identity code is "O" and Rated position indicator is "1."

[l]Total authorized aircraft of any type.

[m]Total authorized quarterly flying hours for all types of aircraft.

[n]Sum across Model/Design/Series (MDSs) of products of quarterly flying hours with average base maintenance cost per flying hour (See Appendix A).

those corresponding to the Air Force specialty codes for officers of 10, 11, 12, and 13, respectively. The rated pilot authorization is taken to be those officers with a rated position indicator of 1, indicating an aircrew pilot; aircrew supervisory and operation control pilots are excluded. As for the last three independent variables, the first is simply the total number of authorized aircraft regardless of type; the second is the authorized quarterly flying hours for those aircraft. The third is the estimated base maintenance cost, computed by summing across weapon systems the products of quarterly flying hours and average base maintenance cost per flying hour for that system.*

UNIT OF ANALYSIS: THE DATA PROCESSING INSTALLATION

The unit of analysis in this study is the data processing installation; more precisely, it is the installation together with the activities it supports. Typically, this is simply the base on which the installation is located. The model is built upon corresponding values of dependent and independent variables observed for a set of installations (the observations of regression analysis); furthermore, it is for such installations that future values of the dependent variable are to be predicted, based upon estimated future values of the independent variables.

Each of the USAF's 116 installations of a Burroughs 3500 could potentially be used as an "observation" to help build the model. Of these, there are 77 A level (configured with 100K bytes of core) and 39 B level (configured with 150K bytes).†

Because of the difference in core size, the two levels cannot be indiscriminately pooled to build a single model. To pool them it would first be necessary to understand the effect of core size on the

---

*Appendix A describes the motivation behind the creation of this variable and the means by which to calculate it.

†These figures are for the period of our data, the last six months of fiscal year 1972. During this period, one B level installation, K. I. Sawyer, actually had a core size of 210K to accommodate the test of the Maintenance Information Control System (MMICS).

dependent variable. If it could be shown that there were no effect, the two could then be pooled; if there were an effect and it could be determined, it might be possible to reduce the dependent variable of the B level installations to A level equivalents in order to pool them. Such an analysis is beyond the scope of this study. Hence, we had the choice of developing models for both levels or for only one. Because of several problems associated with the B level installations, as discussed in Appendix B, we have restricted the analysis to A level installations.

Of the 77 A level installations, we dropped five. No data were available on one, Bergstrom, and four were intentionally omitted because of problems similar to those with the B installations. Two, Robbins and Griffiss, were dropped because B level installations also existed at the base. Newark and Los Angeles were omitted because of their unique missions.

Table 4 lists the remaining 72 installations. Of these, 22 are Strategic Air Command (SAC) bases; 17, Tactical Air Command (TAC); and the remaining 33 mostly Air Training Command, Air Force Europe, Military Airlift Command, and Air Force Systems Command. Corresponding values of the dependent and independent variables for each of these 72 installations are now to be employed as the observations on which to develop our models.

Table 4

USAF A LEVEL BURROUGHS 3500 INSTALLATIONS SELECTED TO BUILD MODEL

Strategic Air Command (SAC)

Anderson
Beale
Blytheville
Carswell
Castle
Davis Monthan
Dyess
Ellsworth
F. E. Warren
Fairchild
Grand Forks
Grissom
Lockbourne
Loring
Malmstrom
March
McCoy
Minot
Pease
Plattsburgh
Whiteman
Wurtsmith

Tactical Air Command (TAC)

Cannon
England
Forbes
George
Holloman
Homestead
Hurlburt
Little Rock
Luke
MacDill
McConnell
Mountain Home
Myrtle Beach
Nellis
Pope
Seymour Johnson
Shaw

Air Training Command (ATC)

Columbus
Craig
Laredo
Laughlin
Mather
Moody
Reese
Webb
Williams

Air Force Europe (AFE)

Aviano
Bentwaters
Bitburg
Incirlik
Lakenheath RAF
Rhein-Main
Torrejon
Upper Heyford RAF

Military Airlift Command (MAC)

Altus
Charleston
Dover
Lajes Field
McChord
McGuire

Air Force Systems Command (AFSC)

Brooks
Edwards
Kirtland
L. G. Hanscom
Patrick

Other

Hamilton, Air Defense Command (ADC)
Tyndall, ADC
Maxwell, Air University (AU)
Ching Chuan Kang, Pacific Air Force (PACAF)
Albrook, Southern Command (SC)

III. DETERMINING CANDIDATE INDEPENDENT VARIABLES FOR GENERAL MODELS

Having selected total direct time as the dependent variable of primary interest, we now determine a set of base characteristics to serve as candidate independent variables by which to develop a general model of total direct time. We begin by determining the major functional systems supported on the Burroughs 3500. We then ascertain their specific functions and select those base characteristics thought to be correlated with the generated workload. Intermediate models of the same form discussed previously are then built, with direct time charged to the functional system as the dependent variable and these base characteristics as candidate independent variables. In Sec. IV we use the best predictors for the individual systems to model total direct time and total number of I/Os.*

Table 5 lists the systems supported on the 3500.† About half are systems software or utility programs, and are disregarded as they are of no help in suggesting candidate independent variables by which to model the total load. Inasmuch as they are basically the overhead of the functional systems, it is reasonable to presume that a set of independent variables that predicts the total load of the functional

---

*Note that we use the best predictors of direct time charged to the major systems to model both total direct time and total number of I/Os. Instead, we could, of course, independently obtain the best predictors of number of I/Os for each of the systems and use these to model the total number of I/Os. We have declined to do so inasmuch as the high correlation between the direct time and total number of I/Os (r = .93 for our 72 A level installations in the last half of fiscal year 1972) implies that good predictors of direct time will also be good predictors of number of I/Os. Hence, the best predictors of direct time should serve also to model the number of I/Os. The results of this study bear this out. Though it is possible that even better results for the I/O model might be obtained by employing the alternative procedure, it is thought that any improvement obtained would be slight.

†This list includes individually only those systems for which WAM captures the utilization data elements given in Table 1. The remaining systems supported on the Burroughs 3500 are aggregated under "Other Standard Systems" and "Other Utility and Command-Unique Programs."

Table 5

SYSTEMS SUPPORTED ON THE BURROUGHS 3500

System Type	Code	System
U	NAB	AF Standard Utility System
S	NAC	Data Communications Control System
F	NAE	Base Level Military Personnel System
F	NAT	Base Engineer Automated Management System
F	NAV	Medical Material Management System
F	NAW	Aerospace Vehicle Status Reporting System
F	NBD	Maintenance Data Collection System
F	NBJ	Base Vehicle Reporting Subsystem
F	NBP	Flight Data Management System
F	NBQ	General Accounting and Finance System
F	NBS	Civilian Pay System
F	NBT	Joint Uniform Military Pay System
F	NBU	Accrued Military Pay System
S	NBZ	ADPS Program Management System
S	NCD	Program Distribution System
S	NDV	ADPE Utilization Recording and Reporting System
S	NIW	Hardware Diagnostic System
F	NMY	Civil Engineering Accounting System
F	NRA	Vehicle Integrated Management System
F	OST	Other Standard Systems
S	ASM	The Advanced Assembler
U	BAC	BACKUP (the tape-to-print utility program)
S	COB	The COBOL compiler
S	FOR	The FORTRAN compiler
U	PBD	PBDOUT (the disk-to-print utility program)
U	PCH	PCHOUT (the disk-to-punch utility program)
U	PR1	PRINTD (the new disk-to-print utility program)
U	PR2	PRINIT (the new tape-to-print utility program)
U	PR3	PUNCHD (the new disk-to-punch utility program)
F	OUC	Other Utility and Command-Unique Programs

NOTE: F indicates a functional system; S, systems software; and U, a utility program.

systems will also predict that of the systems software and utility programs and, hence, the total load. Thus, we should not need to concern ourselves with these.

Of the remaining (functional) systems, two are so small they are not worth considering. These are the Base Vehicle Reporting Subsystem, with an observed mean of .10 hours of direct time per month, and the Civil Engineering Accounting System, with a mean of .03 hours. The two "systems" indicated by the codes OST and OUC are not actually systems at all, but rather residual categories for (small) standard, and unique systems, respectively. These too are disregarded because they do little to suggest independent variables and because they are likely to be of negligible import in predicting the total load.

We are left with the eleven functional systems listed in Table 6 with their means and standard deviations of charged direct time. In developing the intermediate models, more attention should be paid to systems with larger standard deviations, since they are likely to contribute more to the variance of total direct time. The systems are listed and addressed in this order.

In building the intermediate models, the direct time charged to each functional system modeled is first regressed upon all of the corresponding candidate independent variables; the R^2 of this regression is the maximum that can be achieved with any combination of these variables.* We then regress the dependent variable individually on each of the candidate independent variables. This provides us with the best *single* predictor among the variables we have selected. It also reveals the degree of correlation of each of the independent variables with the dependent variable. Based upon these correlations and on the specific independent variables involved, regressions are then run with a variety of combinations of the variables. These are then assessed in terms of reduction in the standard error of the estimate obtained by employing

---

*Actually, in order to eliminate multicollinearity among the variables, only a linearly independent subset of these variables is employed. That is, a set of variables which can be expressed as linear combinations (for our purposes, as sums or differences) of others is omitted. The R^2 obtained is still the maximum to be obtained with any combination of the variables.

Table 6

MAJOR FUNCTIONAL SYSTEMS SUPPORTED ON THE BURROUGHS 3500

Code	System	Mean (hours per month)	Standard Deviation (hours per month)
NAE	Base Level Military Personnel System	59.94	16.08
NAT	Base Engineer Automated Management System	24.43	8.49
NBQ	General Accounting and Finance System	20.08	6.97
NRA	Vehicle Integrated Management System	8.89	3.91
NBD	Maintenance Data Collection System	7.01	3.27
NBS	Civilian Pay System	3.86	3.09
NBU	Accrued Military Pay System	2.47	3.08
NAV	Medical Material Management System	3.06	2.52
NBT	Joint Uniform Military Pay System	3.49	1.35
NAW	Aerospace Vehicle Status Reporting System	3.93	1.27
NBP	Flight Data Management System	2.50	1.17

NOTE: The means and standard deviations are based on the selected A level installations listed in Table 4 for the last six months of fiscal year 1972.

additional variables, and the significance of the partial F statistic used to determine whether the coefficient can statistically be considered significantly different from zero. For each functional system, the regressions are based on the 72 A level installations listed in Table 4, except that any for which the direct time charged to the system was zero is omitted. The one or more variables thought best able to predict the dependent variable are then noted (see Table 18 below), later to be used to model the total workload. Appendix C presents, for each system, the actual regression equation with the best single independent variable and, if different, that equation minimizing the standard error among all equations with each coefficient significant at the .10 level.

BASE LEVEL MILITARY PERSONNEL SYSTEM (NAE)

The largest functional system supported on the Burroughs 3500 is a military personnel system that provides a repository of personnel data with variable inquiry and report capabilities. This is the base level military personnel system, to which an average of 60 direct hours per month are charged. Table 7 presents the four independent variables we selected to relate to the direct time charged to this system. Of these, the last three are simply the people on which this system maintains files: the total military population, the airmen, and the officers. We distinguish between airmen and officers because their processing per capita might well differ. The first variable, data control personnel, was included because it is this function's responsibility to manage officer and airmen records.

The top line of the table shows us the best R^2 obtainable from any combination of the independent variables. It employs all of the independent variables, except those that can be expressed as linear combinations of others.* In the case at hand, Total Military is excluded because it is simply the sum of Airmen and Officers, and nothing can be gained by including it. With the other three variables, we obtain an R^2 of .64 and a standard error of 9.9. Equations (2), (3), (4), and (5) present the results of regressions with each of the four independent variables individually in an equation. In Eq. (4), we find that with the variable Airmen, we obtain an R^2 almost as high as that of the first equation and a standard error just slightly higher. In Eqs. (6), (7), (8), and (9) we try the equivalent of all pairs of our independent variables. Although there are actually six such pairs, the Airmen and Officers pair is equivalent both to Total Military and Airmen and to Total Military and Officers, inasmuch as any two of these variables determine the third. All three pairs would then have identical R^2s and standard errors. At any rate, the improvement obtained is not with these, but rather with Eq. (7). Here, with the addition of the Data Control variable, both the R^2 and the standard error are slightly improved, bringing them to about the level obtained with all three variables. Hence, for this military personnel system, Airmen is the best single predictor among our independent variables, with Data Control Personnel adding only a slight improvement. This is noted in the first row of Table 18.

---

*See footnote on p. 18.

Table 7

REGRESSIONS FOR DIRECT TIME CHARGED TO BASE LEVEL MILITARY PERSONNEL SYSTEM (NAE)

Number of Independent Variables in Equation	Equation	Independent Variables: Data Control/ Consolidated Base Personnel Office (165X)	Independent Variables: Total Military	Independent Variables: Airmen	Independent Variables: Officers	R^2	s
All	1	1	(a)	1	1	.643	9.890
One	2	2				.449	12.110
	3		3			.615	10.120
	4[b]			4		.627	9.956
	5				5	.258	14.050
Two	6	6	6			.631	9.979
	7[b]	7		7		.642	9.827
	8	8			8	.478	11.870
	9			9	9	.627	10.030

[a]Total military, being collinear with Airmen and Officers as their sum, is excluded.

[b]This regression equation is presented in Appendix C.

BASE ENGINEER AUTOMATED MANAGEMENT SYSTEM (NAT)

Known by its acronym BEAMS, this is the second largest functional system. It comprises four subsystems: cost accounting, labor, real property, and work control.* As listed in Table 8, the independent variables selected include Civil Engineering staff, all of Civil Engineering as well as six of its major subfunctions, Medical Manpower, and the Total Base Population. The reasons are obvious for including the Civil Engineering manpower categories; Medical Manpower was selected as a proxy for medical facilities, thought to possibly require much civil engineering support; and the Base Population was included as probably the best usable surrogate for utilized area. As to the latter, we would have preferred to use covered acreage or number of buildings, but these would not serve our purposes, since, to the best of our knowledge, future estimates are nonexistent.

We see that with all variables included, and R^2 of .50 and a standard error of 5.8 are obtained. In the runs with the single independent variables, we see that the best predictors of direct time are, as might be expected, the Civil Engineering Manpower functions. Furthermore, we note that by simply using the Civil Engineering variable in Eq. (3) we do slightly better, as measured by standard error, than with all the variables or with any attempted combinations of them.

GENERAL ACCOUNTING AND FINANCE SYSTEM (NBQ)

The General Accounting and Finance System provides the base level accounting and finance operation records and reports. It includes general funds, stock funds, industrial funds, and disbursement and collection control. The system is large, with a mean of 20 direct hours per month. As shown in Table 9, the independent variables we selected are,

---

*This fact was unknown to us at the time of selecting independent variables. It may well be that civil engineering functions more directly related to those subsystems (such as Civil Engineering Cost Account (4444), Civil Engineering Operations and Maintenance (443X), Real Estate Management (4413), and Civil Engineering Work Control (4431)) would provide better predictors. This omission may account for the relatively low R^2 obtained for this system.

Table 8

REGRESSIONS FOR DIRECT TIME CHARGED TO BASE ENGINEER AUTOMATED MANAGEMENT SYSTEM (NAT)

Number of Independent Variables in Equation	Equation	Independent Variables										R^2	s
		Civil Engineering Staff (17XX)	Civil Engineering (44XX)	Pavements and Grounds (444X)	Structures (445X)	Mechanical --Civil Engineering (446X)	Electrical --Civil Engineering (447X)	Electrical Power Production (448X)	Sani-tation (449X)	Medical (5XXX)	Total Base Popu-lation		
All	1	1	1	1	1	1	1	1	1	1	1	.502	5.784
One	2	2										.007	7.604
	3[a]		3									.468	5.566
	4			4								.274	6.504
	5				5							.391	5.954
	6					6						.285	6.456
	7						7					.322	6.283
	8							8				.012	7.587
	9								9			.171	6.950
	10									10		.015	7.574
	11										11	.197	6.841
Two	12		12								12	.469	5.602
Three	13			13	13	13						.412	5.939
Four	14			14	14	14					14	.420	5.946

[a] This regression equation is presented in Appendix C.

Table 9

REGRESSIONS FOR DIRECT TIME CHARGED TO GENERAL ACCOUNTING AND FINANCE SYSTEM (NBQ)

Number of Independent Variables in Equation	Equation	Independent Variables: Accounting and Finance Operation (151X)	Accounts Control (1511)	Military Pay (1512)	Civilian Pay (1513)	Travel (1514)	Commercial Services (1515)	Paying and Collect-ing (1518)	Management Analysis (152X)	Budget (153X)	Data Automation/ Operational (154X)	Audit (155X)	Total Base Popu-lation	R^2	s
All	1	(a)	1	1	1	1	1	1	1	1	1	1	1	.735	3.925
One	2	2												.430	5.335
	3[b]		3											.605	4.441
	4			4										.058	6.859
	5				5									.442	5.278
	6					6								.504	4.978
	7						7							.283	5.983
	8							8						.148	6.521
	9								9					.022	6.998
	10									10				.441	5.282
	11										11			.247	6.133
	12											12		.301	5.910
	13												13	.119	6.631
Two	14	14								14				.496	5.051
	15	15									15			.455	5.256
	16	16										16		.463	5.215
	17		17		17									.639	4.273
	18		18			18								.665	4.118
	19		19				19							.660	4.150
	20				20	20								.590	4.559
	21				21		21							.557	4.740
	22					22	22							.564	4.699
	23	23	23											.640	4.273
	24	24			24									.523	4.916
	25	25				25								.526	4.900
	26		26							26				.631	4.322
	27				27					27				.525	4.906
	28					28				28				.546	4.787
Three	29		29			29	29							.698	3.940
Four	30	30								30	30	30		.530	4.950
	31[b]		31		31	31	31							.713	3.866
Six	32		32		32	32				32	32	32		.708	3.962
Seven	33		33		33	33	33			33	33	33		.722	3.894

[a] This variable, being approximately collinear with the following six, is excluded.

[b] This regression equation is presented in Appendix C.

with one exception, the subfunctions of the General Accounting and Finance operation. Three of these, Materiel Accounting and Finance (functional account 1516), Cost (1517), and Accounting and Finance/Staff (1519), were omitted because their values are so small as to be of little use, the largest having a mean of 3.3 hours. The one other variable included is Total Base Population, which may be useful if the workload on the accounting and finance system is closely related to the size of the base.

In Eq. (1) we find that we can obtain an R^2 of .74 and a standard error of 3.92. Looking at Eqs. (2) through (13), we find that the best single variable is Accounts Control, with an R^2 of .60 and a standard error of 4.44. Adding the variable Travel in Eq. (18) reduces the standard error to 4.12. The standard error achieves its minimum of 3.87 among the regressions run in Eq. (31), by the addition of two further variables, Civilian Pay and Commercial Services. The partial F tests of the four coefficients in this equation are each significant at least at the .07 level. Each of these additional variables is noted in Table 18, though it seems likely that only Travel may be useful in modeling the total load.

