DEVELOPMENT OF A DYNAMIC SIMULATION MODEL FOR PLANNING PHYSICAL DISTRIBUTION SYSTEMS: VALIDATION

Thesis for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY

PETER GILMOUR

1971

This is to certify that the

thesis entitled

DEVELOPMENT OF A DYNAMIC SIMULATION MODEL FOR PLANNING PHYSICAL DISTRIBUTION SYSTEMS: VALIDATION

presented by

Peter Gilmour

has been accepted towards fulfillment of the requirements for

Ph.D. degree in Production

Richard J. Gonzalez
Major professor

Date April 19, 1971

O-7639

ABSTRACT

DEVELOPMENT OF A DYNAMIC SIMULATION MODEL
FOR PLANNING PHYSICAL DISTRIBUTION
SYSTEMS: VALIDATION

By

Peter Gilmour

Only recently has the potential cost saving and the competitive advantage of an integrated physical distribution system been realized. The aim of a recently completed research project at the Michigan State University Graduate School of Business Administration was to develop a general model which would enable the user to evaluate total cost and service capability interactions within the physical distribution system over the long term. This dynamic simulation model has been developed and is named the Long Range Environmental Planning Simulator (LREPS).

Simulation, as a managerial decision making tool, has greatly increased in acceptance and use over the past decade. Problems have been approached which up until this time were considered too large to be manageable. So the extremely complex problem is quite often analyzed through the use of a simulation model. Because of the large

investments of time and money needed to develop a simulation model of a complex situation, little energy is often left to consider the question of the validity of the final model.

This dissertation is a formal study of validation of computer simulation models in general, and in particular an analysis of the performance of the LREPS model.

The concept of validity for a computer simulation model is rather naturally divisable into design validity and output validity. While design validity is the establishment of the reasonableness of the basic underlying processes of the model, output validity is the acceptability of the form of the model's endogenous data streams. The argument that the validity of a theory (or model) is not based on the realism of its assumptions, but on the accuracy of its predictions, is accepted. Although this means concentration on output validity, design validity is not ignored. Testing for design validity can take the form of determining the model's face validity, that is, testing in a rudimentary way to see if the model "makes sense" in relation to the available knowledge of the situation being modeled. This type of testing is a coarse screening device at stages during the model's development and as a test at its initial completion.

Three major procedures are applied to establish a model's output validity:

1. Analysis of the stability of the model over the long term. Stability is the ability of the model to generate endogenous data streams which show persistent behavior.
2. Comparison of the output of the model for some past time period with the actual historical data that was recorded for that time period.
3. Comparative analysis of the data streams generated by the model before and after significant changes in the model's major assumptions. The output of a simulation model should not be related to the nature of specific assumptions contained in the model.

A reasonably comprehensive subset of possible statistical techniques is examined for use in each of these three validation procedures. Considered are (1) Graphical Analysis, (2) Analysis of Variance, (3) Multiple Comparison, (4) Multiple Ranking, (5) The F Test, (6) Correlation, (7) Regression Analysis, (8) Sequential Analysis, (9) The Kolmogorov-Smirnov Test, (10) Response Surface Analysis, (11) The Chi-square Test, (12) Theil's Inequality Coefficient, (13) Spectral Analysis, and (14) Factor Analysis.

Due to the rather stringent assumptions included in many of the other techniques, and also due to the fact that spectral analysis considers the effects of autocorrelation, the results of this technique were relied upon most heavily.

The LREPS model was subjected to the proposed validity testing. Initial face validity testing established the acceptability of a wide range of cost and service data streams to the management of the industrial sponsor. Now the three general procedures could be applied to determine the output validity of LREPS.

1. The model was found to be stable over the long run.
2. The ability of the model to duplicate actual historical data was not established.
3. The model output was not significantly related to the nature of the two major assumptions embodied in the model.

The only unfavorable results for LREPS was the failure to establish the predictive ability of the model. Availability of sufficient historical data obtained at an adequate time increment was a necessary condition for the satisfactory completion of this validation procedure. Data of the required quality was not available.

Establishing the validity of a computer simulation model is a difficult task. These three validation procedures do provide a general method which, together with the particular knowledge required for face validity testing, can be used to perform this task.

DEVELOPMENT OF A DYNAMIC SIMULATION MODEL FOR PLANNING PHYSICAL DISTRIBUTION SYSTEMS: VALIDATION

By

Peter Gilmour

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Management

1971

ACKNOWLEDGEMENTS

This dissertation developed from a need to establish the validity of the LREPS model. Since this model was constructed by a team of doctoral candidates under the direction of Professor Donald J. Bowersox, acknowledgement must be given to O. Keith Helferich, Edward J. Marien, Michael L. Lawrence, Richard T. Rogers, and Fred W. Morgan, Jr., for assistance, both direct and indirect, in the progress of this study.

I would also like to express my appreciation to the Johnson and Johnson Domestic Operating Company for support of the LREPS project.

My dissertation committee was composed of Dr. Richard F. Gonzalez, Professor of Production Management, Dr. Donald J. Bowersox, Professor of Marketing and Transportation and Co-chairman of the committee with Dr. Gonzalez, and Dr. Thomas J. Manetsch, Associate Professor of Electrical Engineering and Systems Science. Dr. Gonzalez, as my academic advisor, has been of great assistance in planning my progress through the doctoral program. His knowledge of the technical aspects of computer simulation and the procedural aspects of writing a doctoral dissertation has speeded my task. Dr. Bowersox negiotated the LREPS project with the industrial sponsor, and for the opportunity to participate in the project I owe him thanks. Dr. Manetsch provided assistance from his experience with continuous simulation model building.

Finally, I would like to thank my parents, the Reverend and Mrs. William F. Gilmour, for the direction they provided, and my wife, Laurie, for typing the drafts, correcting the grammar, and for her patience.

TABLE OF CONTENTS

Chapter | Page

Chapter Page

LIST OF TABLES

LIST OF FIGURES

Figure		Page

Figure | Page

CHAPTER I

VALIDATION

Introduction

The development and use of a mathematical model has become a popular means by which a solution to a problem is attempted. But when the quantitative relationships in the model become so complex that a mathematical solution is not possible or extremely difficult to obtain, computers and numerical methods offer a feasible alternative. This approach is simulation.

The aim of computer simulation can basically be described as system design or system analysis. System design (a normative approach) is an attempt to find the combination of exogenous variables and parameter values that will optimize a specified endogenous variable, possibly subjected to the attainment of specified limits on other endogenous variables. System analysis (a positive approach) is an explanation of the relationship between the endogenous variable and the controllable exogenous variables and parameters.

Simulation allows the analyst, in his drive for greater realism, to develop a much more detailed and complex model than he could using an analytical technique.

But a simulation model is a symbolic or numerical abstraction of the real process, and the danger exists that the limitations and assumptions of the method will become hidden (or not adequately considered) by its complexity. A simulation model may be constructed of a firm's physical distribution system. Sales forecasts and product line at that time form an integral part of the model. If the model is used over a period of years without updating these factors, the output of the model may well be of no value to the firm.

Validation of the operation of a simulation model is as desirable as the validation of the operation of any other scientific experiment. While the basic problem of validation is no different for a simulation experiment, the complexity of the model is such that the processes by which its validity is established are quite different. With most scientific experiments it is rather easy and inexpensive to carry out several independent replications. Due to the complexity of most simulation models, the expense of performing more than one experiment is often prohibitive, while longitudinal observations during this one experiment are autocorrelated.

The time and effort needed to develop and make operational a computer simulation model are at present so great that the problem of its validation has generally been neglected. A common attitude seems to be that crude judgmental and graphic methods[1] are preferable to completely ignoring validation.

Philosophy of Validation

To validate a model in a strict sense means to prove that the model is true. That truth is a rather elusive concept can be seen in the difficulty one has in developing a set of criteria for differentiating between a model which is "true" and one which is "not true." Fortunately most simulations are seldom concerned with proving the "truth" of the model (an exception might be Clarkson's model to simulate the behavior of a bank's investment trust officer).[2] Popper,[3] therefore, suggests that efforts should be concentrated on determining the degree of confirmation rather than verification. Models should be subjected to tests, the results of which could be negative with respect to the aims of the model. Each such test passed will add confidence to our assumption that the model behavior confirms the behavior of the real system. "Thus instead of verification, we may speak of gradually increasing confirmation of the law."[4]

Van Horn describes validation as the "process of building an acceptable level of confidence that an inference about a simulated process is a correct or valid inference for the actual process."[5] The focus for validation should be to understand the input-output relationships in the model and to be able to translate "learning" from the simulation to "learning" about the actual process. Naylor and Finger[6] basically agree and provide some insight as to how this focus can be operationalized. The computer simulation model and its output are based on inductive inferences about behavior of the real system in the form of behavioral assumptions or operating characteristics. The real situation under study is usually so complex that the construction of an exact model is not possible. Another factor besides complexity which makes computer simulation the desirable method of analysis is the random nature of one or more of the exogenous variables. Therefore:

> The validity of the model is made probable, not certain, by the assumptions underlying the model. . . . The rules for validating computer simulation models and the data generated by these models are sampling rules resting entirely on the theory of probability.[7]

Three major methodological positions on validation are summarized by Naylor and Finger: rationalism, empiricism, and positive economics.

