DTIC FILE COPY

AD-A220 183

VALIDATION OF THE TERRAIN EFFECTS
MODEL BY COMPARING AUTOREGRESSIVE
INTEGRATED MOVING AVERAGE MODELS

THESIS

Robert E. Massey
Major, USAF

AFIT/GST/ENS/90M-12

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GST/ENS/90M-12

VALIDATION OF THE TERRAIN EFFECTS
MODEL BY COMPARING AUTOREGRESSIVE
INTEGRATED MOVING AVERAGE MODELS

THESIS

Robert E. Massey
Major, USAF

AFIT/GST/ENS/90M-12

Approved For Public Release: Distribution Is Unlimited

AFIT/GST/ENS/90M-12

VALIDATION OF THE TERRAIN EFFECTS MODEL BY COMPARING AUTOREGRESSIVE INTEGRATED MOVING AVERAGE MODELS

THESIS

Presented to the Faculty of the School of Engineering

of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Operations Research

Robert E. Massey, B.S., M.S.A.

Major, USAF

March 1990

Approved For Public Release: Distribution Is Unlimited

Preface

The purpose of this study is to evaluate the ability of comparing time series models produced by the Terrain Effects Model to those derived from field testing. I wish to thank AFOTEC for their support in providing this topic and the data. I am especially indebted to Captain James Bennett, my contact at AFOTEC, for his aid and support for this effort.

My thanks go out to my advisor, Lieutenant Colonel James Robinson, and my reader, Lieutenant Colonel Thomas Schuppe, for their guidance, enthusiasm, and comments. Many members of my class deserve recognition for the countless hours they saved me in some difficult classes and with various software packages which gave me the time to complete this project.

Finally, I am especially grateful to my wife, Andrea, and children, Brian, Stephanie, and Lynn, who graciously returned my neglect and inattention with understanding and patience.

Robert Massey

Table of Contents

List of Figures

List of Tables

AFIT/GST/ENS/90M-12

Abstract

This thesis examines the capability of validating the Terrain Effects Model (TEM) by comparing a time series output of the model to one produced during field testing. The TEM is a simulation of an air-to-air missile's flight path while being subjected to electronic countermeasures when launched from an altitude above the target aircraft.

Data from the field and simulation tests were characterized and fit with time series models using Box-Jenkins' Autoregressive Integrated Moving Average methodology. The models had less explanatory power than that which is usually associated with a time series representation, most probably caused by large deterministic noise intrusions in the field data. Several recommendations for other simulation validation techniques are offered.

Validation of the Terrain Effects Model By Comparing Autoregressive Integrated Moving Average Models

I. Introduction

Background

The B-1B has had a history of problems with its electronic countermeasure system, the ALQ-161. Due to its high cost and the political controversy surrounding any strategic weapons system, the aircraft's deficiencies have been highly publicized and have drawn close scrutiny from Congress. The Air Force has been developing and testing many improvements to the ALQ-161 system carried by the B-1B. If the Air Force wishes to field any of these improvements, thorough and convincing evidence of their capabilities will be needed to overcome Congressional skeptics and win funding.

One of the Air Force's major testing organizations is the Air Force Operational Test and Evaluation Center (AFOTEC) at Kirtland Air Force Base, New Mexico. AFOTEC performs independent operational testing and evaluation for the Chief of Staff. "The primary purpose of operational test and evaluation is to reduce risk in the acquisition process by determining how well systems perform when operated and maintained by Air Force personnel in a realistic operational environment." (Air Force Magazine, 1989:116). AFOTEC serves as an impartial third party since they are not involved in the actual designing, procurement, and deployment of weapons systems in the same way as the command gaining the weapons system or Systems Command.

As with many weapons systems, it is not possible to directly test how the system will perform in combat. Yet, with all that is at stake in systems such as the B-1B, some kind of testing and confidence building must take place. When direct testing is not possible, other surrogate measures are often used. One way to test a weapon system is to test how well it performs against another weapon it is likely to be pitted against in battle. AFOTEC has chosen a particular missile which is believed to be representative of the Soviet air-to-air threat against the B-1B. Flown in a captured configuration on an F-15 with data recording capabilities, this is known as a "Golden Bird" system (Bennett, 1989). The Golden Bird system was used as a surrogate to gather air-to-air threat data against the B-1B in field testing to be used in simulation model validation.

Field tests are very expensive. It is frequently not possible or affordable to run field tests under all the conditions which may be encountered in combat. Simulation is often used as the preferable technique to answer questions about weapon system performance.

The Terrain Effects Model (TEM) is a simulation of an air-to-air missile's flight path when launched from an altitude above the target. Since present and foreseeable future doctrine is for bombers to penetrate enemy airspace at low level, this will be the most likely scenario of air-to-air threat from which the bombers can expect to be put at risk. The Air Force has awarded a contract for over $300,000 to update the Terrain Effects Model to simulate the B-1B's newer electronic counter measures techniques (ECM) against the more modern Soviet threats (Bennett, 1989). This simulation evaluates the ALQ-161's performance against air-to-air threats over various types of terrain. Some field test data has been provided to the contractor to aid in model calibration. The remainder of the data sets from the initial field test, and a future set from a field test over a different type of terrain, will be used to validate the simulation model. If the simulated model can emulate the behavior of the field test, then some confidence can be placed in its ability to predict the B-1B's ECM under other conditions.

AFOTEC plans to use time series analysis to fit time series models to the field test data sets. Each model derived from a field test will be applied to a corresponding simulated data set. The residuals from both sets will be statistically compared for similarity. If the simulated data compares favorably with the field test data the simulation model will be considered valid to predict the ECM technique in question over other types of terrain.

Research Problem

The capability of comparing time series to evaluate the ability of the Terrain Effects Model (TEM) to predict the outcome of the B-1B Electronic Counter Measures (ECM)/Golden Bird Field testing will be determined. If the contractor is able to build a simulation capable of predicting the results of the field test, then some confidence can be placed in the simulation's ability to predict the B-1B's ECM under other conditions.

Research Objective

The research objective is to characterize the field test data, comparing that data to the contractor's simulation output, and to evaluate methodologies for accomplishing this process.

Scope, Limitations, and Assumptions

A major assumption is that the measure of effectiveness AFOTEC has chosen (angular error between seeker head and actual aircraft position) is a valid measure of the underlying process. AFOTEC feels that so many physical phenomena contribute to angular error, the model of angular error has to capture the underlying process to be able to predict this measure over an entire test event (Bennett, 1989). Data is not available to test other measures.

Another major assumption is that the Golden Bird system is a valid surrogate. This question cannot be investigated in an unclassified report and AFOTEC already has another project investigating this question (Bennett, 1989).

Organizational Overview

Chapter Two contains a review of literature concerning the major aspects of this research. Information on why a validation effort for a simulation model is important, what work others have accomplished to validate anti-aircraft missile simulation models, and how the Box-Jenkins methodology of time series analysis is accomplished are reviewed. Chapter Three contains the planned methodology for evaluating the Terrain Effects Model's ability to simulate an actual missile. Chapter Four explains the outcomes and adjustments as the methodology was applied. The conclusions and recommendations of the thesis follow in Chapter Five.

II. Literature Review

Introduction

This literature review consists of three major parts. The first area of concern is the validation of simulation models. The reasons simulation models must be validated are covered, followed by a review on how validation is accomplished. The second major area is a review of anti-aircraft missile simulation validation techniques developed by the U.S. Army Missile Command. The last major section of this review covers the basics of time series analysis concentrating on the Box-Jenkins Autoregressive Integrated Moving Average (ARIMA) methodology. These topics will explain the basis for this thesis and the fundamentals of the techniques related to this effort.

Simulation Validation

A simulation is the mechanical manipulation of a model on a computer to numerically estimate the true characteristics of that model (Law and Kelton, 1987:1). This section of the literature review explains why a simulation model must be validated, and then contains a description of how a simulation model validation process should be performed. Validation is determining if the simulation model is an accurate representation of the real-world system of interest (Law and Kelton, 1982:33). Models are constructed to present objective and reliable information for use by decision makers (GAO, 1973:5). The essence of this process is to gain the ability to make predictions of some real-world systems.

The Need For Simulation Validation. There are many reasons why the United States Military uses simulation models to make predictions of the output of real-world systems. New weapons testing and acquisitions are among the most important of these.

> Weapon system technology is rapidly producing more complex, more costly, and more lethal weapons. This technology is the marginal hedge on which U.S. defense depends to offset the numerical advantages that the Soviets have in almost every category of weaponry. Successful development and deployment of state-of-the-art weapons places a premium on thorough testing of these new weapon systems. (Mann, 1983:1).

As the weapon systems grow more expensive and complex the actual testing grows dramatically more costly and difficult. It is impossible to test many weapon systems under conditions which replicate actual combat as might be desirable. The sophistication of the weapons, the restraints on the reality of the tests, and the costs often require the tester to gather data at a less detailed level than desired and derive results which answer the real issues in question (Mann, 1983:15). Data is usually collected in field tests on some measures of effectiveness. However, the sample sizes are usually too small to allow testers to confidently predict beyond the test conditions directly observed (Mann, 1983:44). Without enough empirical test data, simulation is frequently the technique used to gather the needed information. Since simulations are only approximations of reality, their credibility is always open to question (GAO, 1987:2).

The General Accounting Office (GAO) has been asked by Congress several times to investigate the credibility of military simulation models used in studies, doctrinal development, systems evaluation and testing, and during the acquisition process. In Simulation Modeling and Analysis, Law and Kelton say, "If a model is not a 'valid' representation of a system under study, the simulation results, no matter how impressive they may appear, will provide little useful information about the actual system" (Law, and Kelton, 1982:9). Models are only approximations of reality and must be assumed invalid until proven that no difference significant enough to affect any decisions exists between them and the underlying system they represent (Law and Kelton, 1982:341). Until validated, a simulation has no inherent credibility and the results should be considered no better than speculation (Mann, 1983:37). The lack of validation effort has been identified as a consistent weakness in the simulation models used throughout the Department of Defense (GAO, 1987:2-3). In many instances the problem was merely a lack of documentation of the validation effort performed, but in far too many others the resources were never available to attempt validation and establish any credibility.

Simulation Validation Process. The burden of proof lies on the modeler to prove the simulation should be used for the purpose intended. A simulation model could be valid under one set of conditions and not under another (Sargent, 1988:33). Unfortunately, there is no generally accepted standard for model validation (GAO, 1979:3). As such, the level of confidence in a simulation's results are not absolute, but measured on a continuum (GAO, 1987:13). Many people in the field of operations research have a number of techniques to accomplish the goal of validating simulation models. Recognizing that there is no ultimate test, the General Accounting Office recommends putting a model, "to enough appropriate tests so that qualified researchers would say that it appears to be valid or that the results are credible." (GAO, 1987:20). Researchers outside the government also recognize that confidence increases gradually as the model passes more tests and shows its correspondence with empirical reality (Forrester, 1980: 209).

Without an ultimate test to perform, most authors have several recommendations to follow in building a simulation model's credibility. They generally fall into areas which can be classified as: checking the face validity of the model, comparing the correspondence to real-world results, and the disclosure of the validation results. Robert Sargent, a leading authority on validating models, recommends the following steps as a minimum in the validation process (Sargent, 1988:38):

> (1) An agreement between the user and model builder on a minimum set of validation techniques prior to model development.
>
> (2) The assumptions and theories underlying the model be tested when possible.
>
> (3) Face validity be performed on the conceptual model.
>
> (4) The model's behavior be explored.
>
> (5) The model should be compared with the system for at least two sets of experimental conditions.
>
> (6) Validation effort be included in documentation.

After the GAO studied the credibility of DOD simulation models, they recommended the factors listed in Table 1 on the next page be considered when attempting to evaluate the credibility of any simulation model. Many of the tests for

GAO Framework For Simulation Validation		
Area of Concern		**Factor**
Theory model design, and input data	1.	Match between theoretical and simulated events
	2.	Choice of measures of effectiveness
	3.	Portrayal of weapon's immediate combat environment
	4.	Representation of performance
	5.	Depiction of Critical aspects of broad-scale battle environment
	6.	Appropriateness of mathematical and logical representation
	7.	Selection of input data
Correspondence between the model and the real world.	8.	Verification effort
	9.	Attention to statistical quality of results
	10.	Sensitivity testing effort
	11.	Validation effort
Management issues.	12.	Organization support
	13.	Documentation
	14.	Full disclosure of results

Table 1 GAO Framework For Validation (GAO, 1987:3)

building credibility are common to these and many other researchers.

There are several techniques to perform tests in many of the areas recommended. Face validity can be ascertained by methods ranging from intuitive evaluation, through agreement by experts in the field, up to performing Turing tests (Law and Kelton, 1982:341). Face Validity and correspondence to the real world are often tested simultaneously by graphical comparisons of the system and simulation.

Extreme condition testing is another popular means of determining a model's reasonableness and relationship to the actual system. A properly structured model will have no output if the inputs are reduced to zero, or if the inputs arrive faster than a system's capabilities, there should be a bottleneck building somewhere within the system. The extreme condition test enhances the simulation's credibility to predict outside the range of historical observations (Forrester, 1980:214).

The techniques which confirm that the model is representative of the real-world remain the ultimate measures for building credibility. If a problem can be solved analytically a good comparison is to see if the model arrives at the known results (Sargent, 1988:33). The best corroboration is determined by actual comparison between simulation and real-world results (GAO, 1973:20). Field tests are often the best, or only source of data to substitute for real-world results to make these comparisons.

Because the correspondence between real-world and the simulation is so important it is common for a model to be calibrated to insure this test is passed. The model's parameters or structure are manipulated until the model output and test data agree (Law and Kelton, 1982:34). It becomes absolutely critical to test the model on an independent data set to validate this method and protect against merely modeling the input received (Law and Kelton, 1982:34).

Applying all the tests recommended would be a massive undertaking. Money, time, and staff available must be weighed against the impact the simulation model will make (GAO, 1979:26). It makes little sense to follow a validation procedure that will cost more than a wrong decision derived from a simulation model. All simulations must have some form of validation, but the level of effort must be balanced against the use of the simulation results.

The costs are not the only problem associated with validation through comparisons between the system and its simulation. The output of a stochastic simulation consists of random variables and therefore is only one realization of the true characteristics of the model in question (Law and

Kelton:3). Many classical statistical techniques are used to account for the variability in the output data to allow comparisons. Unfortunately, as Law and Kelton point out, "the output processes of almost all real-world systems and simulations are nonstationary (the distributions of the successive observations change over time) and autocorrelated (the observations in the process are correlated with each other). Thus, classical statistical tests based on Independent Identically Distributed observations are not directly applicable" (Law and Kelton, 1982:341). There is no universally accepted technique for output analysis when classical statistics do not apply (Law and Kelton, 1982:279). The methods which do exist are often quite complicated to apply.

Many of these validation recommendations are being followed by AFOTEC both within and outside the context of this thesis. The builder is aware the simulation results will be compared to field test data and that they will have to predict the results of a second field experiment yet to take place (Bennett, 1989). Techniques will be explored to compare the field test and simulated data within this thesis. Some verification and evaluation of the measure of effectiveness (seeker head angle error) will be performed outside the bounds of this document.

Missile Simulation Validations

The US Army Missile Command has been very active in validating simulations of anti-aircraft missiles. The effects of electronic countermeasures have been included in many of the past simulations of the flight of anti-aircraft missiles. Validating a simulation of a missile's flight while being jammed is the same as validating a simulation of that ECM's effect on the anti-aircraft missile.

The purpose in building any simulation is to produce an output similar to the real system if it were subjected to the same input conditions. Consequently, data from actual flight tests compared with computer simulation data is the most commonly used method for validating this type of simulation (Greene and Montgomery, 1981:3). Both static and dynamic performance measures have been compared in validation efforts.

Static measures such as kill probabilities and miss distance can be analyzed using classical statistical techniques. The more dynamic measures vary continuously with time during the missile's flight producing correlated data which is usually expressed as a time series (Greene and Montgomery, 1981:3). These time series are usually highly autocorrelated, often nonstationary, and may exhibit deterministic components within their internal structure (Greene and Montgomery, 1981:5). Examples of time series variable data include missile trajectory, roll position, roll rate, various guidance parameters, and seeker line-of-sight-rates (Army, 1987:8).

Many techniques have been used for the comparisons of the time series data produced by the simulations and in the field tests. One of the most common methods has beer to apply identical inputs to the simulation as existed in a field test and then overlay the data plots from each (Kheir and Holmes, 1978:119). An example of this method appears in Figure 1.

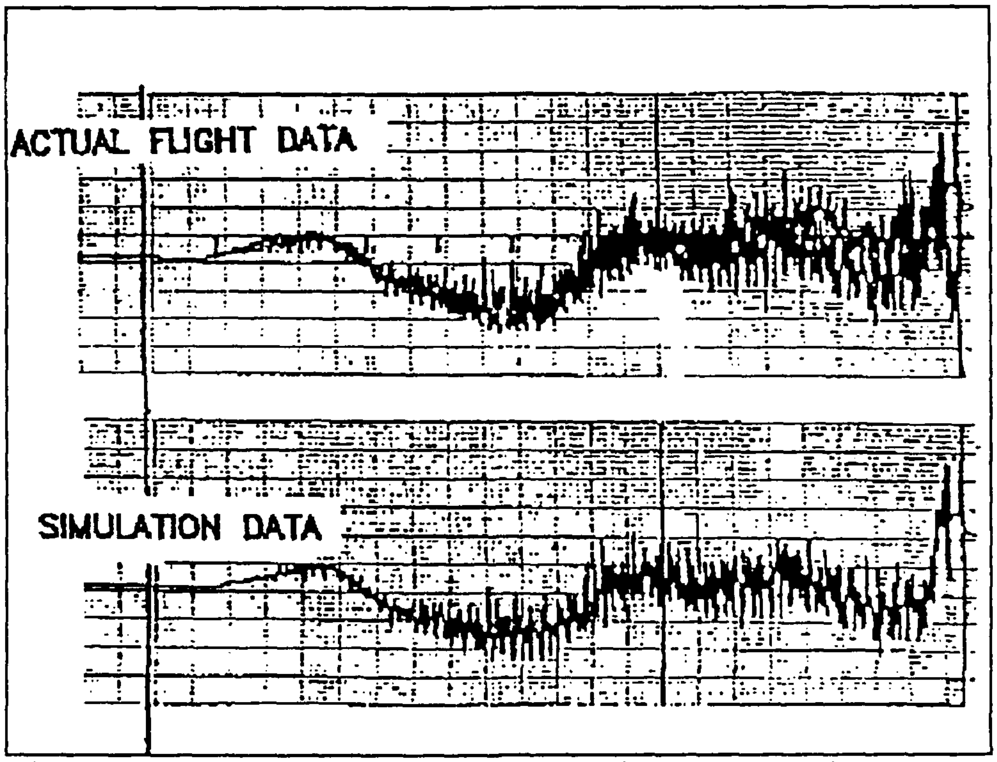


Figure 1 Time Series Plot Comparisons (Gravitz and Waite, 1988:780)

This method obviously carries some subjectivity in which different experts may disagree on how closely the results correspond. Still it is a fast and easy way to gain immediate confidence in the simulation's credibility or identify the need for improvements.

A more quantitative assessment of the likeness between two time series is the Theil's inequality coefficient. This measure has been used in many past missile simulation validations. Taking paired data points P_i and A_i from two time series, the coefficient U is determined by

$$U = \frac{\sqrt{1/n\sum (P_i - A_i)^2}}{\sqrt{1/n\sum P_i^2} + \sqrt{1/n\sum A_i^2}} \; ; \; 0 \leq U \leq 1$$

where n is the number of sampling points (Kheir and Holmes, 1978:122).

The coefficient ranges between zero and one. At zero $P_i = A_i$ for all i and perfect equality exists. At $U = 1$ the case of maximum inequality exists.

Spectral analysis is another method that has been used many times in missile simulation validations. The spectra of the simulation output is compared to the spectra from the corresponding flight test variable for similarity (Greene and Montgomery, 1981:5). Another related approach has been to fit appropriate stochastic models using Box and Jenkins

Auto Regressive Moving Average Modeling methodology to each of the time series outputs. More on this methodology is presented in the next section of this literature review. The inference is, if the time series are from the same population, the same models should be derived from fitting the series (Greene and Montgomery, 1981:5) Hunter and Hsu proposed an inferential statistic $G(\psi,T)$ where ψ is the difference between an autoregressive parameter in one time series and the corresponding autoregressive parameter in another series and T is the variance of the first series divided by the variance of the second. The inferential statistic G is distributed as a χ^2 to test if the autoregressive parameters are the same (Hunter and Hsu, 1977:182). Unfortunately for the short time duration of the output of a missile system, it is possible to have significantly different models from the same underlying stochastic process due to differences in many factors such as phase angle, gain, or frequency (Greene and Montgomery, 1981:5).