VEHICLE INTEGRATED MANAGEMENT SYSTEM (NRA)

The Vehicle Integrated Management System is designed to provide the functional areas of Vehicle Operations and Maintenance with those products required to manage the base vehicle fleet. The system maintains files and produces summary reports on vehicle use and operating and maintenance costs. The authorizations in Vehicle Operations, Vehicle Maintenance, and Vehicle Maintenance Control were selected as candidate independent variables. As indicated in Table 10, three more aircraft-related variables were included, because the amount of ground transportation activity may be related to the amount of flying activity.

In the first equation, we find the maximum R^2 to be obtained with these variables is .41. As might be expected, the next six equations show that Vehicle Maintenance and Vehicle Operation are the better predictors, the former being best with a standard error of 3.03, smaller than that with all of the variables. The last equation achieves a

Table 10

REGRESSIONS FOR DIRECT TIME CHARGED TO VEHICLE INTEGRATED MANAGEMENT SYSTEM (NRA)

Number of Independent Variables in Equation	Equation	Independent Variables: Vehicle Operations (4210)	Vehicle Maintenance Control (4220)	Vehicle Maintenance (4241)	Aircraft	Flying Hours	Base Maintenance Cost	R^2	s
All	1	1	1	1	1	1	1	.411	3.068
One	2	2						.308	3.203
	3		3					.022	3.807
	4[a]			4				.379	3.035
	5				5			.022	3.808
	6					6		.024	3.805
	7						7	.028	3.797
Two	8	8		8				.399	3.007

[a]This regression equation is presented in Appendix C.

slightly lower standard error, but the coefficient of the Vehicle Operations variable tests as significantly different from zero only at the .13 level. Hence, we disregard the equation and merely note in Table 18 that Vehicle Maintenance is the best single predictor.

MAINTENANCE DATA COLLECTION SYSTEM (NBD)

The Maintenance Data Collection System processes maintenance data collected on aircraft, missiles, munitions, and a variety of other equipment. The system's output consists of production reports, failure data reports, and scheduling reports. For independent variables, we drew primarily from the subfunctions of mission equipment maintenance; additionally, we included Aircrew, the four pilot categories, Rated Pilots, Aircraft, Flying Hours and Base Maintenance Cost.

As can be seen from the first row of Table 11, the maximum R^2 to be obtained is .78. The best single variable is Base Maintenance Cost, with an R^2 of .59 and a standard error of 2.07. Field Maintenance, Chief of Maintenance, Organization Maintenance, and all of Mission Equipment Maintenance (excluding Depot Maintenance), each do almost as well. A slight improvement is obtained by employing Mission Equipment Maintenance as well as Maintenance Cost in Eq. (24), achieving a standard error of 1.97, with the coefficients of both variables being highly significant. Of the many regressions run, none provides a smaller standard error and has each of its coefficients significant at the .10 level. A multitude of combinations being possible,* it is likely that a better fit could be obtained, but it would undoubtedly be only slightly better. We know, for example, that the R^2 cannot exceed the .78 of the first equation. For our purposes, the Base Maintenance Cost variable alone, perhaps with the addition of Mission Equipment Maintenance, will probably suffice.

CIVILIAN PAY SYSTEM (NBS)

Using time and attendance reports, the Civilian Pay System computes civilian pay and leave statements. To model the workload on this system,

---

*There are 2^{20} = 1,048,576 possible combinations.

Table 11

REGRESSION FOR DIRECT TIME CHARGED TO MAINTENANCE DATA COLLECTION SYSTEM (NBD)

Number of Independent Variables in Equation	Equation	Independent Variables								
		Mission Equipment Maintenance (2XXX)[a]	Chief of Maintenance (21XX)	Organization Maintenance (22XX)	Flight Line/Site Maintenance (2210)	Periodic Mobile Maintenance (222X)	Field Maintenance (23XX)	Avionics Maintenance (24XX)	Munitions Management (25XX)	Ground Communications/ Electronics Maintenance (26XX)
All	1	(b)	1	(c)	1	1	1	1	1	1
One	2	2								
	3		3							
	4			4						
	5				5					
	6					6				
	7						7			
	8							8		
	9								9	
	10									10
	11									
	12									
	13									
	14									
	15									
	16									
	17									
	18									
	19									
	20									
	21[d]									
Two	22	22								
	23	23								
	24[d]	24								
	25			25			25			
	26									
	27									
	28									
Three	29			29			29			
	30			30			30			
	31			31			31			
	32				32	32	32			
	33				33	33				
Four	34	34		34			34			
	35			35			35	35	35	
Five	36	36	36		36	36	36			
	37	37			37	37	37			
	38	38								
	39			39			39	39	39	
	40				40	40	40	40	40	
	41									
Six	42	42	42		42	42	42			
	43	43								
	44			44			44			
Seven	45			45			45			
	46				46	46				
Eight	47	47		47			47			
Nine	48	48	48		48	48	48			
	49	49			49	49	49			
Ten	50	50	50		50	50	50			

[a]Depot Maintenance (27XX) is excluded.

[b]This variable, being approximately collinear with the next ten variables, is excluded.

[c]This variable, being approximately collinear with the two variables to its right, is excluded.

[d]This regression equation is presented in Appendix C.

Ground Launched Missile Electronics Maintenance (28XX)	Ground Support Equipment Mainte-nance (29XX)	Aircraft Crew (3110)	Transport Pilots	Fighter Pilots	Bomber Pilots	Reconnais-sance and Trainer Pilots	Rated Pilots	Aircraft	Flying Hours	Base Mainte-nance Cost	R^2	s
1	1	1	1	1	1	1	1	1	1	1	.783	1.735
											.479	2.333
											.515	2.251
											.489	2.311
											.415	2.472
											.274	2.754
											.576	2.106
											.254	2.792
											.206	2.880
											.012	3.213
11											.016	3.207
	12										.004	3.225
		13									.178	2.930
			14								.059	3.136
				15							.260	2.781
					16						.012	3.213
						17					.006	3.223
							18				.087	3.089
								19			.157	2.969
									20		.146	2.987
										21	.592	2.066
								22			.504	2.294
									23		.548	2.189
										24	.636	1.965
		26									.576	2.119
										26	.601	2.058
								27		27	.592	2.081
									28	28	.606	2.044
								29			.588	2.105
									30		.589	2.104
										31	.626	2.007
											.636	1.979
										33	.604	2.065
										34	.639	1.985
											.609	2.068
											.662	1.936
										37	.675	1.898
			38	38	38	38					.589	2.135
										39	.650	1.970
											.679	1.888
			41	41	41	41				41	.676	1.896
										42	.676	1.912
			43	43	43	43				43	.682	1.894
			44	44	44	44					.638	2.019
			45	45	45	45				45	.688	1.889
			46	46	46	46				46	.708	1.827
			47	47	47	47				47	.688	1.905
			48	48	48	48					.706	1.866
			49	49	49	49				49	.723	1.823
			50	50	50	50				50	.720	1.818

we selected only the three independent variables in Table 12. The first is that subfunction of the accounting and finance operations responsible for civilian pay. The second is a subfunction of Personnel, responsible for conducting civilian personnel programs. The third is obvious.

Table 12

REGRESSIONS FOR DIRECT TIME CHARGED TO CIVILIAN PAY SYSTEM (NBS)

Number of Independent Variables in Equation	Equation	Independent Variables			R^2	s
		Civilian Pay (1513)	Base Civilian Personnel (1680)	Civilian Population		
All	1	1	1	1	.801	1.354
One	2	2			.625	1.828
	3		3		.742	1.517
	4[a]			4	.758	1.468
Two	5	5	5		.759	1.478
	6[a]	6		6	.795	1.362
	7		7	7	.761	1.470

[a]This regression equation is presented in Appendix C.

With all three variables, we obtain an R^2 of .80 and a standard error of 1.4. The best single independent variable is the Civilian Population, which does almost as well with an R^2 of .76 and a standard error of 1.47. The addition of Civilian Pay makes a slight improvement, achieving approximately the levels obtained with all three variables. The coefficients of both variables test as significantly different from zero at levels less than .002. Hence, we note that the best single predictor is Civilian Population and that adding Civilian Pay yields a small improvement.

ACCRUED MILITARY PAY SYSTEM (NBU)

The Accrued Military Pay System computes and processes military pay data; its output consists of pay lists and payment vouchers, and general and expense ledger data. As independent variables, we employ the Military Pay subfunction of accounting and finance, Data Control, the Total Military authorization, and the authorizations for Airmen and Officers.

In the first equation of Table 13, we see that an R^2 as high as .90 can be obtained, with a standard error of 1.01. By simply employing the variable Military Pay in Eq. (2), we obtain an R^2 of .88 and a standard error of 1.07. All equations with smaller standard errors have the coefficient of at least one variable not testing as significant at the .10 level. Hence, we simply note Military Pay as the best predictor of direct time charged to this system.

MEDICAL MATERIAL MANAGEMENT SYSTEM (NAV)

This functional system provides for the maintenance of accountable medical stock records for base medical supply accounts and in-use stock records for all medical facilities. We selected as one of our independent variables the authorization for Medical Material, that subfunction responsible for operation and management of the medical supply accounts. We also chose the entire Medical function, the Hospital/Dispensary Services subfunction, Physicians, and Total Base Population.

As indicated in Table 14, Medical Material is the best single predictor with an R^2 of .60. With all of the variables in an equation, an R^2 of .70 and a standard error of 1.41 are achieved, but a smaller standard error is obtained by simply employing Medical Material and Physicians, the coefficients of both variables being significant at the .0001 level. Hence, we note in our table that Medical Material is the best single variable, but the inclusion of Physicians provides a smaller standard error.

JOINT UNIFORM MILITARY PAY SYSTEM (NBT)

The interface with a central site system to update pay and leave accounts is provided by the Joint Uniform Military Pay System, known as JUMPS. Pay checks, leave and earning statements, W-2 forms, and base level management reports concerning pay and leave are all products of this system. As independent variables to relate to the load generated

Table 13

REGRESSIONS FOR DIRECT TIME CHARGED TO ACCRUED MILITARY PAY SYSTEM (NBU)

Number of Independent Variables in Equation	Equation	Independent Variables: Military Pay (1512)	Data Control/ Consolidated Base Personnel Office (165X)	Total Military	Airmen	Officers	R^2	s
All	1	1	1	(a)	1	1	.903	1.011
One	2[b]	2					.884	1.067
	3		3				.310	2.606
	4			4			.253	2.712
	5				5		.226	2.760
	6					6	.310	2.607
Two	7	7	7				.889	1.056
	8	8		8			.885	1.078
	9	9			9		.884	1.079
	10	10				10	.891	1.049
	11		11		11		.322	2.614
	12		12			12	.389	2.481
Three	13	13	13	13			.895	1.043

[a] Total military, being collinear with airmen and officers as their sum, is excluded.

[b] This regression equation is presented in Appendix C.

Table 14

REGRESSIONS FOR DIRECT TIME CHARGED TO MEDICAL MATERIAL MANAGEMENT SYSTEM (NAV)

Number of Independent Variables in Equation	Equation	Independent Variables					R^2	s
		Medical (5XXX)	Medical Material (5110)	Hospital/ Dispensary Services (52XX)	Physicians (5201)	Total Base Population		
All	1	1	1	1	1	1	.704	1.415
One	2	2					.010	2.394
	3[a]		3				.604	1.588
	4			4			.070	2.434
	5				5		.006	2.515
	6					6	.006	2.516
Two	7	7	7				.693	1.409
	8	8		8			.143	2.354
	9	9			9		.406	1.960
	10		10	10			.692	1.411
	11[a]		11		11		.701	1.389
	12			12	12		.431	1.917

[a]This regression equation is presented in Appendix C.

by this system, we employ the same variables as for the Accrued Military Pay System. Whereas for the latter, the authorized manpower in military pay is by far the best predictor, here, as can be seen in Table 15, Airmen and Total Military are both much better, the latter being slightly preferable with a standard error of 1.00, less than that of the first equation. As no improvement in standard error is obtained by using two variables in Eqs. (7)-(13), we list only Total Military as a predictor for this system.

AEROSPACE VEHICLE STATUS REPORTING SYSTEM (NAW)

The primary function of this system, active at all bases possessing aircraft or missiles, is to report inventory changes and status and operational data. The base characteristics chosen to relate to the load generated by this system are all aircraft-related: number of aircraft, flying hours, rated pilots, pilots by type of aircraft, flight line maintenance personnel, and base maintenance costs. As can be seen from Table 16, none of these is very highly correlated, the best being number of aircraft in Eq. (9), with an R^2 of .44. In Eq. (1), we find that the maximum R^2 to be obtained with these variables is .61, with a corresponding standard error of .81. We do as well in Eq. (31), for which the standard error is also .81, by using the four pilot groups and flying hours. In our summary table, we note both Number of Aircraft as the best single predictor and these five variables, though the latter are likely to be of little benefit in predicting the total load.

FLIGHT DATA MANAGEMENT SYSTEM (NBP)

The Flight Data Management System generates, for various Air Force activities, current files on the flying experience of each person assigned or attached for flying. The processing requirements to support this system might reasonably be thought to be correlated both with the amount of flying at the base and the number of people assigned for flying. As independent variables, we chose the total authorized Flying Hours and the authorizations for Aircraft Crew and Rated Pilots, as well as for the four pilot categories. We also included the Flight Line/Site Maintenance crew and the Maintenance Cost. In all, these are the same

Table 15

REGRESSIONS FOR DI[illegible] TIME CHARGED TO JOINT UNIFORM MILITARY PAY SYSTEM (NBT)

Number of Independent Variables in Equation	Equation	Independent Variables					R^2	s
		Military Pay (1512)	Data Control/ Consolidated Base Personnel Office (165X)	Total Military	Airmen	Officers		
All	1	1	1	(a)	1	1	.474	1.018
One	2	2					.145	1.270
	3		3				.315	1.136
	4[b]			4			.466	1.003
	5				5		.459	1.010
	6					6	.257	1.184
Two	7	7	7				.316	1.144
	8	8		8			.466	1.010
	9	9			9		.459	1.017
	10	10				10	.294	1.162
	11		11	11			.472	1.005
	12		12		12		.466	1.010
	13		13			13	.372	1.096
Three	14	14	14	14			.474	1.011

[a] Total military, being collinear with airmen and officers as their sum, is excluded.

[b] This regression equation is presented in Appendix C.

Table 16

REGRESSIONS FOR DIRECT TIME CHARGED TO AEROSPACE VEHICLE STATUS REPORTING SYSTEM (NAW)

Number of Independent Variables in Equation	Equation	Independent Variables										R^2	s
		Flight Line/Site Maintenance (2210)	Aircraft Crew (3110)	Transport Pilots	Fighter Pilots	Bomber Pilots	Reconnaissance and Trainer Pilots	Rated Pilots	Aircraft	Flying Hours	Base Mainte-nance Cost		
All	1	1	1	1	1	1	1	1	1	1	1	.610	0.810
One	2	2										.190	1.089
	3		3									.014	1.201
	4			4								.043	1.184
	5				5							.236	1.058
	6					6						.005	1.206
	7						7					.144	1.120
	8							8				.053	1.177
	9[a]								9			.442	0.904
	10									10		.234	1.059
	11										11	.135	1.125
Two	12	12	12									.352	0.981
	13	13						13				.194	1.094
	14	14							14			.452	0.902
	15	15								15		.291	1.026
	16	16									16	.192	1.095
	17		17					17				.089	1.163
	18		18						18			.445	0.908
	19		19							19		.286	1.030
	20		20								20	.350	0.983
	21							21	21			.447	0.906
	22							22		22		.241	1.062
	23							23			23	.137	1.132
	24								24	24		.448	0.905
	25								25		25	.443	0.910
	26									26	26	.242	1.061
Five	27	27		27	27	27	27					.523	0.861
	28		28	28	28	28	28					.496	0.885
	29			29	29	29	29	29				.481	0.898
	30			30	30	30	30		30			.526	0.858
	31[a]			31	31	31	31			31		.574	0.814
	32			32	32	32	32				32	.509	0.874

[a]This regression equation is presented in Appendix C.

as those independent variables employed for the Aerospace Vehicle Status Reporting System.

In Table 17, we see that Rated Pilots is the best predictor, with an R^2 of .45 and a standard error of .86. Using all the variables in the equation, the maximum R^2 obtained is .60. In Eq. (28), and R^2 of .57 and a standard error of .77, lower than that with all of the variables, is obtained with the three variables, Bomber Pilots, Reconnaissance and Trainer Pilots, and Flying Hours. Both the best predictor and these three are noted in the table, though again the slight improvement with the latter is likely to be of little value.

SUMMARY

Table 18 compiles the results for the eleven major functional systems. The second column presents, for each system, the best single predictor of charged direct time among the candidate independent variables we selected. The next column lists variables that improve the model when used jointly with the best single variable. The final column lists variables that improve the model when used in lieu of the best single variable.

Table 17

REGRESSIONS FOR DIRECT TIME CHARGED TO FLIGHT DATA MANAGEMENT SYSTEM (NBP)

Number of Independent Variables in Equation	Equation	Independent Variables: Flight Line/ Site Maintenance (2210)	Aircraft Crew (3110)	Transport Pilots	Fighter Pilots	Bomber Pilots	Reconnaissance and Trainer Pilots	Rated Pilots	Aircraft	Flying Hours	Base Mainte- nance Cost	R^2	s
All	1	1	1	1	1	1	1	1	1	1	1	.600	0.782
One	2	2										.268	0.986
	3		3									.186	1.040
	4			4								.354	0.926
	5				5							.009	1.147
	6					6						.171	1.050
	7						7					.011	1.146
	8[a]							8				.449	0.856
	9								9			.071	1.111
	10									10		.226	1.014
	11										11	.275	0.982
Two	12	12	12									.440	0.869
	13	13						13				.304	0.969
	14	14							14			.462	0.851
	15	15								15		.268	0.993
	16	16									16	.338	0.944
	17		17					17				.304	0.968
	18		18						18			.511	0.812
	19		19							19		.283	0.983
	20		20								20	.339	0.944
	21					21				21		.290	0.979
	22							22	22			.450	0.861
	23							23		23		.456	0.856
	24							24			24	.484	0.834
	25								25	25		.260	0.998
	26								26		26	.275	0.989
	27									27	27	.311	0.963
Three	28[a]					28	28			28		.569	0.768
Five	29	29		29	29	29	29					.496	0.843
	30		30	30	30	30	30					.512	0.829
	31			31	31	31	31	31				.477	0.858
	32			32	32	32	32		32			.491	0.847
	33			33	33	33	33			33		.575	0.774
	34			34	34	34	34				34	.519	0.824

[a]This regression equation is presented in Appendix C.