Rationalism . . . Models or theory are a system of logical deductions from a series of synthetic premises of unquestionable truth. Validation is the search for the basic assumptions underlying the behavior of the system.

Empiricism . . . The opposite view to rationalism is that empirical science is the ideal form of knowledge. The model should be constructed with facts, not assumptions. So any postulates or assumptions which cannot be independently verified should not be considered.

Positive Economics . . . This view championed by Milton Friedman is that the validity of a model depends upon its ability to predict the behavior of the dependent variables and not on the validity of the assumptions on which the model rests.

These three positions are combined by Naylor and Finger into a multi-stage verification procedure, each stage of which is necessary but not sufficient. Stage 1 is the formulation of a set of postulates or hypotheses describing the behavior of the system. This involves specification of components, selection of variables, and formulation of functional relationships using observation, general knowledge, relevant theory, and intuition. Stage 2 is the attempt to verify the assumptions of the model by statistical analysis, and the final stage is to test the degree to which data generated by the model conforms to observed data. The multi-stage verification procedure attempts to include all major ways in which to build confidence in a model.

A final view on validation is that of Fishman and Kiviat[8] which is a narrower concept because they divide simulation testing into three parts.

> (1) Verification insures that a simulation model behaves as an experimenter intends. (2) Validation tests the agreement between the behavior of the simulation model and a real system. (3) Problem analysis embraces statistical problems relating to (the analysis) of data generated by computer simulation.[9]

Experimental Design and Validation

It is difficult to distinguish where experimental design ends and validation begins. The process of computer simulation experimentation is interative: model construction, model operation, validation, and experimental design. If the validation criteria are not satisfied, the process is repeated making adjustments until validity is indicated.

The aim of a simulation experiment may be stated as the desire to explore and describe the response surface over some region in the factor space (system analysis) or to optimize the response over some feasible region in the factor space (system design). In order to achieve this aim in the most economical manner, careful attention must be paid to experimental design.[10] The types of experiments for which the model is used will depend upon the particular requirements that the model was designed to meet.[11] But the types of problems that can be associated with experimental design are universal.

A single run of a computer simulation provides an estimate of population parameters. Because the model contains exogenous random variables, this estimation, or sample of one, will not exactly equal the population parameter. However, the larger the sample or the more runs that are made, the greater is the probability that the sample averages will be very close to the population averages. The convergence of sample averages to population averages with increasing sample size is called stochastic convergence. Because stochastic convergence is slow, methods other than increasing the sample size may be required.

Another problem is that of size. The number of cells required for a full factorial experiment becomes very large even with few levels of a moderate number of factors. If a complete investigation of all factors is not essential, fractional factorial designs can ameliorate the problem.

Yet another common problem associated with experimental design arises from the desire to observe many different response variables in a given experiment.

> It is often possible to bypass the multiple response problem by treating an experiment with many responses as many experiments each with a single response. Or several responses could be combined (e.g. by addition) and treated as a single response. However, it is not always possible to bypass the multiple response problem; often multiple responses are inherent to the situation under study. Unfortunately, experimental design techniques for multiple response experiments are virtually nonexistent.[12]

This dissertation will be concerned only with validation. The other elements of the interative process of computer simulation experimentation are discussed in detail elsewhere.[13]

Scope and Method

From the rather diverse views on validation examined earlier, a position must be taken. The validity of a computer simulation model can be shown by the model's ability to satisfy three distinct validation procedures.

The output of a simulation model is in the form of a time path for each of the endogenous variables. The first validation procedure is to determine if these time series are statistically under control. Being under control broadly means that over the long run the time path will show convergence properties or else the rate of change of the endogenous variable under study will be proportional to or acceptable to the rate of change in all other endogenous variables.

Simulation models can be broadly classified as positive or normative. Positive models must by definition show reasonable correspondence to the real system, while normative models indicate a desirable level of operation for the real system which may or may not be currently achieved. But is it reasonable for a model to show the desired state and not to indicate how to reach this state from the current real state? If the normative model was built by changing starting conditions and parameter values of the positive model, management would be provided with the means to move from the current actual position to the more desirable normative position. The normative model should then be built from the basis of the positive model. For the positive simulation model, then, the second validation procedure is to compare the model output over a past time period to the actual historical data from the same time period.

The assumptions upon which a model is based often cannot be examined beyond the level of face validity. But the sensitivity of these assumptions can be examined, and this is the third validation procedure. If the values of the key endogenous variables are sensitive to the nature of the assumption, then managerial knowledge and intuition must be applied to confirm the assumption, or else the model must be restructured to eliminate or replace the assumption.

Organization

A research project to develop a long-range planning model for physical distribution has been established at the Michigan State University Graduate School of Business Administration. The project has two broad aims: to develop the model, which has been done, and to use the model and adaptions to it to provide management with information about the physical distribution system over the long run. Five dissertations will describe in detail the project development as shown in Figure 1.1. The scope of this dissertation is delineated in the figure although other aspects of the project will be briefly described for the sake of continuity.

Attitudes towards the validation of computer simulation models and the general position to be taken in this

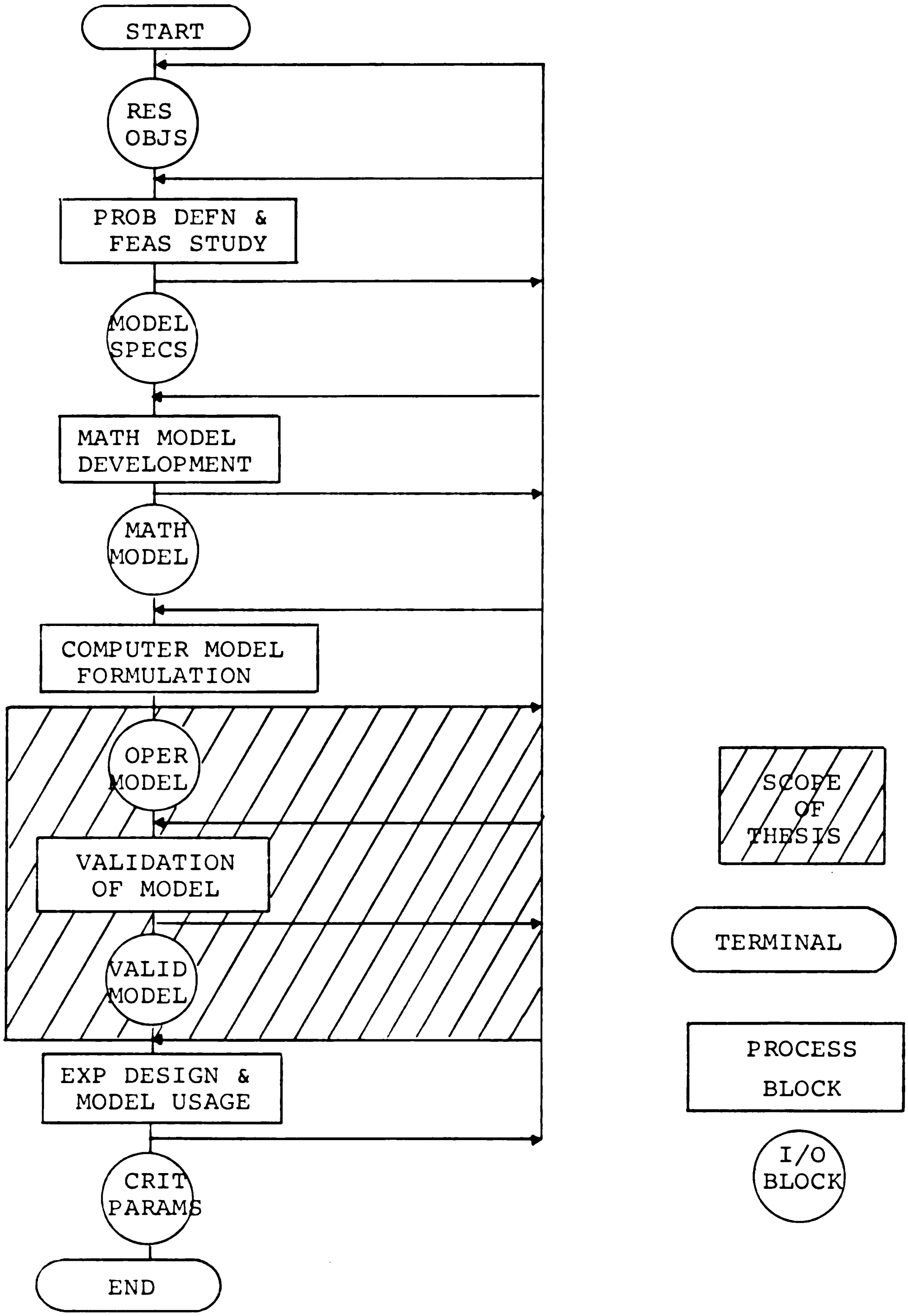


Figure 1.1.--LREPS Systems Design Procedure.[1]

[1]D. J. Bowersox, et al., Dynamic Simulation of Physical Distribution Systems, Monograph (East Lansing, Michigan: Division of Research, Michigan State University, Forthcoming).

dissertation were discussed in this introductory chapter. Many different statistical methods can be used in order to establish the validity of a computer simulation model. A reasonable subset of these statistical techniques is discussed in Chapter II without attempting at this point to establish the relative merit. Chapter III is a brief description of several celebrated simulation models and an evaluation of the attempts made by the model builders to validate their models.