A method used in a recent Chaparral missile simulation validation was to draw input parameters from representative populations and do Monte Carlo simulations. Because each run of a simulation is only one stochastic realization of the outcome, several realizations of the simulation were used to produce a mean simulation value with its standard deviation. The actual field test results were then graphically overlayed on the simulation results for comparison. This process is shown in Figure 2 below.

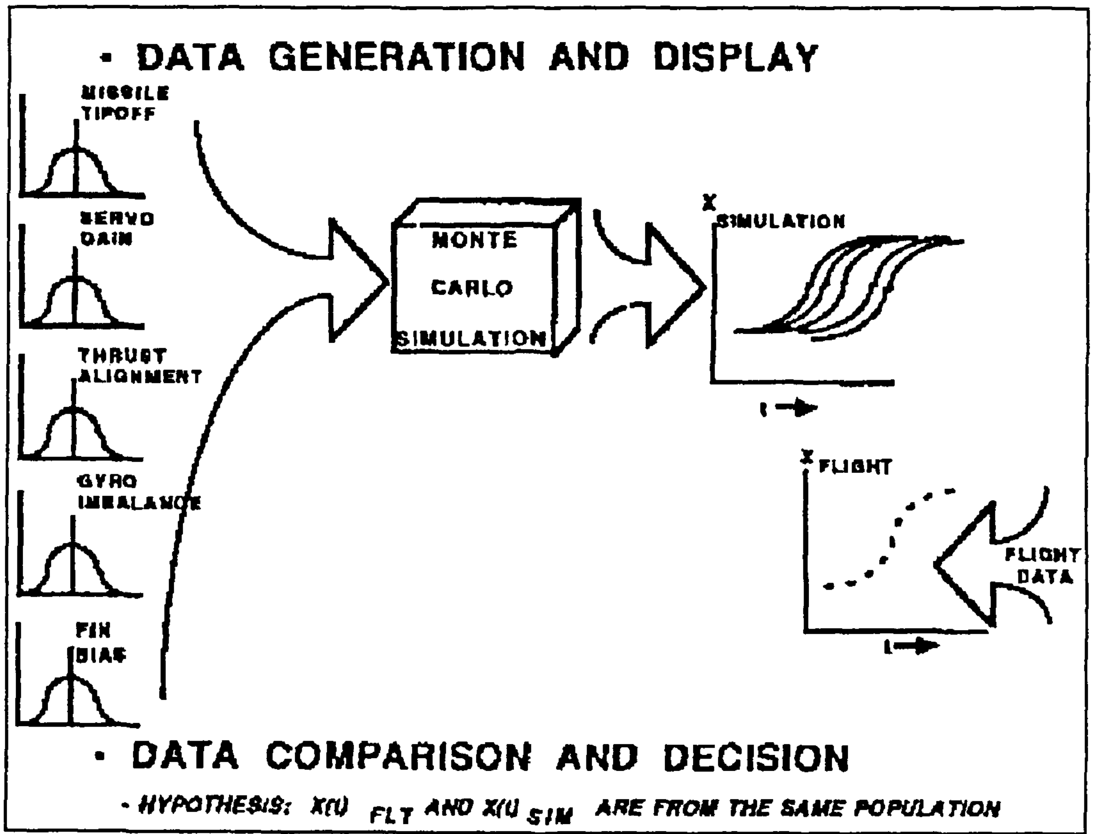


Figure 2 Simulation Validation Strategy (Gravitz and Waite, 1988:775)

Several stochastic realizations of the simulation were used as shown in Figure 3 on the following page to derive mean values with the standard deviations for each output being measured for comparison. To better test specific characteristics, nominal overlays were also created by holding the variance of the independent input parameters to zero.

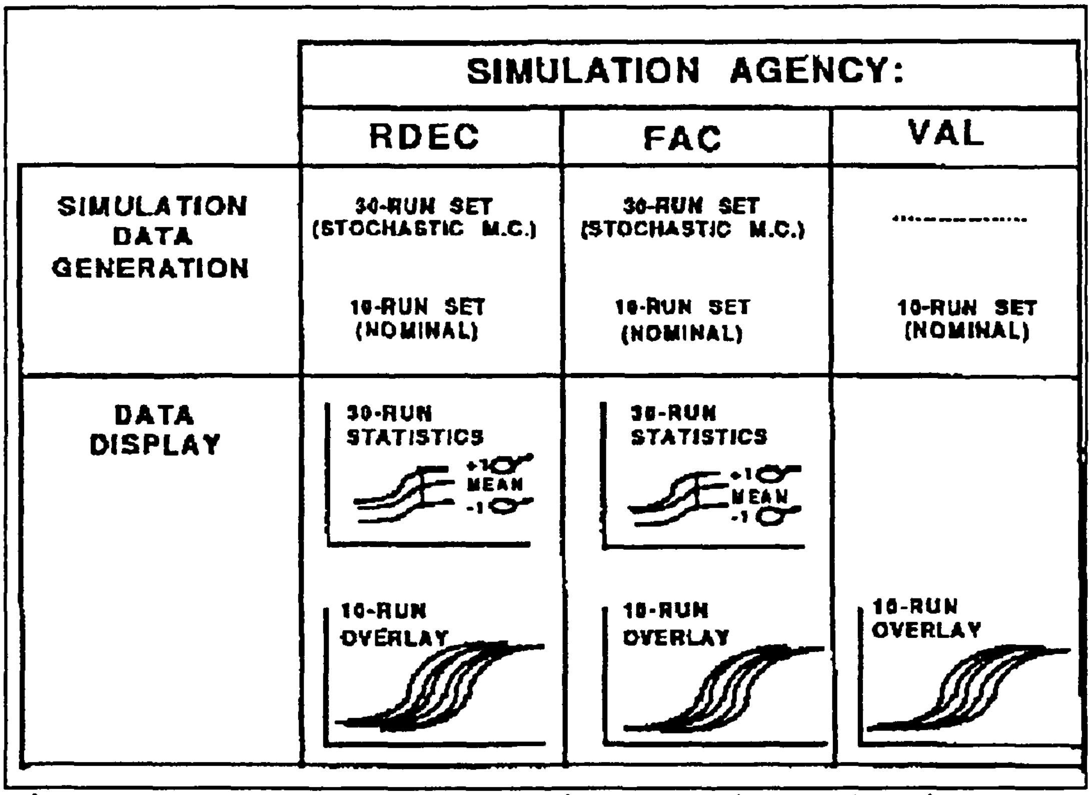


Figure 3 Overlay Data Generation (Gravitz and Waite, 1987:779)

If the flight test output fell within one standard deviation of the mean simulation data output at least sixty-eight percent of the time the simulation was judged to be acceptable (Army, 1987:12). This comparison is graphically demonstrated in Figure 4 below. This procedure takes into account the stochastic nature of simulations, has intuitive appeal, and is easy to interpret.

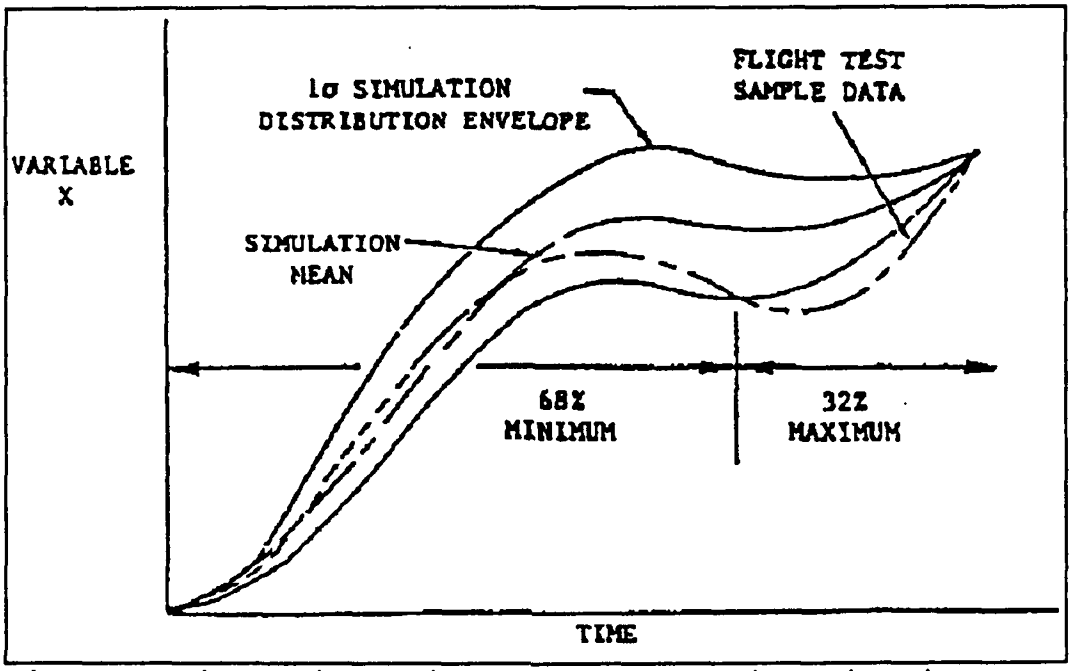


Figure 4 Field/Simulation Goodness-of-Fit Criteria (Army, 1987:15)

Another helpful test used in the validation process for the Chaparral missile was the calculation of the Pearson Product Moment of Correlation Coefficients for each pairwise set of corresponding time series variables (Gravitz and Waite, 1988:780). The equations for this calculation appears in Figure 5 on the next page. A correlation

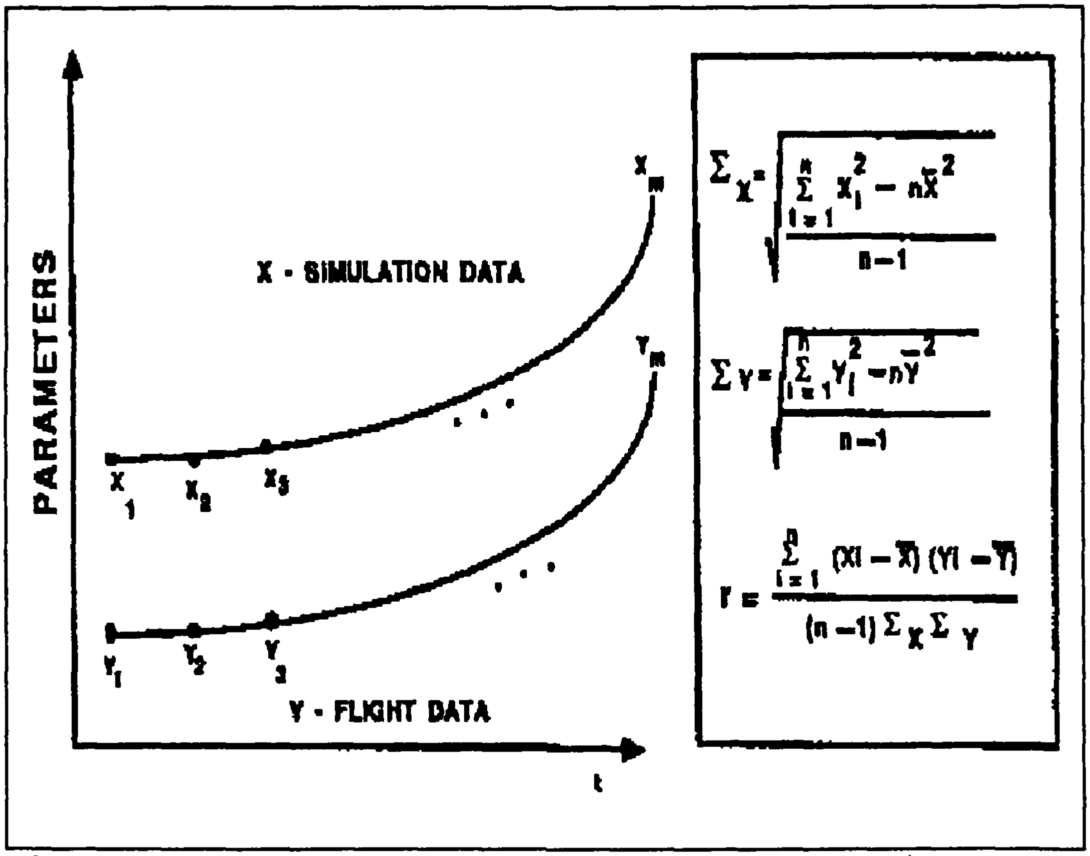


Figure 5 Pearson Correlation Coefficients (Gravitz and Waite, 1988:780)

coefficient of one would be perfectly correlated while a negative one would be perfectly negatively correlated. A zero coefficient would indicate no correlation between the two data streams. Typical results of this test were above .95 between each data set (Gravitz and Waite, 1988:780). Since each time series is normalized, the difference in means will not affect the statistic.

The use of these validation tests has been twofold. During model development the tests become part of an iterative process to help refine and improve the model. For actual validation, it is essential that the data used to validate has not been used in developing the model (Kheir and Holmes, 1978:118).

Time Series Analysis

Time series analysis is a systematic approach to answering mathematical and statistical questions about a stream of time correlated data (Shumway, 1988:1). A time series is a sequentially ordered set of observations of a process taken at constant time intervals (Box and Jenkins, 1976:23). Methodologies for time series analysis include harmonic analysis, which deals in the frequency domain, or regression analysis, which analyzes the series in the time domain (McCleary, 1980:17). Harmonic analysis assumes a time series is composed of sine and cosine waves of different frequencies (Box and Jenkins, 1976:36). Regression analysis assumes the output can be determined by the inputs to a system.

Harmonic analysis is known under many names such as spectral analysis, Fourier analysis, and periodograms. In general it requires a more advanced mathematical background and the use of computers has made the regression approach much easier and more popular (McCleary, 1980:17). Some spectral analysis is usefully applied in the identification of appropriate models in the regression analysis of a time series (Makridakis, 1983:372). Because AFOTEC's proposed methodology calls for model fitting using a time series software package, this review will focus on the time domain analysis techniques dominated by the Box-Jenkins Autoregressive Integrated Moving Average methodology.

Normally the regression technique would develop a mathematical model of a system by determining the relationship between input (independent) variables and output (dependent) variables. The regression would draw a correspondence between the n^{th} observation of the input stream(s) to the n^{th} observation of an output stream. The variability in the output data would be explained by the variability in the input data. Since this form of a regression is causal in nature, it is built on research and theory (McCleary, 1980:20).

A time series model, however, is built solely on empirical output data (McCleary, 1980:20). The time series regression technique remains much the same, but the input variables are replaced by previous outputs instead of independent quantities. The primary statistical problem becomes identifying the number of coefficients and estimating their values (Shumway, 1988:2). There are two major reasons for using a time series. Sometimes the system of interest is not understood or extremely difficult to measure. A second major reason to use a time series is when the output is the only item of interest, and not why the output occurred (Makridakis, 1983:18).

Box-Jenkins Time Series Analysis Box and Jenkins combined and integrated previous forms of regression based time series with their methodology using Autoregressive Integrated Moving Average (ARIMA) models. This review will cover the basics of that approach as presented by them and expanded or clarified by other authors.

According to the Box-Jenkins methodology a time series can be represented by an Autoregressive (AR) process, a Moving Average (MA) process or a mixture of the two. The Autoregressive Integrated Moving Average model notationally is designated an ARIMA (p,d,q) where the p is the order of the AR process, the d is the degree of differencing involved, and the q is the order of the MA process (Makridakis, 1983:362).

One of the basic principles of the ARIMA approach is that of parsimony. Parsimony is employing the smallest number of parameters possible while still presenting a suitable representation (Box and Jenkins, 1976:17). This principle is made easier by the most important tenet of the ARIMA model, that the latest input will have a greater impact than any earlier input (McCleary, 1980:19). Any output of a time series model should be determined by a few of the immediately preceding inputs.

The distinguishing characteristics between autoregressive and moving average models is in the nature of how long an input affects the output of the process. The white noise process a_t is the series of random shocks which drive the system (Box and Jenkins, 1976:46). An autoregressive model is an exponentially weighted sum of all past shocks, meaning each shock persists indefinitely at a diminishing rate (McCleary, 1980:61).

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + a_t$$

where ϕ = Autoregressive coefficient, a = error term, and p = order of auto regressive parameters.

In a moving average process a random shock has a finite persistence of no more than q observations (McCleary, 1980:61).

$$Y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

where θ = moving average coefficient, a = error term, and q = order of the moving average parameters

Another basic tenet in the Box-Jenkins methodology is the iterative stages in selecting a model (Box and Jenkins, 1976:18). Figure 6 on the next page describes this process. The three main stages consist of: identification, estimation, and diagnostic checking (Box and Jenkins, 1976:171).

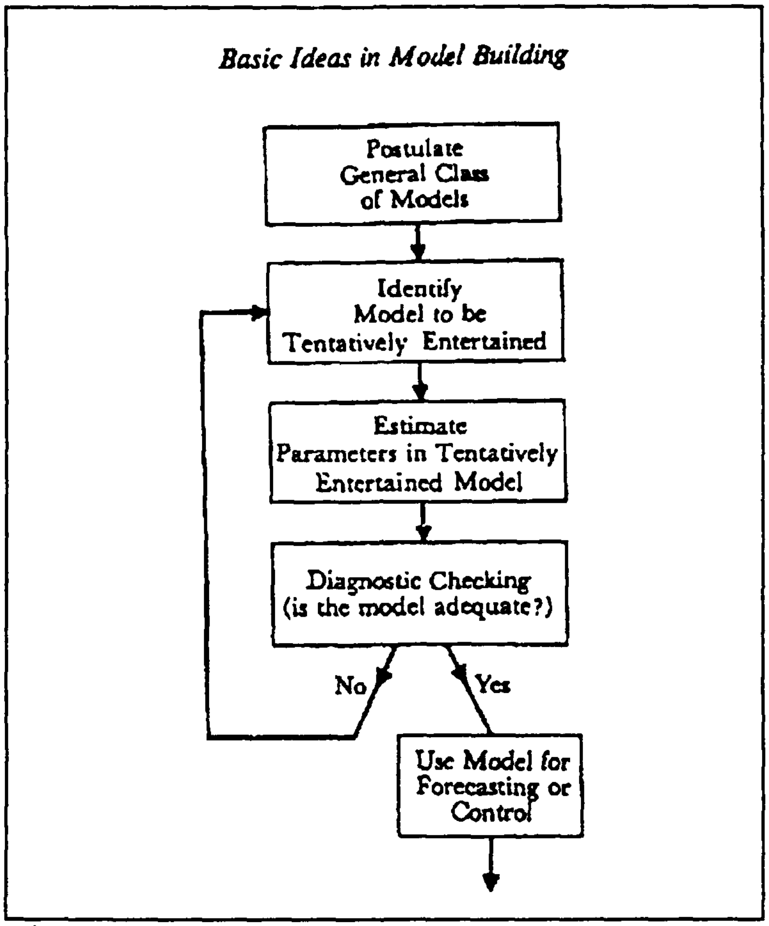


Figure 6 Box-Jenkins Methodology (Box: 19)

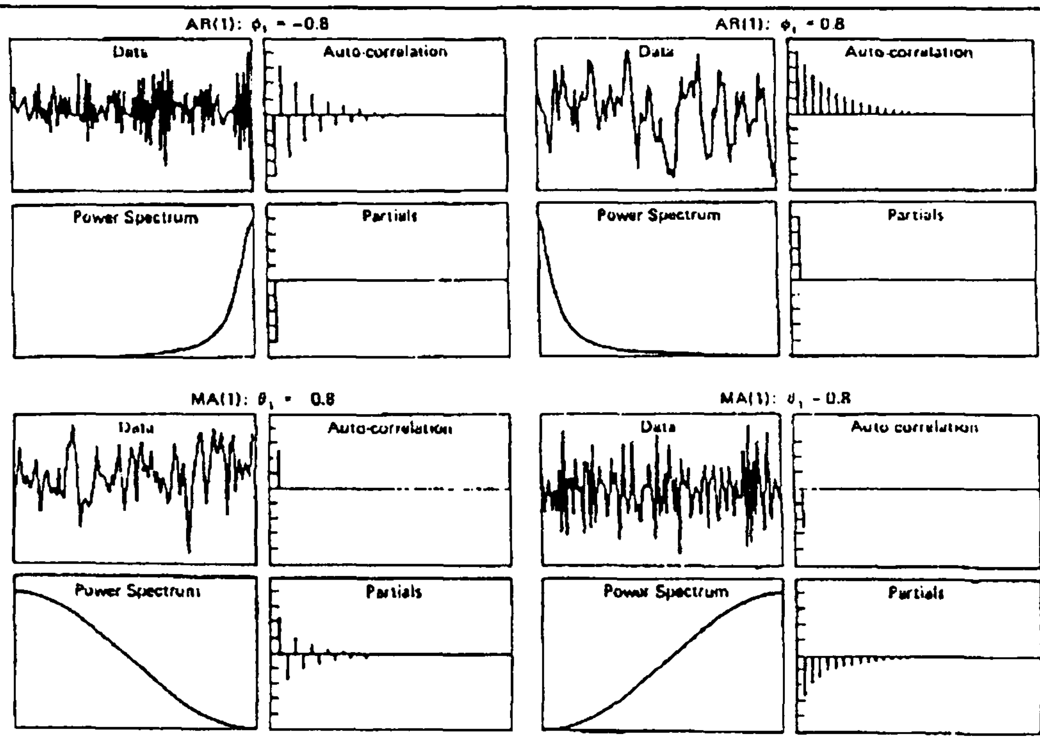
<u>Identification</u> The Identification stage is the longest and most difficult. Computers can rapidly produce decision criteria for the other two stages while identification often requires subjective judgements. This subjectivity is then removed throughout the rest of the model building process.