Table 18

SUMMARY OF BEST INDEPENDENT VARIABLES FOR MAJOR FUNCTIONAL SYSTEMS

System Code	Best Variable	Additional Variables	Alternative Variables
NAE	Airmen	Data Control/CBPO (165X)	
NAT	Civil Engineering (44XX)		
NBQ	Accounts Control (1511)	Civilian Pay (1513) Travel (1514) Commercial Services (1515)	
NRA	Vehicle Maintenance (4241)		
NBD	Base Maintenance Cost	Mission Equipment Maintenance (2XXX)[a]	
NBS	Civilian Population	Civilian Pay (1513)	
NBU	Military Pay (1512)		
NAV	Medical Material (5110)	Physicians (5201)	
NBT	Total Military[b]		
NAW	Aircraft		
			Flying Hours Transport Pilots Fighter Pilots Bomber Pilots Reconnaissance and Trainer Pilots
NBP	Rated Pilots		Bomber Pilots Reconnaissance and Trainer Pilots Flying Hours

[a]Depot Maintenance (27XX) is excluded.

[b]This variable is omitted in modeling the total processing requirements in the next section, since its extremely high correlation with airmen (r = .99) implies little can be gained by the inclusion of both.

IV. DEVELOPING GENERAL MODELS

Having selected predictors for each major functional system, we now use these as candidate independent variables in modeling the total processing requirements, where this total is the load not only from the major functional systems of Table 6 but from all of the systems listed in Table 5. They are first used to model our primary measure of load, total direct time, and then to model the total number of I/Os.*

The selection from these variables is made by stepwise regression.† This procedure enters variables into a regression equation one at a time, at each step introducing the next variable making the largest contribution among those not yet entered. Because a variable inserted at one step may become superfluous after new variables are entered, the procedure reexamines the equation at each step and eliminates any variables no longer making a significant contribution.

Statistically, the procedure begins by computing the simple correlation of each independent variable with the dependent, placing that with the highest correlation in the regression first. It then adds variables one at a time, at each stage entering that which has the highest partial correlation with the dependent variable given those already in the equation or, equivalently, that which has the largest partial F statistic. It then calculates partial F statistics for all of the variables thus far included and removes from the model any for which the statistic is not significant. The procedure terminates when none of the partial F statistics of the variables not yet entered are statistically significant. We have arbitrarily set the level of significance for termination at .10, though we will frequently mention the points at which the procedure terminates for higher levels of significance.

---

*See first footnote on p. 16.

†Draper and Smith, pp. 171-172.

MODELING TOTAL DIRECT TIME

In developing a model for direct time, we use all the variables in Table 18 as candidate independent variables in a stepwise regression. Table 19 presents the results. The variable Airmen is entered at the first step. It is the single best predictor of our dependent variable, achieving an R^2 of .54 and a standard error of 30.9. This standard error is only 12.5 percent of the 248 mean monthly direct hours. Table 19 lists the partial F statistics of the variables in the equation and tabulates their degrees of freedom.* For the first variable, the partial F is necessarily the same as the overall F; it is here equal to 81. Travel is entered in the second step, raising the R^2 to .67 and lowering the standard error to 26.2. At this point, the stepwise procedure terminates if the criterion for entry of variables is set at a significance level of .05 or less for the F statistic. If we allow a slightly less significant term to enter, the variable Civil Engineering comes into the equation, increasing the R^2 to .69 and decreasing the standard error to 25.6. The partial F statistics for both Airmen and Travel are still very high; the statistic for Civil Engineering is 3.8, significant at the .06 level. Mission Equipment Maintenance enters at Step 4, producing an equation with an R^2 of .72 and a standard error of 24.4 With the inclusion of this variable, Airmen no longer contributes to the model, its partial F statistic being an insignificant 0.3. Step 5 therefore eliminates Airmen, giving us an equation with Travel, Civil Engineering, and Mission Equipment Maintenance. The coefficient of each variable is significant at the .002 level,† and the equation achieves an R^2 of .72 and a standard error of 24.3.

---

*Each partial F statistic is distributed as the F-distribution with one degree of freedom for the numerator and the indicated degrees of freedom for the denominator. The latter equals the number of observations minus the number of parameters estimated. Since the single constant term and one coefficient for each variable are estimated, the (indicated) degrees of freedom for the denominator simply equals the number of observations minus the quantity one plus the number of independent variables.

†Note that we have obtained an equation with each of the coefficients significant at least at the .002 level, even though in one step

Table 19

STEPWISE REGRESSION FOR TOTAL DIRECT TIME

Step	Partial F Statistics of Independent Variables Entered (Removed): Airmen	Travel (1514)	Civil Engineering (44XX)	Mission Equipment Maintenance (2XXX)[a]	Degrees of Freedom	Number of Variables	R^2	s
1	81.0				70	1	.536	30.9
2	68.9	28.9			69	2	.673	26.2
3	35.1	22.5	3.8		68	3	.691	25.6
4	0.3	28.7	9.2	8.1	67	4	.724	24.4
5	(0.3)	29.0	11.1	47.2	68	3	.723	24.3

[a]Depot maintenance (27XX) is excluded.

Several steps of this procedure provide useful models of total direct time. The first equation, with only the Airmen variable, will suffice for any purpose requiring only gross estimation. The Airmen variable is very satisfying as a predictor, since one would expect it to be a reasonable surrogate for overall base activity, and it is the best predictor for the functional system for which the variability of direct time is largest.

Typically, however, the improvement obtained by using more variables would be worthwhile. Since the Airmen variable in the equation of Step 4 is of little use, we are left to choose among the equations represented by Steps 2, 3, and 5. Step 3 is intuitively attractive because two variables in the equation, Airmen and Civil Engineering, are the best predictors for the two systems with largest variation in direct time, and the third variable, Travel, is the second best predictor of the system for which direct time variation is third largest.

---

we allowed a variable whose coefficient was significant at only the .06 level to enter the equation. Such occurrences are frequent with the stepwise selection procedure, and occur often in the applications of the procedure contained in this report. They result from the deletion, or even the addition, of variables, which may raise the partial F statistics of the variables remaining, or previously included, in the equation.

The equation of Step 5 has both the Travel and Civil Engineering variables, but includes Mission Equipment Maintenance rather than Airmen. The Maintenance variable is logically a good predictor because it is very highly correlated with Airmen,* the best single predictor, and with the direct time charged to the maintenance data collection system.

We believe the equations of Steps 3 and 5 provide the best general models of the direct time. Being somewhat at a loss to choose between them--since the first is extremely satisfying intuitively, while the second is almost as much so and provides a slightly better fit to the data--we simply present both in Table 21 at the end of this section.

MODELING TOTAL NUMBER OF I/Os

In modeling the total number of I/Os we again use the stepwise regression procedure with all the candidate independent variables selected in Sec IV.† In Table 20, we find that Step 1 again selects Airmen as the best single independent variable. The R^2 for this regression is .53; the standard error indicated in the last column is 4.11×10^6, the mean value of total I/Os being 22.7×10^6. Adding an Accounts Control term in the second step improves the fit immensely, increasing the R^2 to .76 and decreasing the standard error to 2.93, less than 13 percent of the mean. The procedure terminates at this stage if we restrict ourselves to terms whose coefficients are significant at the .01 level, as measured by the partial F statistic. Allowing a slightly less significant variable, the procedure next enters Civil Engineering, which increases the R^2 to .78 and lowers the standard error to 2.82. The coefficient of the added variable has a partial F of 6.2, significant at the .02 level. The fourth step adds Medical Material, bringing the R^2 to just under .80 and reducing the standard error of the previous equation by 2 percent. The least significant coefficient, here that of the variable just entered, has a partial F of 4.1, significant at the .05 level. The procedure would stop at this point if the level for entry were set at .05. If less

---

*The sample correlation coefficient equals .94.

†See first footnote on p. 16.

Table 20

STEPWISE REGRESSION FOR TOTAL NUMBER OF I/O'S

Step	Partial F Statistics of Independent Variables Entered (Removed)							Degrees of Freedom	Number of Variables	R^2	$s : 10^6$
	Airmen	Accounts Control (1511)	Civil Engineering (44XX)	Medical Material (5110)	Data Control/ CBPO (165X)	Fighter Pilots	Travel (1514)				
1	78.7							70	1	.529	4.11
2	164.9	69.0						69	2	.765	2.93
3	73.4	53.0	6.2					68	3	.784	2.82
4	71.0	50.0	7.2	4.1				67	4	.797	2.76
5	20.8	34.8	9.2	4.6	3.9			66	5	.808	2.70
6	11.9	36.7	12.1	3.2	4.6	3.2		65	6	.817	2.66
7	10.8	11.8	12.3	2.7	1.6	5.1	3.6	64	7	.827	2.61
8	24.1	11.7	11.0	2.4	(1.6)	5.3	6.6	65	6	.822	2.62
9	23.2	12.1	10.8	(2.4)		7.1	7.2	66	5	.816	2.65

significant terms are again allowed to enter, Steps 5, 6, and 7 add Data Control, Fighter Pilots, and Travel. Each brings a slight increase in R^2 and decrease in standard error. Steps 8 and 9 then remove the Medical Material and Data Control terms, leaving us with a five-variable equation achieving an R^2 of .82 and standard error of 2.65. In fact, each term of this equation is significant at the .01 level. The procedure terminates at this step, even if we require a significance level of only .10 for entry.

Again we find that several steps of the selection procedure provide useful models, the equations of Steps 2, 3, 4, 5, 6, and 9 each being reasonable. We prefer the equation of Step 4. Three of its variables are the best single predictors for the three systems with largest variability, and the fourth variable is the best predictor of another major system. Furthermore, the partial F statistics of the coefficients of the variables are all significant at the .05 level, all but one being significant at the .01 level. The equation itself is presented in Table 21.

EXAMINATION OF THE MODELS

The two direct time equations and the single I/O equation we have selected are given in Table 21. Each equation is presented with its R^2, the standard error of the estimate, the standard error as a percent of the mean, the F statistic (with the degrees of freedom for its numerator and denominator, respectively), and the significance level of the F statistic (denoted P). All three equations appear reasonable. The variables included in each, as discussed above, are all intuitively very satisfying. Furthermore, all of the coefficients are positive. Hence, an increase in any variable, which we would expect to result in a larger processing requirement, also results in a larger forecasted requirement.

For each model, plots were made of each independent variable versus the residuals* and of the fitted dependent variable versus the residuals,

---

*The residuals are the differences between the actual and the fitted values of the dependent variable.

Table 21

GENERAL MODELS OF TOTAL DIRECT TIME AND TOTAL NUMBER OF I/Os

Direct Time Models

Model 1 $Y = 137.5 + .01586X_1 + 4.231X_2 + .05574X_3$

where Y = total direct time	$R^2 = .691$
X_1 = airmen	$s = 25.6$
X_2 = travel (1514), and	s as % of mean = 10.3
X_3 = civil engineering (44XX)	$F(3,68) = 50.6$
	$P = .000000$

Model 2 $Y = 137.3 + 4.522X_1 + .08127X_2 + .02489X_3$

where Y = total direct time	$R^2 = .723$
X_1 = travel (1514)	$s = 24.3$
X_2 = civil engineering (44XX), and	s as % of mean = 9.8
X_3 = mission equipment maintenance	$F(3,68) = 59.2$
(2XXX), excluding depot	$P = .000000$
maintenance (27XX)	

I/O Model

$$Y = 3419000 + 2513X_1 + 826300X_2 + 8427X_3 + 114900X_4$$

where Y = total I/Os	$R^2 = .797$
X_1 = airmen	$s \div 10^6 = 2.76$
X_2 = accounts control (1511)	s as % of mean = 12.2
X_3 = civil engineering (44XX), and	$F(4,67) = 65.6$
X_4 = medical material (5110)	$P = .000000$

NOTE: The estimated variance-covariance matrices corresponding to each of these regressions are presented in Appendix D.

to check for any nonlinearity or heteroscedasticity,* that is, a non-constant variance about the regression. Neither could be detected.

SUMMARY

The equations of Table 21 provide credible models of direct time and number of I/Os for the 72 A level installations. One direct time model achieves an R^2 of .72 and a standard error of 24 hours, only 10 percent of the mean direct time. The I/O model has an R^2 of .80 and a standard error equal to 12 percent of the mean number of I/Os. The high degree of fit obtained with these models is depicted in Fig. 2, which plots the fitted versus the actual values of the dependent variables. The fit would be "perfect" if all the points lay precisely on the diagonal line.

---

*The fact that the dependent variables were measured as sample means based upon a *varying* number of months implies that some heteroscedasticity must exist. The sampling error in the estimates of the means, which decreases as a function of the number of months on which the estimate is based, contributes to the variance of the error term. Hence, those observations based upon larger numbers of months must have smaller error variance. Since the variation in the number of months is small (all but eight observations were based upon from four to six months), and since the contribution to the variance of the error term from the sampling error is thought to be small, it is felt that this heteroscedasticity is, in all likelihood, negligible. An analysis of the residual terms from the second direct time model showed that, for this model at least, no heteroscedasticity could be detected as a function of the number of months on which the dependent variable was measured.

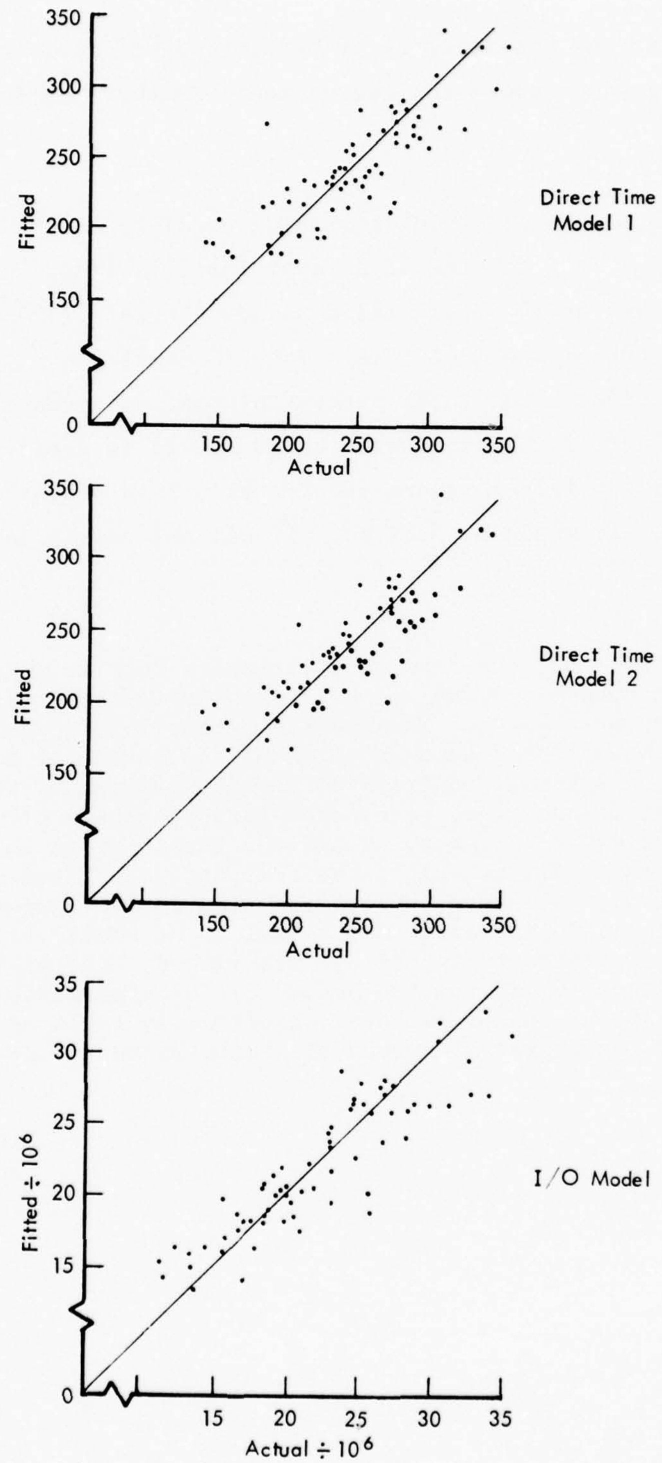


Fig. 2 — Plots for the general models

V. DEVELOPING COMMAND MODELS

Having built general models of direct time and number of I/Os in the last section, we now develop command specific models to see if an improved fit can be obtained. The regression analysis model is again employed to model both direct time and number of I/Os, the distributions of which are given by command in Table 22. As before, in selecting independent variables we first determine good predictors of direct time for the major systems and then use stepwise regression to select those to model the total load. The 72 A level installations are partitioned into three sets of observations: the 22 owned by SAC, the 17 by TAC, and the 33 owned by other commands.* Each set is used to model both dependent variables for the corresponding command.

Table 22

DEPENDENT VARIABLES BY COMMAND

Command	Total Direct Time (hours per month)		Total Number of I/Os (millions per month)	
	Mean	Standard Deviation	Mean	Standard Deviation
SAC	249	36.9	22.7	4.29
TAC	270	36.7	25.3	5.12
Other Commands	235	50.1	21.3	6.91
All	248	45.1	22.7	5.95

DETERMINING CANDIDATE INDEPENDENT VARIABLES

We again use the variables chosen in Sec. III as likely to be correlated with the loads from the major functional systems. Here, however, we restrict our choice to the single best variable for each

---

*We will frequently use the term "commands" to refer loosely to these three owning command categories.

command. We do not attempt the combinations of variables used previously, as it was thought that little would be gained by so doing.

Table 23 presents the variables found to be most highly correlated with the direct time charged to each major system for each command. To the direct time charged to the Military Personnel System (NAE) at the SAC installations, for example, Total Military was found to correlate most highly among all variables listed in Table 7. As indicated, the R^2 is .71. Among the variables listed in Table 8, Civil Engineering best predicts BEAMS (NAT) workload for both SAC and Other Commands, achieving R^2s of .44 and .41, respectively. The three columns of variables in this table are now input into the stepwise selection procedure to model both direct time and number of I/Os for the three commands.* The regression equations obtained with the procedure are presented at the end of this section.

MODELING TOTAL DIRECT TIME

We begin by developing the direct time models.

SAC Installations

To build a model for the SAC installations, the procedure selects from among the variables listed in the first column of Table 23. As indicated in Table 24, the first step enters Airmen, the variable included first in building the general model of the last section. Here the R^2 obtained is .59 and the standard error only 24.2. This one variable provides a regression for the SAC installations for which the standard error is as small as for the general model previously obtained. The procedure would terminate at this step, if we allowed only variables with partial F statistics significant at the .05 level to be included. Letting a slightly less significant term enter, we add Vehicle Maintenance, raising the R^2 to .66. Chief of Maintenance is next inserted,

---

*A comparison of the best command predictors with the overall predictors in Table 18 shows that they are often identical. Further, an analysis showed that, if not identical, the overall predictors typically perform almost as well in predicting direct time charged for the three commands. This suggests that we might do almost as well by selecting from the variables of Table 18 with a stepwise procedure to build the command models.