The simulation model is described in Chapter IV. The degree to which the model's face validity has been established is discussed. Also given in this chapter is the manner in which the model and its output will be used in order to satisfy the general validation procedures outlined in Chapter I.

The next three chapters deal in detail with each of these three general validation procedures. From the set of statistical techniques detailed in Chapter II are selected those most suitable for stability analysis (Chapter V), for the comparison of simulation output and actual data (Chapter VI), and for sensitivity analysis of the model's major assumptions (Chapter VII). After a technique is found to be suitable for a particular validation procedure, the results of its application will be analyzed in the light of the assumptions inherent in the technique.

The final chapter (Chapter VIII) is a summary statement of the validity of the simulation model. The question of establishing a general validation procedure for computer simulation models is also explored.

CHAPTER I--FOOTNOTES

[1]One basic procedure is to determine the model's face validity. This is a necessary, but not sufficient, condition for validation, which is discussed at some length in Chapter IV.

[2]G. P. E. Clarkson, Portfolio Selection: A Simulation of Trust Investment (Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1962).

[3]K. R. Popper, The Logic of Scientific Discovery (New York: Basic Books, 1959).

[4]R. Carnap, "Testability and Meaning," Philosophy of Science, Vol. 3, No. 4 (October, 1936).

[5]R. Van Horn, "Validation," The Design of Computer Simulation Experiments, ed. by T. H. Naylor (Durham, N.C.: Duke University Press, 1969), pp. 232-251.

[6]T. H. Naylor and J. M. Finger, "Verification of Computer Simulation Models," Management Science, Vol. 14 (October, 1967), pp. 92-101.

[7]Ibid., p. 93.

[8]G. S. Fishman and P. J. Kiviat, Digital Computer Simulation: Statistical Considerations (Santa Monica, Calif.: The Rand Corporation) RM-3281-PR, 1962.

[9]Ibid.

[10]T. H. Naylor, D. S. Burdick, and W. E. Sasser, Jr., "The Design of Computer Simulation Experiments," The Design of Computer Simulation Experiments, ed. by T. H. Naylor (Durham, N.C.: Duke University Press, 1969), pp. 3-35.

[11]R. T. Rogers, "Development of a Dynamic Simulation Model for Planning Physical Distribution Systems: Experimental Design and Analysis of Results"(unpublished Ph.D. dissertation, Michigan State University, Forthcoming).

[12]Naylor, Burdick and Sasser, p. 30.

[13]D. J. Bowersox, *et al.*, *Dynamic Simulation of Physical Distribution Systems*, Monograph (East Lansing, Michigan: Division of Research, Michigan State University, Forthcoming).

CHAPTER II

VALIDATION TECHNIQUES

Introduction

Three types of analysis for the validation of the simulation model are to be performed:

1. Stability testing.
2. The comparison of actual historical data with the simulation output for the same time period.
3. The comparison of two simulation data streams in order to test the sensitivity of some major assumptions made during model development.

Many statistical and graphical techniques have been proposed and used in an attempt to validate the output of computer simulation models.[1] In order to determine which of these techniques will be most suitable for each of the three types of analysis, the nature of the techniques must be examined. This chapter presents what is hopefully a large subset of all possible validation techniques.

Sequential Analysis

Most decision-making procedures are carried out with the sample size predetermined and fixed. It is possible that this sample size is larger than it need be resulting in superflous information and unnecessary expense. But this can be avoided if after each observation is examined the decision is made to:

1. Accept the hypothesis.
2. Reject the hypothesis.
3. Postpone a decision on the hypothesis and make another observation.

Together with this variable sample size are required managerially determined values of α (producers risk) and β (consumers risk) to make the system operational. The decision rules for testing $H_o:\mu=\mu_o$ and $H_1:\mu=\mu_1$ are:

1. If $$\frac{\prod_{i=1}^{y} f(x_i,\mu_1)}{\prod_{i=1}^{y} f(x_i,\mu_o)} \leq \frac{\beta}{1-\alpha}$$; Accept $H_o:\mu=\mu_o$ (Reject H_1),

2. If $$\frac{\prod_{i=1}^{y} f(x_i,\mu_1)}{\prod_{i=1}^{y} f(x_i,\mu_o)} \geq \frac{1-\beta}{\alpha}$$; Reject $H_o:\mu=\mu_o$ (Accept H_1),

3. If $$\frac{\beta}{1-\alpha} < \frac{\prod_{i=1}^{y} f(x_i,\mu_1)}{\prod_{i=1}^{y} f(x_i,\mu_o)} < \frac{1-\beta}{\alpha}$$; Take another observation,

where y is the number of observations taken.[2]

This general statement of the procedure for sequential analysis provides a method for deciding at the ith observation whether to stop sampling and accept or reject the hypothesis under consideration or whether to continue sampling by making the (i+1)th observation. At observation i the division of the i-dimensional space of all possible observations into the three mutually exclusive and exhaustive sets is the basic problem of sequential analysis.

The method has several applications for the analysis of the results of computer simulations. Procedures for testing the position of the true mean in relation to a hypothesized mean and for comparing the means of k experiments with a control mean have been developed by Paulson.[3] A heuristic approach to Bechhofer and Blumenthal's method[4] of selecting the population with the largest mean is described by Sasser, Burdick, Graham, and Naylor.[5]

The Chi-Square Test

The Chi-square statistic can be used to measure the discrepancy between observed and expected frequencies.

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

If $\chi^2=0$, perfect agreement between observed and expected frequencies exists while the larger the value of χ^2, the greater the discrepancy between the two.

The sampling distribution of χ^2 is approximated by the Chi-square distribution $Y = Y_o(\chi^2)^{\frac{1}{2}(v-2)} e^{-\frac{1}{2}\chi^2} = Y_o\chi^{v-2} e^{-\frac{1}{2}\chi^2}$ where v is the number of degrees of freedom and y_o is a constant related to v such that the total area under the curve is unity.

When using the Chi-square Test, expected frequencies are developed from a hypothesis H_o. It is reasonable to expect the calculated Chi-square value to be less than a critical value such as $\chi^2_{.95}$ which is the critical value at the .05 significance level. If this turns out to be the case, H_o is accepted at this level of significance. Otherwise it is rejected.

Caution should be exercised if the correspondence between observed and expected is too close. If χ^2 is less than $\chi^2_{.05}$ at the .05 significance level, the agreement is too great for the degree of significance chosen.

Regression Analysis[6]

It is often meaningful to be able to express the relationship between the variable under study (the dependent variable) and other variables which have influence over it (the independent variables). The most commonly accepted method of determining this relationship is that of least squares. A line, curve or plane is fitted to the data in such a manner so as to minimize the vertical squared difference between the plotted data value and the value determined by the function being fitted. The result is then the "best fitting" line, curve or plane. While this function shows the relationship between the independent variables and the dependent variables, it also enables predictions of the dependent variable to be made.

The approach is illustrated by the most simple example of fitting a straight line to n pairs of values of two variables x and y. Let ε_i be the error or difference between the true sample value of y and the value of y ($\hat{y}$) determined by the function of the straight line being fitted ($\hat{y} = a + bx$), i.e., $\varepsilon_i = y_i - \hat{y}(i=1,\ldots,n)$. For all observations ε_i^2 must be minimized.

$$\text{MIN} \quad \Sigma\varepsilon_i^2 = \Sigma(Y_i - \hat{Y})^2 = \Sigma(Y_i - a - bX_i)^2$$

Take the partial derivatives with respect to a and b, set them equal to zero, and obtain the normal equations:

$$\Sigma Y_i = n\hat{a} + \hat{b}\Sigma X_i$$

$$\Sigma X_i Y_i = \hat{a}\Sigma X_i + \hat{b}\Sigma X_i^2$$

Solve for $\hat{a}$ and $\hat{b}$

$$\hat{b} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$

$$\hat{a} = \overline{Y} - \hat{b}\overline{X}$$

which are the least-squares point estimators of a and b. The least squares line fitted to the data is then $\hat{y} = \hat{a} + \hat{b}x$.