Identification means using the data and any information on how the series was generated to pick a process to begin model generation (Box and Jenkins, 1976:171). A typical key to identification of an AR or MA process lies within the patterns found in the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) (McCleary, 1980:93). These patterns are only reliable once the time series is stationary. Most time series are not stationary in their original state (Makridakis, 1983:413). A series can be considered stationary when the mean and variance remain constant over time (Makridakis, 1983:359).

Graphical methods are very useful in the identification stage (Box and Jenkins, 1976:173). Nonstationarity can be recognized by examining either the time series plot, or more commonly, by the graph of the ACF. The ACF of stationary data should statistically drop to zero very rapidly (Makridakis, 1983:379). If the data indicates the series is nonstationary, it must be transformed or differenced until it becomes a stationary series (McCleary, 1980:52). Care must be taken not to over difference the data as this will cause an overly complicated and cumbersome model with an increase in variance (Abraham and Ledolter, 1983:233). It is seldom necessary to difference data more than twice to achieve nonstationarity in the real-world (Makridakis, 1983:384).

Characteristic patterns in ACF, PACF, and Power

Spectrum for different AR (1) and MA (1) processes are shown in Figure 7. Typical patterns for an AR (2), MA (2), and ARIMA (1,0,1) are shown in Figures 8 through 10. The primary characteristics shown are summarized in Table Two.



A SUMMARY SHOWING (a) THE KINDS OF DATA SERIES, AND (b) THE THEORETICAL PROPERTIES OF PROCESSES THAT CAN BE MODELED AS AR(1) OR MA(1)

Figure 7 AR (1) & MA (1) MODEL PROPERTIES (Makridakis, 1983:453)

The patterns in the actual data may not be as clearly displayed as the theoretical ones in these figures. The expected patterns are for infinitely long realizations

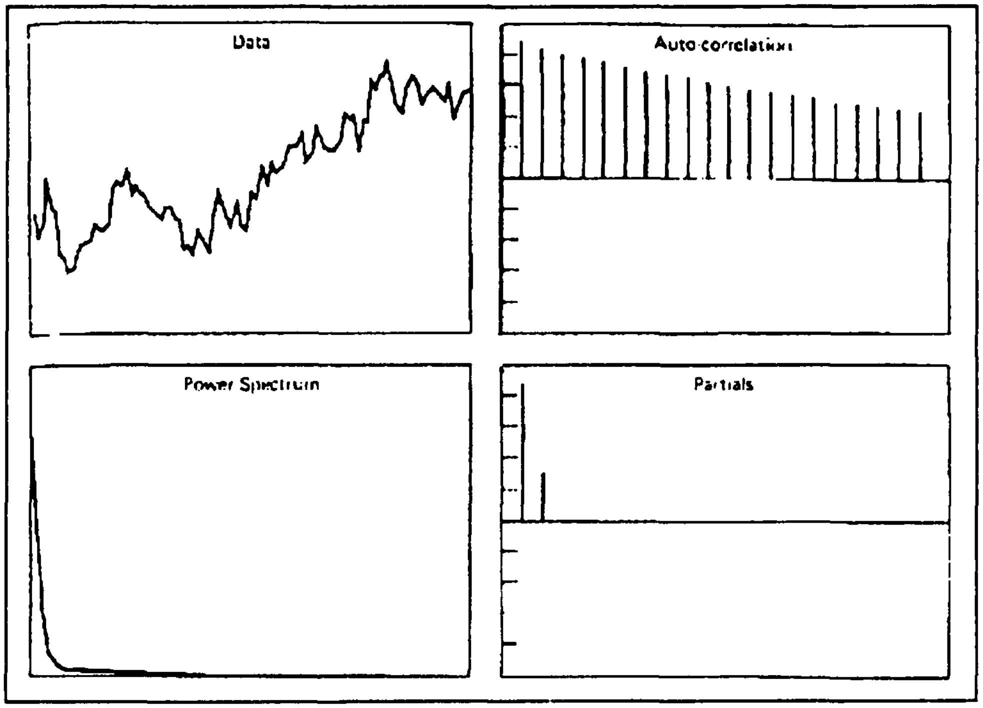


Figure 8 ARIMA (2,0,0) MODEL PROPERTIES
(Makridakis, 1983:423)

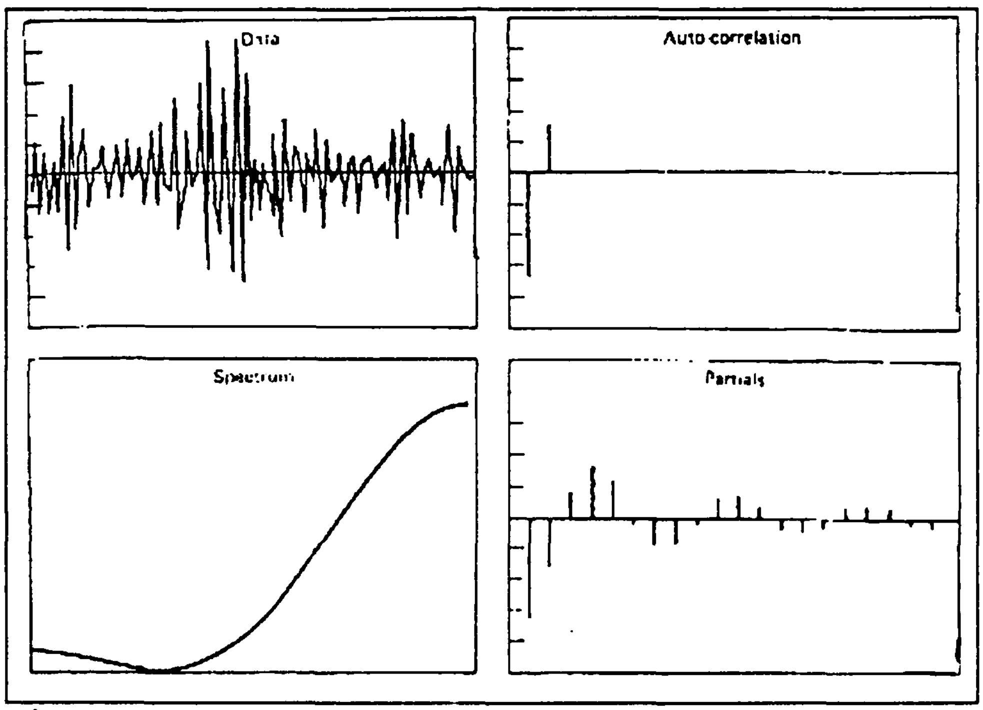


Figure 9 ARIMA (0,0,2) MODEL PROPERTIES
(Makridakis, 1983:427)

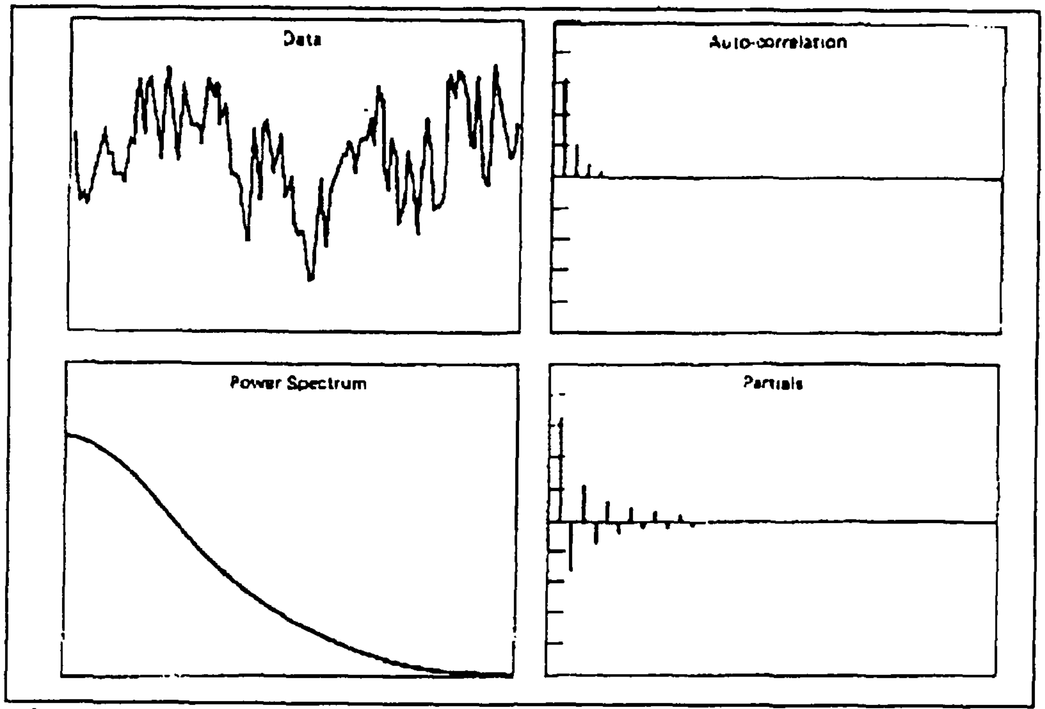


Figure 10 ARIMA (1,0,1) MODEL PROPERTIES (Makridakis, 1983:428)

Table 2 Arima Model Characteristics Summary (Makridakis, 1983:439)

AR(1)	Exponentially decaying autocorrelations, one significant partial, and line spectrum support (low frequencies if the AR coefficient ϕ_1 is positive, higher frequencies if ϕ_1 is negative).
MA(1)	Exponentially decaying partials, one significant autocorrelation, and line spectrum support.
AR(2)	Damped sine wave decay of autocorrelations and two significant partials.
MA(2)	Damped sine wave decay of partials and two significant autocorrelations.
ARMA(1,1)	Exponentially decaying autocorrelations and partials.

(McCleary, 1980:94). All the authors suggest a relatively long series of data is required for time series analysis. Box and Jenkins say at least 50 observations, and preferably over 100 observations, should be required (Box and Jenkins, 1976:18).

Estimation Having tentatively identified the process, the parameters of the model are usually estimated using a nonlinear technique such as the Marquardt algorithm. Each of the Autoregressive and Moving Average parameters should be statistically significant and each should lie within the stationary-invertibility bounds (McCleary, 1980:98).

Diagnostic Checking Diagnostic checking is the final step to insure the model assumptions are satisfied and that the model is adequate. It generally amounts to examining the residuals for any remaining pattern and studying the present model for possible improvement (Makridakis, 1983:446).

A good model will leave only white noise and have no remaining pattern in the residuals. The ACF and PACF will all be insignificant and the line spectrum will consist of high amplitudes across the whole range of frequencies (Makridakis, 1983:446). At the .05 significance level a chance for two or three significant spikes does exists for 20-30 lags by chance alone (McCleary, 1980:99). The Box-Pierce Q statistic can be used to test whether the entire residual ACF is different from zero (Makridakis, 1983:390).

The statistic is computed as follows:

$$Q = n\sum_{k=1}^{m} r_k^2$$

where m=maximum lag considered
n=number of observations
r_k=autocorrelation for lag k (Makridakis: 390)

Since the Q statistic is distributed approximately chi-square with (m - p - q) degrees of freedom, it is used to test the null hypothesis that the residuals are white noise (McCleary, 1980:99). Box and Jenkins likened this process to the "goodness of fit" test performed with classical statistical distributions (Box and Jenkins, 1976:385). They also found the use of the cumulative periodogram useful in insuring there was no periodic characteristics remaining within the residuals (Box and Jenkins, 1976:294).

Having found an adequate model, the second half of Box's diagnostic stage calls for over fitting that model in a search for a better model. If the original model was the best, no significant parameters should be found (Box and Jenkins, 1976:286). McCleary calls this stage metadiagnosis which he describes as playing the "devil's advocate" (McCleary, 1980:103). He suggests both over and under fitting the chosen model. McCleary additionally tests for the residual mean square (RMS) statistic in this stage. It's formula is

$$RMS = \frac{1}{N}\sqrt{\sum_{T=1}^{N} \hat{a}_t^2}$$

> where the smaller the RMS the better the model for the set of N residuals (McCleary, 1980: 101).

Having failed to find a better model the present model can safely be assumed to be the best representation of the underlying process available using this technique.

Summary

The first part of this chapter established the need to validate the simulation models used in military decisions. Techniques, balanced against the cost and impact of the model, were suggested. Many techniques which have been used in similar missile simulation validations were then discussed. The last part of the chapter reviewed the basic mechanics of Time Series Analysis using the Box-Jenkins Autoregressive Integrated Moving Average methodology.

III. Methodology

Introduction

In his research on simulations in Operational Test and Evaluation, Lieutenant Colonel Mann states:

> If there is to be any confidence in the results of evaluation conducted with computer simulations or in the hardware-in-the-loop or man-in-the-loop simulators, then there must be a reasonable validation between those results and the results of field testing conducted under the same conditions (Mann, 1983:58).

Not only is confidence gained, but the GAO concluded the benefits of simulation used in conjunction with field experimentation and other analytical methods will likely result in a synergistic effect where the benefits of the combination exceed the benefits of the individual methods (GAO, 1987:10).

AFOTEC wishes to validate that the upgraded Terrain Effects Model can accurately simulate the B-1B electronic counter measures' capability to defeat enemy air-to-air threats. This chapter will describe several techniques used in comparing the output data produced by the Terrain Effects Model and the output gathered from the Golden Bird system during field tests.

Good data is the starting point for any model fitting analysis. The chapter begins with a description of the data sets provided and the techniques used to describe characteristic patterns discovered in examining them. Patterns within the data are shown both before and after model fitting.

Model fitting for a representative matched pair of field and simulated data will be demonstrated in a stepwise fashion. The AFOTEC proposed methodology for comparing the field and simulated data will be presented. The chapter will end by explaining the non-use of the other missile simulation validation techniques described in Chapter Two which may be of interest under other conditions.

Data Description & Analysis Techniques

The data of interest used in this research effort is the angular elevation error of the anti-aircraft missile seeker head. This is a measure of distance between where the missile seeker head is looking and the actual straight line directly to the target aircraft. A zero would represent the missile seeker head looking directly at the target aircraft. A positive value is returned when the seeker head is looking above the direct line to the target aircraft while a negative number is looking below the line.

The data is sequentially recorded at constant time intervals.

The missile seeker head error is a dynamic measurement which impacts the air-to-air missile throughout its flight. AFOTEC chose this value as the measure of effectiveness based on the assumption so many factors influence this value, if the simulation can reproduce this data vector it must have captured the underlying processes which determine the flight path of the missile.

Each data set was highly autocorrelated. As discussed in Chapter Two, this was expected. This characteristic is typical of time sequential information that can normally be represented by a time series.

Data arrived at three separate times with varying degrees of background information associated with each set. Each of the first two data groups prompted request for additional information and detail. Each iteration allowed a more thorough examination of the possible informational value stored within each number stream. This section will describe the techniques used to discover the characteristic patterns and observations made on each group of data.

Each examination begins with some general observations on the raw data sets. Several graphical techniques are used to display the results making similarities and differences easy to distinguish.

A preliminary data set consisting of sixty sets of field test data averaging one thousand points per set was analyzed to scope the validation problem and identify significant characteristics of the data sets. Each set was accompanied with the associated above ground altitude of the bomber, the direction of flight, and whether the aircraft carrying the Golden Bird system was approaching from the nose or tail. Data sets were available for both with and without jamming in effect. No simulated data was available with this preliminary group.

With no simulation data available, the primary comparisons were accomplished between data sets with the ECM in effect and those without. Many observations on the similarities and differences between data files were evident by plotting the time series. Each series plots time along the horizontal axis against the angular seeker head error on the vertical axis. Throughout this paper, comparison type data is normally displayed side by side in a graphical form. Figure 11 on the next page depicts the distinctive pattern (a downward concave curve) common for all data sets in the ECM environment. No discriminating features could be found in the initial examination to distinguish direction, altitude, or whether the jamming aircraft was approaching head on or from the tail.

Figure 12 displays a typical non-jamming scenario. The result is far different than with the ECM on. The pattern is less distinctive among data files with the ECM off as variations appear in the different runs. The variations still lack any patterns which can be used to discriminate direction, altitude, or the approaching aircraft's relative position.

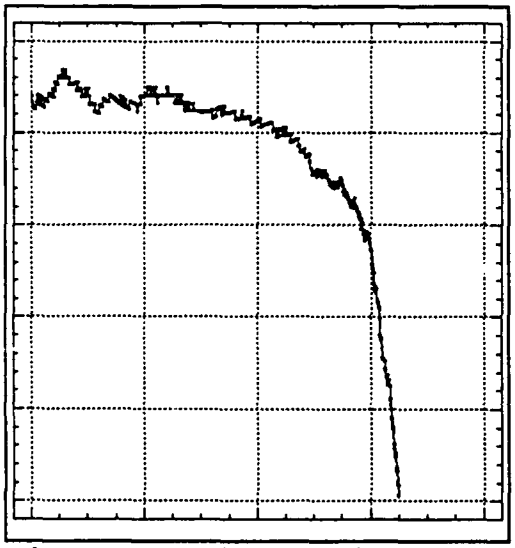


Figure 11 Time Series Plot ECM On

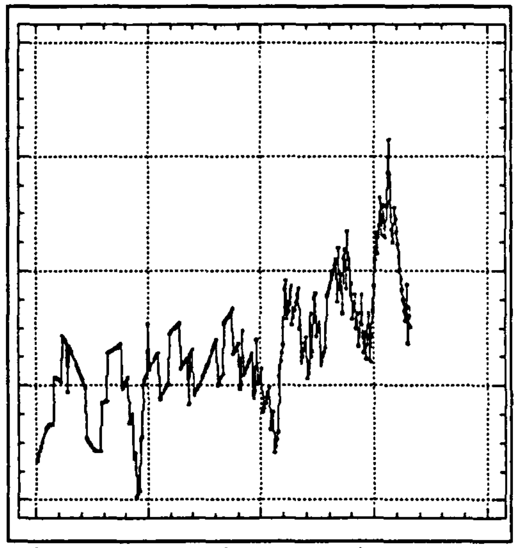


Figure 12 Time Series Plot ECM Off

While the general patterns between jamming and non-jamming are quite different, there are many similarities between the data sets. Both types of data have small smooth areas following each jump in the level of the data values. As evidenced by a changing mean, all data sets are nonstationary. When differenced to achieve stationarity, spikes in the data become much more obvious. The differenced data sets for ECM on and off appear in Figures 13 and 14. The discrete jumps in data values have a deterministic value appearing at random intervals. Each jump has an average value of 8.3 units with a variance of approximately plus or minus .2 units. Occasionally a jump appears to be much larger. These larger jumps are, in fact, two jumps which occur within one time increment.

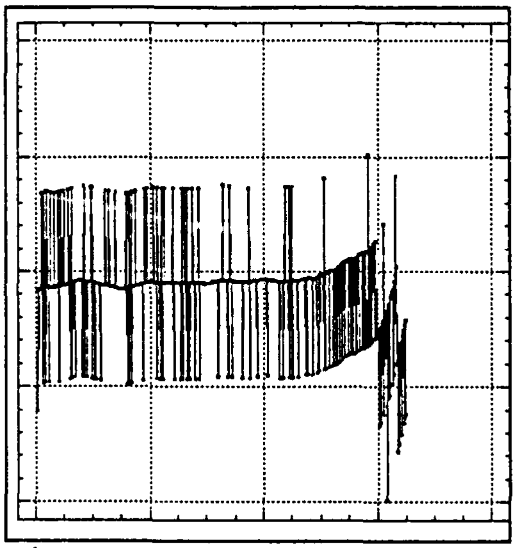


Figure 13 Differenced Time Series ECM On

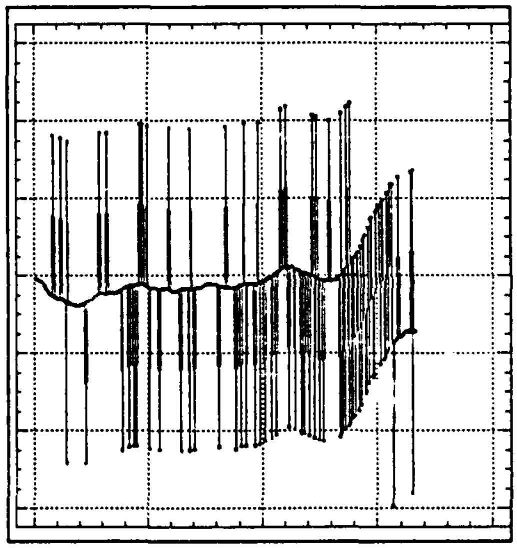


Figure 14 Differenced Time Series ECM Off

AFOTEC confirmed they were aware of the existence of these intrusions into the data output. The exact cause of the data jumps is undetermined, although the conversion from analog to digital data is suspect (Bennett, 1989). The analog to digital conversions and its associated precision can not be discussed in an unclassified report. The important point is that the discrete jumps have nothing to do with the ECM. These perturbations appear both with the ECM on or off and are therefore not an object of this research. Because these jumps indicate added noise in the recorded process, they are of concern in this study. The problem is not unique to this validation effort.