Table 23

BEST INDEPENDENT VARIABLES FOR THE MAJOR FUNCTIONAL SYSTEMS BY COMMAND

System Code	SAC		TAC		Other Commands	
	Best Variable	R^2	Best Variable	R^2	Best Variable	R^2
NAE	Total Military	.707	Total Military	.700	Airmen	.522
NAT	Civil Engineering (44XX)	.438	Total Base Population	.422	Civil Engineering (44XX)	.411
NBQ	Travel (1514)	.501	Travel (1514)	.489	Accounts Control (1511)	.642
NRA	Vehicle Maintenance (4241)	.556	Base Maintenance Cost	.114	Vehicle Maintenance (4241)	.439
NBD	Chief of Maintenance (21XX)	.364	Mission Equipment Maintenance (2XXX)[a]	.694	Mission Equipment Maintenance (2XXX)[a]	.766
NBS	Base Civilian Personnel(1680)	.362	Total Civilian	.661	Total Civilian	.846
NBU	Military Pay (1512)	.970	Military Pay (1512)	.891	Officers	.841
NAV	Total Base Population	.119	Hospital/Dispensary Services (52XX)	.681	Medical Material (5110)	.792
NBT	Airmen	.508	Total Military	.485	Total Military	.391
NAW	Base Maintenance Cost	.265	Fighter Pilots	.544	Aircraft	.620
NBP	Bomber Pilots	.547	Flying Hours	.150	Base Maintenance Cost	.556

[a]Depot Maintenance (27XX) is excluded.

Table 24

STEPWISE REGRESSION FOR SAC TOTAL DIRECT TIME

Step	Partial F Statistics of Independent Variables Entered (Removed): Airmen	Vehicle Maintenance (4241)	Chief of Maintenance (21XX)	Degrees of Freedom	Number of Variables	R^2	s
1	29.0			20	1	.592	24.2
2	13.9	3.8		19	2	.660	22.6
3	0.03	17.5	14.3	18	3	.811	17.4
4	(0.03)	35.1	39.9	19	2	.810	16.9

raising the R^2 further to .81 and lowering the standard error to 17.4. Finally, Airmen is removed, having been made superfluous with the entry of Chief of Maintenance. We are left with a model based only on Vehicle Maintenance and Chief of Maintenance, achieving an R^2 of .81 and a standard error of 16.9. The partial F statistics of the variables are both significant at the .00001 level. The actual regression equation is presented in the first row of Table 30 and is discussed at the end of this section.

TAC Installations

The stepwise procedure is next applied to the second column of variables in Table 23 to model direct time for the TAC bases. The first variable entered, as indicated in Table 25, is the Total Base Population, for which a remarkable R^2 of .80 and standard error of 16.8 are obtained. The second step includes Maintenance Cost, its partial F being significant at the .05 level. The R^2 is thus increased to .85. On the final step, Mission Equipment Maintenance is inserted, raising the R^2 to just under .90 and lowering the standard error to 13.4; the partial F statistics of all three variables are significant at the .05 level.

Table 25

STEPWISE REGRESSION FOR TAC TOTAL DIRECT TIME

Step	Partial F Statistics of Independent Variables Entered (Removed)			Degrees of Freedom	Number of Variables	R^2	s
	Total Base Population	Base Maintenance Costs	Mission Equipment Maintenance (2XXX)[a]				
1	61.1			15	1	.803	16.8
2	13.5	4.4		14	2	.850	15.2
3	19.5	10.8	5.1	13	3	.892	13.4

[a]Depot Maintenance (27XX) is excluded.

Other Command Installations

Table 26 gives the results of applying the stepwise procedure to model direct time for Other Command installations. Airmen is included first, Vehicle Maintenance second. The equation obtained has an R^2 of .64 and a standard error of 30.9. We found, however, that we obtain a substantially improved fit by using the three variables (Travel, Civil

Table 26

STEPWISE REGRESSION FOR OTHER COMMANDS TOTAL DIRECT TIME

Step	Partial F Statistics of Independent Variables Entered (Removed)		Degrees of Freedom	Number of Variables	R^2	s
	Officers	Vehicle Maintenance (4241)				
1	32.7		31	1	.513	35.5
2	26.3	10.7	30	2	.641	31.0

Engineering, and Mission Equipment Maintenance) of the second direct time model of Table 21.*

The regression equation with these as independent variables, based upon the 33 observations of the Other Command installations, has an R^2 of .72 and a standard error of 24.3. Two of the partial F statistics are significant at the .02 level; the third is significant at the .12 level. This equation is presented in Table 30.

MODELING TOTAL NUMBER OF I/Os

We now model the number of I/Os for the three commands, again using the candidate independent variables listed in Table 23.†

SAC Installations

For the SAC installations, the stepwise procedure requires only the two steps in Table 27. As was the case for SAC direct time, Airmen is entered first, here achieving and R^2 of .69. Civil Engineering is included next, bringing the R^2 to .84 and achieving a standard error of 1.83×10^6, only 8 percent of the mean number of I/Os for the SAC installations. The partial F statistics for both variables are significant at the .001 level.

TAC Installations

For the TAC installations, the selection of variables is made from the second column of Table 23. The stepwise procedure begins with the Total Base as indicated in Population Table 28. An extraordinary R^2 of .926 is obtained with this single variable; the standard error is 1.44×10^6, only 6 percent of the mean for the TAC installations. The procedure would terminate with this one variable in the model if we set the criterion for entry at the .05 level. Allowing the insertion of

---

*For each of the three commands, we ran a regression of direct time on these three independent variables and a regression of total number of I/Os on the four independent variables (Accounts Control, Civil Engineering, Medical Material, and Airmen) of the general I/O model of Table 21. Only in modeling the Other Commands direct time did the variables of the general models provide an improved fit.

†See first footnote on p. 16.

Table 27

STEPWISE REGRESSION FOR SAC TOTAL NUMBER OF I/O'S

Step	Partial F Statistics of Independent Variables Entered (Removed): Airmen	Civil Engineering (44XX)	Degrees of Freedom	Number of Variables	R^2	s : 10^6
1	45.6		20	1	.695	2.43
2	23.7	16.3	19	2	.836	1.83

Table 28

STEPWISE REGRESSION FOR TAC TOTAL NUMBER OF I/O'S

Step	Partial F Statistics of Independent Variables Entered (Removed): Total Base Population	Base Maintenance Cost	Mission Equipment Maintenance (2XXX)[a]	Degrees of Freedom	Number of Variables	R^2	s : 10^6
1	187.5			15	1	.926	1.44
2	57.7	2.9		14	2	.939	1.36
3	49.8	7.9	4.5	13	3	.954	1.21

[a]Depot Maintenance (27XX) is excluded.

less significant terms, Base Maintenance Cost and Civil Engineering are included, raising the R^2 to .95 and lowering the standard error to less than 5 percent of the mean. The coefficients of all three variables in the equation are significant at the .06 level.

Other Command Installations

In Table 29 the stepwise procedure begins by entering Officers and Vehicle Maintenance, and would then terminate if we required the partial F for entry to be significant at the .01 level. Allowing less significant terms, Accounts Control and Military Population are included, and then in Steps 5 and 6 the first two variables entered are removed. In the final step, Base Maintenance Cost is inserted, giving us a three-variable equation achieving an R^2 of .79 and a standard error of 3.3×10^6, which is 15 percent of the mean. The least significant coefficient, that of the Maintenance Cost variable, is significant at the .06 level.

EXAMINATION OF THE MODELS

Table 30 presents the models of direct time and number of I/Os we have selected for each command. Aside from the "other" direct time model not obtained by the stepwise regression procedure, each model selected is the equation obtained in the last step of the procedure. It should be remembered that the "last" step is arbitrary, inasmuch as it is determined by setting a required level of significance for the partial F statistic. We have used the .10 level. The stepwise procedure would continue if we allowed terms whose partial Fs were less significant to enter. Further, there is no a priori reason for not selecting the equation of an earlier step. The choice of those of the last steps as our models is based upon examiniation of the standard errors and the partial F statistics. In each case, the last step provides the equation with the smallest standard error, often much smaller than that of the previous steps. The partial Fs of the coefficients of each variable in these equations are all highly significant, the least being significant at the .06 level.

The independent variables on which the models are built are

Table 29

STEPWISE REGRESSION FOR OTHER COMMANDS TOTAL NUMBER OF I/O'S

Step	Partial F Statistics of Independent Variables Entered (Removed)					Degrees of Freedom	Number of Variables	R^2	$s \div 10^6$
	Officers	Vehicle Maintenance (4241)	Accounts Control (1511)	Total Military	Base Maintenance Cost				
1	32.5					31	1	.512	4.90
2	27.2	14.0				30	2	.667	4.12
3	12.6	13.7	5.6			29	3	.721	3.84
4	0.2	1.0	12.7	6.7		28	4	.775	3.51
5	(0.2)	0.8	30.1	22.1		29	3	.773	3.46
6		(0.8)	34.1	47.1		30	2	.767	3.45
7			24.9	33.1	3.8	29	3	.794	3.30

plausible. Each of the six contains at least a surrogate for base population. Both TAC models include the actual base population. The SAC I/O models include airmen; the Other Commands I/O model contains total military. The SAC and Other Commands direct time models include, respectively, Chief of Maintenance and Mission Equipment Maintenance, which have correlations of .69 and .94 with the total military populations at the corresponding installations. The additional variables appear, in general, quite reasonable. Though it was unexpected to find the (aircraft) maintenance variables appearing so frequently, the frequency itself adds to their credibility. Perhaps the explanation is simply that they relate to several functional systems.

An examination of the coefficients of the variables shows that all are positive, aside from Mission Equipment Maintenance in the two TAC models and Base Maintenance Cost in the Other Commands I/O model. We would have to reject these models if an increase in the maintenance variables would result in an estimated decrease in workload on the Burroughs 3500. An increase in mission equipment maintenance personnel, however, would be accompanied by an identical increase in the total base population, as well as an almost definite increase in the maintenance cost variable. Inasmuch as the coefficient in the TAC I/O model for Total Population is larger than that of the maintenance personnel variable, an increase in the latter with its accompanying increases in the other variables would increase the load estimated by this model. Although the case is not as evident for the other two models, a quick analysis showed that an increase in the maintenance variables with negative coefficients would likely increase other variables in the equations enough to result in increased estimates of workload.

Plots were again made, for each model, of the independent variables versus the residuals and of the fitted dependent variable versus the residuals to check for nonlinearity or heteroscedasticity.* Again, neither could be detected.

---

*The heteroscedasticity resulting from the dependent variables having been measured as sample means based upon a *varying* number of months is thought to be negligible. See footnote on p. 47.

SUMMARY

The equations of Table 30 provide very credible command models of both direct time and number of I/Os.

Overall, they substantially improve the fits obtained with the general models developed in Sec. IV. The general direct time model has a standard error of 24 hours, whereas the SAC model has a standard error of 17, and the TAC a standard error of 13. Similarly, the general I/O model has a standard error of 2.8×10^6, whereas the SAC and TAC models have standard errors of 1.8×10^6 and 1.2×10^6, respectively. Only the Other Commands models do less well than the general models,* and they do only slightly less well. Figures 3 and 4, which plot the fitted versus the actual values of dependent variables, portray the extremely close fit achieved by these models.

---

*Undoubtedly, this results precisely because they seek to generalize for a variety of commands. As discussed in Sec. VIII, a decomposition of this model into several command-specific models would likely decrease the standard errors obtained.

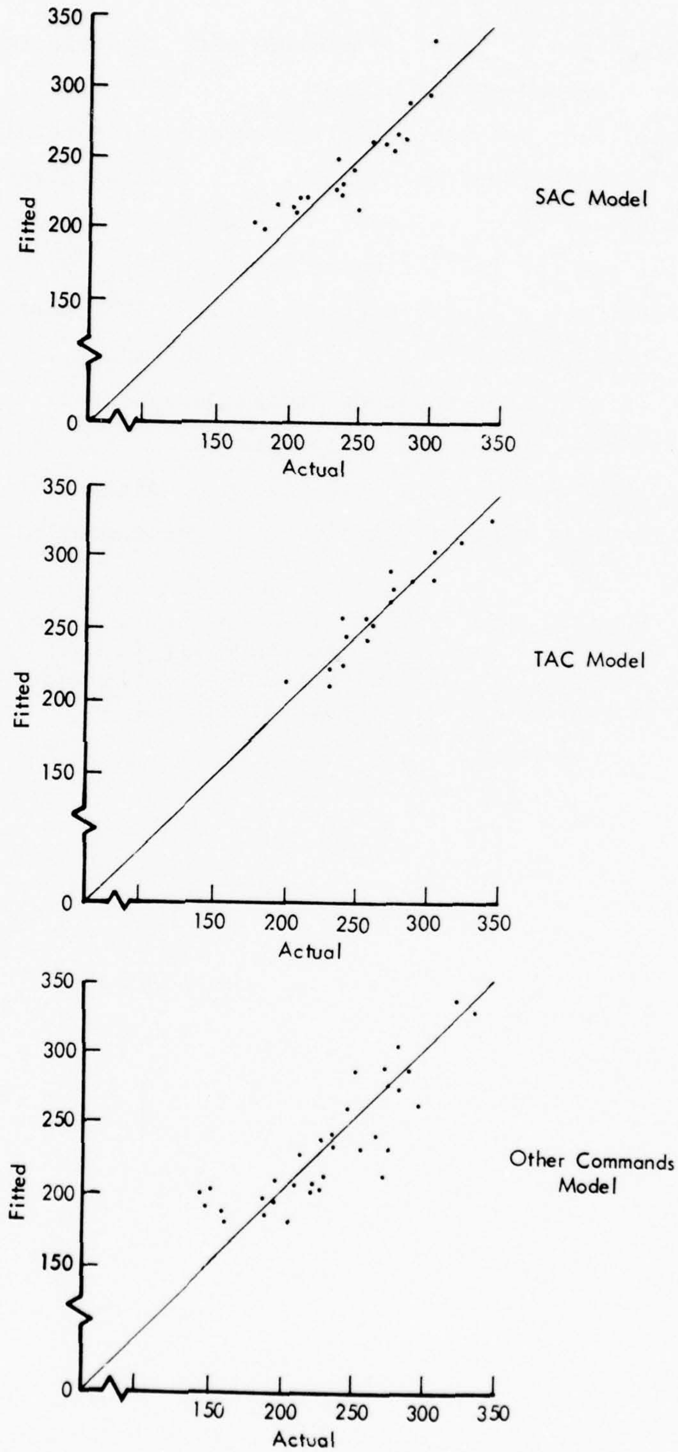


Fig. 3—Plots for the command direct time models

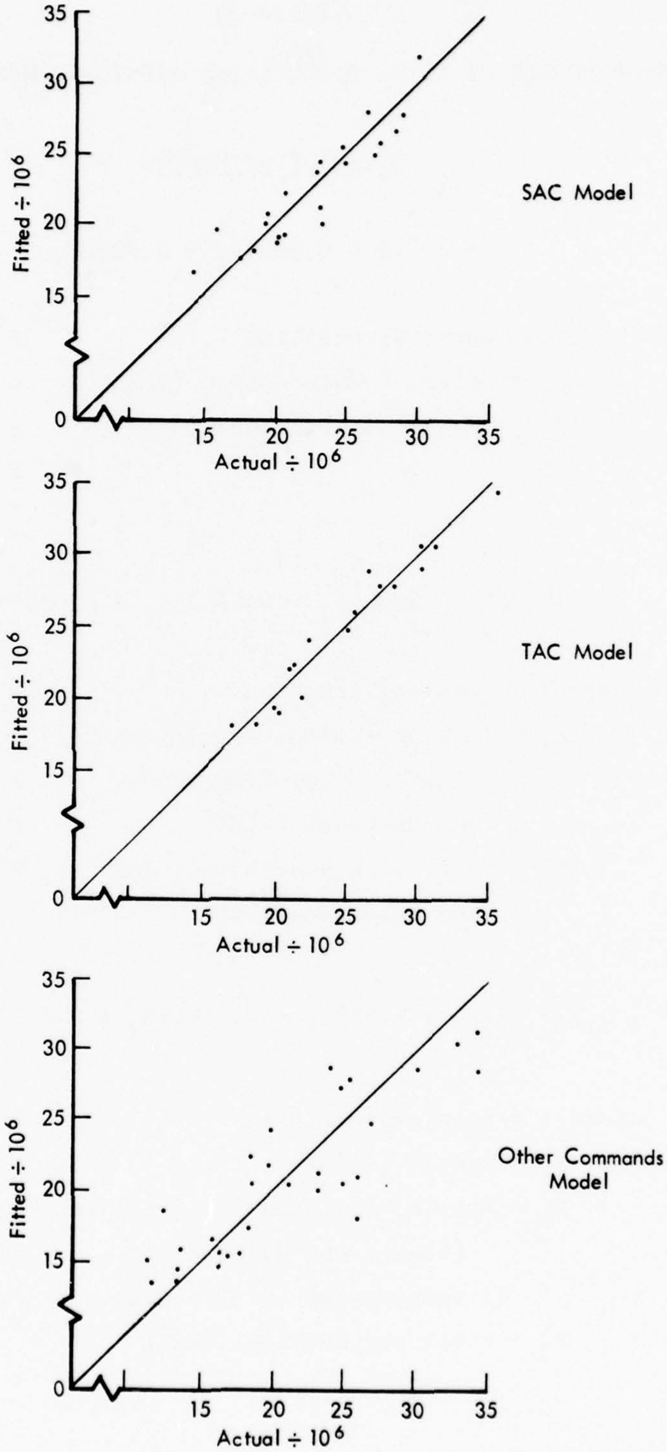


Fig. 4—Plots for the command I/O models

Table 30

COMMAND MODELS OF TOTAL DIRECT TIME AND TOTAL NUMBER OF I/O'S

Direct Time Models

SAC

$$Y = 152.0 + 0.3830X_1 + 0.7878X_2$$

where Y = total direct time	$R^2 = .810$
X_1 = chief of maintenance (21XX),	s = 16.9
X_2 = vehicle maintenance (4241)	s as % of mean = 6.8
	F(2,19) = 40.6
	P = .000000

TAC

$$Y = 134.8 - 0.02807X_1 + 0.02476X_2 + 0.00001854X_3$$

where Y = total direct time	$R^2 = .892$
X_1 = *mission equipment maintenance*	s = 13.4
(2XXX), excluding depot	s = % of mean = 4.9
maintenance (27XX)	F(3,13) = 35.9
X_2 = total base population, and	P = .000003
X_3 = base maintenance cost	

Other Commands

$$Y = 141.8 + 5.652X_1 + 0.01965X_2 + 0.06224X_3$$

where Y = total direct time	$R^2 = .723$
X_1 = travel (1514)	s = 27.7
X_2 = mission equipment maintenance	s as % of mean = 11.8
(2XXX), excluding depot	F(3,29) = 25.2
maintenance (27XX), and	P = .000000
X_3 = civil engineering (44XX)	

I/O Models

SAC $$Y = 5277000 + 19550X_1 + 1934X_2$$

where Y = total I/Os	$R^2 = .836$
X_1 = civil engineering (44XX)	$s \div 10^6 = 1.83$
X_2 = airmen	s as % of mean = 8.0
	$F(2,19) = 48.4$
	$P = .000000$

TAC $$Y = 5749000 - 2380X_1 + 3596X_2 + 1.446X_3$$

where Y = total I/Os	$R^2 = .954$
X_1 = mission equipment maintenance	$s \div 10^6 = 1.21$
(2XXX), excluding depot	s as % of mean = 4.8
maintenance (27XX)	$F(3,13) = 90.5$
X_2 = total base population, and	$P = .000000$
X_3 = base maintenance cost	

Other Commands $$Y = 3889000 + 829200X_1 + 4123X_2 - 0.9834X_3$$

where Y = total I/Os	$R^2 = .794$
X_1 = accounts control (1511)	$s \div 10^6 = 3.30$
X_2 = total military	s as % of mean = 15.5
X_3 = base maintenance cost	$F(3,29) = 37.2$
	$P = .000000$

NOTE: Appendix E presents the estimated variance-covariance matrices corresponding to each of the regressions.