Analysis of Variance

Analysis of variance is used to test if two or more samples differ significantly with respect to a particular (usually qualitative) property. If observations are classified on the basis of a single property, the ratio of the variance between the groups and the average variances within the groups (the F ratio) is used to determine if a significant difference does exist between the groups with respect to this property.

To test the null hypothesis, H_o, that the expected profit from each of a number of plans is equal, this decision rule is set up. If $F \geq F_{\alpha.k-1.k(n-1)}$, where α is the significance level, k is the number of plans considered and n is the number of replications per plan, reject H_o, otherwise accept it. While if H_o is accepted, the differences in expected profit between the plans is only due to random fluctuation, if H_o is rejected, further analysis (such as multiple comparison or multiple ranking) is needed to quantify this significant difference between plans.

Given[7]

X_{ij} = Total profit from the ith replication of the jth plan
$\overline{X}_{.j}$ = Average profit for jth plan over all replications
$\overline{X}_{..}$ = Grand average profit for all plans over all replications.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Between Plans	SS plans $= n\sum_{j=1}^{k}(X_{.j}-X_{..})^2$	k-1	$MSp=\frac{SS\ plans}{k-1}$
Error	SS error $= \sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij}-\bar{X}_{.j})^2$	k(n-1)	$MSe=\frac{SS\ error}{k(n-1)}$
TOTAL	SS total $= \sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij}-\bar{X}_{..})^2$	nk-1	

The value of F obtained $(\frac{MSp}{MSe})$ is compared to the appropriate value from the F table in the manner indicated.

The F Distribution

To compare the variances of small samples, the F distribution is used. The function is

$$f(F) = C\frac{F^{\frac{1}{2}(n_1 - 2)}}{(n_2 + n_1F)^{\frac{1}{2}(n_1 + n_2)}}$$

where

$$F = \frac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2} \quad ,$$

n_1 is the number of degrees of freedom for the $\chi^2_{n_1}$ distribution, and n_2 is the number of degrees of freedom for the $\chi^2_{n_2}$ distribution.

The F statistic is equal to the ratio of the sample variances. Given a level of significance and the two sample sizes (from which can be determined the degrees of freedom), the critical value of F can be read from a table of the F distribution. By comparing the value of the F statistic to the critical value of F, the hypothesis that the variances are significantly different can either be accepted or rejected.

Multiple Comparison

Analysis of variance uses the F Test to determine if a significant difference exists between a statistic from different samples. If homogeneity does not exist, the method of multiple comparison quantifies the difference, while the method of multiple ranking[8] (to be discussed) directly identifies the "best" sample or plan on the basis of the measured statistic. Both multiple comparison and multiple ranking must follow analysis of variance for another reason--the computational reason that both these methods use the mean square of the error.

Use of confidence intervals rather than tests of hypotheses is a characteristic of the method.

Tukey[9] developed simultaneous confidence intervals for the differences between all pairs. Continuing with the notation used in the section on analysis of variance, the confidence intervals are:

$$(\bar{X}_{\cdot j} - \bar{X}_{\cdot J}) \pm q_{k,v} \sqrt{\frac{MSe}{n}} \qquad j,J=1,2, \ldots , k$$

where the q statistic can be obtained from tables and v is the number of degrees of freedom. If Student's t statistic is used, the intervals are not all simultaneously true at

$$(\bar{X}_i - \bar{X}_j) \pm t\sqrt{\frac{2MSe}{n}} \qquad j,J=1,2, \ldots , k.$$

A somewhat different approach is taken by Dunnett.[10] Instead of taking all possible pairs, he compares the control statistic (usually a result of the present operation of the system under study) to all alternative values of this statistic.

$$(\bar{X}_{\cdot j} - \bar{X}_{\cdot c}) \pm d \sqrt{\frac{2MSe}{n}} \qquad j=2, \ldots , k$$

where $\bar{X}_{\cdot c}$ is the control sample statistic (mean) and d is Dunnett's t statistic with k(n-1) degrees of freedom for a one factor experiment.

Multiple Ranking

This is a method to find the "best" plan.[11] It is a more direct method than multiple comparison, answering questions such as, "With what probability can it be said that a ranking of sample means represents the true ranking of the population means?"

Bechhofer, Dunnett, and Sobel describe a two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. Take a first sample of N_1 observations from each of the k populations or plans under investigation. Calculate the mean square of the error (MSe) which is an unbiased estimator of the population variance having k(n-1) degrees of freedom for $n = N_1$. Now take a second sample of $N_2 - N_1$ observations from each of the k populations.

$$N_2 = \text{SUP} \mid N_1, \ [2\text{MSe}(h/\delta^*)^2] \mid$$

where $[2\ \text{MSe}(h/\delta^*)^2]$ is equal to the smallest integer greater than or equal to the rational number $2\text{MSe}(h/\delta^*)^2$. The values of h are tabulated, and δ^* is the smallest difference between expected values that is acceptable. So if $2\text{MSe}(h/\delta^*)^2$ is less than or equal to N_1, a second sample is not taken, and N_2 is set to N_1. The next step is to calculate the overall sample mean ($\overline{X}_j$) for each population.

$$\overline{X}_i = \frac{1}{N_2} \sum_{i=1}^{N_2} X_{ij} \qquad j=1,2, \ldots , k$$

denote ranked values of $\overline{X}_j$ by $\overline{X}_{[1]} < \overline{X}_{[2]} < \ldots < \overline{X}_{[k]}$. Rank populations according to observed $\overline{X}$; and select that with the largest $\overline{X}_{[x]}$.

Theil's Inequality Coefficient

When comparing predicted results against actual outcomes, it is desirable to be able to establish the quality of the prediction. One way to do this is to calculate Theil's Inequality Coefficient.[12]

The mean-square prediction error for a set of n observations is equal to

$$\frac{1}{n} \sum_{i=1}^{n} (P_i - A_i)^2$$

where (P_i, A_i) stands for a pair of predicted and observed values. Theil calls its square root the root-mean-square prediction error (RMS). This term is expressed in the same dimensions as the predictions and realizations. If the RMS prediction error is divided by the square root of the mean square successive difference of the realizations, the result is the inequality coefficient (U) of the n pairs (P_i, A_i).

$$U = \sqrt{\frac{\Sigma (P_i - A_i)}{\Sigma A_i^{\,2}}}$$

If $U = 0$, the forecasts are perfect, as $P_i = A_i$ for all i. While it should be observed that $U = 1$ indicates a prediction error equal to that obtained by the naive method of no-change extrapolation, it should also be noted that U has no finite upper bound. Worse methods of forecasting than simple extrapolation are possible. Comparison of the technique being used and extrapolation provide valuable information.

Because the denominator of the inequality coefficient is a factor only to provide the proper unit of measurement, attention can be centered on the numerator. The square of the numerator can be decomposed into three terms, each of which expresses the extent to which a particular kind of prediction error is present.

$$\frac{1}{n}\Sigma(P_i - A_i)^2 = (\bar{P} - \bar{A})^2 + (S_P - S_A)^2 + 2(1 - r)S_P S_A$$

where $\bar{P}$ and $\bar{A}$ are the means:

$$\bar{P} = \frac{1}{n}\ \Sigma P_i \qquad \bar{A} = \frac{1}{n}\Sigma A_i$$

S_p and S_a are the standard deviations:

$$S_P{}^2 = \frac{1}{n}\Sigma(P_i - \bar{P})^2 \qquad S_A{}^2 = \frac{1}{n}\Sigma(A_i - \bar{A})^2$$

and r is the correlation coefficient of the predicted and realized changes:

$$r = \frac{\frac{1}{n}\Sigma(P_i - \overline{P})(A_i - \overline{A})}{S_P S_A} .$$

Errors leading to positive values for the first term of the decomposition are errors of central tendency; errors leading to positive values for the second term are errors of unequal variation; and errors due to incomplete covariation result in positive values for the decomposition's final term. If each of these three terms is divided by their sum, the resulting inequality proportions--U^m the bias proportion, U^s the variance proportion, and U^c the covariance proportion--provide additional information as to the quality of the prediction and an indication as to the direction in which effort should be applied for improvement.

$$U^m = \frac{(\overline{P} - \overline{A})^2}{\frac{1}{n}\Sigma(P_i - A_i)^2}$$

$$U^s = \frac{(S_P - S_A)^2}{\frac{1}{n}\Sigma(P_i - A_i)^2}$$

$$U^c = \frac{2\ (1 - r)\ S_P S_A}{\frac{1}{n}\ (P_i - A_i)^2}$$

And $U^m + U^s + U^c = 1$.

The Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov Test[13] is a nonparametric test to determine if a given sample is a sample from a particular distribution function. A Chi-square Test can also be developed to supply the same information.