> A potentially frustrating problem for the data analyst is dealing with wild or unusual observations in either the observed flight data or the simulation data. These wild or aberrant observations may severely distort the sample spectrum or estimate of the parameters of the underlying distribution. Often we find that some type of data editing or preliminary screening of the data is necessary (Greene and Montgomery, 1981:115).

Attempts to add or subtract the deterministic component without removing the informational quality of the data were unsuccessful. The smoothing nature of the programs developed destroyed the time series nature of the data and left behind only white noise with no informational value to the validation process.

A request for matched simulated and field test data led to the arrival of a second group of data files. The second group of data sets included thirteen pairs of field and simulated data. An immediate problem evident in the data sets was the difference in sampling rates. Throughout this project the field data streams have a much greater sampling rate than do those of the simulation output streams.

No common reference allowing alignment of the data sets for comparisons could be found. Sampling the field data vectors to match their sampling rate to those of the simulated data was of little use. Even after reducing both data streams to the same sampling rate the files were of unequal size. Trying to match beginning, ending points, or some apparent similar points along the curve was too subjective and arbitrary to provide a credible comparison for validation purposes.

A significant observation from this data was the simulation output followed the same general pattern previously shown in Figure 11 for the field tests with jamming on. No simulations with ECM off were provided. A second observation was the lack of deterministic jumps within the simulation data. The discrete jumps are strictly a function of field system dynamics and probably related to the test system.

A request for matched simulated and field test data with common reference points resulted in the arrival of the final group of data used in this research. The data was given with time space position indicator (TSPI) markings. This information allows exact pairing between field and simulation data. The overlapping times to be matched between the field data points and the simulated data points can be determined. Ideally, the TSPI data recorded during field tests means the same input conditions such as terrain, ECM signals, etc., are used as inputs into the simulation.

Two attempts were tried at matching the field to simulated data. The first matching scheme averaged the field data centering on the time exactly corresponding to a simulation time. The intention of this effort was to negate the effects of the deterministic jumps by averaging them out. However, time series models fitted to the data averaged in this fashion have only statistically insignificant coefficients. Interestingly, when the field data sampled without averaging and taken at times exactly matching those with the simulated data, a time series model with significant coefficients can be fit. The correlation between averaged field data and exact field data is always high (above .95) and both the averaged and exact field data have virtually identical correlation to the simulated data.

Each data file within the last group is quite large, but the overlapping times between paired sets is limited. Each simulation contained two thousand observations, but the number of common points between field and simulated data range from a low of eighty-four to a high of only three hundred seventeen. No information on the relation between the times and where the missile was along its flight path was available. Consequently, there is no way of knowing if the matched points of time fall within the area of real interest for which the simulation validation is required. The pattern in the data shows ECM is on, but there is no way of telling if the matched points cover enough time to tell if the ECM is effective and predict the results of the engagement.

The availability of exactly matched data pairs allows significant comparisons for commonality. One of the first techniques used is the Box and Whisker plots. Appearing in Figure 15 on the next page this method provides a quick and easy means to depict a lot of information for comparisons between the data vectors. The box covers the middle fifty percent of the data values with the line marking the median value. The whiskers extend out to cover the predominant range while values far out from the bulk of data are plotted as separate points. Figure 15 shows the bulk of data points for the field and simulated outputs fall within the same range. Both data vectors are skewed in the negative direction and both have similar median values.

Another comparison can be graphically made with a scatter plot such as the one in Figure 16. Like the Box and Whisker plots it shows the bulk of data in the higher numbers. Possible patterns within the data may be detected by examining this type of plot. The forty five degree line indicates high cross correlation between the data vectors.

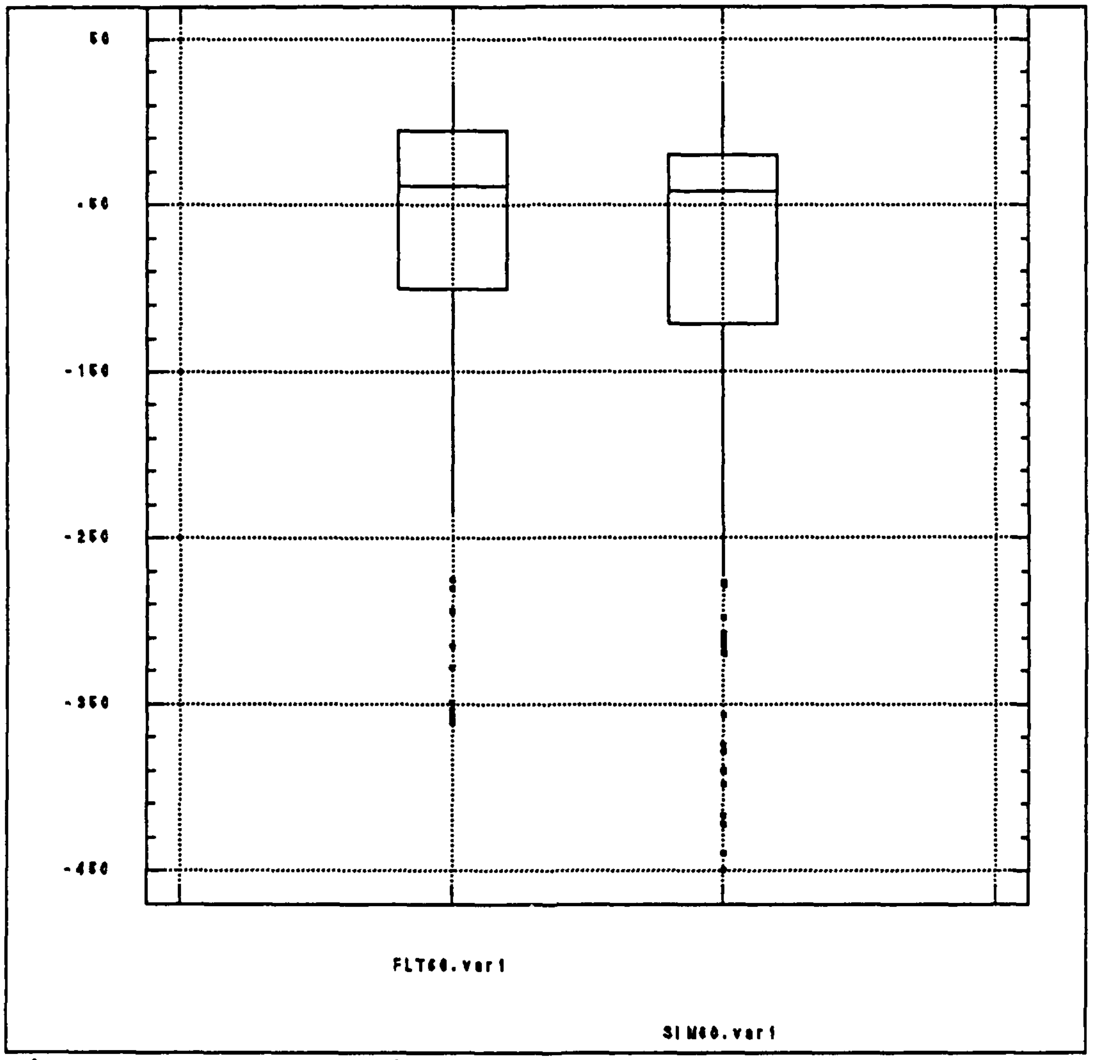


Figure 15 Box-and-Whiskers Data Set 60

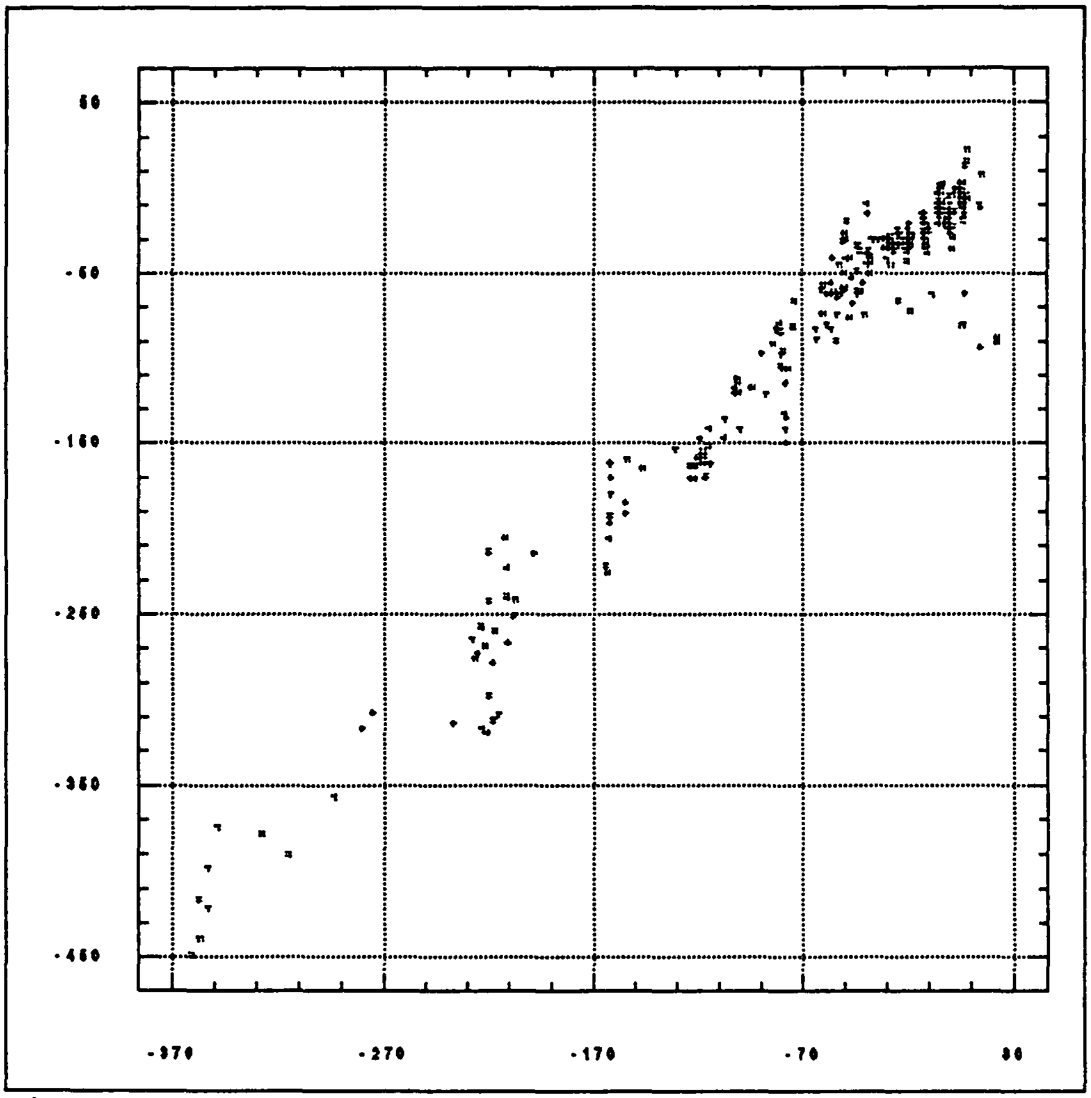


Figure 16 Scatter Plot Data Set 60

The last method of comparing the data pairs is the calculation and plots of the cross-correlation. As explained in the literature review this method has been used as an indicator in past missile simulation validations. A good simulation could be expected to produce a high correlation with the field test data. Since both time series are nonstationary the cross correlation of the differenced series would be expected to be high also. These plots will be shown in the next chapter with their results.

Model Fitting

Time series model fitting on a representative paired data set is graphically demonstrated in this section. The flight and simulated data are examined simultaneously side by side in a stepwise fashion to aid in comparison. Time series analysis as developed by Box and Jenkins and described in Chapter Two is used to model each data set. Each of the six paired data sets fitted in this research are from the final data group received. The final models for each data pair are presented in Chapter 4 and the graphical displays for the last five model sets is available in Appendix 1. As discussed in the data description, the simulation was created with input data recorded during the flight test. The points were matched by the time marks to insure the simulation is as closely related to the flight information as possible.

The original time series for flight and simulated data set number sixty are presented in Figures 17 and 18. They both have the general characteristic shape identified with the jamming in effect. The temporary departure from the overall pattern apparent in the field data is unique to this one data set and not considered characteristic.

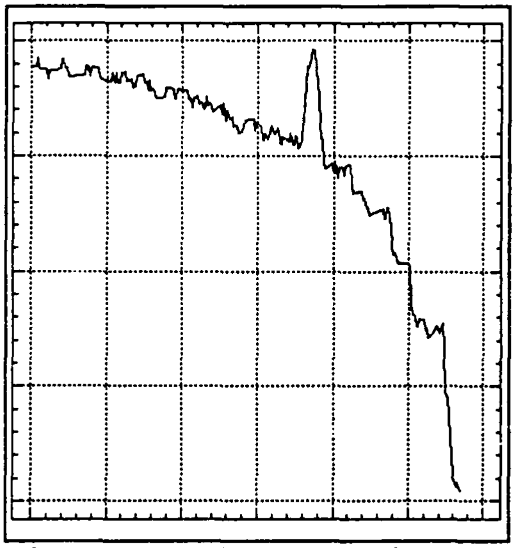


Figure 17 Field 60 Time Series Plot

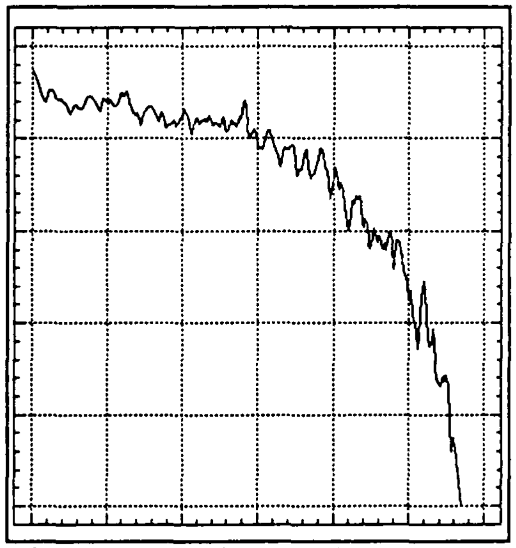


Figure 18 Simulation 60 Time Series Plot

Both sets are non-stationary. The differenced time series appears in Figures 19 and 20. The deterministic jumps are apparent in the flight data. The lowering of the sampling rate to match the simulation did not decrease the relative frequency of this occurrence. The simulated series in Figure 20 appears to have a slight non-stationarity in the variance which may also be present to some extent in the field test output. Models fitted on the second difference and a logarithmic transformation of the data prove to be inferior to fitting the series as it appears.

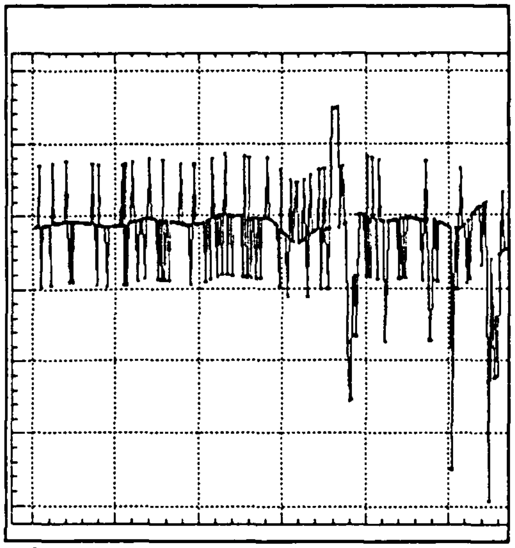


Figure 19 Field 60 Differenced Time Series Plot

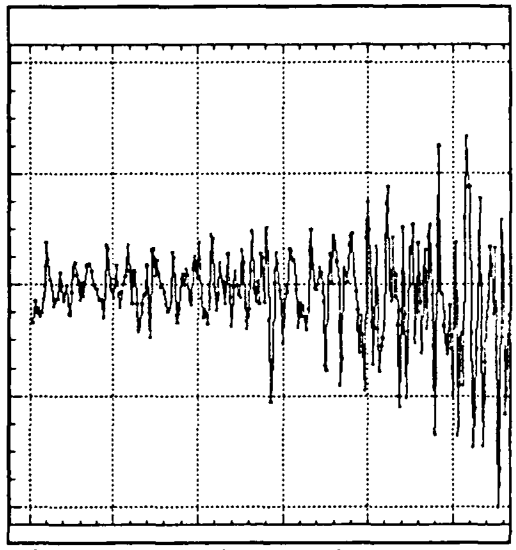


Figure 20 Simulation 60 Differenced Time Series

The problem with the discrete jumps may have an undefined effect on the model identification. It is assumed for this preliminary analysis the data continues to represent a stochastic series.

The autocorrelation functions for both the flight and simulated data are shown in Figures 21 and 22 on the next page. Each drops rapidly to an insignificant level with occasional spikes which are barely significant at distant lags.

The partial autocorrelation functions appearing on the following page in Figures 23 and 24 were used to determine the order of autoregressive model to fit. Significant spikes at lags 1,3,and 7 for the field test can be seen in Figure 23. Model estimation began by fitting coefficients

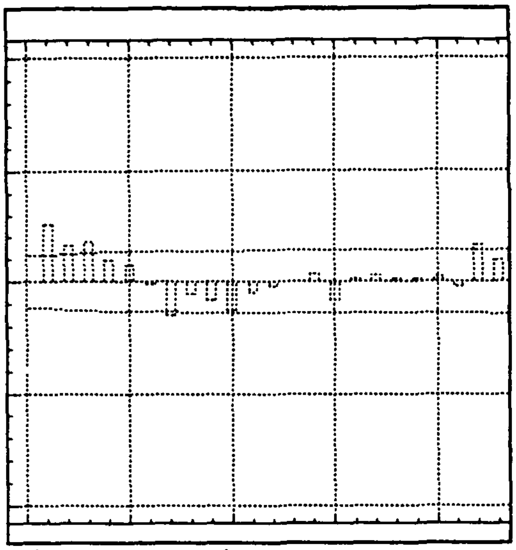


Figure 21 Field 60 ACF

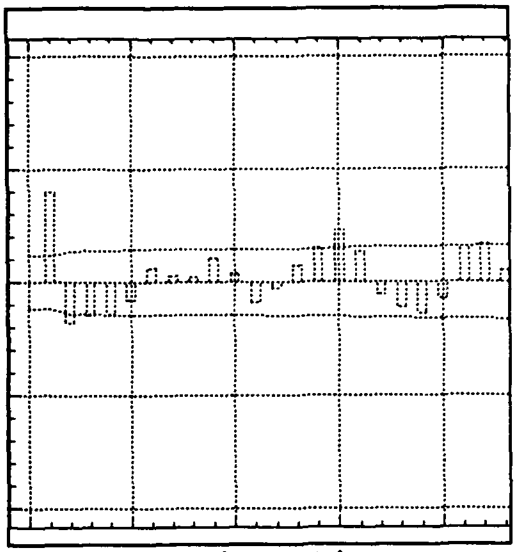


Figure 22 Simulation 60 ACF

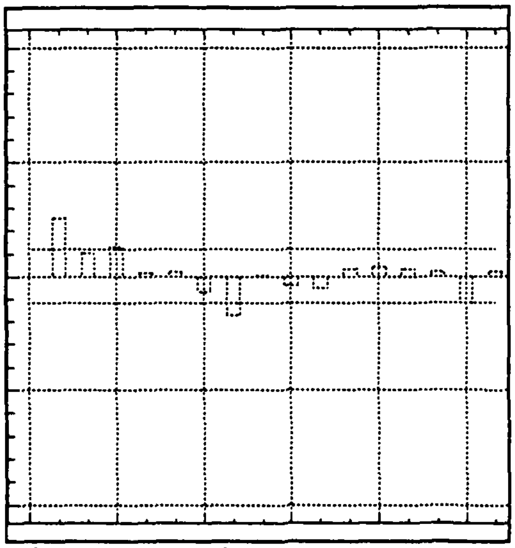


Figure 23 Field 60 PACF

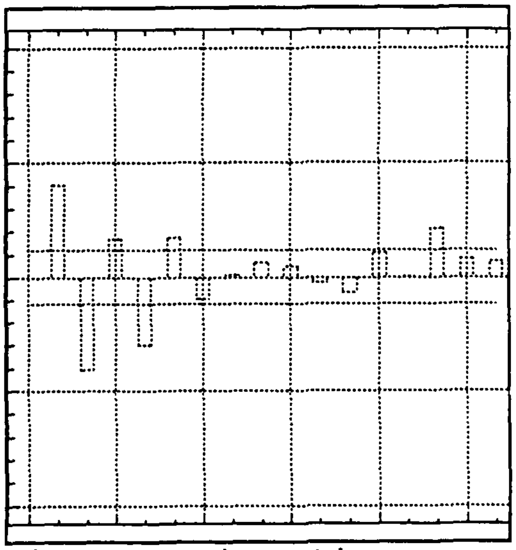


Figure 24 Simulation 60 PACF

with lags of these orders. In Figure 24 the first five lags were significant. This determined the order of the coefficients of an autoregressive model fit.