VI. PREDICTING WITH THE MODELS

Having developed models of the total processing requirements, we can now use these models to forecast future load. We begin by indicating the method by which predictions are made, and then present and compare the levels of precision obtainable with the general and command models. We then discuss forecasting with a model that takes into account a likely correlation between observations at a single installation.

METHOD OF PREDICTION

In modeling the requirements, we began by assuming the existence of a theoretical relationship of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon ,$$

where Y is the measure of load and X_i are base characteristics. We then obtained least squares estimates b_i of the β_i, which, when substituted for the β_i, gave us an estimate of this relationship:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + \varepsilon .$$

To predict future workload at an installation, we then substitute planned values of the base characteristics, say $X_1^0, \ldots, X_p^0$, into this equation to provide

$$\hat{Y} = b_0 + b_1 X_1^0 + b_2 X_2^0 + \ldots + b_p X_p^0$$

as an unbiased estimate of future processing requirements. Furthermore, by virtue of the normality assumption discussed at the beginning of Sec. II, $(1 - \alpha)$ confidence limits* for this prediction are given by the formula†

$$\hat{Y} \pm t(n - p - 1, 1 - \tfrac{1}{2}\alpha) \sqrt{s^2 + X^{0'} \hat{V} X^0} ,$$

where $t(n - p - 1, 1 - \frac{1}{2}\alpha)$ = the $(1 - \frac{1}{2}\alpha)$ percentage point of the t-distribution with $(n - p - 1)$ degrees of freedom,

n = number of observations used to build the regression,

p = number of independent variables in the equation,

s = standard error of the estimate,

$X^{0'}$ = vector of values of the independent variables $(1, X_1^0, X_2^0, \ldots, X_p^0)$,

$\hat{V}$ = estimated variance-covariance matrix of the estimators of the coefficients

---

*It is important to note that these confidence bounds take into account only the variation about the regression and the sampling variation in the estimates of the coefficients; the derivation of the bounds assumes perfect knowledge of the values of the independent variable corresponding to which the dependent variable is to be estimated. Inasmuch as our independent variables are planned authorizations for the future, the assumption is not entirely realistic. Confidence intervals taking into account the uncertainty in our estimates of the independent variables would, of course, be larger. No attempt to derive such intervals is made in this study.

†The computationally much simpler, approximate formula $\hat{Y} \pm t(n - p - 1, 1 - \frac{1}{2}\alpha) \times s\sqrt{1 + 1/n}$ may suffice for many purposes. It ignores the contribution to the width of the exact interval from the variances and covariances of the estimators of the coefficients, aside from the variance of the constant term. It coincides with the exact interval only when each of the independent variables is at its mean and is otherwise narrower than the exact interval. For the models developed here, with each of the independent variables shifted by up to 50 percent, it is at most only 12 percent narrower than the exact interval.

$(\hat{V} = s^2(X'X)^{-1}$ where X is the matrix of observations,* and

X^0 = transpose of $X^{0'}$.†

GENERAL MODELS

The precision of estimation obtainable with the general models can be seen in Table 31. The first column indicates a percentage difference of the values of the independent variables from their corresponding means over the observations on which the model was built. Each independent variable is taken to have the identical percentage difference from its corresponding mean. The next column indicates the prediction that would be made for such a set of independent values. The third column indicates the percentage difference of the predicted value from the predicted value corresponding to a zero percentage change in the independent variables. The final column gives the width of the 90 percent confidence interval, which is a measure of the precision of the estimation.

Direct Time Model

In the middle row of Table 31, where the percentage difference is zero, each independent variable has as its value the means given in Table 3. The corresponding predicted direct time is 248 hours, obtained by calculating

$$\hat{Y} = 137.3 + 4.522(7.67) + .08127(428.26) + .02489(1651.50) = 248 .$$

The 90 percent confidence interval for the predicted value is [207, 289]. That is, we are 90 percent certain that this interval would include the actual value corresponding to such a forecast. It is also true that the upper bound provides a 95 percent upper confidence bound. That is, we have 95 percent certainty that this bound will be above

---

*The estimated variance-covariance matrices for the general and command models are given in Appendixes D and E, respectively.

†Draper and Smith, pp. 121-122.

Table 31

PRECISION OF ESTIMATION WITH THE GENERAL MODELS

Direct Time Model[a]

Percentage Difference in Each Independent Variable	Predicted Total Direct Time (hours per month)	Percentage Difference in Predicted Total Direct Time	90 Percent Confidence Interval (hours per month)		
			Lower Bound	Upper Bound	Width
-50	193	-22	151	234	83
-40	204	-18	162	245	83
-30	215	-13	174	256	82
-20	226	-9	185	267	82
-10	237	-4	196	278	82
0	248	0	207	289	82
+10	259	+4	218	300	82
+20	270	+9	229	311	82
+30	281	+13	240	322	82
+40	292	+18	251	333	82
+50	303	+22	262	345	83

I/O Model

Percentage Difference in Each Independent Variable	Predicted Total I/Os (millions per month)	Percentage Difference in Predicted Total I/Os	90 Percent Confidence Interval (millions per month)		
			Lower Bound	Upper Bound	Width Bound
-50	13.1	-42	8.3	17.8	9.5
-40	15.0	-34	10.3	19.7	9.4
-30	16.9	-26	12.2	21.6	9.4
-20	18.8	-17	14.2	23.5	9.3
-10	20.8	-8	16.1	25.4	9.3
0	22.7	0	18.1	27.3	9.2
+10	24.6	+8	20.0	29.3	9.3
+20	26.6	+17	21.9	31.2	9.3
+30	28.5	+26	23.8	33.2	9.4
+40	30.4	+34	25.7	35.1	9.4
+50	32.3	+42	27.6	37.1	9.5

[a]The model employed here is the second direct time model given in Table 21.

the corresponding actual value. The last column gives the width of the interval as 82 hours.

Similarly, the predicted direct time with each variable 10 percent above its mean is 259 hours, as calculated from the equation

$$\begin{aligned}\hat{Y} &= 137.3 + 4.522\ [7.67 + .10(7.67)] \\ &\quad + .08127\ [428.26 + .10(428.26)] \\ &\quad + .02489\ [1651.50 + .10(1651.50)] \\ &= 259.\end{aligned}$$

The third column shows us that the predicted value has increased by only 4 percent over that predicted with the independent variables at their means, even though the independent variables are 10 percent greater. The 90 percent confidence interval is given by [218, 300].

As the independent variables increase 20, 30, 40, and 50 percent above their means, the predicted direct time increases by 9, 13, 18, and 22 percent, respectively.* The width of the confidence interval remains almost constant at approximately 82 hours.

The results with decreases in the independent variables are entirely symmetric to those with the increases, aside from occasional apparent deviation due to rounding. The predicted value corresponding to a 10 percent decrease is 9 hours below that corresponding to the zero percentage difference; the predicted value corresponding to a 10 percent increase, as discussed above, is 9 hours above. The percentage differences in the predicted values for the decreases, consequently, are simply the negatives of the differences for the corresponding increases. The confidence bounds are symmetric about the bounds for the

---

*As will be seen, forecasts with *all* of the direct time models imply that increases in the independent variables result in substantially smaller percentage increases in charged direct time. For example, with a 50 percent increase in the independent variables, the direct time increases by only 20 to 25 percent. This results from "overhead" direct time estimated by the constant term in the models. An interesting implication of this, irrelevant to this study, is that total computer processing requirements would likely be reduced with larger, but fewer, base installations.

zero percentage difference; the lower bound corresponding to a 10 percent decrease in the independent variables is 11 hours below that for the zero difference, and the lower bound corresponding to a 10 percent increase is 11 hours above. The width of the intervals for each decrease is, consequently, indentical to that for the corresponding increase.

I/O Model

The lower half of Table 31 relates to the general I/O model. The predicted value with each of the independent variables at its mean is 2.27×10^7. An increase in the independent variables results in a comparable increase in the predicted value; raising the independent variables by 10, 20, and 30 percent causes the predicted number of I/Os to increase, respectively, by 9, 17, and 26 percent. The confidence intervals widen only slightly as the independent variables shift away from their means.

COMMAND MODELS

Tables 32 and 33 present analogous results for the command models.

Direct Time Models

Looking at Table 32, we find each level of increase in the independent variables resulting in a substantially smaller increase in predicted direct time, as was the case for the general direct time model. A 20 percent increase in the variables, for example, causes increases of less than 10 percent for each command. The widest confidence interval with the SAC model is only 61 hours; with the TAC model, only 52; and with the Other Commands model, 96.

I/O Models

Turning to Table 33, we see that increases in the independent variables result in only slightly smaller increases in the predicted number of I/Os. With 20 percent increases in the independent variables, the predicted values are about 15 percent higher; with 50 percent increases, about 40 percent. The widths of the confidence

Table 32

PRECISION OF ESTIMATION WITH THE COMMAND DIRECT TIME MODELS

Percentage Difference in Each Independent Variable	Predicted Total Direct Time (hours per month)	Percentage Difference in Predicted Total Direct Time	90 Percent Confidence Interval (hours per month)		
			Lower Bound	Upper Bound	Width
SAC					
-50	201	-19	170	231	61
-40	210	-16	181	240	59
-30	220	-12	191	250	59
-20	230	-8	201	259	58
-10	240	-4	211	269	58
0	249	0	221	278	57
+10	259	+4	230	288	58
+20	269	+8	240	298	58
+30	279	+12	249	308	59
+40	288	+16	259	318	59
+50	298	+20	268	328	60
TAC					
-50	203	-25	177	229	52
-40	216	-20	191	241	50
-30	230	-15	206	254	48
-20	243	-10	220	267	47
-10	257	-5	234	280	46
0	270	0	247	293	46
+10	284	+5	261	307	46
+20	298	+10	274	321	47
+30	311	+15	287	335	48
+40	325	+20	300	350	50
+50	338	+25	312	364	52
Other Commands					
-50	188	-20	140	236	96
-40	198	-16	150	245	95
-30	207	-12	160	254	94
-20	216	-8	169	264	95
-10	226	-4	179	273	94
0	235	0	188	282	94
+10	244	+4	198	291	93
+20	254	+8	207	301	94
+30	263	+12	216	310	94
+40	272	+16	225	320	95
+50	282	+20	234	330	96

Table 33

PRECISION OF ESTIMATION WITH THE COMMAND I/O MODELS

Percentage Difference in Each Independent Variable	Predicted Total I/Os (millions per month)	Percentage Difference in Predicted Total I/Os	90 Percent Confidence Interval (millions per month)		
			Lower Bound	Upper Bound	Width
		SAC			
-50	14.0	-38	10.5	17.5	7.0
-40	15.8	-30	12.4	19.1	6.7
-30	17.5	-23	14.2	20.8	6.6
-20	19.2	-15	16.1	22.4	6.3
-10	21.0	-7	17.8	24.1	6.3
0	22.7	0	19.6	25.9	6.3
+10	24.5	+8	21.3	27.6	6.3
+20	26.2	+15	23.0	29.4	6.4
+30	28.0	+23	24.7	31.2	6.5
+40	29.7	+31	26.3	33.1	6.8
+50	31.5	+39	27.9	35.0	7.1
		TAC			
-50	15.5	-39	13.2	17.9	4.7
-40	17.5	-31	15.2	19.7	4.5
-30	19.4	-23	17.2	21.6	4.4
-20	21.4	-15	19.3	23.5	4.2
-10	23.3	-8	21.2	25.4	4.2
0	25.3	0	23.2	27.4	4.2
+10	27.2	+8	25.2	29.3	4.1
+20	29.2	+15	27.1	31.3	4.2
+30	31.2	+23	29.0	33.3	4.3
+40	33.1	+31	30.8	35.4	4.6
+50	35.1	+39	32.7	37.4	4.7
		Other Commands			
-50	12.6	-41	6.9	18.4	11.5
-40	14.4	-32	8.7	20.1	11.4
-30	16.1	-24	10.5	21.8	11.3
-20	17.8	-16	12.2	23.5	11.3
-10	19.6	-8	14.0	25.2	11.2
0	21.3	0	15.7	26.9	11.2
+10	23.1	8	17.5	28.7	11.2
+20	24.8	16	19.2	30.4	11.2
+30	26.6	25	20.9	32.2	11.3
+40	28.3	33	22.6	34.0	11.4
+50	30.1	41	24.3	35.8	11.5

intervals with no percentage change are 28, 17, and 52 percent of the corresponding means for the SAC, TAC, and Other Commands models, respectively.

COMPARATIVE PRECISION OBTAINABLE WITH GENERAL AND COMMAND MODELS

Table 34 contrasts the precision of estimation with the general and command models. It presents, for each, the approximate half-width of the 90 percent confidence interval, expressed both in absolute terms and as a percentage of the overall mean of the dependent variable. The half-width is the distance between the predicted value and the upper bound, which, as mentioned before, is a 95 percent upper confidence bound. Hence, we can be 95 percent certain that the predicted value will not underestimate the actual value by more than the half-width.

Table 34

COMPARATIVE PRECISION OF ESTIMATION OBTAINABLE
WITH GENERAL AND COMMAND MODELS

Model	Direct Time Models		I/O Models	
	Approximate Half-Width of 90 Percent Confidence Interval (hours per month)	Percent of Overall Mean	Approximate Half-Width of 90 Percent Confidence Interval (millions per month)	Percent of Overall Mean
General	41	17	4.7	21
SAC	29	12	3.2	14
TAC	24	10	2.1	9
Other Commands	47	19	5.6	25

NOTE: The approximate half-widths presented are the half-widths of the intervals corresponding to a 30 percent shift in the independent variables.

Direct Time Models

Looking at the direct time models first, we find that the general model has a half-width of 41 hours. The SAC and TAC models, however, have substantially smaller half-widths of 29 and 24 hours, respectively, each of which, with the addition of prorated time,* would correspond to about two 24-hour days of processing. Only the Other Commands model does less well than the general model. It does only slightly less well, however, with a half-width of 47 hours as compared to 41. This corresponds to only about three and one-half full days of processing. Hence, we think that, overall, the command direct time models substantially improve upon the precision of estimation with the general model. Further, we judge the levels of precision obtainable with the command models to be excellent.

I/O Models

The results for the I/O models are almost identical. The SAC and TAC models appreciably improve on the precision of forecasting obtainable with the general model, and the Other Commands model does only slightly less well than the general model. Consequently, the command models are again thought overall to provide a higher level of precision, and the levels obtainable are judged to be excellent.

PREDICTIONS BASED ON A MODEL WITH AN AUTOREGRESSIVE STRUCTURE

Having thus far disregarded the possibility of autocorrelation, we now discuss forecasting with a model taking it into account.†

Autocorrelation is defined as correlation between the error terms of observations on the dependent variable. It occurs frequently with the use of longitudinal data, rarely when the data are cross-sectional.

In building our models, we ignored autocorrelation since the data were entirely cross-sectional. With only one observation from each installation, we could safely assume that the residuals were mutually independent.‡ It seems reasonable to presume that the residual charged direct time at one installation is independent of that at another.

---

*See p. 7.

†J. Johnston, *Econometric Methods*, McGraw-Hill Book Company, Inc., New York, 1963, pp. 177-200.

‡This is equivalent to the assumption of independent observations made on p. 7.

It would not be so reasonable to assume that the residual for an installation in one period is independent of that in a subsequent period. It seems quite likely, in fact, that an installation with a positive residual in one period will typically have a positive residual in a subsequent period. If this is so, the residual errors from a single installation for different periods of time would be autocorrelated.*

Though we had no need to address the issue of autocorrelation in *building* the models, by virtue of the use of cross-sectional data, the issue is raised in *forecasting* with the models by the need to predict requirements for the same installations employed to build the models. For each installation for which we wish to make a prediction, we know the residual difference between the actual and fitted values in the data on which the model was developed. If the autocorrelation is nonzero, this residual is correlated with the residual corresponding to the forecast value. By incorporating any such autocorrelation into the model, we could employ the observed residuals to improve the forecasting.

Without longitudinal data, however, we cannot verify the presence of autocorrelation. Furthermore, if we postulate a model with "autoregressive" structure incorporating the autocorrelation, we cannot estimate the autocorrelation coefficient.† What we can do is formulate a model and base our forecasts on "bounding" assumptions regarding the value of this coefficient.

We postulate such a model as follows:

$$Y_{jt} = \beta_0 + \beta_1 X_{jt}^{(1)} + \beta_2 X_{jt}^{(2)} + \ldots + \beta_p X_{jt}^{(p)} + \epsilon_{jt} , \qquad (1)$$

---

*Consequently, if longitudinal data were to be employed in building models, the existence of autocorrelation must be checked, and if it exists, as is likely, an autoregressive model as is presented on pp. 72-73 should be built.

†We could check for autocorrelation by breaking each of our observations for a six-month period into observations for two three-month periods. The autocorrelation between the observations would likely be much higher, however, than that for periods of time more distant from one another.

where $\mathrm{Cov}\,(\varepsilon_{jt}\varepsilon_{j't'}) = 0$ for $j \neq j'$, for all t, t' , (2)

and $\varepsilon_{jt} = \rho_{(t-t')}\varepsilon_{jt'} + \gamma_{jt}$ for all j, for $t > t'$, (3)

with $0 \leq \rho_{(t-t')} \leq 1$ for all t, t' and (4)

$E(\gamma_{(t-t')}) = 0$ for all $t > t'$. (5)

The linear form in line (1) is identical to that of the basic model presented at the beginning of Sec. II, except that here we use the subscript "j" to index installations, the subscript "t" to index the time period for which the observation is made, and a superscript notation to label the independent variables. Line (2) specifies that the covariance, and hence the correlation, between the residual error terms for observations at different installations are all zero. Line (3) postulates the autoregressive structure as a linear relationship (without constant) between the error terms of different time periods at a single installation. The coefficient $\rho_{(t-t')}$ is the autocorrelation between observations at any single installation taken (t - t') units of the time apart. In line (4), we assume the autocorrelation coefficient to be a (presumably decreasing) function, bounded by zero and one, of the difference between time periods corresponding to the two error terms. A value of zero for this coefficient reduces this model to that previously discussed. Line (5) specifies that the expected value of the error term in the error model is zero.