Order the given sample, X_i, in ascending order. Find $F(X_i)$ for each X_i as the area below X_i in the theoretical distribution being considered. Where

$$Fn(t) = \frac{\text{Number of } X_i \leq t}{n}$$

, $Fn(X_i^+)$ is the right-hand limit at X_i of $Fn(t)$ and $Fn(X_i^-)$ is the left-hand limit at X_i of $Fn(t)$. Dn is then equal to the maximum of the absolute values of $Fn(X_i) - Fn(X_i^+)$ or $Fn(X_i) - Fn(X_i^-)$. Now, given that $X = nDn$ and where n is the sample size, look up P which is tabulated. Finally the null hypothesis that the sample is a sample from this theoretical distribution is rejected if P is no larger than a preassigned number α.

Response Surface Analysis[14,15]

When the response y is a continuous function of a single factor x, the method of response surface analysis is relatively easily applied to find the maximum or minimum of this function in the practical range of interest. Several conditions must be satisfied before this technique can be effective. It must be assumed that the response function can be approximated by a simple polynomial over the range of interest and that the function has only a single maximum (or minimum) within this range. So the key to this method is seen to be the managerial skill with which the relevant area of interest is selected. The general area of the extreme point must be known.

The general aims of the procedure are to find the extreme point and also to determine the sensitivity of the response function in the area of the extreme point.

Make several observations of y for different values of x within a selected subregion. Within this subregion, if it is assumed that the response function can be approximated by a straight line, the slope of this line can indicate in which direction x should change for the next observation of y. If the slope is relatively steep, the indication is that the extreme point of the function is still a reasonable distance (in terms of x) away, while if the slope is small, the extreme point is either very near or very far. So if the slope is small, several more observations of y are taken for a given change in x. If this new slope declines, the optimum is indeed close by; but if the new slope increases, the optimum is still some distance away. When the area of the extreme point is reached, a second-degree polynomial ($y = a + bx + cx^2$) is fitted to the observations made in the region. The first derivative of this function will provide the extreme point, and the second derivative will provide the relative sensitivity of the function in the area of the extreme point.

The size of the change in x used is important. This change is determined from a general knowledge of the process being examined. But at the same time, the change in x must be such that the resulting change in y is greater than can be explained by experimental error, otherwise, a poor estimate of the slope will result. When fitting the polynomial, the size of the change in x should be held constant.

When the response y is dependent on more than one factor, the principles of the method remain the same, but now more than one path to the extreme point exists. The question now becomes how to reach the region of the optimum most economically.

Considering two factors, one method is to hold the first factor constant and vary the second until the response is at an optimum. Hold the second factor constant at this level and vary the first until the response to it is optimal. Continue this procedure until the response is optimal for both factors simultaneously. This method and a response surface for which the method would not work are shown in Figures 2.1 and 2.2 respectively.

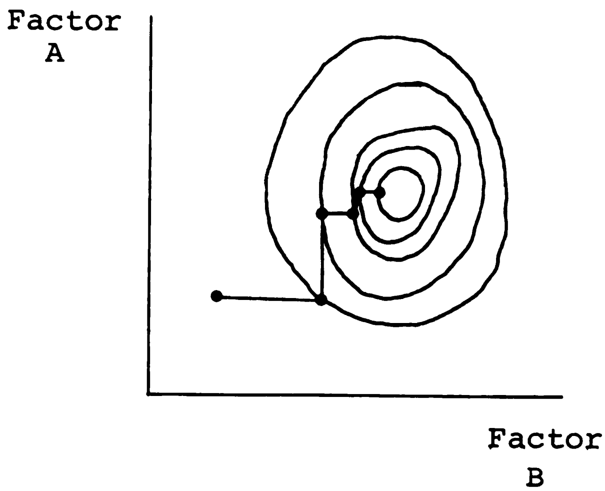


Figure 2.1.--Successful Application of Response Surface Analysis.

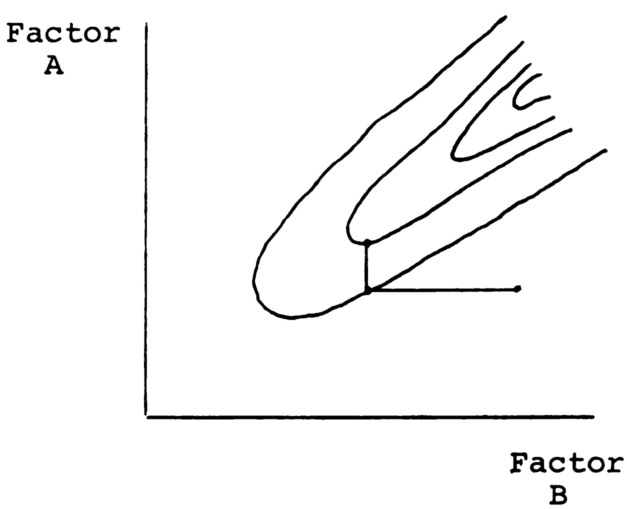


Figure 2.2.--Unsuccessful Application of Response Surface Analysis.

The Box-Wilson Method[16] of steepest ascent overcomes this disadvantage to the one-at-a-time method. The greatest ascent at any point is obtained if movement is made in a direction perpendicular to the contour line through that point. To find the contour, a small number of observations must be made in a subregion which is considered near the maximum and to these points is fitted a linear function or plane. Movement is made in a perpendicular direction, and if a marked gain in the response function is observed, further observations are made in this new region and a new plane is fitted. This procedure is repeated until the fitted plane levels out (the increase in the response along the path of steepest ascent is diminishing) at which point the response surface is mapped with a second degree equation. Classical methods then determine the extreme point and its sensitivity. The method is illustrated in Figure 2.3.

Spectral Analysis

Because all data generated by time series is autocorrelated to some degree, a method of analysis which will account for this autocorrelation is desirable. After transforming the data from the time domain to the frequency domain, spectral analysis[17,18] is a method by which the autocorrelation can be quantified and evaluated. Information about the magnitude of deviations from the average level of a given activity and information

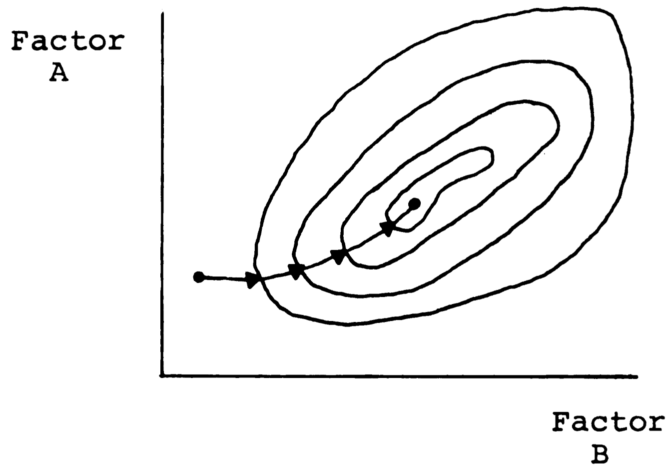


Figure 2.3.--The Box-Wilson Method of Steepest Ascent.

about the period or length of these deviations can be obtained.

Let $\{X_t, t \varepsilon T\}$ be a generating process or ensemble from which a sample time series $\{X_t, t=1,2, \ldots ,n\}$ is taken. Due to the stochastic nature of the system, analysis of $\{X_t\}$ cannot determine exactly the value of the series at any particular time, but the approximate structure of the generating process can be determined. This is done by obtaining estimates of the parameters which describe the generating process:

Mean of the Process	$\mu_t = E[X_t]$
Variance of the Process	$\sigma_x^2 = E[(X_t - \mu_t)^2]$
Autocovariance of the Process between observations at times t and s	$\gamma(t,s) = E[(X_t-\mu_t)(X_s-\mu_s)]$

These parameters can be estimated from M independent samples from $\{X_t\}$ i.e., $\{X_t^k, k=1,2, \ldots , M\}$. One great advantage of computer simulation is that in order to cut across the ensemble at a particular t in this fashion all that is required is an alteration in the value of the seed of the pseudorandom number generator. As an example, cut across the ensemble at $t = t_o$ in order to calculate the ensemble average estimating

$$\mu_{t_o} : \overline{X}_{t_o} = \frac{1}{M} \sum_{k=1}^{M} X_{t_o}^k .$$

Estimates of σ_x^2 and $\gamma(t,s)$ are obtained in the same way.