After identifying and estimating an initial model, diagnostic testing as described in Chapter Two is performed. Good models should leave white noise residuals with no correlation and distributed normally with a mean of zero. Figures 25 and 26 below depict the autocorrelation function of the residuals.

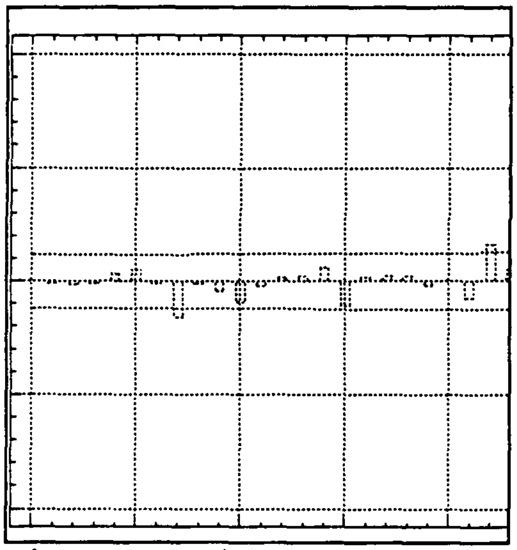


Figure 25 Field 60 Residual ACF

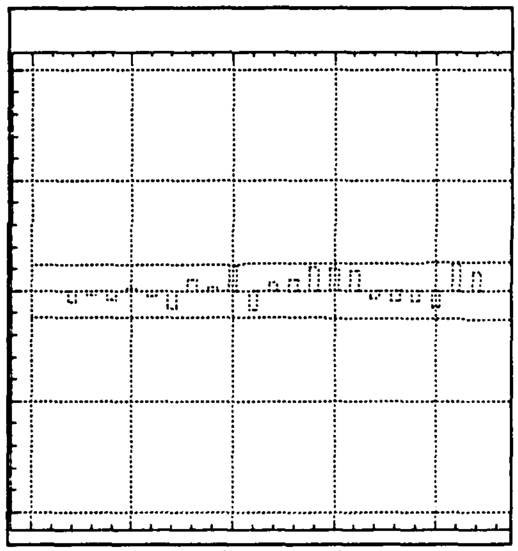


Figure 26 Simulation Residual ACF

The flight residuals are essentially white with only minor significant spikes at distant lags. All the residuals for this simulation sample have statistically insignificant autocorrelations. Individually each autocorrelation is statistically insignificant. However, when combined within the Box-Pierce test for significant residual autocorrelations, they are large enough to imply unfavorable results from each model. Many of the best time series model fits possible with these data sets retained more autocorrelation than desired. The time series models fit for this study have far less informational value than that which would normally be associated with a time series model fit.

Each model was over and underfitted to insure a model with the least residual autocorrelation and most explanatory power had been discovered. Coefficients which proved significant were added while any which were insignificant were dropped. For this particular paired data set the original autoregressive models attempted proved to be the best available.

Periodograms of each data set appear in Appendix 2. Figures 27 and 28 on the next page contain the periodogram for this example data set. The power in the lower frequencies matches those of the theoretical models shown in Chapter 2 and tends to confirm the order of the models chosen. Since the power spikes are not repetitive at a fixed interval, further confirmation of the lack of any seasonality is gained.

Each of the five other data pairs were modeled in this manner. The results and comments are included in the next chapter.

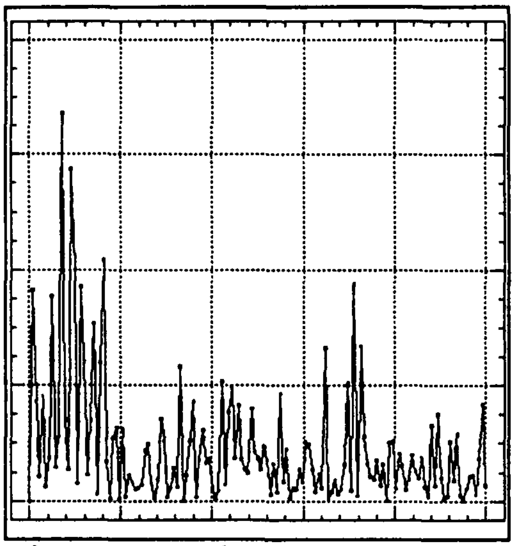


Figure 27 Field 60 Periodogram

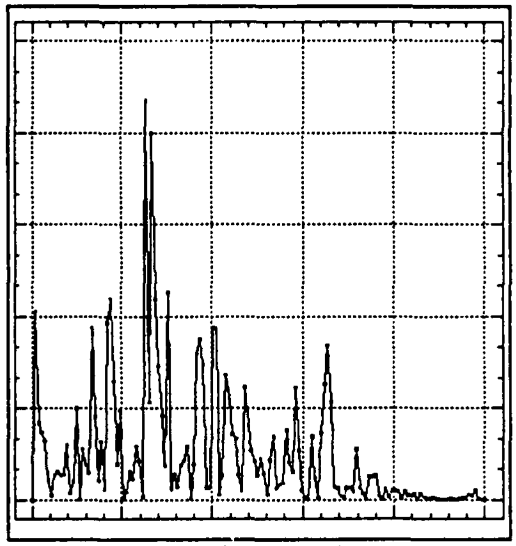


Figure 28 Simulation 60 Periodogram

<u>AFOTEC Proposed Methodology</u>

AFOTEC's proposed methodology is to compare the output data from each paired field and simulation test to see if the two are statistically the same. The planned procedure is to fit the best possible time series model to each field data set. Once a model is fit which explains the data well (meaning the residuals are white), the paired simulated output data will be filtered through that same field model.

The residuals are then to be statistically compared (Bennett, 1989).

If the simulation is valid, the residuals left by the filtered output data through the model derived from the corresponding field test should be white. This is equivalent to testing if the two output vectors came from the same population. The results of this methodology on the six data pairs is discussed in the next chapter.

Rejected Techniques

The non-use of other significant techniques described in the literature review is explained in this section. Spectral analysis is the major technique for analyzing time series data streams which was not used in this research. This has been a common technique in older missile simulation validations. AFOTEC was interested in comparisons based on actually modeling the data output, and so this research focussed in that direction. Theil's inequality coefficient has been used for many years in numerous past missile simulations. It is very similar to the Pearson's Cross Correlation Coefficient. Pearson's test was accomplished because it distinguishes both positive and negative cross correlation while Theil's coefficient is an absolute value of cross correlation. The ability to distinguish positive from negative cross correlation proved useful as will be described in the next chapter.

The statistical procedure for comparing autoregressive parameters developed by Hunter and Hsu could not be attempted. Their procedure was exclusively applied to models of the same order (Hunter and Hsu, 1975:3-18, 3-19).

The most interesting procedure which could not be applied in this effort was overlaying the field test output stream on an average simulation mean and testing the time that the field test fell within one standard deviation of the mean. Since each simulation is but one stochastic realization of the process, this method has a lot of appeal. Unfortunately, only one realization of each simulation run was available.

Summary

Several similarities exist between the field and simulation data files provided for this research. Information is stored within the general characteristics of each group. There are problems fitting time series models to the data streams which have high explanatory value. Problems with the model fitting effort are potentially caused by the deterministic spikes which have been unintentionally added to the field test data sets, most probably during the analog to digital transformation.

The simulated data is as closely matched with the field data as possible. Since the information stored during the field test was used as input factors in the simulation, each point in the simulation should experience the same factors which affected the missile at that corresponding time.

Each pair of data was fitted with the best autoregressive model possible. The procedure was demonstrated on the first pair of data files available. AFOTEC plans to fit models to each data set in the same way. After model fitting, each simulated data stream is filtered through its corresponding field test model. A decision on validity is based on an assessment of how many times the residuals of this procedure are random white noise.

Several interesting techniques were not used in this effort. Some were not used because the data was not available, while others would not work because of the structure of the models.

The next chapter will present the results of the techniques which were applied along with some observations and interpretations of their meaning.

IV. Results

Introduction

The results of the data analysis comparison techniques used to characterize and test similarities are discussed in this chapter. The majority of all results reported in this project are drawn from the last data group because its information allowed the closest possible association between field and simulated test data. The first comparisons and observations are derived from the raw data files. Box and Whisker plots and Cross-Correlation analysis are used to examine external similarities in the data for face validity. The assumption implied is if the simulation is really valid it should match the field data in as many ways as possible.

Following the raw data observations, the results of Box-Jenkins ARIMA model fitting are presented. The models derived are used to predict the test results AFOTEC would find when using their proposed methodology. The chapter closes with a summary of the results found during this research.

Data Comparison

Some general characteristics and comparisons of the data sets has already been reported in Chapter Three during the course of describing the techniques used. This section examines some more specific comparisons between the field and simulation data sets.

Table Three displays some summary statistics on each of the paired data sets. Examination of this table shows the field data files always have a higher average, median, and usually have a lower variance and range.

	FLD60	SIM60	FLD62	SIM62	FLD63	SIM63
Sample size	285	285	317	317	84	84
Average	-65.8	-86.5	-57.9	-71.9	-38.4	-48.6
Median	-38.3	-41.1	-34.3	-46.6	-37.2	-40.3
Variance	7098	9698	5071	5515	288	440
Minimum	-361	-449	-298	-356	-88	-92
Maximum	22	22	15	22	-13	-21
Range	383	471	313	378	75	71
	FLD66	SIM66	FLD76	SIM76	FLD77	SIM77
Sample Size	151	151	189	189	270	270
Average	-11.4	-13.5	12.9	2.7	-170.6	-204.7
Median	-1.2	-5.7	15.9	7.4	-131.4	-170.4
Variance	651	587	144	181	11710	9998
Minimum	-97	-86	-18	-35	-458	-495
Maximum	20	25	31	26	-54	-82
Range	117	111	49	61	404	413

Table 3 Summary Statistics

The Box and Whisker plots of the data sets, such as the one shown in Figure 15 in the previous chapter confirm the simulation average is lower. The plots show considerable overlap in the data streams, but the distribution of the simulated data points tends to be more negatively skewed than those of the field test. The whiskers also graphically depict the difference in ranges between data sets showing the field normally has the smaller range.

The next comparison between field and simulated data is a measure of the lagged cross-correlation as is demonstrated in Figure 29 on the next page. This plot is representative of all the data pairs. Correlation between each successive lag is very high and decays very slowly.

Table Four shows the zero lag coefficient for each paired data set. The results show high cross-correlation between the field and simulated data. This should be expected from a valid simulation. The validity of the data could easily be questioned if this had not been so.

Since both data vectors are nonstationary, the Pearson's cross-correlation on the differenced data streams is calculated. If the data vectors represent the same process, the cross-correlation would be expected to remain high after differencing. Unfortunately, as shown in Figure 30, the cross-correlation between the differenced field and

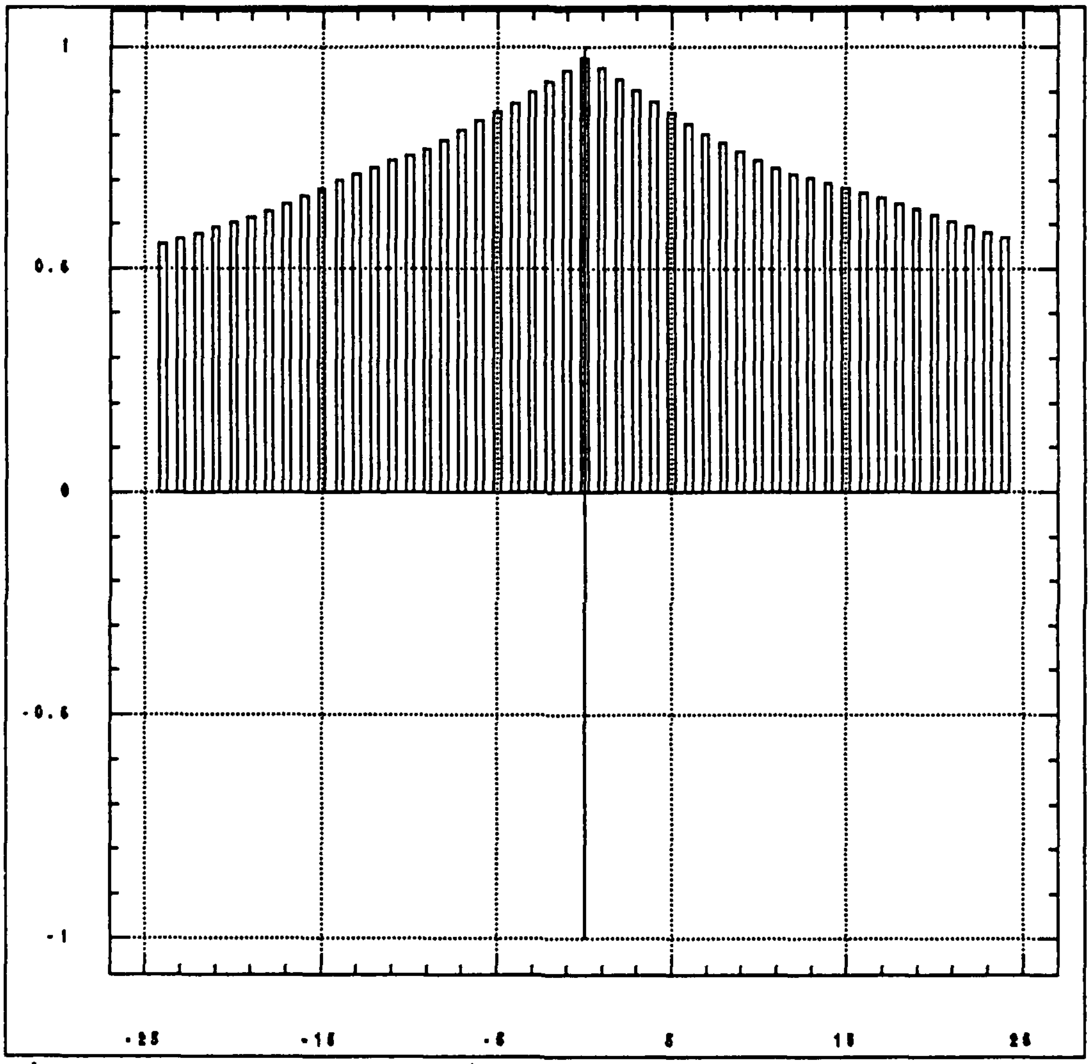


Figure 29 Cross Correlations Data Set 60

DATA SET #	FLIGHT TO SIMULATED CORRELATION	DIFFERENCED FLIGHT TO SIMULATED CORRELATION
60	.974	.085
62	.989	.112
63	.883	.130
66	.951	.019
76	.924	.077
77	.963	.963

Table 4 Pearson Correlation Coefficients

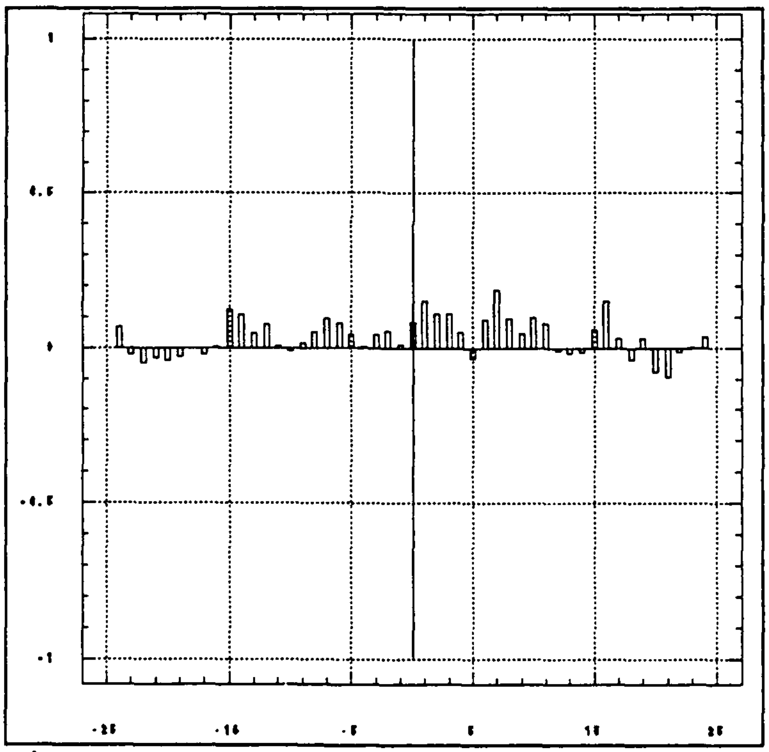


Figure 30 Correlation Differenced Field 60 & Sim 60

differenced simulated data is very low. The general pattern could be the factor causing the high cross-correlation and not any internal similarities in the models. The zero lag coefficient for each of the differenced data pairs is also shown in Table Four.

The real question is how much can the cross-correlation information add or subtract from the hypothesis that the simulation data came from the same population as the actual field test. To test this idea the cross-correlation was measured between one field data set and a simulated data set other than the one paired with it. Figure 31 on the next page shows significant cross-correlation. This is unexpected. Little or no cross-correlation should exist between independent data sets. One possible explanation for this discrepancy is the fact that each simulation and field test is for ECM on over a fairly uniform terrain test range. This would make each data vector very similar and possibly explain the unexpected correlation. A second explanation is the underlying pattern is very dominant and overshadows the other differences between each run.

Another factor in comparing data sets which are not paired is the editing of the data to make each vector an equal length. This is necessary for the computation of cross-correlation. Which values are edited to make the data sets the same length can have a great impact on the amount of cross-correlation. The cross-correlation in Figure 31 is only .66, but the same two data sets have a cross-correlation of .92 when matched at a different starting point.

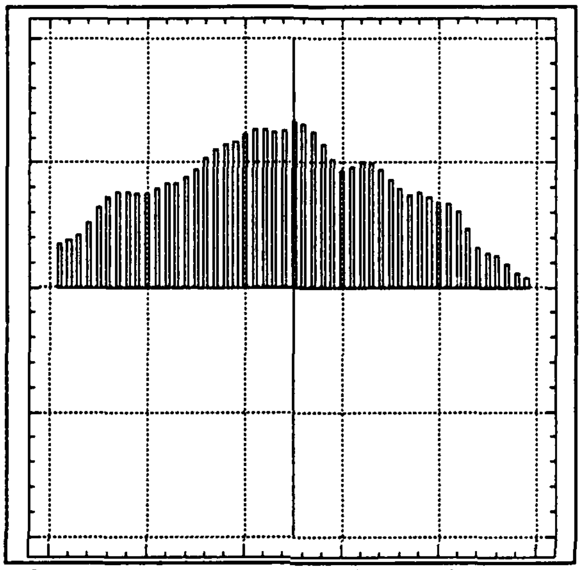


Figure 31 Correlation Field 66 with Simulation 63

Cross-correlation was measured again between a simulated data set with the ECM on and a field test without ECM. The results expected would be even more independent than the mismatched field and simulated data cross-correlation. The results are shown in Figure 32 on the following page. This time the cross-correlation was negative. Unfortunately, the magnitude was still high. Data streams with the ECM on would be expected to be independent and without cross-correlation to those with the ECM off.

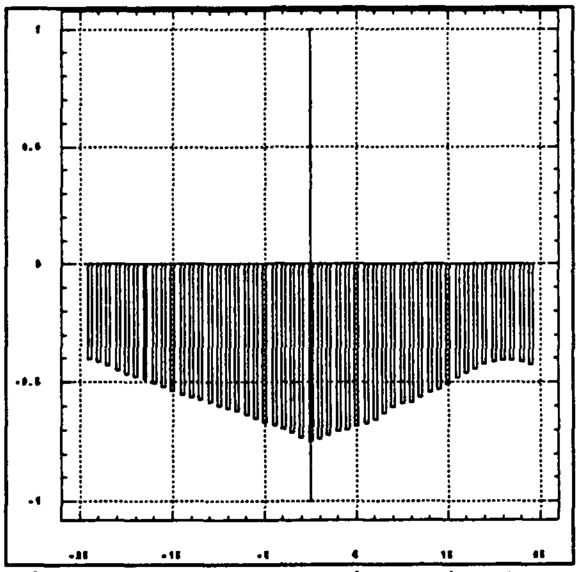


Figure 32 Correlation Field 16 (ECM Off) & Simulation 60

High cross-correlation is easy to obtain in this data. The high cross-correlation of the raw data vectors is expected and can be considered as adding to the face validity of the simulation model. The lack of cross-correlation between the differenced data raises questions about how well the simulation model represents reality. With high cross-correlation so easy to achieve, the Pearson correlation coefficient is more a test not to fail than a definitive way to add credibility to models with output as highly correlated as these.