Under an assumption of this model, unbiased forecasts of the dependent variables are given by

$$\hat{Y}_{jt} = b_0 + b_1 X_{jt}^{(1)} + b_2 X_{jt}^{(2)} + \ldots + b_\rho X_{jt}^{(\rho)} + \hat{\rho}_{(t-t')} \varepsilon_{jt'} , \qquad (8)$$

where $\hat{Y}_{jt}$ = forecast value for installation j for time period t,

b_i = least squares estimate of β_i in Eq. (1),

$X_{jt}^{(i)}$ = (planned) values of the independent variables at installation j in time period t,

$\hat{\rho}_{(t-t')}$ = estimated autocorrelation between residual terms of observations taken (t - t') time periods apart, and

$\varepsilon_{jt'}$ = value of the residual in period t'.

With only cross-section data, the least squares estimates b_i for this model are identical to those for our earlier model. The autocorrelation coefficient $\rho_{(t-t')}$, however, cannot be estimated. With a value of zero for this parameter, the last term in the expression for the forecast value is dropped, so that the observed residual value ε_{jt}' is ignored. This reduces the forecast to that made with our earlier model, as it should, since a value of zero for this parameter reduces this model to our earlier one. With $\rho_{(t-t')}$ equal to one, the full value of the residual is added to the forecast based upon an assumption of no autocorrelation.

Figure 5 illustrates these forecasts based upon a model with only a single independent variable, the case with more independent variables being analogous. The observation for the j^{th} installation, taken at time t', is indicated by the point $(X_{jt'}, Y_{jt'})$. The line plotted is the estimated regression line $Y = b_0 + b_1 X$. Hence, the fitted value of Y corresponding to $X_{jt'}$ is the indicated value $\hat{Y}_{jt'}$. The residual for the j^{th} installation is then given by $\varepsilon_{jt'} = Y_{jt'} - \hat{Y}_{jt'}$. Suppose that the independent variables were to be increased by time t to X_{jt}. Under our earlier model or, equivalently, under an assumption of zero autocorrelation, the forecast value would necessarily lie on the estimated

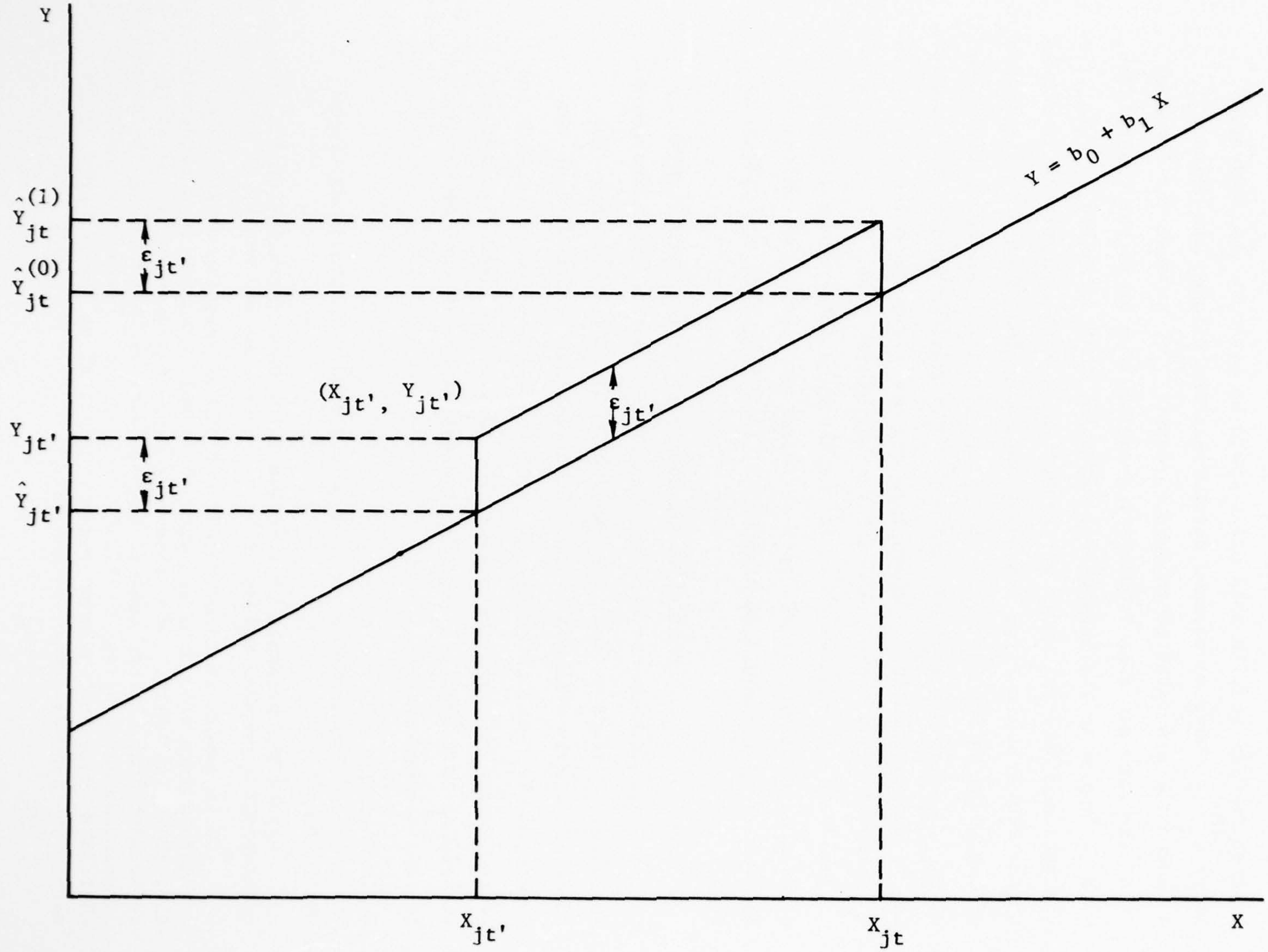


Fig. 5— Predictions with model with autoregressive structure

regression line and, in this case, would be given by the indicated value $\hat{Y}^{(0)}_{jt}$. Under an assumed autocorrelation of one, the value $\varepsilon_{jt'}$ is added to this value to instead forecast $\hat{Y}^{(1)}_{jt'}$, so that the forecast is at the same distance from the regression line as is the observation. With no change in the independent variable, that is with $X_{jt'} = X_{jt}$, the unbiased forecast under an assumption of zero autocorrelation would be the fitted value $\hat{Y}_{jt'}$, and under an assumption of an autocorrelation of one would be the observed value $Y_{jt'}$. In all likelihood, the value of $\rho_{(t-t')}$ lies between zero and one.* Unbiased forecasts corresponding to autocorrelations between zero and one lie between the two values $\hat{Y}^{(0)}_{jt'}$ and $\hat{Y}^{(1)}_{jt'}$ and, hence, lie off the regression line, but closer to it than does the observation. As the distance between the time period of data base and period of forecast increases, the actual values for the future period would be expected to fall closer to the regression line.

Lacking any information about the value of the autocorrelation coefficient, perhaps the best procedure is to obtain forecasts corresponding to values of both zero and one. As discussed above, the former is simply obtained by use of our earlier models without the autoregressive structure, and the latter by simply adding to this the corresponding observed residual.† The correct unbiased *forecast* corresponding to the actual value of the parameter can then be assumed to fall between these.‡

---

*The estimation of $\rho_{(t-t')}$ with longitudinal data is discussed on pp. 88-89.

†Appendix F presents the residuals for each installation for the corresponding command models of both direct time and number of I/Os.

‡It is important to note that, if autocorrelation exists, the confidence intervals obtained with the models assuming no autocorrelation are still valid. Such intervals are, of course, centered at the regression line, rather than at the unbiased estimate taking the autocorrelation and observed residual into account. Confidence intervals that take these into account would be centered at the unbiased estimate, and typically would be narrower.

PLANNING FOR PEAKS

The dependent variables were measured, as will be recalled, as mean monthly utilizations, the means typically being based on five or six months.* The purpose in so doing was to eliminate seasonal variation to the extent possible with our data. The variance about the regression for such sample means is, of course, smaller than the variance for individual observations of monthly utilization. Inasmuch as the confidence intervals obtained herein are based on estimates of the former variance, the intervals do *not* provide bounds for utilization during a specific month. That is, a 90 percent confidence interval does not indicate that we are 90 percent certain that this interval will cover the actual value obtained for a single month. The variance about the regression for the sample means is, however, *larger* than the variance about the regression for *theoretical* mean monthly utilization. Hence, the confidence intervals as presented herein provide somewhat conservative bounds for the theoretical utilization rate, that is we are *at least* 90 percent certain that the intervals obtained will cover the theoretical mean monthly utilization at a random installation.

In determining required capacity, there is no need to address variation about the theoretical mean utilization, if workload can be shifted from one period to another when necessary. If workload cannot be so smoothed, however, it is necessary to plan for peak loads. To do so, one need only measure the variation from one period to another and then provide sufficient excess capacity over that required to support the mean load.

PITFALLS IN PREDICTION

In predicting with these as with any models, one must use reason and care. Two particular hazards lie in (1) violation of the assumed invariance of the coefficients across time, and (2) extrapolation beyond the range of the data on which the models were built.

Predictions with regression models assume that the coefficients of each of the variables in the theoretical regression equation remain

---

*See footnote on p. 8.

unchanged from the period of the data base to the period of prediction. In all likelihood, this assumption would be invalidated if changes occurred in the relationships between any of the variables included in the equations or between any of these and others, not included, that affect the dependent variable. A major civilianization of the Air Force, for example, would drastically alter some relationships. The relatively small load that civilian activities currently generate is undoubtedly represented, at least in part, by the variables Airmen and Total Military in models incorporating these variables. Since large-scale civilianization would change the relationship between civilian manpower and military manpower, the coefficients of the military variables would inadequately represent the load generated by civilians and the models would grossly underestimate the processing requirements. A subtler example would be a change in the derivation of manpower authorizations. Suppose the current formulas were changed to increase all vehicle maintenance authorizations by 25 percent, simply because the current authorizations were judged insufficient. The model incorporating this variable would increase its predicted requirement, even though no increase is expected in the activities this variable represents nor, therefore, in the requirements these activities generate.

Another hazard lies in extrapolating beyond the region of the data on which the models were built. Within that region, the models may simply represent a good approximation to a much more complex function not at all well represented outside the region. One must take a cautious view of both predictions and confidence intervals corresponding to points lying outside this region. The farther from the region, the greater the uncertainty. Appendix G provides an approximation to the regions of data on which each model was built; for each model, it gives the minimum and maximum of each included independent variable over the values in the data on which the model was built. The actual regions are, of course, subsets of the regions so defined. Consequently, if any of the values of the independent variables falls outside its indicated range, the corresponding predicted value is extrapolated.* All such extrapolated predictions should be used with caution.

---

*The converse is not true, however. That is, the values of each variable can be within its range, and yet the vector of values be such

AD-A063 821 RAND CORP SANTA MONICA CALIF F/G 9/2

FORECASTING PROCESSING REQUIREMENTS OF STANDARD USAF BASE LEVEL--ETC(U)

JUL 78 E C POGGIO, N AINSLIE F49620-77-C-0023

UNCLASSIFIED RAND/R-2334-AF NL

2 OF 2

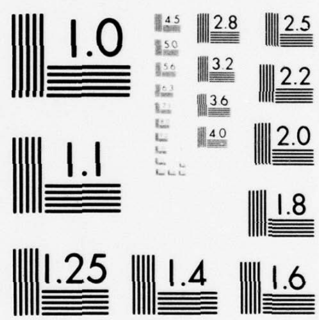
AD
A063 821

END

DATE
FILMED

3-79

DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

SUMMARY

This section has shown that the command models substantially improve upon the already high level of precision in forecasting obtainable with the general models. The improvement is large enough to recommend the use of the command models over the corresponding general model. The command direct time models had 90 percent confidence intervals with half-widths of only 29, 24, and 47 hours for SAC, TAC, and Other Commands, respectively. The respective I/O models had half-widths equal to 14, 9, and 24 percent of the overall mean. Each of these is judged to represent a very high level of precision.

---

as to lie outside the actual range of the data. Because of the difficulty in representing multidimensional regions, no attempt to indicate the actual region is herein made. It is thought that the somewhat larger regions in Appendix G provide sufficient guidelines.

VII. PREDICTING THE PROCESSING REQUIREMENTS OF A REGIONAL COMPUTER SYSTEM

In addition to being used to forecast changed base-level processing requirements due to changes in the activities and composition of a base, the methodology developed herein can be used to estimate the processing requirements of regional (or even central) USAF base-level computer systems.

The utility of these models for regional systems stems from consideration of the several bases within a region as a single hypothetical base of the same size and composition as the several bases combined. If one can assume that the processing requirements to support the several bases are identical to those of the hypothetical base, then the models can be used directly to predict the requirements for the regional computer system. Otherwise, an adjustment to the prediction obtained under this assumption would have to be made.

The actual prediction for a region would be obtained by simply substituting for each independent variable the sum of the corresponding variable across each base to be included in the region. For example, if one were interested in estimating the direct time requirements for a regional computer for three SAC bases, one would use the SAC direct-time model and substitute for X_1 the total number of personnel assigned to chief of maintenance (21XX) at all three bases and for X_2 the total number assigned to vehicle maintenance (4241).

Some notion of the benefits to be gained from regionalization, under the above assumption, can be gleaned from Tables 32 and 33. We find that 50 percent increases (above the observed sample means) in each of the independent variables result in only 25 percent increases in predicted direct time at TAC bases and only 20 percent increases at other bases. Performing the same calculation for 200 percent increases in the independent variables results in 100 percent increases at TAC bases and 80 percent increases at other bases. Since under our assumption, the formation of a region composed of three bases (each with values of the independent variables identical to our sample means) corresponds to a 200 percent increase, we predict that such a region would require only 100 or

80 percent more direct time than any one of the bases, certainly a very substantial savings in processing time. Using the same procedure with the I/O models, we find that the three-base region would require about a 160 percent increase in I/O capacity over that of one of the bases. These models suggest then the possibility of a very substantial savings with regionalization. Of course, there are other costs and benefits that also must be taken into account in order fully to compare a regional computer system with a base-level system.

It is important to note several potential hazards in using these models to forecast the requirements of a regional system. First of all, to do so will typically require extrapolation far beyond the range of the data. This is risky since though the linear form of the models may provide a perfectly fine approximation within the range of observed data, we can have no assurance that it will do so outside this range. We can take some comfort in the fact that we did check for curvilinear effect and found none, suggesting that the linear form may be the form of the true relationship. Further, as discussed on page 4, there is some theoretical basis for believing this to be so.

Secondly, the prediction of the requirements for a regional system would likely be of interest when considering the installation of a new computer system. No data would be available on the new system and one would have to use data from an old system. In order to use the models to make predictions for the new system, one would have to assume that the benefits would be the same as for the old system, or to adjust the estimates made under this assumption. This assumption should be carefully examined by consideration of differences between the systems including the hardware, software, and applications.

Finally, the assumption of equivalence between the several bases within a region and a hypothetical base of the same size and composition must also be carefully examined. The validity will depend in large part on the way in which the processing is handled. If, for example, the military pay system were run three times each month, once for each base, instead of once each month for the three bases together, the benefits from regionalization would be lost. If this assumption is not found to be completely valid, an adjustment to the predictions obtained under this assumption must be made.

In spite of these hazards, these models probably provide the best available means of predicting the processing requirements for a regional computer system.

VIII. CONCLUSIONS

In summary, we have established that the problem of forecasting future USAF base level computer processing requirements to support currently existing functional systems can be solved by developing regression models that relate such requirements to base characteristics for which future planning figures are available. Using planned base characteristics as inputs, future processing requirements can be forecast. Further, we have developed sets of command specific models for both direct time and total number of I/Os that can be used to make such forecasts with high precision. We now discuss verification and maintenance, use, improvement, and extensions of these models.

VERIFYING AND MAINTAINING THE MODELS

It is recommended that the command models herein developed be verified and then maintained on a periodic basis. The simplest means of accomplishing the verification is to compare a set of model forecasts with a set of actual values. A simple plot of the two is helpful. The frequency with which the confidence intervals cover the actual values also should be checked. Additionally, the coefficients can be verified by comparing past estimates with new ones based on an independent set of data.

In maintaining the models, the first step is to perform the periodic verifications discussed above. At a minimum, it is recommended that forecasts and actuals be compared annually. Whenever a model is found to forecast requirements inadequately a new model should be built. It is likely that this can be done simply by again applying the stepwise regression procedure to the candidate independent variables listed in Table 23, to build a new set of command models. Should this prove inadequate, the best approach would probably be to use the complete procedure employed herein in developing the command models. First, determine the major functional systems; then select the base characteristics likely to be correlated with the corresponding workload. Next, determine for each command those base characteristics most highly correlated with the workload of each major system. Finally, use these characteristics as candidate independent variables in a stepwise regression procedure to build command models. The set of models can thus be kept current.

USING THE MODELS

It is recommended that forecasts be made annually for each of the five subsequent years so the Air Force can assess the need for alternative computer systems. The forecasts should also be revised immediately following any *major* change in planned authorizations, to allow the Air Force maximum lead time if a change in hardware is required.

In addition to their use in predicting the processing requirements for a regional computer system (of primary interest in this study and the subject of Section VII), the techniques of this study should also be used in addressing other alternative systems. Suppose, for instance, we were considering the purchase of an additional computer system at each installation, and a division between the two of the workload now supported solely on the Burroughs 3500. Perhaps we wished to consider placing the military personnel system, the two military pay systems, the civilian pay system, and the general accounting and finance system on the new computer, and leave all other systems on the 3500. We could then use the candidate independent variables for each system as obtained in Section III to build models of direct time charged to the two sets of systems, the direct time charged to the software systems being allocated as appropriate. We could then use these models to forecast processing requirements for the two sets of systems. Similarly, in considering a computer dedicated to a single functional area, the candidate independent variables obtained in Section III could be used to model the requirements at each installation to support the systems in that functional area.

IMPROVING THE MODELS

There are several possibilities for improving our models.

Alternative Independent Variables

Perhaps the first thought is that a different set of independent variables might lead to greater precision. Our approach to *the selection* of independent variables, by building models of the individual systems, allows us to pinpoint weak areas. Looking at the tables of Sec. III, we see, for instance, that we have not obtained very good predictors for BEAMS. It may be profitable to look for better ones, this being the system with the second largest variability. In retrospect, the only specific alternative variables that have occurred to us as potential predictors of the individual systems are number of missiles and those suggested for BEAMS in the footnote on p. 22. We do not believe, however, that simply adding or substituting other independent variables will appreciably improve the precision of estimation.

Additional Observations

Another possible way to improve the models is to increase the number of observations on which each is built. Doing so increases the precision of the estimators of the regression coefficients and, hence, shortens the confidence interval about a predicted value.

The confidence intervals are, however, a function of the standard error of the estimate, as well as the variances and covariances of the coefficients.[*] In fact, as the number of observations increases, the contribution from the variances and covariances to the length of the interval approaches zero. Inasmuch as the contribution from these to the lengths of the intervals obtained with our models is quite small, the improvement to be obtained by simply increasing the number of observations would be marginal.[†]

Moreover, to increase the number of observations would require the use of longitudinal, as well as cross-sectional, data. Autocorrelation

---

[*] See p. 65.