But spectral analysis is usually performed on time series which have first and second moments that are not a function of time.[19] There is no trend in the mean or variance of the series, and the autocovariance is a function of the time lag only. Such a series is called stationary. From a single time series can be obtained

$$\overline{X} = \frac{1}{n} \sum_{t=1}^{n} X_t$$

$$s^2 = \frac{1}{n} \sum_{t=1}^{n} (X_t - \overline{X})^2$$

and

$$C_t = \frac{1}{n-\tau} \sum_{t=1}^{n-\tau} (X_k - \overline{X})(X_{t+\tau} - \overline{X})$$

where $\gamma(o) = \sigma^2$ and $C_o = s^2$ which can be used as estimators for

$$E[X_t] = \mu$$

$$E[(X_t - \mu)^2] = \sigma^2$$

and

$$E[(X_t - \mu)(X_s - \mu)] = \gamma(t-s)$$

for all $t,s = \gamma_\tau$ where $\tau = t - s$.

The power spectrum is defined as the Fourier cosine transformation of the autocovariance

$$\phi(\omega) = \gamma_o + 2 \sum_{\tau=1}^{\infty} \gamma_t \cos(\omega t) \qquad o \leq \omega \leq \pi$$

The spectrum can be regarded as the "decomposition" of the variance of a time series. This is because the auto-covariance is recovered from the spectrum by the inverse transformation

$$\gamma_t = \frac{1}{\pi}\int_0^{\pi} \phi(\omega)\cos(\omega\tau)\,d\omega \qquad \tau = 0,1,2,\ldots$$

and in the special case when $\tau = 0$, γ_o is equal to the variance (σ^2). From the power spectrum is obtained the squared amplitude associated with oscillations at different frequencies ω, and the process is thus characterized in terms of independent additive contributions to the variance from each ω. So in order to obtain this information, an estimate of the power spectrum must be obtained. Estimators of power spectra usually have the form

$$f(\omega_j) = \lambda_o C_o + 2\sum_{\tau=1}^{m} \lambda_\tau C_\tau \cos(\omega_j \tau)$$

where $f(\omega_j)$ is an estimate of the power spectrum averaged over a band of frequencies centered at ω_j, and $\omega_j = \frac{\pi j}{m}$ $j = 0,1,2,\ldots,m$, λ_τ are weights, and m is the number of frequency bands to be estimated. The values of m and n should be selected with care in order to balance the conflicting requirements of resolution and statistical stability. Granger and Hatanaka[20] recommend a sample size of at least one hundred.

The spectrum is analyzed by plotting $f(\omega_j)$ against ω_j. Two important statistical properties are associated with the spectrum if X_t is normal. The first is that spectral estimates at nonadjacent frequencies are statistically independent. So confidence intervals can be used. The second is that if the control or theoretical spectrum $(\phi(\omega_j))$ is reasonably smooth, the distribution of $\frac{f(\omega_j)}{\phi(\omega_j)}$ is approximately $\frac{\chi^2{}_k}{k}$ with $K = \frac{2n}{m}$ degrees of freedom. With this knowledge, confidence intervals can be constructed around $\phi(\omega_j)$[21], the succession of which at frequency points ω_j $(j=0,1, \ldots , m)$ combine to form a confidence band. Now the question, does the spectrum for any plan under consideration lie within the confidence band of the control spectrum, can be answered.

An extension of this type of analysis is the comparison of two spectra. The ratio $P_j = \frac{\phi_1(\omega_j)}{\phi_2(\omega_j)}$ is used instead of $\frac{f(\omega_j)}{\phi(\omega_j)}$. Define R_j to be equal to $\frac{f_1(\omega_j)}{f_2(\omega_j)}$ and obtain the F statistic $F_{k_1,k_2} = \frac{R_j}{P_j}$ where $k_1 = k_2 = \frac{2n}{m}$ degrees of freedom. The 95% confidence interval for P_j is then $P(F_{.975,k_1,k_2} < \frac{R_j}{P_j} < F_{.025,k_1,k_2}) = .95$ and solving for P_j $P(\frac{R_j}{F_{.025,k_1,k_2}} < P_j < \frac{R_j}{F_{.975,k_1,k_2}}) = .95$ sets up the simultaneous confidence band $P(\frac{R_j}{F_{.001,k_1,k_2}} < P_j < \frac{R_j}{F_{.999,t_1,k_2}})$

= .95. If P = 1 lies within the desired simultaneous confidence band for P for all values of $o \leq \omega \leq \pi$, the hypothesis that the two spectra under consideration are not significantly different can be accepted.

Spectral analysis has been used to decompose the variance of a time series into its frequency components. A rather different application of the technique is to obtain an estimate of the variance as a whole for a given time series. Because of the autocorrelation S^2 does not have a Chi-square distribution with (n-1) degrees of freedom, but as σ^2 can be expressed in terms of ϕ, so can S^2 be expressed in terms of f. Blackman and Tukey[22] state that $S^2 = C_o = \frac{1}{m}\left[\frac{f(o)}{2} + \sum_{j=1}^{n-1} f(\omega_j) + \frac{f(n)}{2}\right]$ follows a Chi-square distribution with K degrees of freedom, where

$$K = \frac{\left[\frac{f(o)}{2} + \sum_{j=1}^{m-1} f(\omega_j) + \frac{f(\pi)}{2}\right]^2}{\frac{[f(o)]^2}{2} + \sum_{j=1}^{m-1} [f(\omega_j)]^2 + \frac{[f(\pi)]^2}{2}} \cdot \frac{n}{m}$$

For the comparison of two time series, the F statistic is $\frac{n_1 s_1{}^2/\sigma_1{}^2 k_1}{n_2 s_2{}^2/\sigma_2{}^2 k_2}$. A confidence interval can be set up for any desired level of significance about this statistic, and then statements about the two variances can be made after solving for $\frac{\sigma_1{}^2}{\sigma_2{}^2}$.

Spectral analysis is a significant method of analysis of the output of computer simulations because it does account for autocorrelation.

Correlation

Correlation theory can most easily be examined in terms of regression analysis. When all observations fall on the regression line developed from the data, perfect correlation exists between these variables. For two variables x and y, direct correlation exists if as y increases so also does x, while inverse correlation exists when x increases with a decrease in y. Perfect correlation occurs when both the amount and direction of change is identical for both variables, or the regression equation of x on y is identical to the regression of y on x.

The standard error of estimate is a measure of dispersion about the regression line. This statistic has the same properties as the standard deviation. The standard error of y on x is

$$S_{y.x} = \sqrt{\frac{(Y - Y_{est})^2}{N}} \; .$$

A good measure of linear correlation is the coefficient of correlation (r). The total variation in y can be expressed as the sum of the unexplained variation and the explained variation:

$$\text{Total Variation in } y = \Sigma(Y - \overline{Y})^2 = \Sigma(Y - Y_{est})^2 + \Sigma(Y_{est} - \overline{Y})^2.$$

From this expression r is developed as plus or minus the square root of the explained variation as a fraction of the total variation. An advantage of r is that it is dimensionless.

$$r = \pm \sqrt{\frac{\Sigma(Y_{est} - \overline{Y})^2}{\Sigma(Y - \overline{Y})^2}} \qquad -1 \leq r \leq 1$$

The linear correlation coefficient measures the departure of the regression lines for each variable from each other. The slope of the regression line of y on x is equal to the slope of the regression line of x on y only if r is equal to plus or minus one.

When considering time series, the degree of correlation between the present value of a variable and its value a fixed number of time periods prior to the present time is of concern. Correlation between members of a series (k-1) units apart is called autocorrelation of order k:

$$e_k = \frac{\text{COV }(\mu_t, \mu_{t+k})}{\sqrt{\{\text{VAR}\mu_t \text{ VAR}\mu_{t+k}\}}}.[23]$$

Non-linearity and multiple variables add computational complexity, but do not alter the logic of this type of analysis.

Factor Analysis

Correlation theory provides the basis for Factor Analysis (as it does for spectral analysis). Because the literature is extensive[24] and the method is of primary use when considering qualitative change, a detailed description will not be given.

Using the matrix of all correlations between the variables under consideration, the resolution of the set of variables linearly in terms of a small number of factors is possible. If this process is carried out satisfactorily, the factors will convey as much information about the system as did the original set of variables. The main aim of Factor Analysis then is to provide the most economical description of the observed data.

A given matrix of correlations can be factored in an infinite number of ways. Factor solutions are usually generated according to statistical considerations, such as attempting to account for a maximum amount of the total variance, or according to the meaningfulness of the solution to the particular experimental context. It should be emphasized that Factor Analysis does not produce an exhaustive set of fundamental factors which are a complete description.

Graphical Techniques

A graphical description of a time series has the advantage that it is easily developed. But this technique must be considered only in the sense that used alone it is better than no attempt at validation at all. Together with the preceeding forms of analyses, graphical measures provide a small marginal contribution to the analyst's confidence in the validity of his model. While the real value of this procedure is dubious, its high visual content does make it readily acceptable to general management.

Among the many possible graphical measures for comparing two time series are:[25] number, timing, and direction of turning points; amplitude of the fluctuations for corresponding time segments; average amplitude over the entire series; simultaneity of turning points for different variables; average values, probability distributions, and variation about the mean (variance, skewness, kurtosis) of variables; and exact matching of variables.