ARIMA Results

Table Five shows the coefficients and order of models fit to each data set. Many observations can be made from the results of the model fitting. The most obvious result seen in the table is no matching field and simulation data set are fit to the same order model. Indeed, there are more similarities within each group of field and group of simulated models than there are between the matched pairs. The coefficients within each group are far more likely to be of the same sign. For instance, the first coefficient for every simulated set is negative. A more general observation is the tendency for the simulated models to be of a higher order than those of the field test.

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	SAR 10	SAR 20
F60	.26		.18				-.15		
S60	-.77	-.73	.52	-.44	.21				
F62	-.14								
S62	.86	-.69	.16						
F63	-.33							-.75	-.41
S63	.45	-.37							
F66	-.23								
S66	.89	-.88	.58	-.35					
F76	-.37								
S76	.43	-.57							
F77	-.09	-.08		-.27					
S77	.61	-.77	.27						

Table 5 Fitted ARIMA Models

Each field model was used to filter the simulated data with AFOTEC's proposed validation strategy. Figures 33 and 34 show the residual autocorrelation and partial autocorrelation functions are not white and, therefore, the simulation fails to accurately match the outcome of the Golden Bird field test with respect to model form. The answer as to validity is more complex. Figures 33 and 34 are not only representative of the results that almost every data set portrays, but almost exact duplicates. This common pattern aspect in the residuals is so close it leads one to ask if some common factor is missing from the field models. No modeling factors could be found. This raises the question once again, are the deterministic noise spikes in the field data corrupting the models in such a way as to invalidate this test? It appears the significance of some higher order factor in the field models is possibly being masked by the added noise.

The result, that the simulations fail to pass this filter test, is not surprising when examining the models in Table Five. The simulated data vectors require higher order models to reduce them to white noise. Additionally, in each case where the field model does have a coefficient of the same order as the simulated model the magnitude is much smaller and is often of the opposite sign.

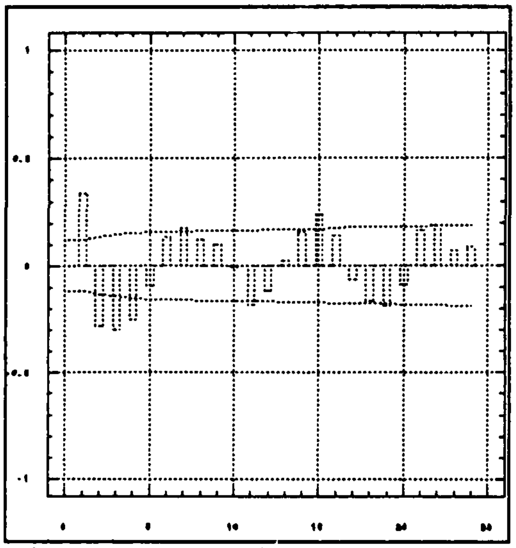


Figure 33 Residuals ACF Set 60 Filtered Set 60

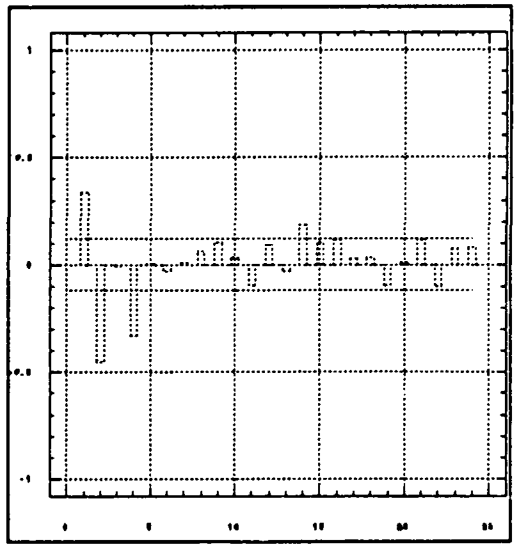


Figure 34 Residual PACF Filtered Set 60

Summary

The output data sets produced by the Terrain Effects Model have similar features to those of the field test, but are distinctly different. Classical statistical comparisons on the raw data show the TEM's output consistently has a lower average, median, and usually has a higher variance and range. The cross-correlation between simulated and field data is high. Unfortunately high cross-correlation exists between data sets where it would not be expected. This leaves cross-correlation as a means to disprove validity, but can do very little to add to the credibility of the model. The cross-correlation of the differenced simulated and differenced field test data is low when it should be high. The cross-correlation measurements tend to confirm gross pattern similarities, but fail to support the hypothesis the models are the same.

Time series analysis using the ARIMA model fitting methodology produces vastly different models for the field and simulated data sets. The field models are lower order models. The simulated output streams cannot be filtered through their corresponding field model to produce white noise and there by imply the models equivalent.

V. Conclusions & Recommendations

Introduction

The objectives of this research and a description of how the objectives were met is included in this final chapter. The conclusions from this effort are then presented based on the results of the techniques applied. The thesis concluded that given the data available, the Terrain Effects Model could not be validated with the methodologies attempted within this research. As was pointed out in the first section of Chapter Two, it is very important that some level of validation be performed to allow the simulation to be accredited for its intended purpose. Several recommendations are offered here for use in validating the Terrain Effects Model, and for future simulations used for weapons testing.

Objectives

The objective of this research was to characterize the field and simulation output data comparing and contrasting them, and to evaluate methodologies for doing so. Several general and specific characteristics from each group of data were discovered. The comparisons tend to lend credence to the face validity aspects of the simulation while raising questions about its true ability to predict the outcome of an air-to-air engagement for the ECM technique in question. The evaluation of the methodologies planned with the data available cast doubt on their ability to provide AFOTEC with a definitive answer as to whether the TEM is a valid representation of the jamming technique under study.

Conclusions

The Terrain Effects Model cannot be validated with the data set provided. While the simulation output appears to have some face validity, the internal stochastic structure of the model appears to be very different. There is no significant evidence the simulation models the real world as portrayed in the field test output. This conclusion ignores the obvious fact that the models are estimated from different types of data. The deterministic spikes in the field data could have potentially drastically altered the model that would have been estimated without their presence. To answer the validation question with any mathematically based scheme will require data unhampered by noise of this magnitude.

There does appear to be evidence the models are reasonably consistent. The residual autocorrelation patterns in the simulated data after it has been filtered through its corresponding field model are definitely consistent.

Recommendations

Several possibilities still exist to add credibility to the TEM. One of the most appealing methods which could be used for this problem and future validation efforts is the mean simulation overlay method described in Chapter Two and shown in Figures 2 through 4. This would require the contractor to produce several realizations of the same simulation from which the mean simulation output and standard deviations could be derived. The field test is then overlaid and should fall within one standard deviation of that mean. The method has intuitive appeal, is easy to interpret, and is far less expensive than repeated field testing. This method appears worthy of investigating for many future validation efforts as well.

Improved data collection, which eliminates the deterministic spikes in the field data, may substantially change the field test models and allow the proposed methodology originally intended to be successfully used. Improved data would at least add confidence that the field models are not biased towards the lower order models which were unable to filter the simulation outputs to white noise.

A concern must be raised about staking the general validity of the system on the validity of one engineering measure such as the seeker head angle error. AFOTEC presented strong rationale as to why this measure is believed to capture the underlying process. But, it must still be recognized that there are risks in claiming validity based on only one measure. The missile validations examined in this research each had several measures of effectiveness. The validation could easily be expanded into a multivariate pattern recognition problem by including such time series variables as azimuth error, control surface deflections, and others.

Of course, more measures of effectiveness require that more resources be expended in the validation process. Time, manpower, and money to validate could be scarce in this period of declining defense spending. In light of past failures to consider the validity of the simulations used, the forethought AFOTEC has given to this validation effort is commendable. Perhaps a more effective means of validating future simulations would be to require them from the builder as part of the contract. AFOTEC could then evaluate the contractors' validation efforts which should require less manpower than doing the actual validations. The Army Missile Command has had good results from overseeing the contractors' validation efforts, rather than doing the job themselves as an acceptance test. The concurrent verification and validation could very well improve the quality of the simulations that the Air Force receives. As part of a contractual obligation, the contractors have incentive to produce the closet approximations to reality possible. Good results accrue to the builder with this obligation as well. The validation results are accomplished sooner and any problems with the model can be corrected at an earlier stage of development. The simulation builders do many of the verification and validation requirements in thee normal course of developing the simulation. The buyer just does not receive the benefits and accreditation that could be gained if these efforts were known.