[†] The contribution to the length of the interval from the variances and covariances varies as a function of the values of the independent variables; with each independent variable increased by as much as 50 percent, the variances and covariances account for a maximum of 15 percent of the length of the intervals obtained with our models.

between observations at the same installation would likely exist and have to be taken into account in building the models.*

Additional Command Models

As seen in Secs. V and VI, the SAC and TAC models substantially improve over the general models. The models for the Other Commands, however, perform less well, precisely because, it would seem, they seek to generalize for many commands. If so, it is reasonable to expect that decomposing them into several command specific models would increase the precision of estimation, probably to about the levels obtainable with the SAC and TAC models.

Inasmuch as the other commands have but a small number of installations, one should obtain longitudinal data, as well as cross-sectional, to develop the models. Autocorrelation between observation at a single installation would then likely have to be accounted for in developing the models.

Estimation of the Autocorrelation Coefficient

In Sec. VI, we defined a model with an autoregressive structure, taking into account a likely correlation between the residual errors from a single installation at different points in time. With only cross-sectional data, we could not estimate the autocorrelation coefficient and so suggested simply making forecasts corresponding to the bounding values of zero and one for this coefficient. The correct unbiased forecast, taking into account the observed residual for the installation, could then be assumed to fall between zero and one.

By actually estimating the autocorrelation coefficient, forecasting might well be substantially improved; the higher the autocorrelation, the greater the improvement. As indicated in Sec. VI, the autocorrelation is thought to be a decreasing function of the difference between the time period of the forecast and that of the observed residual. To estimate this function, one should collect the actual mean monthly utilizations for each installation and make corresponding

---

*See first footnote on p. 74.

forecasts (based upon current, rather than planned, authorizations) for each six-month period subsequent to the period of the data base (the last half of FY 1972). Ideally, this should be done for a length of time equal to the farthest distance into the future for which forecasts are to be made, perhaps for the five years for which planned authorizations are now made.* By subtracting each forecast from the corresponding actual value, residuals can be obtained. These can then be directly employed to estimate the autocorrelation coefficient function. To estimate the value of the function for a difference in time periods between forecast and observed residual of six months, one would simply calculate the (Pearson or product-moment) correlation between all pairs of residuals for which both elements of the pair are for the same installation and for which the second element is the residual for the six-month period immediately subsequent to the period of the first. Similarly, to estimate the value of the function for a difference in time periods of one year, one would calculate the correlation between all pairs of residuals for which both elements are for the same installation but for which the second element is the residual for the second six-month period subsequent to that of the first. In such a manner, the autocorrelation function can be estimated at each six-month interval of difference between forecast and observed residual.

The values of the function so obtained could be substituted directly for $\rho_{(t-t')}$ in Eq. (8) to provide unbiased forecasts of future requirements; preferably, a curve can be fit to the values, and the fitted values instead employed to make the predictions. In this manner, if the autocorrelation is high between observations at a single installation taken several years apart, the precision of estimation may be substantially improved.

---

*If estimates of the autocorrelation between residuals only six months or one year apart are very close to zero, little improvement in precision is to be gained with the model incorporating the autoregressive structure. Hence, the estimation procedure should be discontinued, and this model should be discarded in favor of the simpler model assuming a zero autocorrelation.

EXTENSIONS OF THE MODELS

The models built in this study are for all systems currently operational on the Burroughs 3500 at 72 A level installations listed in Table 3. We now discuss extensions of these models to the other A level installations, to the B level installations, to the Univac 1050, to those currently operational systems still to be operational in the future, and, finally, to systems not yet operational.

Other A Level Installations

During the period for which we obtained our data (the last half of fiscal year 1972), there were 77 A level installations; as of January 1973, there were 81. Our models are explicitly applicable only to the 72 on which they were built. Obviously, we would like to extend the applicability to include the 9 additional bases and any others subsequently established. We omitted 2 of the A level installations on which we had data, since B level installations existed at the same base. These require special treatment, as is discussed below regarding extension to the B level installations. Two others were omitted as they were thought possibly to be unique. Another was excluded for lack of data. The latter, and any new installations established subsequent to the period of our data, are likely to be well represented by the models herein developed. That is, forecasts made for these installations with the specific models of this study would likely be close to actual values. It is also possible that the models would well represent the two bases omitted for their uniqueness. In any case, all of these should be checked to see if, in fact, the developed models are appropriate estimators of their load. This can be done simply by comparing predicted and actual current workloads. Alternatively, it can be accomplished while verifying the coefficients, as discussed previously, by including the additional installations in the data base and checking for significant deviations. We expect the models will be able to represent the workload from most of these additional installations.

B Level Installations

As discussed in Appendix B, there are a number of difficulties in modeling B level installations, which require detailed analyses beyond the scope of this report. It is felt, however, that models for most of the installations can be developed, though perhaps the precision obtainable will not be as high as with the A level models.

The first problem is the high level of support to nonstandard systems. The load from a system unique to an installation obviously cannot be modeled with cross-sectional data. Theoretically, longitudinal data from each single installation could be used, but the time it would take to obtain the necessary variation in the independent variables is likely to make this recourse infeasible. Probably the best approach is to model the processing requirements from all but the major nonstandard systems, estimate the load from the nonstandard systems with separate analyses, and sum these estimates to forecast the total load.

Two sources of difficulty that frequently arise with B level installations are the presence of two installations at a single base and relatively heavy satelliting. In both cases, the problem is the determination of appropriate values of the independent variables. When a base has two installations we should, if possible, partition the base into two segments, each with its own machine supporting for it all the systems being modeled. The values of the independent variables should, of course, be those corresponding to each segment. The problem becomes much more complex if each computer supports some systems for the whole base and others for only portions of the base. The solution requires a detailed analysis of the workload supported by each machine. It may be that some of these installations cannot be incorporated into a general model.

The problem with satelliting is similar: the host supports its satellites *at most* for the three largest functional systems. For example, a computer may support the military personnel system for the military population of both host and satellites, but support the military pay systems only for the military population of the host. An analysis of the processing requirements from the satellites is required to decide how they should be treated.

The diversity of the B level installations is another potential source of difficulty. If it proves to be so, the best solution may be to build a number of command-specific models: perhaps one for the five SAC installations, one for the ATC bases, another for the five AFLC depots, one for the PACAF bases, and perhaps another for the three MAC installations. Again, longitudinal data would need to be employed and any autocorrelation taken into account. With such a small number of installations for each, these models should probably be based upon only a single independent variable, probably either Total Base Population, Military Population, or Airmen. Attempting to handle the ten remaining installations with a single model would likely provide less precise estimation, though it may well suffice.

The final difficulty mentioned for the B level installations is the smaller number of them with which to build models. This can be handled, as discussed above, by employing longitudinal data, that is, several observations from different periods of time for each installation.

The Univac 1050

The other current base level computer, the Univac 1050-II, should also be amenable to the methods of this study. Inasmuch as it is a system "dedicated" to supply, one would expect it to be more readily modeled than the Burroughs 3500 with the wide variety of functional areas that it supports. The processing requirements cannot be modeled directly, however, since there are no hardware utilization data for this machine. Nevertheless, there are available several surrogates such as number of inputs and number of transactions that could be modeled; of course, forecasts of these would have to be translated into measures of hardware utilization.

In simply correlating number of transactions with the authorized manpower in Base Supply, we obtained a correlation coefficient of .84. This, of course, implies the existence of a regression model, with these as dependent and independent variables, respectively, that achieves an R^2 of $(.84)^2 = .70$. Hence, we feel confident that a model can be built that can estimate future workload on the 1050 with high precision.

Specifically, it may well be that such a model can be built based solely upon the manpower authorizations for Base Supply (functional account code 41XX) or its subfunctions. Additional variables that may help in building such a model include the weapon system authorizations (both number of aircraft and flying hours) and the manpower authorizations for Mission Equipment Maintenance, Civil Engineering, Ground Communications (38XX), and Transportation (42XX). Finally, our Base Maintenance Cost variable may be useful, or perhaps the analogous variable based simply on the base material support cost.

Currently Operational Systems Still to be Operational in the Future

Models like those developed in this study can be employed to predict the workload only from functional systems currently operational; it is important to note further that the specific models built in this study predict workload from *all* functional systems currently operational. No attempt is made here to deduct the load from any systems that may be planned for phase-out in the future. To take these into account, one can either deduct an estimate of the load from these systems from forecasts made with the existing models, or build new models that include only systems that will be operational in the period for which forecasts are to be made. The former approach would likely suffice if the workload for systems to be phased out were small; the latter approach would otherwise be preferable.

Systems Not Yet Operational

The prediction of load from functional systems yet to be implemented requires an entirely different analysis. The techniques herein discussed have a potential application to this problem, however, as a complement to this other analysis. In trying to analyze the processing requirements for a new system, the first analysis would likely use current data on such measures as number of transactions, and then would transform these into estimates of hardware utilization. The problem in so doing is that the number of transactions, and hence the corresponding hardware workload, may well be different in the future. The methods of this study can be used at either end of this analysis.

Applying them beforehand, they can be used to predict the number of transactions, which can then be transformed by the first analysis into a measure of hardware utilization. Alternatively, having first transformed current transaction data into estimated "current" utilization by the first analysis, the technique of this study can be applied to predict utilization directly. In the first case, number of transactions would be the dependent variable; in the second, it would be the measure of utilization. In both cases, the independent variables would be base characteristics.

SUMMARY

The command models as herein developed should be verified and then maintained on a periodic basis. They should be used annually to forecast processing requirements at each installation for each of the five subsequent years. The most promising ways to improve these models would be decompositions of the Other Commands models into several command-specific models, and the estimation of the autocorrelation coefficient. The most profitable future endeavor would be extension of the models to the other A level installations, to the B level installations, and to the Univac 1050. As needed, extensions can be made to include only those current systems still to be operational in the future, and systems not yet operational.

Appendix A

THE BASE MAINTENANCE COST VARIABLE

This Appendix describes both the motivation behind and the means by which to calculate the Base Maintenance Cost variable. We desired a measure of aircraft activity because we thought it might be closely related to the processing requirements to support the Maintenance Data Collection System, the Aerospace Vehicle Status Reporting System, and the Flight Data Management System.

Total Flying Hours aggregated across all weapon systems provides one possible measure, but it has the disadvantage of weighting equally the flying hours of T-41s and F-111s. Obviously, the F-111 generates more maintenance transactions and, hence, requires more processing to support the Maintenance Data Collection System. Another alternative would be to use the flying hours for each Model/Design/Series, or for aggregations of MDSs, perhaps using the groups we employed for pilots (Transports, Fighters, Bombers, and Reconnaissance and Trainers). The disadvantage here is that too many independent variables are created. Having one independent variable for each MDS is completely infeasible; having one for each of several categories is to be avoided, if possible.

Hence, we have instead defined a single independent variable, which is simply a weighted average of the flying hour authorizations for each MDS, the weights being the base maintenance cost per flying hour for that MDS. In this manner, flying hours for F-111s are weighted twelve times as heavily as those for T-41s.*

The base maintenance cost variable is defined algebraically as follows:

$$\text{Base Maintenance Cost} = \sum_{i \in S} c_i f_i ,$$

*The F-111 has a base maintenance cost per flying hour of $550, whereas the T-41 has a cost of only $43.

where c_i = total base maintenance cost per flying hour given in Table 35 for the i^{th} MDS,

f_i = authorized flying hours for the i^{th} MDS, and

S = set of distinct MDSs.

Hence, the value of the variable at a base equals the sum, across MDSs at the base, of the products of the total base maintenance cost per flying hour for an MDS with the corresponding total quarterly flying hours authorization. Table 35 presents the total base maintenance cost per flying-hour factors for each MDS. These were obtained from the 10 May 1972 update of Table 12A ("Aircraft Maintenance Cost Per Flying Hour Factors") in AFM 172-3.

Consider a base that has 20 B-52Gs and 40 KC-135s, each with a quarterly authorization of 100 flying hours. This makes a quarterly total of 2000 and 4000 flying hours for the two MDSs. The value of the cost variable is obtained by computing

$$\text{Base Maintenance Cost} = (496)(2000) + (224)(4000) = 1{,}888{,}000,$$

since the total base maintenance costs per flying hour for the B-52Gs and the KC-135s as presented in Table 35 are \$496 and \$224.

In forecasting with the models of this report, it is imperative to use the weights of Table 35. It is strictly inappropriate to use those from any updated version that may be released. If this table becomes obsolete, then new models should be developed, using the techniques of this report, to replace those that include the Maintenance Cost variable.*

Furthermore, it is necessary that the authorizations for each MDS be expressed in terms of *quarterly* flying hours. The authorized

---

*To incorporate into this variable the activity of new weapon systems for which no factor is now included in this table, it is reasonable, however, to use factors newly derived for these systems. But is necessary to base them on the same factor prices, such as the cost per man-hour of labor, used to derive the base maintenance cost for the old systems.

quarterly flying hours for a given MDS at a given base should be computed from the PA and PD by the following methods:*

Bases Other Than Forward Operating Bases

1. Using the PA, divide by four the total of the four quarterly flying hour authorizations, corresponding to the fiscal year of interest, the appropriate MDS, and the command to which the aircraft are assigned (excluding any authorizations to F.O.B. units), in order to obtain the average quarterly flying hour authorization.
2. Again using the PA, total the operating active aircraft, corresponding to the same fiscal year and MDS as in #1, for all units in the command (excluding any for F.O.B. units) for the four quarters, and divide by four to obtain the average quarterly operating active aircraft.
3. Divide the average quarterly flying hour authorization (from #1) by the average quarterly operating active aircraft (from #2) to obtain the average utilization rate.
4. Multiply this average utilization rate times the number of aircraft of this type authorized at the base of interest (as obtained from the PD) to obtain the authorized quarterly flying hours for that aircraft type at that base.

Forward Operating Bases

1. Same as above, using instead the total of the four quarterly flying hour authorizations to all F.O.B. units.
2. Same as above, using instead the operating active aircraft for all F.O.B. units in the command.

3,4. Same as above.

---

*This method (see reference in final footnote on p. 9) reproduces the flying hour figures on the basis of which the models were developed to within an average absolute difference of about 6 percent. An analysis of the effect of this discrepancy on predictions with models requiring these figures (i.e., those using the base maintenance cost variables) indicated that the effect would typically be very small (less than one percent).

Table 35

TOTAL BASE MAINTENANCE COST PER FLYING HOUR FACTORS

MDS	Total Base Maintenance Cost per Flying Hour ($)
Attack	
A-1	118
A-7	363
A-37	166
A-X	233
Bomber	
B-1	697
B-52C,D,E	484
B-52F,G	496
B-52H	496
B-57A,B,C [WB-57C]	204
B-57G	434
RB-57F [WB-57F, B-57E]	307
EB-66	487
FB-111	502
Cargo/Transport/Recon	
C-5	594
C-7	137
C-9	201
C-47	131
C-54	207
C-97	296
KC-97	313
C-118	243
AC-119K	398
C-119	168
C-121	273
EC-121	335
C-123J	133
C-WC/VC-123K	181
C-124	239
C-130A,B,C	264
C-130E	287
AC-130A [DC-130A]	546
AC-130E [DC-130E]	450
HC-130H,N,P	293
RC-130A	357
WC-130A,B,E	313
C-131	150
C-133	264
C-135B	246
EC-135	298
KC-135	224
C/RC-135	280
C-140	191
C-141	316
AABNCP	618
AWACS (E-3A)	244

MDS	Total Base Maintenance Cost per Flying Hour ($)
Fighter/Recon	
RF/F-4C	500
F-4D,E	438
F-5	179
F-15	449
F-84	245
F-86	243
F-100	350
F-101	461
RF-101	530
F-102	367
F-104	284
F-105	535
F-106	572
F-111	550
Helicopter	
H-1	104
H-1N	123
H-3	223
H-19	131
H-21	217
H-34	128
H-43	126
CH-47	134
H-53	231
Observation	
O-1	40
O-2	59
OV-10	71
Trainer	
T-28	98
T-29	137
ET-29	177
T-33	140
T-37	72
T-38	130
T-39	110
T-41	43
T-43 (T-X)	150
Utility	
U-3	60
U-4	91
U-6	79
U-10	69
HU-16	133
U-17	70
QU-22	127

SOURCE: U.S. Department of the Air Force, *USAF Cost and Planning Factors* (U), AFM-172-3, Washington, D.C., October 1970 (Confidential), Table 12A updated May 10, 1972. The table is unclassified.

NOTE: To those MDSs enclosed in brackets, we assigned the base maintenance cost per hour of the MDS on the same line.

Appendix B

DIFFICULTIES IN MODELING B LEVEL INSTALLATIONS

This appendix discusses problems posed by the B level installations, which convinced us that modeling their processing requirements would require detailed analyses.

One difficulty is the relatively high level of support given to nonstandard systems. At two installations more than half the total load is from such systems; at another five, it is at least one-fifth. In contrast to an average A level base with only 5 percent, the average B level receives 13 percent of its load from such systems.* This difference probably makes it more difficult to devise a general model for these bases.

Another problem arises when a single Air Force base has two installations; there are five such bases.† Of the two Wright-Patterson installations, for example, that belonging to the Logistics Command supports the entire base level military personnel system, whereas both support the general accounting and finance system. With such a division of workload, a very careful, detailed analysis is called for to determine the values of the independent variables for each machine.

A third difficulty is relatively heavy satelliting. Whereas the A level bases have only 8 satellites supported by the 77 installations, the 39 B level bases host a total of 29 satellites (the one at Bolling alone supports 7). Again, the problem is to determine the appropriate values of the independent variables; these must be selected to correspond to the workload generated, be it from host, satellite, or both. If the military population were used as a predictor of total direct time, one could then add the military population of the satellite to

---

*These figures are based on utilization figures for February 1972.

†These are Andrews (Headquarters Command and Headquarters Systems Command); Griffiss (Systems Command (A level) and Strategic Air Command); Kelly (Special Services and Logistics Command); Robins (Headquarters Reserves (A level) and Logistics Command); Wright-Patterson (Headquarters Logistics Command and System Command).

that for the host. The problem is complicated, however, by the fact that only some of the functional systems are supported for the satellite; hence, the military population of the satellite is only partially supported by the host, and it is then inappropriate to either include or exclude it.

Furthermore, it may be more difficult to provide a general model for the B level installations simply because of their diversity. While most of the A level installations have primarily an operational mission, the B level have functions ranging from headquarters to logistics. This diversity may cause the utilization of even the standard systems supported on the 3500 to differ markedly.

The final problem is the small number of B level installations with which to build a model. The problem is still worse if we eliminate those for which the above problems are particularly bad.

Together, these problems convinced us that the B level installations could be modeled only with detailed analyses beyond the scope of this report.*

---

*A preliminary analysis did in fact suggest that the B level installations could not be modeled as readily as is done herein for the A level.