Application

The process of selecting the most suitable techniques for each of our three purposes and their application to simulation output will be described in Chapters V, VI, and VII.

CHAPTER II--FOOTNOTES

[1]For example: T. H. Naylor and J. M. Finger, "Verification of Computer Simulation Models," Management Science, Vol. 14 (October, 1967), pp. B92-101.

[2]K. Chu, Quantitative Methods For Business And Economic Analysis (Scranton, PA: International Textbook Co., 1969), p. 156.

[3]E. Paulson, "Sequential Estimation And Closed Sequential Decision Procedures," The Annals of Mathematical Statistics, Vol. 35 (September, 1964), pp. 1048-1058.

[4]R. E. Bechhofer and S. Blumenthal, "A Sequential Multiple-Decision Procedure for Selecting the Best One of Several Normal Populations with a Common Unknown Variance, II: Monte Carlo Sampling Results and New Computing Formulae," Biometrics, Vol. 18 (March, 1962), pp. 52-67.

[5]W. E. Sasser, D. S. Burdick, D. A. Graham, and T. H. Naylor, "The Application of Sequential Sampling to Simulation: An Example Inventory Model," Communications of the ACM, Vol. 13 (May, 1970), pp. 287-296.

[6]For an excellent treatment of regression analysis see: N. R. Draper and H. Smith, Applied Regression Analysis (New York: John Wiley and Sons, Inc., 1966).

[7]T. H. Naylor, K. Wertz, and T. H. Wonnacott, "Methods for Analyzing Data from Computer Simulation Experiments," Communications of the ACM, Vol. 10 (November, 1967), pp. 703-710.

[8]Ibid., p. 707.

[9]J. W. Tukey, "The Problem of Multiple Comparisons" (Princeton, N. J.: Dittoed Manuscript, Princeton University, 1965). Also in: H. Scheffe, The Analysis of Variance (New York: John Wiley and Sons, Inc., 1959).

[10]C. W. Dunnett, "A Multiple Comparison Procedure for Comparing Several Treatments With a Control," Journal of the American Statistical Association, Vol. 50 (1955), pp. 1096-1121.

[11] Naylor, Wertz and Wonnacott, p. 708.

[12] H. Theil, Applied Economic Forecasting (Amsterdam: The North-Holland Publishing Co., 1966), pp. 26-32.

[13] C. H. Kraft and C. VanEeden, A Nonparametric Introduction to Statistics (New York: The MacMillan Co., 1968), pp. 167-169.

[14] A. J. Duncan, Quality Control and Industrial Statistics (Homewood, Illinois: Richard D. Irwin, Inc., 1965), Chapter XXXVII.

[15] G. E. P. Box, "The Exploration and Exploitation of Response Surfaces: Some General Considerations and Examples," Biometrics, Vol. X (1954), pp. 16-60.

[16] G. E. P. Box and K. B. Wilson, "On the Experimental Attainment of Optimum Conditions," Journal of the Royal Statistical Society, Series B, Vol. XIII (1951), pp. 1-38.

[17] G. S. Fishman and P. J. Kiviat, Spectral Analysis of Time Series Generated by Simulation Models (Santa Monica, Cal.: The Rand Corporation, RM-4393-PR, 1965).

[18] T. H. Naylor, K. Wertz and T. H. Wonnacott, "Spectral Analysis of Data Generated by Simulation Experiments with Economic Models," Econometrica, Vol. 37 (April, 1969, pp. 333-352.

[19] If the time series is not covariance stationary, transformations must be made, such as polynomial regression to remove a trending mean, so as to obtain stationarity.

[20] C. W. J. Granger and M. Hatanaka, Spectral Analysis of Economic Time Series (Princeton, N. J.: Princeton University Press, 1964).

[21] For example, the 95% confidence interval for $\phi(\omega_j)$ is:

$$P\left(\frac{\chi^2_{.975,k}}{k} < \frac{f(\omega_i)}{\phi(\omega_j)} < \frac{\chi^2_{.025,k}}{k}\right) = .95$$

$$P\left(\frac{f(\omega_j)}{\chi^2_{.025,k/k}} < \phi(\omega_j) < \frac{f(\omega_j)}{\chi^2_{.975,k/k}}\right) = .95$$

[22] R. B. Blackman and J. W. Tukey, The Measurement of Power Spectra (New York: Dover Publications, Inc., 1958).

[23] G. U. Yule and M. G. Kendall, An Introduction to the Theory of Statistics (London: Charles Griffin and Co., Ltd., 1950), p. 638.

[24] A large bibliography is provided in: H. H. Harman, Modern Factor Analysis (Chicago: The University of Chicago Press, 1960).

[25] R. M. Cyert, "A Description and Evaluation of Some Firm Simulations," Proceedings of the IBM Scientific Computing Symposium on Simulation Models and Gaming (White Plains, N.Y.: IBM, 1966).

CHAPTER III

VALIDATION OF RECENT COMPUTER SIMULATION EXPERIMENTS

Introduction

As indicated in Chapter I, the analyst's approach to the question of the validity of the results of his simulation experiment is fundamentally determined by his basic point of view as to the aim and method of execution of his experiment. The type of model built which is a function of the analyst's outlook and training is a primary factor in the nature and extent of the validation procedure employed for the results of the model. This chapter will examine the procedures used to validate the results of some of the better known and better documented simulation experiments of the recent past.

Computer Models of the Shoe, Leather, Hide Sequence

Cohen (1960) constructed two simulation models to describe the aggregate behavior of shoe retailers, shoe manufacturers, and cattlehide leather tanners between 1930 and 1940. This aggregate behavior was described in terms of selling price, production or sales, and receipts. While the first model (Model II) was a "one-period-change" model determining values for these endogenous variables only one time period in advance, the second model (Model IIE) was a "process" model which determines endogenous variable values for an arbitrarily large number of future time periods. Cohen's model is discrete and dynamic with a time increment of one month.

Visual comparison of the time paths of the model predictions of selling price, production, and receipts with the actual historical time paths of these variables comprised the only validation of the model.

> The simulation runs for both Models II and IIE generate time paths for the endogenous variables which, although not in complete agreement with observed time paths, indicate that our models may incorporate some of the mechanisms which determine behavior in the shoe, leather, hide sequence.[1]
>
> Both models produce time paths which fluctuate around the observed time paths. For most variables, the amplitude of the oscillations is greater for Model IIE than for the actuals, with Model II having the largest amplitude. However, none of the time paths for either Model seem to be either explosive or overly damped.[2]
>
> The findings are also similar for average price. The time paths of both Models II and IIE are reasonably on course with observed values, although Model II shows even wider fluctuations about the actuals than for the preceding prices.[3]

Simulation of Information and Decision Systems in the Firm

Continuing a research effort started principally by Cyert and March,[4] Bonini (1963) constructed a computer model of the behavioral theory of the firm. In order to show the effects of organizational, informational, and environmental factors upon the firm's decision making process, Bonini decided that price, level of inventory, cost, sales, profit, and amount of pressure would be an adequate endogenous variable set to represent the behavior pattern of the organization. The model was used as an exploratory device to describe the relationship between various informational flow patterns and the firm's decision process. From these relationships design changes for the firm could be recommended.

Bonini was not concerned with modeling an actual firm. He was concerned with a comparison of the behavior of his theoretical firm after a proposed change with the original behavior. This comparison involved analyzing two sets of six time series (one time series for each of the variables price, level of inventory, cost, sales, profit, and amount of pressure before and after the proposed change). Because these time series did not exhibit any tendency to obtain steady-state or equilibrium values over time, Bonini settled for a measure of central tendency (the arithmetical mean), a measure of dispersion (the standard deviation), and a measure of trend (the least-squares regression coefficient) to describe the output time-series of his model.

Bonini determined the requirements for the length of these time series in the following fashion:

> On the one hand, the run should extend over sufficient simulated periods so that extreme values in the time series can be averaged out (that is, so there will be relatively small sampling error associated with the above three measures). On the other hand, limitations on computation time would argue for keeping a reasonably short number of periods. In addition, if we are going to apply our results to real organizations, we would be more interested in the immediate and short-run effects (of particular changes) than in what might be the average level over, say 20 or 30 years. In view of these considerations, I have chosen 108 time periods . . . as the length for the simulation runs.[5]

Portfolio Selection: A Simulation of Trust Investment

Clarkson (1962) developed a simulation model to duplicate the procedure by which a trust officer in a bank selected stock for any particular client's portfolio. The model combines a set of decision rules which are selected on the basis on information available about the client's financial situation and requirements.

The output of the model is not a data stream but a selection of a variable number of shares of a variable number of stocks, given the client's position. Clarkson applies two types of testing procedures to his model: those pertaining to the output of the model alone and those pertaining to the decision processes incorporated in the model.