Appendix 1: ARIMA Model Fittings

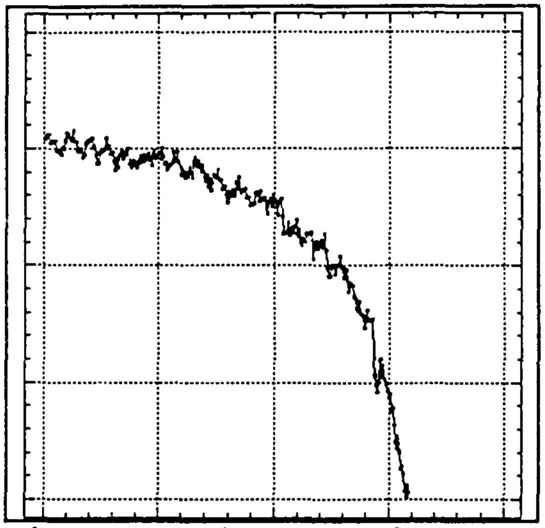


Figure 35 Field 62 Time Series Plot

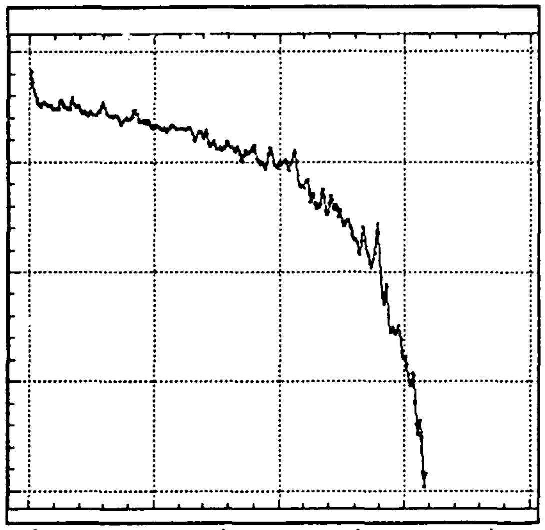


Figure 36 Sim 62 Time Series Plot

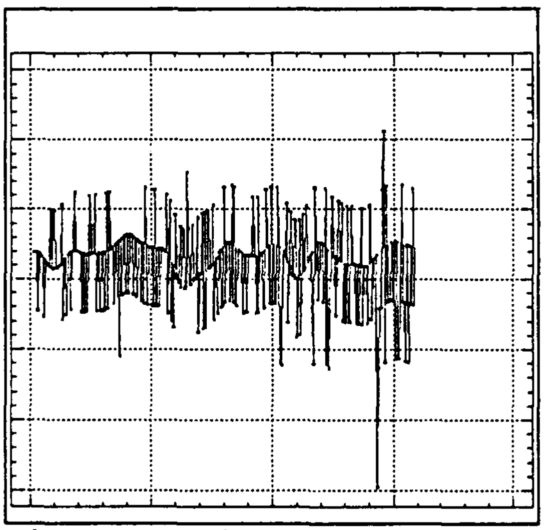


Figure 37 Field 62 Differenced Time Series Plot

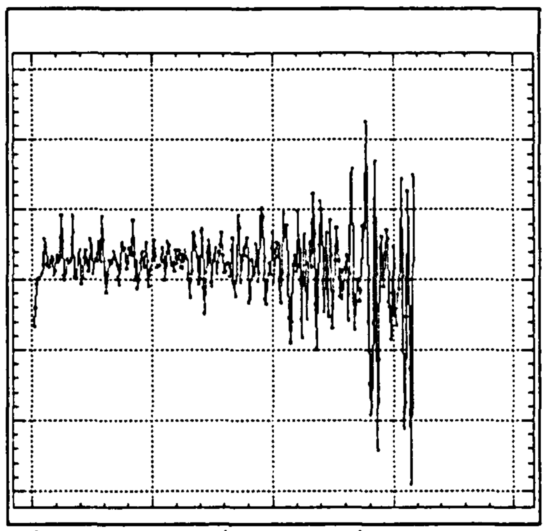


Figure 38 Sim 62 Differenced Time Series Plot

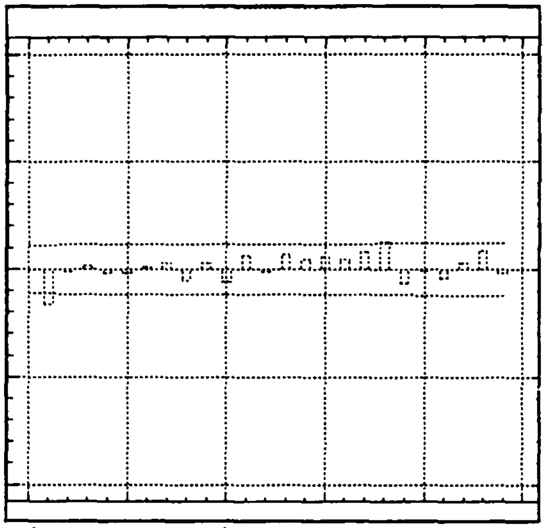


Figure 39 Field 62 ACF

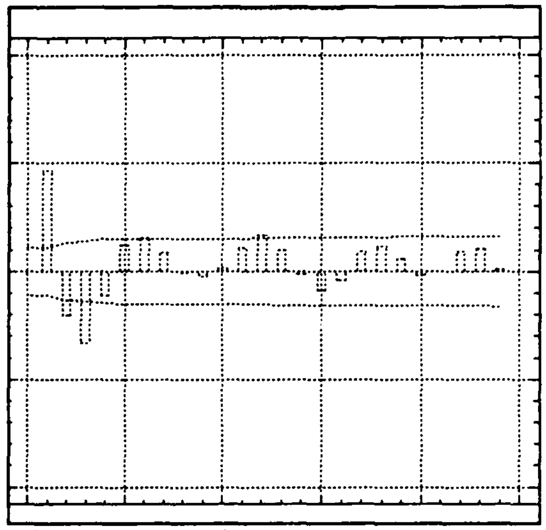


Figure 40 Sim 62 ACF

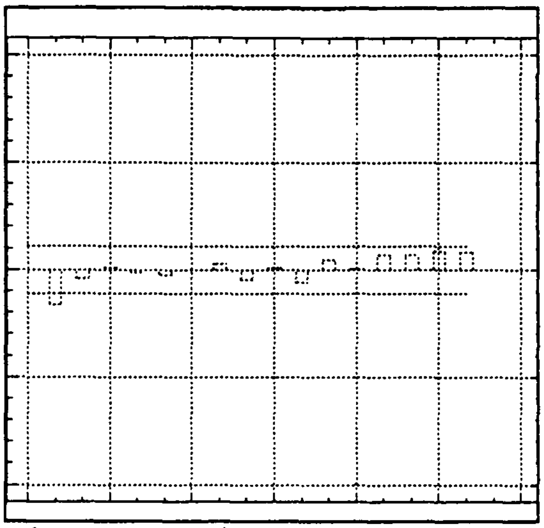


Figure 41 Field 62 PACF

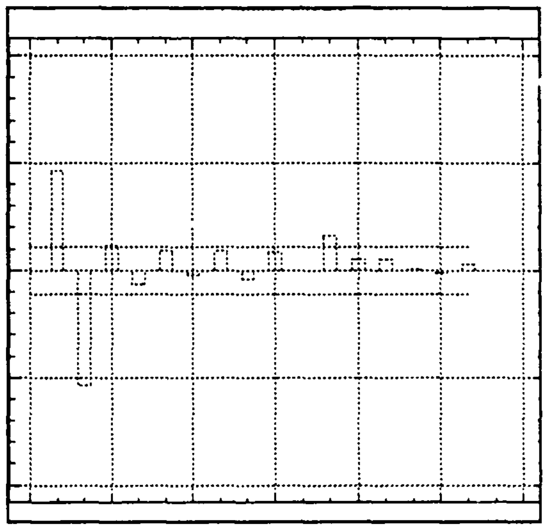


Figure 42 Sim 62 PACF

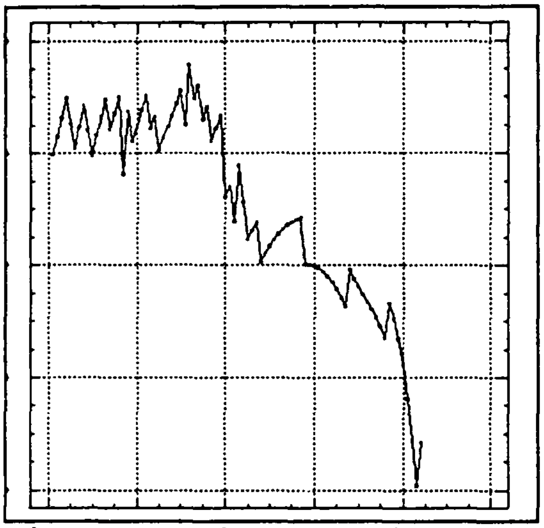


Figure 43 Field 63 Time Series Plot

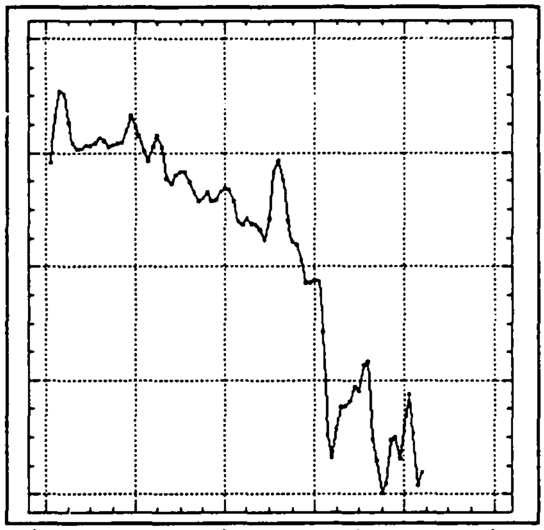


Figure 44 Sim 63 Time Series Plot

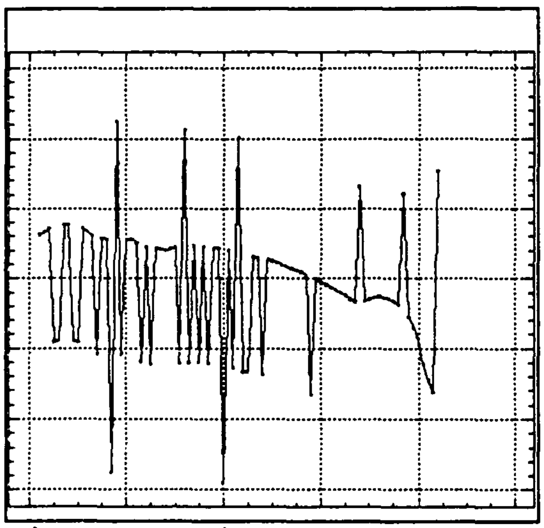


Figure 45 Field 63 Differenced Field Time Series

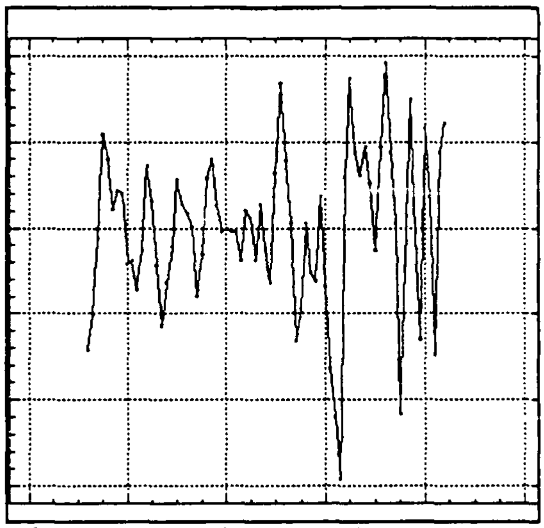


Figure 46 Sim 63 Differenced Time Series Plot

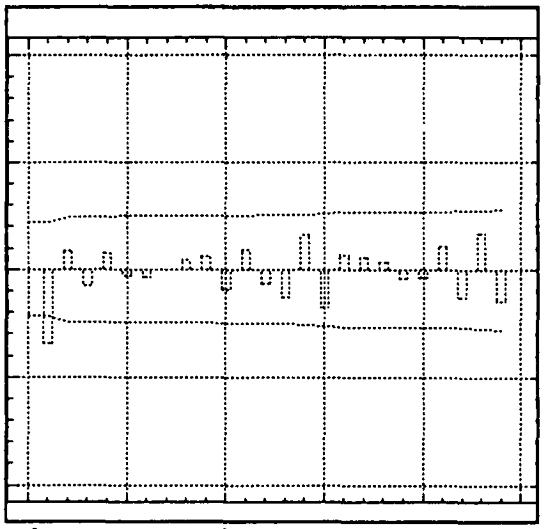


Figure 47 Field 63 ACF

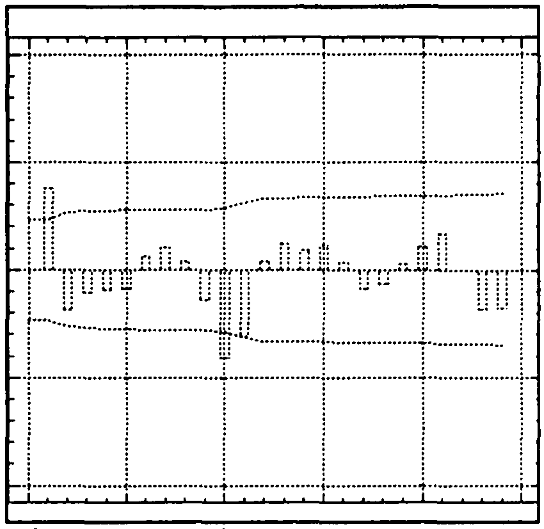


Figure 48 Sim 63 ACF

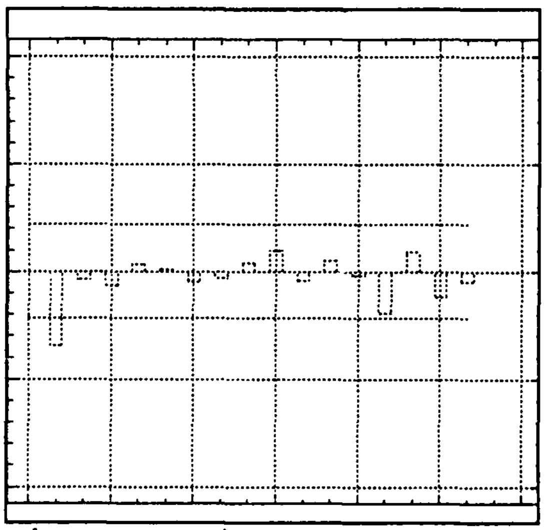


Figure 49 Field 63 PACF

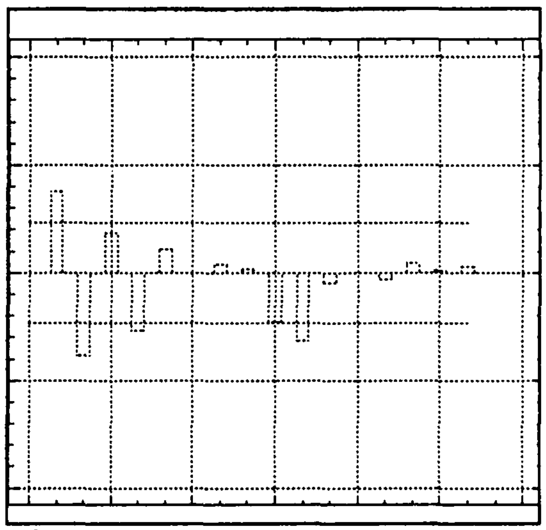


Figure 50 Sim 63 PACF

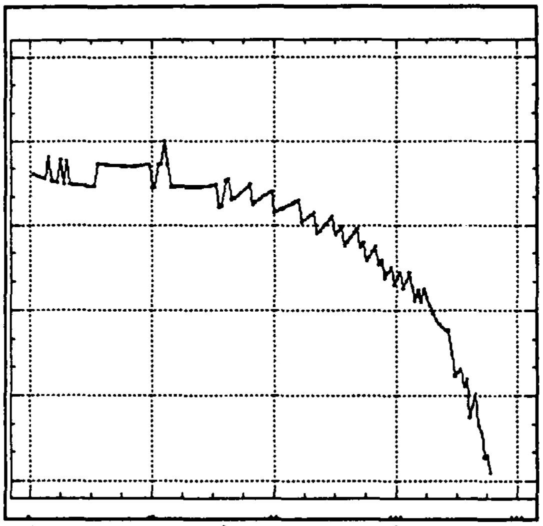


Figure 51 Field 66 Time Series Plot

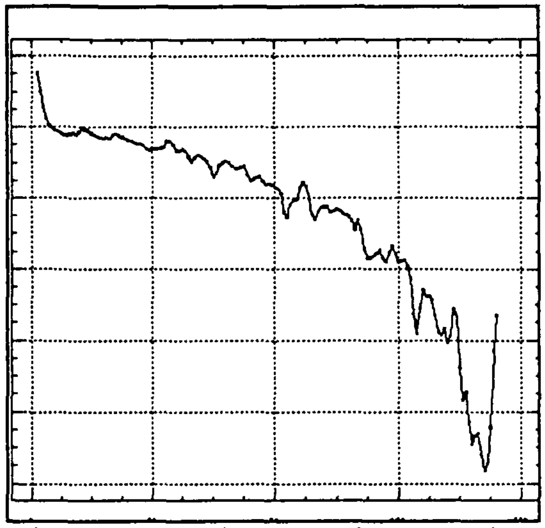


Figure 52 Sim 66 Time Series Plot

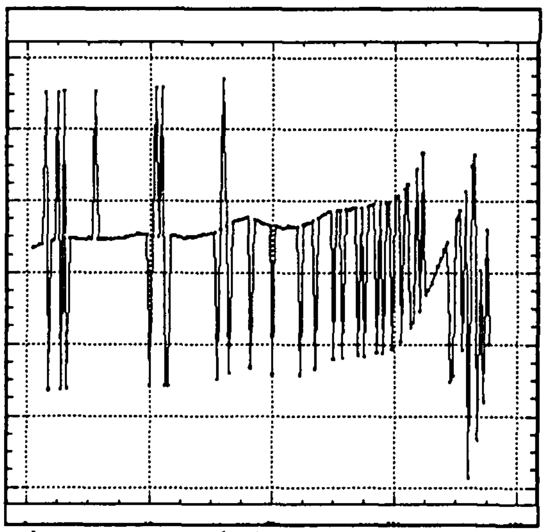


Figure 53 Field 66 Differenced Time Series Plot

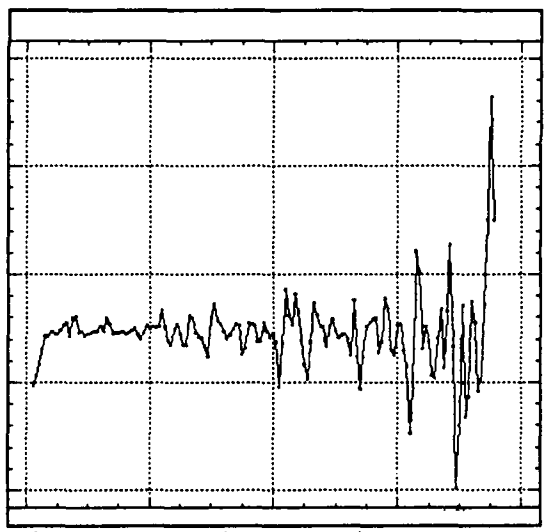


Figure 54 Sim 66 Differenced Time Series Plot

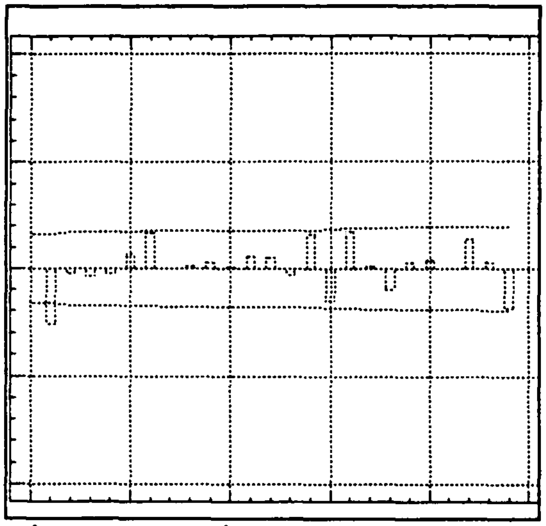


Figure 55 Field 66 ACF

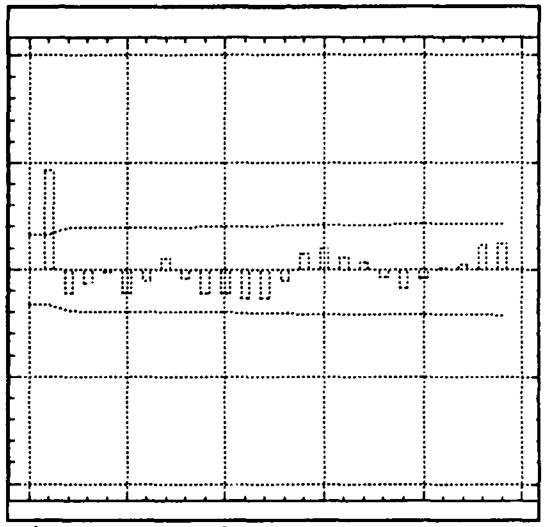


Figure 56 Sim 66 ACF

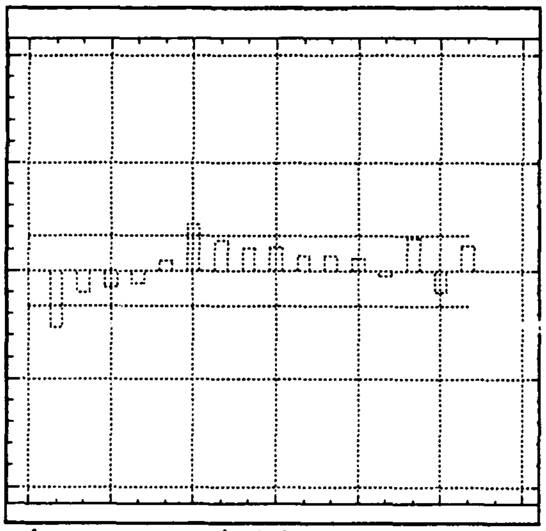


Figure 57 Field 66 PACF

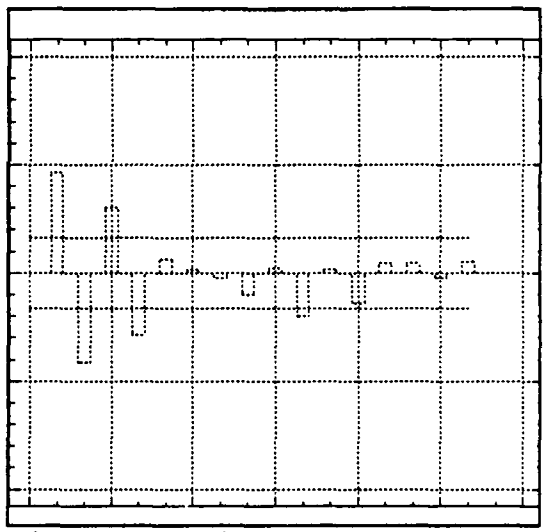


Figure 58 Sim 66 PACF

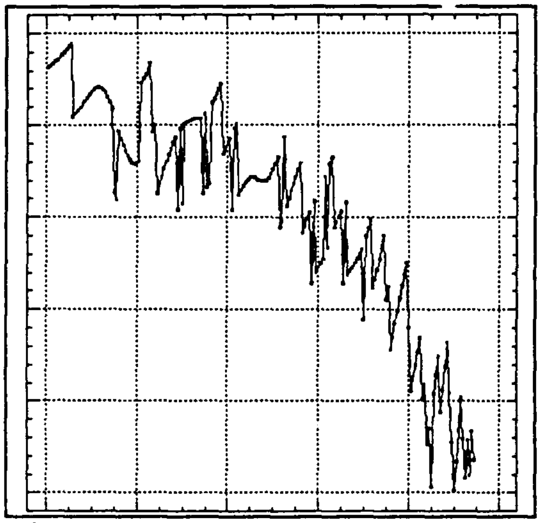


Figure 59 Field 76 Time Series Plot

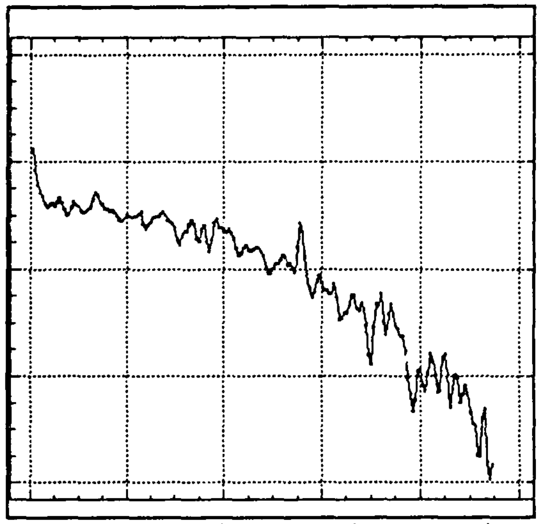


Figure 60 Sim 76 Time Series Plot

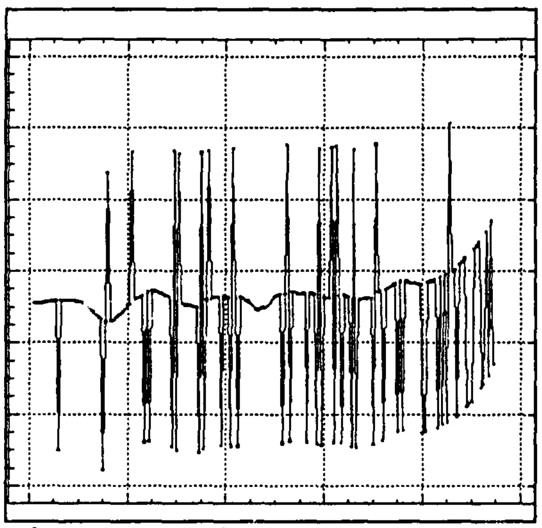


Figure 61 Field 76 Differenced Time Series Plot

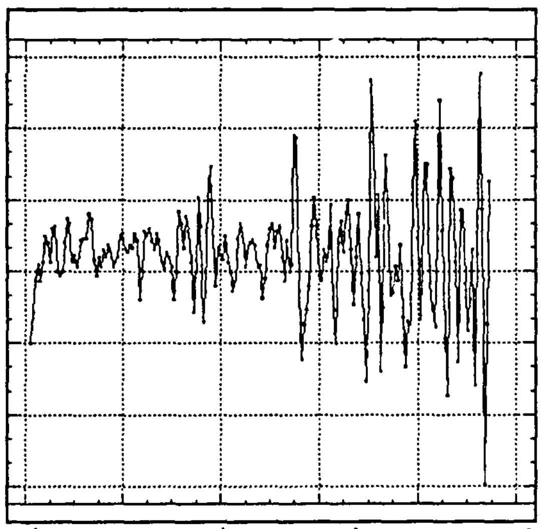


Figure 62 Sim 76 Differenced Time Series Plot

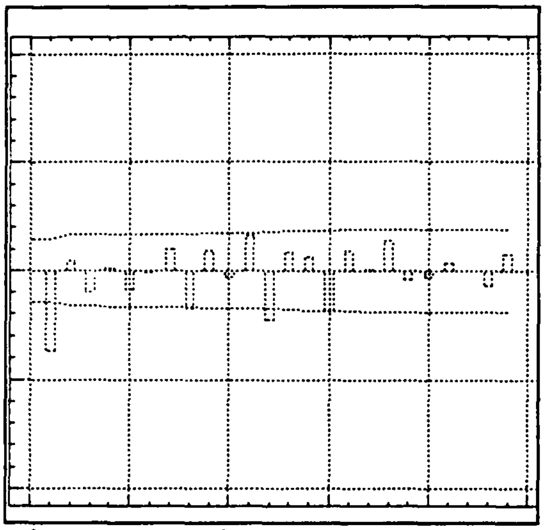


Figure 63 Field 76 ACF

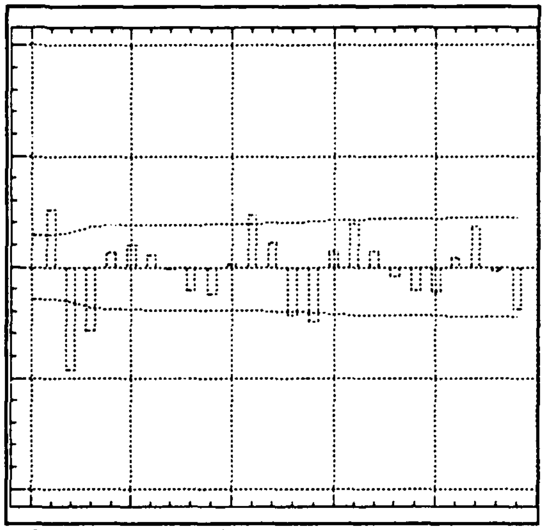


Figure 64 Sim 76 ACF

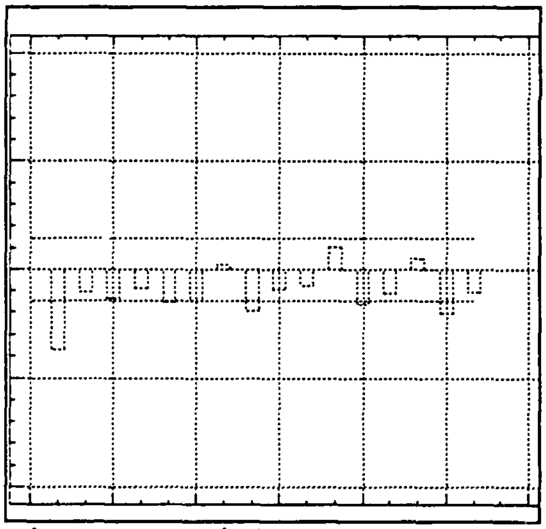


Figure 65 Field 76 PACF

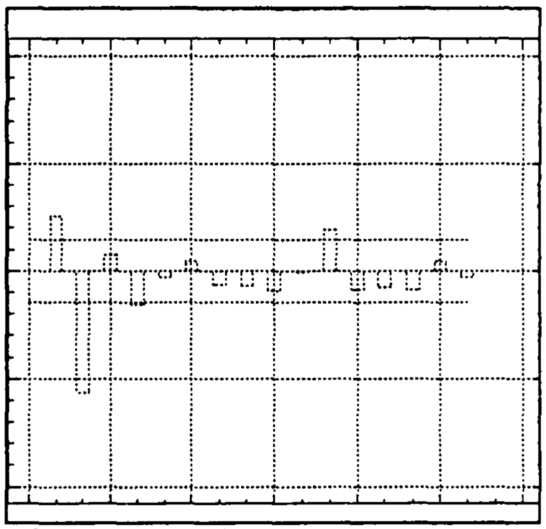


Figure 66 Sim 76 PACF

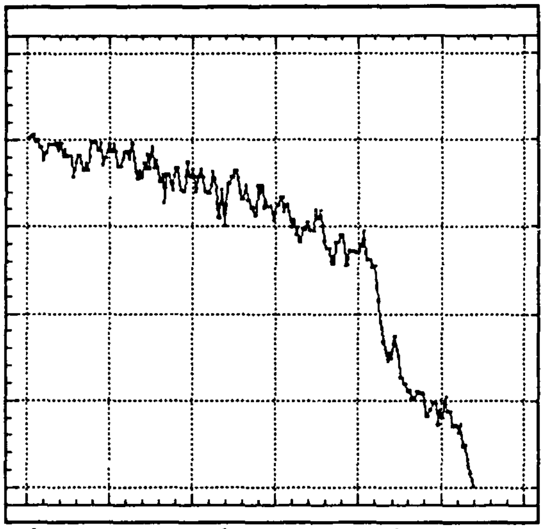


Figure 67 Field 77 Time Series Plot

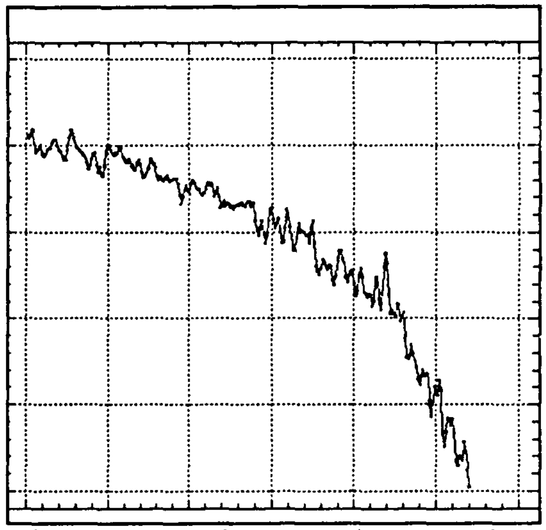


Figure 68 Sim 77 Time Series Plot

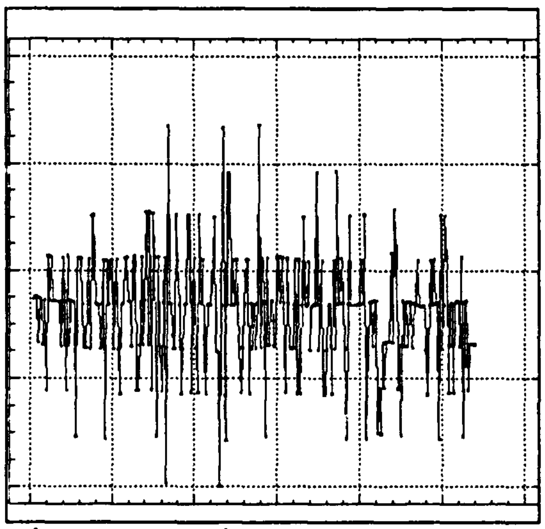


Figure 69 Field 77 Differenced Time Series Plot

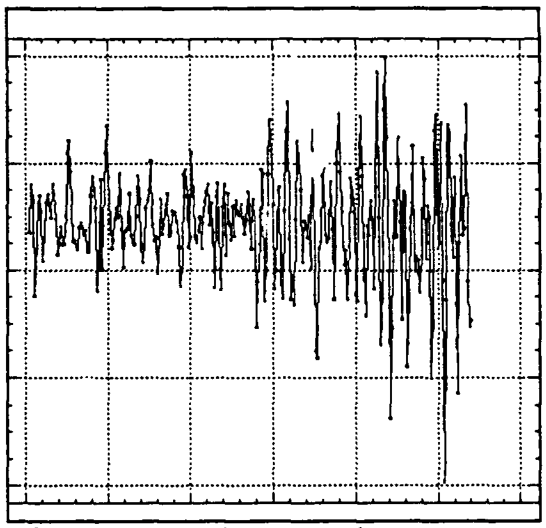


Figure 70 Sim 77 Differenced Time Series Plot

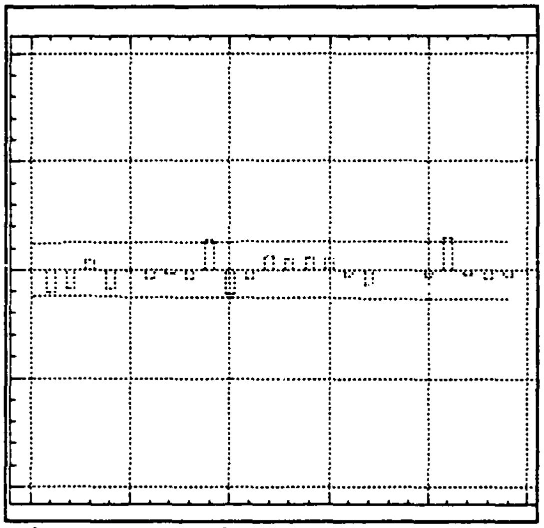


Figure 71 Field 77 ACF

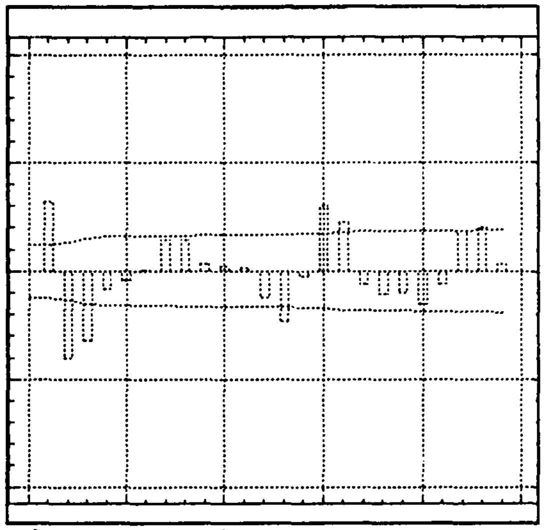


Figure 72 Sim 77 ACF

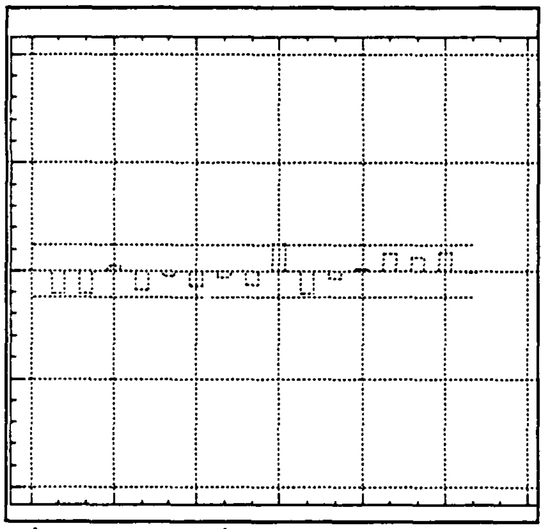


Figure 73 Field 77 PACF

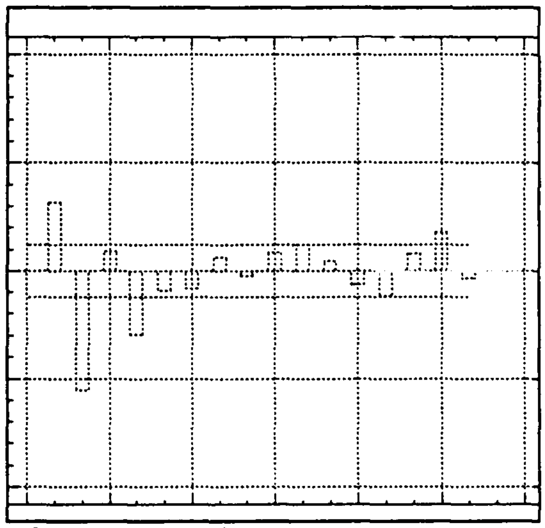


Figure 74 Sim 77 PACF

Appendix 2: Periodograms

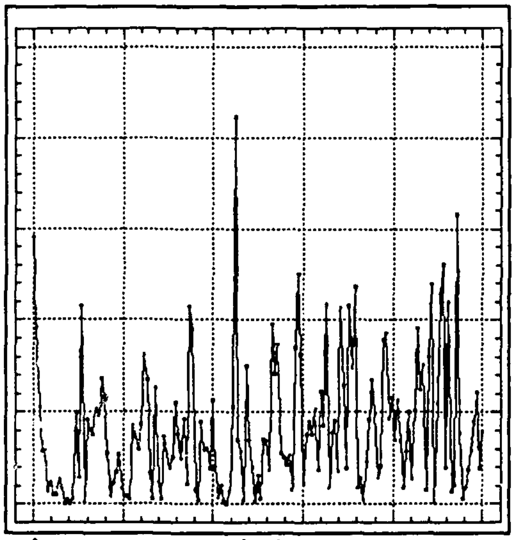


Figure 75: Field 62 Periodogram

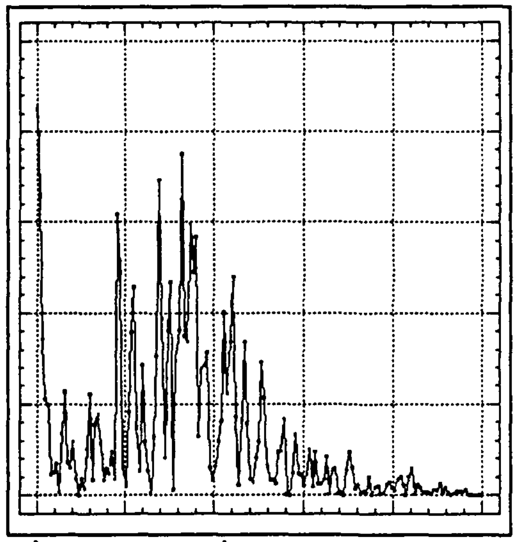


Figure 76 Sim 62 Periodogram

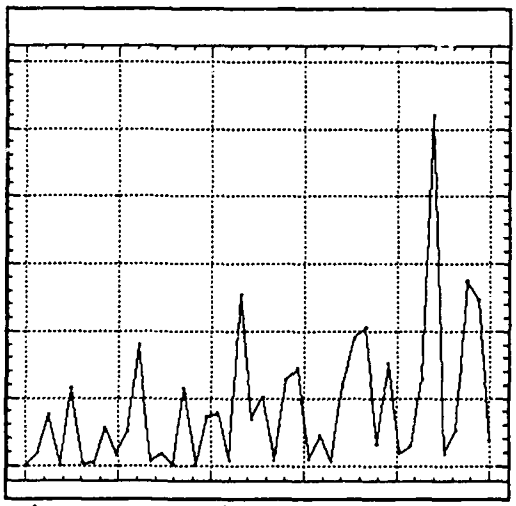


Figure 77 Field 63 Periodogram

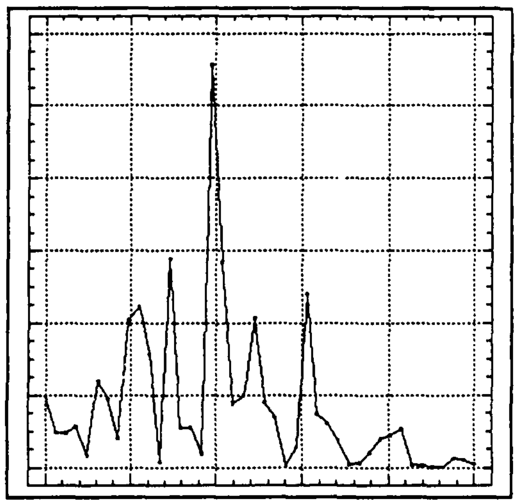


Figure 78 Sim 63 Periodogram

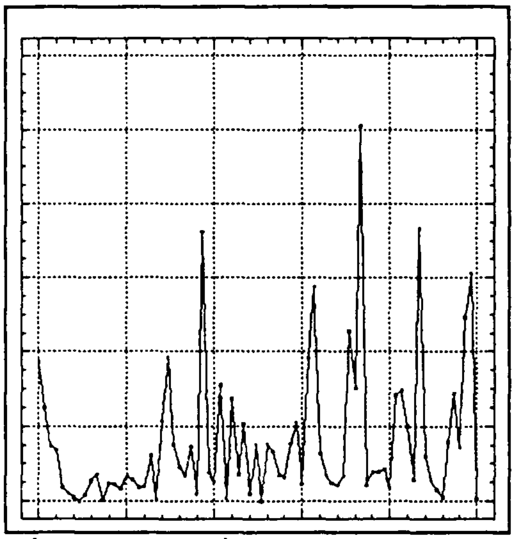


Figure 79 Field 66
Periodogram

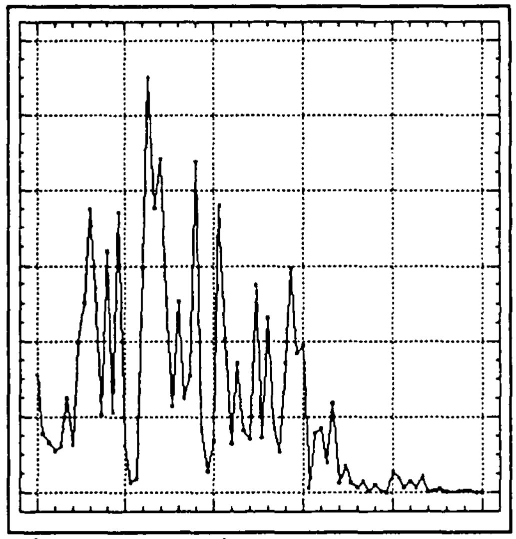


Figure 80 Sim 66
Periodogram

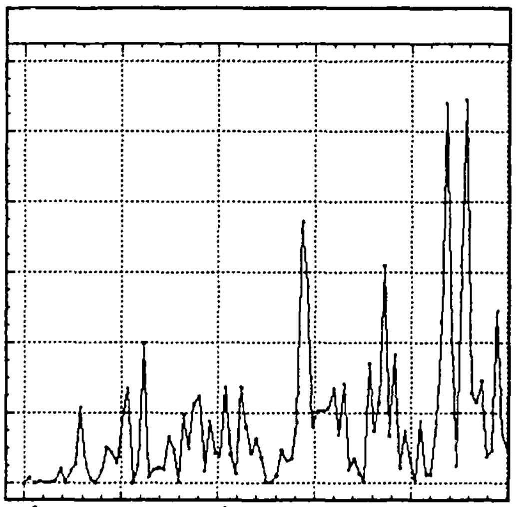


Figure 81 Field 76
Periodogram

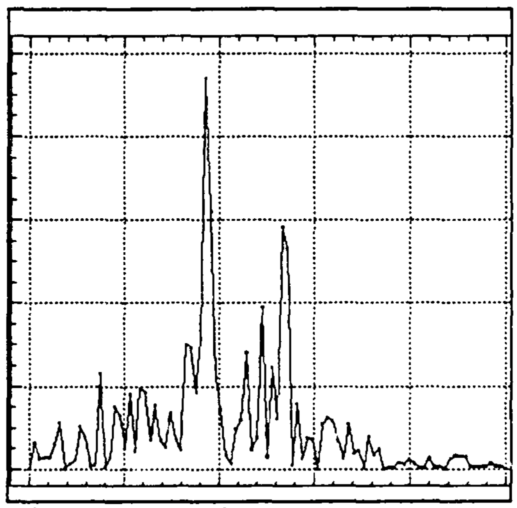


Figure 82 Sim 76
Periodogram

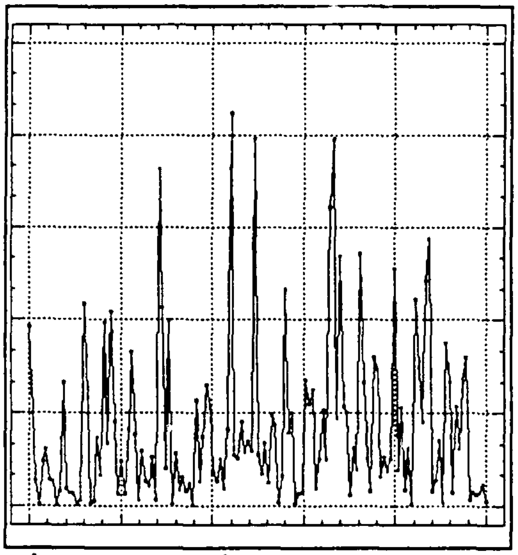


Figure 83 Field 77
Periodogram

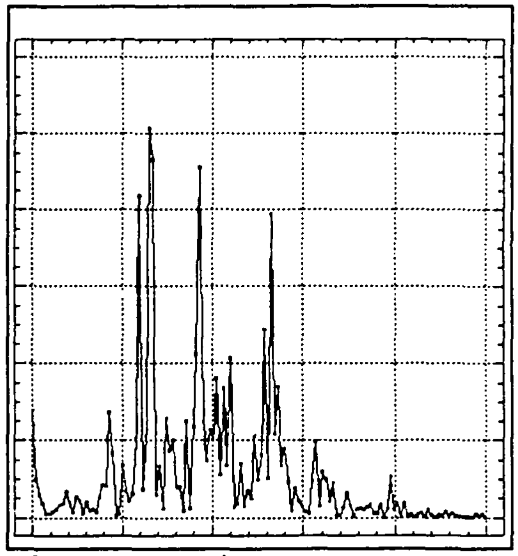


Figure 84 Sim 77
Periodogram

Bibliography

Abraham, Bovas and Johannes Ledolter. Statistical Methods for Forecasting. New York: John Wiley & Sons, 1983.

"Air Force Operational Test and Evaluation Center," Air Force Magazine: 116 (May 1989).

Bennett, Capt James, Chief of Modeling Simulation Branch. Personal Interviews. HQ AFOTEC/OANT, Kirtland AFB, NM. August 1989 through February 1990.

Box, George P. and Gwilym M. Jenkins. Time Series Analysis forecasting and control (Revised Edition). London: Holden-Day, 1976.

Department of the Army. Chaparral RSS Missile System Simulation Validation Plan. Redstone Arsenal: United States Army Missile Command, 23 April 1987.

Forrester, Jay W. and Peter M. Senge. "Test For Building Confidence in System Dynamics Models," IIMS Studies in the Management Science, 14:209-228 (1980).

Greene, Latricha and Dr. Douglas Montgomery. Methods for Validating Computer Simulation Models of Missile Systems. Technical Report RD-81-17. U.S. Army Missile Command, Redstone Arsenal, AL. August 1981 (AD-A117407).

Gravitz, Robert M. and William F. Waite. "Validation Methodologies For Complex, Hybrid, HWIL, 6DOF Missile Simulations --A Structured Approach--," Proceedings of the 1988 Winter Simulation Conference. 775-782. San Diego: The Society of Computer Simulation, 1988.