Appendix C

MODELS FOR THE MAJOR FUNCTIONAL SYSTEMS

The best regressions for the eleven major functional systems as obtained in Sec. III are here presented in detail. For each system, the equation with the best single independent variable is given; if a different equation achieves the minimum standard error among all regressions run and has each of its coefficients significant at the .10 level (except as otherwise noted), it too is presented. In each case, the estimated regression equation is given, together with its R^2, the standard error of the estimate, the standard error as a percent of the mean, the F statistic (with the degree of freedom for its numerator and denominator, respectively), and the significance level of the F statistic (denoted P).

REGRESSION EQUATIONS FOR DIRECT TIME OF BASE LEVEL MILITARY PERSONNEL SYSTEM (NAE)

Best independent variable, Eq. (4):

$$Y = 29.81 + .008837\ X$$

where Y = NAE Direct Time,
X = Airmen.

$R^2 = .627$
s = 9.956
s as % of mean = 16.6
$F(1,70) = 117.8$
P = .000000

Minimum standard error with all coefficients significant, Eq. (7):

$$Y = 25.20 + .5245\ X_1 + .007332\ X_2$$

where Y = NAE Direct Time,

X_1 = Data Control/Consolidated Base Personnel Office (165X),

X_2 = Airmen.

$R^2 = .642$

$s = 9.827$

s as % of mean = 16.4

$F(2,69) = 61.9$

$P = .000000$

<u>REGRESSION EQUATION FOR DIRECT TIME OF BASE ENGINEER AUTOMATED MANAGEMENT SYSTEM (NAT)</u>

Best independent variable, Eq. (3):

$$Y = 7.767 + .03967\ X$$

where Y = NAT Direct Time,

X = Civil Engineering (44XX).

$R^2 = .468$

$s = 5.566$

s as % of mean = 22.1

$F(1,68) = 59.9$

$P = .000000$

<u>REGRESSION EQUATIONS FOR DIRECT TIME OF GENERAL ACCOUNTING AND FINANCE SYSTEM (NBQ)</u>

Best independent variable, Eq. (3):

$$Y = 7.097 + 1.812\ X$$

where Y = NBQ Direct Time,

X = Accounts Control (1511).

$R^2 = .605$

$s = 4.411$

s as % of mean = 22.1

$F(1,70) = 107.2$

$P = .000000$

Minimum standard error with all coefficients significant, Eq. (31):

$$Y = 4.325 + .9782\ X_1 + .5163\ X_2 + .4223\ X_3 + .2280\ X_4$$

where Y = NBQ Direct Time,

X_1 = Accounts Control (1511),

X_2 = Civilian Pay (1513),

X_3 = Travel (1514),

X_4 = Commercial Services (1515).

$R_2 = .713$

$s = 3.866$

s as % of mean = 19.2

$F(4,67) = 41.7$

$P = .000000$

REGRESSION EQUATION FOR DIRECT TIME OF VEHICLE INTEGRATED MANAGEMENT SYSTEM (NRA)

Best independent variable, Eq. (4):

$$Y = 4.610 + .09248\ X$$

where Y = NRA Direct Time,

X = Vehicle Maintenance (4241).

$R_2 = .379$

$s = 3.035$

s as % of mean = 33.7

$F(1,69) = 42.0$

$P = .000000$

REGRESSION EQUATIONS FOR DIRECT TIME OF MAINTENANCE DATA COLLECTION SYSTEM (NBD)

Best independent variable, Eq. (21):

$$Y = 3.541 + .000001476\ X$$

where Y = NBD Direct Time,
X = Base Maintenance Cost.

$R^2 = .592$
s = 2.066
s as % of mean = 29.0
F(1,69) = 99.9
P = .000000

Minimum standard error with all coefficients significant, Eq. (24):

$$Y = 2.698 + .001086\ X_1 + .000001073\ X_2$$

where Y = NBD Direct Time,
X_1 = Mission Equipment Maintenance (2XXX),*
X_2 = Base Maintenance Cost.

$R^2 = .636$
s = 1.965
s as % of mean = 27.6
F(2,68) = 59.4
P = .000000

REGRESSION EQUATIONS FOR DIRECT TIME OF CIVILIAN PAY SYSTEM (NBS)

Best independent variable, Eq. (4):

$$Y = .3829 + .004372\ X$$

*Depot Maintenance (27XX) is excluded.

where Y = NBS Direct Time,

X = Civilian Population.

R^2 = .758

s = 1.468

s as % of mean = 33.8

F(1,62) = 194.3

P = .000000

Minimum standard error with all coefficients significant, Eq. (6):

$$Y = -.5775 + .4163\ X_1 + .003221\ X_2$$

where Y = NBS Direct Time,

X_1 = Civilian Pay (1513),

X_2 = Civilian Population.

R^2 = .795

s = 1.362

s as % of mean = 31.4

F(2,61) = 118.4

P = .000000

REGRESSION EQUATION FOR DIRECT TIME OF ACCRUED MILITARY PAY SYSTEM (NBU)

Best independent variable, Eq. (2):

$$Y = .5908 + .1404\ X$$

where Y = NBU Direct Time,

X = Military Pay (1512).

R^2 = .884

s = 1.067

s as % of mean = 27.6

$F(1,44) = 336.4$

$P = .000000$

REGRESSION EQUATIONS FOR DIRECT TIME OF MEDICAL MATERIAL MANAGEMENT SYSTEM (NAV)

Best independent variable, Eq. (3):

$$Y = -.4642 + .3429\ X$$

where Y = NAV Direct Time,
X = Medical Material (5110).

$R^2 = .604$

$s = 1.588$

s as % of mean = 49.8

$F(1,67) = 102.3$

$P = .000000$

Minimum standard error with all coefficients significant, Eq. (11):

$$Y = .1141 + .4137\ X_1 - .09605\ X_2$$

where Y = NAV Direct Time,
X_1 = Medical Material (5110),
X_2 = Physicians (5201).

$R^2 = .701$

$s = 1.389$

s as % of mean = 43.6

$F(2,66) = 77.5$

$P = .000000$

REGRESSION EQUATIONS FOR DIRECT TIME OF AEROSPACE VEHICLE STATUS REPORTING SYSTEM (NAW)

Best independent variable, Eq. (9):

$$Y = 2.825 + .01619\ X$$

where Y = NAW Direct Time,
X = Aircraft.

$R^2 = .442$
s = .904
s as % of mean = 22.7
F(1,69) = 54.7
P = .000000

Minimum standard error with all coefficients significant,* Eq. (31):

$$Y = 2.904 - .005870\ X_1 + .007848\ X_2 + .007098\ X_3 - .004489\ X_4 + .0001480\ X_5$$

where Y = NAW Direct Time,
X_1 = Transport Pilots,
X_2 = Fighter Pilots,
X_3 = Bomber Pilots,
X_4 = Reconnaissance and Trainer Pilots,
X_5 = Flying Hours.

$R^2 = .574$
s = .814
s as % of mean = 20.4
F(5,65) = 17.5
P = .000000

---

*Here one coefficient, that of the Reconnaissance and Trainer Pilots variable, is significant only at the .16 level.

REGRESSION EQUATION FOR DIRECT TIME OF JOINT UNIFORM MILITARY PAY SYSTEM (NBT)

Best independent variable, Eq. (4):

$$Y = 1.217 + .0005663\ X$$

where Y = NBT Direct Time,
X = Military Population.

$R^2 = .466$
s = 1.003
s as % of mean = 28.8
$F(1,70) = 61.2$
P = .000000

REGRESSION EQUATIONS FOR DIRECT TIME OF FLIGHT DATA MANAGEMENT SYSTEM (NBP)

Best independent variable, Eq. (8):

$$Y = 1.564 + .005809\ X$$

where Y = NBP Direct Time,
X = Rated Pilots.

$R^2 = .449$
s = .856
s as % of mean = 33.8
$F(1,69) = 56.2$
P = .000000

Minimum standard error with all coefficients significant, Eq. (28):

$$Y = 1.325 + .009629\ X_1 - .006863\ X_2 + .001457\ X_3$$

where Y = NBP Direct Time,

X_1 = Bomber Pilots,

X_2 = Reconnaissance and Trainer Pilots,

X_3 = Flying Hours.

R^2 = .569

s = .768

s as % of mean = 30.3

$F(3,67)$ = 29.4

P = .000000

Appendix D

ESTIMATED VARIANCE-COVARIANCE MATRICES FOR THE GENERAL MODELS

Presented below are the three estimated variance-covariance matrices corresponding to the general models given in Table 21. The matrices are obtained from the equation

$$\hat{V} = s^2 \; (X'X)^{-1} \; ,$$

where $\hat{V}$ = estimate of the variance-covariance matrix to be calculated,

s = standard error of the estimate,

X = matrix of observations.

The general form of the matrices is given by

$$\hat{V} = \begin{pmatrix} \widehat{V(b_0)} & \widehat{\mathrm{Cov}\ (b_0 b_1)} & \cdots & \widehat{\mathrm{Cov}\ (b_0 b_p)} \\ \widehat{\mathrm{Cov}\ (b_0 b_1)} & \widehat{V(b_1)} & \cdots & \widehat{\mathrm{Cov}\ (b_1 b_p)} \\ \vdots & \vdots & \vdots & \vdots \\ \widehat{\mathrm{Cov}\ (b_0 b_p)} & \widehat{\mathrm{Cov}\ (b_1 b_p)} & \cdots & \widehat{V(b_p)} \end{pmatrix}$$

Hence, the value of the first element of the principal diagonal is the estimated variance of b_0, the constant term, and the second element of this diagonal is the estimated variance of the coefficient of the first independent variable. The off-diagonal elements are, as indicated, the estimated covariances of the coefficients.

ESTIMATED VARIANCE-COVARIANCE MATRICES FOR THE GENERAL DIRECT TIME MODELS

Model 1

1.042×10^{2}	-4.234×10^{-3}	-2.259	-1.478×10^{-1}
-4.234×10^{-3}	7.157×10^{-6}	-2.579×10^{-4}	-4.249×10^{-5}
-2.259	-2.579×10^{-4}	7.948×10^{-1}	-6.899×10^{-3}
-1.478×10^{-1}	-4.249×10^{-5}	-6.899×10^{-3}	8.071×10^{-4}

Model 2

9.297×10	-2.102	-1.409×10^{-1}	-5.041×10^{-3}
-2.102	7.047×10^{-1}	-7.138×10^{-3}	-1.477×10^{-4}
-1.409×10^{-1}	-7.138×10^{-3}	5.963×10^{-4}	-3.615×10^{-5}
-5.041×10^{-3}	-1.477×10^{-4}	-3.615×10^{-5}	1.311×10^{-5}

ESTIMATED VARIANCE-COVARIANCE MATRIX FOR THE GENERAL NUMBER OF I/Os MODEL

1.721×10^{12}	-8.800×10^{7}	-6.587×10^{10}	-1.378×10^{9}	-2.461×10^{10}
-8.800×10^{7}	8.898×10^{4}	9.238×10^{6}	-6.052×10^{5}	-2.187×10^{6}
-6.587×10^{10}	9.238×10^{6}	1.365×10^{10}	-1.251×10^{8}	-9.586×10^{8}
-1.378×10^{9}	-6.052×10^{5}	-1.251×10^{8}	9.813×10^{6}	1.319×10^{7}
-2.461×10^{10}	-2.187×10^{6}	-9.586×10^{8}	1.319×10^{7}	3.239×10^{9}

Appendix E

ESTIMATED VARIANCE-COVARIANCE MATRICES FOR THE COMMAND MODELS

The estimated variance-covariance matrices for the six command models given in Table 30 are presented below. The equation from which these are derived and the general form of the matrices are shown in Appendix D.

ESTIMATED VARIANCE-COVARIANCE MATRICES FOR THE COMMAND DIRECT TIME MODELS

SAC

$$\begin{matrix} 1.304 \times 10^{2} & -4.349 \times 10^{-1} & -1.003 \\ -4.349 \times 10^{-1} & 3.677 \times 10^{-3} & -6.179 \times 10^{-4} \\ -1.003 & -6.179 \times 10^{-4} & 1.768 \times 10^{-2} \end{matrix}$$

TAC

$$\begin{matrix} 2.277 \times 10^{2} & 9.394 \times 10^{-2} & -6.564 \times 10^{-2} & -1.991 \times 10^{-5} \\ 9.394 \times 10^{-2} & 1.536 \times 10^{-4} & -5.335 \times 10^{-5} & -4.586 \times 10^{-8} \\ -6.564 \times 10^{-2} & -5.335 \times 10^{-5} & 3.145 \times 10^{-5} & 3.835 \times 10^{-9} \\ -1.991 \times 10^{-5} & -4.586 \times 10^{-8} & 3.835 \times 10^{-9} & 3.189 \times 10^{-11} \end{matrix}$$

Other Commands

$$\begin{matrix} 1.683 \times 10^{2} & -1.463 & -3.244 \times 10^{-2} & -2.567 \times 10^{-1} \\ -1.463 & 1.497 & -1.942 \times 10^{-3} & -2.341 \times 10^{-2} \\ -3.244 \times 10^{-2} & -1.942 \times 10^{-3} & 6.189 \times 10^{-5} & -7.663 \times 10^{-5} \\ -2.567 \times 10^{-1} & -2.341 \times 10^{-2} & -7.663 \times 10^{-5} & 1.501 \times 10^{-3} \end{matrix}$$

ESTIMATED VARIANCE-COVARIANCE MATRICES FOR THE COMMAND I/O MODELS

SAC

4.092×10^{12}	-7.452×10^{9}	-4.185×10^{7}
-7.452×10^{9}	2.342×10^{7}	-1.131×10^{6}
-4.185×10^{7}	-1.131×10^{6}	1.575×10^{5}

TAC

1.879×10^{12}	7.754×10^{8}	-5.418×10^{8}	-1.643×10^{5}
7.754×10^{8}	1.267×10^{6}	-4.403×10^{5}	-3.784×10^{2}
-5.418×10^{8}	-4.403×10^{5}	2.595×10^{5}	3.165×10^{1}
-1.643×10^{5}	-3.784×10^{2}	3.165×10^{1}	2.632×10^{-1}

Other Commands

3.095×10^{12}	-1.560×10^{11}	-5.303×10^{8}	7.896×10^{4}
-1.560×10^{11}	2.760×10^{10}	-4.647×10^{7}	3.019×10^{4}
-5.303×10^{8}	-4.647×10^{7}	5.142×10^{5}	-2.942×10^{2}
7.896×10^{4}	3.019×10^{4}	-2.942×10^{2}	2.534×10^{-1}

Appendix F

RESIDUALS FOR THE COMMAND MODELS

Table 36 presents, for the 72 A level installations listed in Table 4, the residuals for the corresponding command models of both direct time and number of I/Os as given in Table 30. The use of these residuals in forecasting is discussed in Sec. VI under the heading, "Predictions Based on a Model with an Autoregressive Structure."

Table 36

RESIDUALS FOR COMMAND MODELS

Base	Residual for Direct Time Model (hours per Month)	Residual for I/O Model (millions per Month)
SAC		
Anderson	18.3	2.11
Beale	9.5	-0.87
Blytheville	-25.3	-2.35
Carswell	37.5	3.02
Castle	8.9	1.99
Davis Monthan	-28.5	-1.84
Dyess	1.6	-1.53
Ellsworth	11.2	-1.35
F. E. Warren	-22.5	-1.31
Fairchild	16.9	-1.22
Grand Forks	4.2	1.63
Grissom	-10.8	0.34
Lockbourne	7.1	1.24
Loring	-11.8	-0.92
Malmstrom	19.8	1.93
March	-8.8	0.40
McCoy	-4.8	-0.04
Minot	-2.1	1.21
Pease	-3.9	1.07
Plattsburgh	-8.2	-0.88
Whiteman	-14.6	-3.87
Wurtsmith	6.1	1.25
TAC		
Cannon	11.6	1.29
England	-16.6	-0.98
Forbes	1.3	-0.94
George	-22.2	-1.56
Holloman	0.3	-0.22
Homestead	-4.3	0.44
Hurlburt	16.8	0.38
Little Rock	0.8	0.46
Luke	6.9	0.32
MacDill	13.7	1.28
McConnell	-21.3	-2.30
Mountain Home	11.0	1.67
Myrtle Beach	4.6	0.56
Nellis	13.8	1.19
Pope	-6.2	-0.94
Seymour Johnson	-6.8	-0.47
Shaw	-3.3	-0.17

Base	Residual for Direct Time Model (hours per Month)	Residual for I/O Model (millions per Month)
Other Commands:		
ATC		
Columbus	18.6	0.94
Craig	2.7	-1.11
Laredo	-9.4	-0.03
Laughlin	25.9	-0.09
Mather	36.3	1.22
Moody	-56.8	-3.73
Reese	17.0	2.17
Webb	3.4	0.53
Williams	-13.8	-0.85
AFE		
Aviano	-41.5	-2.04
Bentwaters	-50.9	-6.17
Bitburg	26.3	4.00
Incirlik	-18.1	-2.25
Lakenheath RAF	60.2	3.02
Rhein-Main	-4.4	-2.04
Torrejon	23.8	-2.69
Upper Heyford RAF	18.2	0.54
MAC		
Altus	-14.2	-2.38
Charleston	-15.2	-2.82
Dover	-33.1	-4.95
Lajes Field	-28.4	-0.41
McChord	-10.5	2.09
McGuire	10.7	2.59
AFSC		
Brooks	3.3	1.72
Edwards	3.6	5.51
Kirtland	-21.5	-1.47
L. G. Hanscom	-11.5	-4.54
Patrick	3.2	1.63
Other		
Hamilton, ADC	45.9	7.54
Tyndall, ADC	-8.2	-4.00
Maxwell, AU	11.6	1.98
Ching Chuan Kang, PACAF	0.6	4.55
Albrook, SC	26.2	1.54

Appendix G

RANGES OF THE INDEPENDENT VARIABLES FOR THE COMMAND MODELS

This appendix presents, for each of the command models, the minimum and maximum of each incorporated, independent variable over the values in the data on which the model was built. It is to be used to check for extrapolation, as discussed under the heading "Pitfalls in Prediction," Sec. VI.

	Minimum	Maximum
SAC Direct Time		
21XX (Chief of Maintenance)	18	290
4241 (Vehicle Maintenance)	30	116
TAC Direct time		
2XXX (Mission Equipment Maintenance)[a]	1,203	4,705
Total Base Population	3,707	8,685
Base Maintenance Cost ($)	1,263,500	6,224,800
Other Direct Time		
1514 (Travel)	2	21
2XXX (Mission Equipment Maintenance)[a]	33	2,689
44XX (Civil Engineering)	12	698
SAC I/O		
44XX (Civil Engineering)	363	656
Airmen	2,220	7,716
TAC I/O		
2XXX (Mission Equipment Maintenance)[a]	1,203	4,705
Total Base Population	3,707	8,685
Base Maintenance Cost ($)	1,263,500	6,224,800
Other I/O		
1511 (Accounts Control)	4	24
Military Population	1,039	6,618
Base Maintenance Cost ($)	0	8,118,500

[a]Depot Maintenance (27XX) is excluded).