HSU, D.A. and J.S. Hunter. Simulation Model For New York Air Traffic Control Communications. Report No. FAA-RD-74-203. Washington: Government Printing Office, February, 1975.

-----, "Analysis of Simulation-Generated Responses Using Autoregressive Models," The Institute of Management Sciences 24: 181-190 (October 1977).

Kheir, Naim A. and Willard M. Holmes. "On Validating Simulation Models of Missile Systems," Simulation 30: 117-128 (April 1978).

Law, Averill M. and W. David Kelton. Simulation Modeling and Analysis. New York: McGraw Hill Book Company, 1982.

Makridakis, Spyros and others. Forecasting: Methods and Applications (Second Edition). New York: John Wiley & Sons, 1983.

Mann, Lt Col Greg A. The Role of Simulation In Operational Test And Evaluation. Research Report No. AU-ARI-83-10. Airpower Research Institute, Air University Press, Maxwell AFB AL, August 1983.

McCleary, Richard and Others. Applied Time Series Analysis for the Social Sciences, London: Sage Publications, 1980.

Sargent, Robert G. "A Tutorial On Validation and Verification of Simulation Models," Proceedings of the 1988 Winter Simulation Conference. 33-39. San Diego: The Society of Computer Simulation, 1988.

Shumway, Robert H. Applied Statistical Time Series Analysis, Englewood Cliffs: Prentice Hall, 1988.

U.S. General Accounting Office. DOD Simulations Improved Assessment Procedures Would Increase the Credibility of Results, GAO/PEMD-88-3. Washington, D.C.: 1987.

-----, Guidelines for Model Evaluation, GAO-79-17. Washington, D.C.: 1979.

-----, Advantages and Limitations of Computer Simulation in Decisionmaking, GAO/71092. Washington, D.C. 1973.

Vita

Major Robert E. Massey [redacted] He graduated from Warner Robins High School in Warner Robins, Georgia in 1973 and entered the Air Force Academy. Major Massey received a Bachelor of Science degree in Economics and was commissioned in the Air Force on 1 June 1977. After attending Undergraduate Navigator Training at Mather AFB, California, he served four years as an EC-130 navigator at Keesler AFB,, Mississippi. Major Massey was then selected to attend Undergraduate Pilot Training at Williams AFB, Arizona. Following training Major Massey was assigned to the 5th Bombardment Wing at Minot AFB, North Dakota. During his five years at Minot he served as a pilot, the Wing Flight Safety officer, and earned a Master of Administration degree from Central Michigan University. Major Massey is married to the former [redacted] of Colorado Springs, Colorado. They have three children: [redacted] He entered the School of Engineering, Air Force Institute of Technology, in August, 1988.

[redacted]

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION: UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Distribution is unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): AFIT/GST/ENS/90M-12	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION: School of Engineering; 6b. OFFICE SYMBOL (If applicable): AFIT/ENS	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code): Air Force Institute of Technology, Wright-Patterson AFB, OH 45433	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: AFOTEC; 8b. OFFICE SYMBOL (If applicable): AFOTEC/OANT	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code): Kirtland AFB, NM 87117	10. SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)

Validation of the Terrain Effects Model by Comparing Autoregressive Integrated Moving Average Models (U)

12. PERSONAL AUTHOR(S)

Robert E. Massey, Major, USAF

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Masters Thesis	FROM ______ TO ______	1990 March	99

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Validation, Time Series, Simulation, ECM
12	03		
17	04	03	

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

This thesis examines the capability of validating the Terrain Effects Model (TEM) by comparing a time series output of the model to one produced during field testing. The TEM is a simulation of an air-to-air missile's flight path while being subjected to electronic countermeasures when launched from an altitude above the target aircraft.

Data from the field and simulation tests were characterized and fit with time series models using Box-Jenkins' Autoregressive Integrated Moving Average methodology. The models had less explanatory power than that which is usually associated with a time series representation, most probably caused by large deterministic noise intrusions in the field data. Several recommendations for other simulation validation techniques are offered.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION: UNCLASSIFIED	
22a. NAME OF RESPONSIBLE INDIVIDUAL: James N. Robinson, Lt Col, USAF	22b. TELEPHONE (Include Area Code): 513-255-3362	22c. OFFICE SYMBOL: AFIT/ENS

DD Form 1473, JUN 86 Previous editions are obsolete. SECURITY CLASSIFICATION OF THIS PAGE