For testing output Clarkson notes,

> Since the problem of determining the type of error when comparing generated to actual output has not yet

> been solved, statistical tests on the goodness of fit of the generated output are not very meaningful. The only statistical test that has much meaning is to test whether the generated data give a significantly 'better fit' than that which would be produced by some random or naive mechanisms.[6]

He tested the model against a "random selector" from the total population. Stocks were being selected at random without replacement from the list of total stocks available. This list contains M stocks of which W have been selected by the trust officer for the particular portfolio under consideration. Z is defined as the number of these stocks selected by the trust officer which occur in a sample of n stocks drawn at random from the list without replacement. Z is called the hypergeometric random variable.

$$Pz(k) = \frac{\binom{W}{k}\binom{M-W}{n-k}}{\binom{M}{n}} \qquad k = 0,1,2,\ .\ .\ .\ ,n$$

$$\text{where } \binom{b}{a} = 0 \text{ for } a>b$$

Clarkson rejected the hypothesis that this probability was equal to the percentage of matching or "correct" responses generated by the model. The size of the list was reduced to include only those issues which displayed the characteristics desired by the client, and the hypothesis was still rejected.

Naive decision rules replaced the random selection procedure, and the hypothesis was still rejected. The decision rules considered were:

1. Rank growth stocks on the basis of growth in price over the last 10 years.
2. Rank growth stocks on the basis of growth in earnings over the last 10 years.
3. Rank growth stocks on the basis of growth in sales over the last 10 years.
4. Rank growth stocks on the basis of growth low yield over the last 10 years.
5. Rank yield stocks on the basis of growth high yield over the last 10 years.

His objective, Clarkson contends, is to simulate investment behavior, to select the correct portfolios with the same processes and for the same reasons as the investment officer. Therefore, the need to test the decision processes exists. Turing's test[7] was used: Can an impartial observer discriminate between the output of the model of human behavior and the output of the actual human behavior?

Simulation of Market Processes

Balderston and Hoggatt (1962) constructed a computer simulation model of the West Coast lumber industry. The emphasis of the model is not to describe the real firms making up this industry, but to study the dynamic behavior of firms in a two-stage market from the viewpoint of an economic theorist. The model is driven by wholesalers to whom suppliers provide and from whom customers purchase. While flows of information, material, and money move vertically through the market, no horizontal movement is allowed. At the end of each market period decisions about output and price and entry and exit to the industry are made.

Concern for the validity of the model centered on the question of viability. Viability, as used by Balderston and Hoggatt, does not require equilibrium of the endogenous time paths, but only requires that "behavior should persist over a significant time interval."[8] Persistent behavior means that the time path is stable--stable in the sense that it settles into a state which exhibits properties of convergence or stable in the sense that change over time is steady with proportional (or acceptable) changes in the other endogenous variables.

This is the extent to which the original study considered the model's validity. Hoggatt in a later article[9] applied G. E. P. Box's[10] method of system analysis to the model. At this time more sophisticated validation techniques were introduced. Hoggatt states that he would consider the model valid if it "duplicated [the] trends and frequency response of [the] real system"[11] rather than aiming to have the model duplicate the time paths of the real system. In order to measure the frequency response of the model, he used the autocorrelation function.

Industrial Dynamics

Industrial Dynamics was developed by Forrester (1962) from his original dynamic simulation model of a firm's production-distribution system. Forrester has tried with limited success to convert his model building techniques into a general management philosophy. He describes Industrial Dynamics as

> the study of the information-feedback characteristics of industrial activity to show how organizational structure, amplification (in policies), and time delays (in decisions and actions) interact to influence the success of the enterprise. It treats the interactions between flows of information, money, orders, materials, personnel, and capital equipment in a company, and industry, or a national economy.
>
> Industrial Dynamics provides a single framework for integrating the functional areas of management--marketing, production, accounting, research and development and capital investment. It is a quantitative and experimental approach for relating organizational structure and corporate policy to industrial growth and stability.[12]

The greatest contribution of the Industrial Dynamics models was to point out the extraordinarily large fluctuations that can occur in the inventory held at the retail level when a change in customer demand is reflected through the lagged order delivery sequence: retailers-distributors-factory warehouse-factory-factory warehouse-distributors-retailers. From this basic production-distribution model many possible changes can be tested: limit factory capacity, eliminate the distributors, add additional sectors such as a market sector, include advertising.

How well the model serves its purpose is Forrester's test of its validity. The purpose of Industrial Dynamics is to design better management systems; therefore, validity can only be tested after an Industrial Dynamics approach has been applied to a situation and the results measured in some concrete terms such as increased profit. Defense of the model prior to use can only be given in terms of an individual defense of each detail of structure and policy so that in sum the total behavior of the model shows performance characteristics associated with the real system. The validity of the model at this stage as a description of a specific system can only be examined relative to the system boundaries (Are the boundaries suitable relative to the objectives of the experiment?), to the interacting variables, and to the values of the parameters. If the similarity of the model output to the actual characteristics of the system is not sufficient, these three factors must be examined and changed. These views on validity can be summarized in the following quotations:

> Validity as an abstract concept divorced from purpose, has no useful meaning.[13]

> The ability of a model to predict the state of the real system at some specific future time is not a sound test of model usefulness.[14]

> Data may serve to reject a grossly wrong decision-making hypothesis, but they can scarcely prove a correct one.[15]

Forrester believes the final test for validity is whether the actual system is being controlled to agree with the model.

Computer Simulation of Competitive Market Response

In order to define and analyze management problems involving the environment of the firm, Amstutz (1967) developed a simulation model of competitive market response.

The objective of the study was to model the firm and the environment external to the firm so that the total effect of changes in variables which can be controlled by management could be measured. Amstutz set up his system structure in terms of three sets of elements. Active elements are human. They can originate and react to signals. The eight active elements involved in the model are the producer, his competitors, distributors and wholesalers, salesmen, retailers, consumers, government officials and research workers. "Elements of flow are the vehicles of interaction between active elements."[16] These are the elements management can manipulate in order to try and achieve his objectives. The elements of flow are product, information and capital. The last set of elements are the passive elements (time delays, dissipators and storage) which describe the channels through which the flow elements move between the active elements. By means of this formulation the dynamic effects of the origination of a signal by management can be examined.

The tests Amstutz carried out in an attempt to analyze the worth of his model were of two types--reliability testing and validity testing.

The purpose of reliability testing is to determine if the results of the model are reproduceable. Are the results obtained on sequential runs sufficiently alike to justify the assumption that they are two samples drawn from the same population of data?

Validity testing is concerned with "truth." As there is no objective measure of truth, Amstutz argues that a subjective evaluation of the consistency of the model's performance with theory and prior knowledge must be made.

> Validity of a model can be established only by examining the realism of the assumptions on which it is based.[17]

Evaluation of the model's performance is possible using the Turing test. If a person knowledgeable in the area to be modeled cannot distinguish the model from the real system when provided with responses from both, then the model is realistic. Other tests for validity can be performed once the validity of the assumptions on which the model is based has been established.

Tests for Viability . . . This is a very gross test which is usually satisfied without explicit consideration. Does the model generate behavior which persists over a significant time interval?

Tests for Stability . . . Variables and processes which are stable in the real world must also exhibit stability when modeled.

Tests for Consistency . . . Consistency between model behavior and behavior observed in the real world. The extent to which the assumptions of the model agree with known facts must be tested as must the internal consistency or "deductive veracity" of the model--does the model "make sense." This testing may be done subjectively as "face validity" testing (does the model appear to be satisfactory), or analytically with sensitivity analysis.

Duplication of Historical Conditions . . . The fourth set of tests proposed by Amstutz.

Prediction of Future Conditions . . . The ability of the model to predict cannot be tested until after the passage of time over which the predictions were made unless "pseudo predictions" are made of past results.

Amstutz carried out these tests in the following manner. Reliability was tested by calculating "interrun deviations" when changing the seed in the random number generator. Subjectivity and "eyeball" testing confirmed viability, stability, and consistency requirements. To determine the extent to which the simulated exogenous time paths matched historical data the absolute error between simulated and actual was summed and averaged. The predictive ability of the model was not examined.

Model Classification

In order to summarize the views on validation of these seven model builders, it might prove instructive to classify their models. The models will be classified as discrete or continuous, positive or normative, and behavioral or physical. A discrete time model is structured using difference equations while a continuous time model is built with differential equations. A positive or descriptive model is one which attempts to replicate a real system. But no consideration is given as to the adequacy or value of this real system. A normative model attempts to produce the optimal conditions for the system under study. Explorative models generate solutions in search of this goal. Positive is to normative as "what is" is to "what ought to be." The last classification dichotomy is behavioral-physical. If any part of the model is an attempt to duplicate human behavior, the model is classified as behavioral, otherwise it is physical (see Table 3.1).

The next task is to use this classification scheme to determine if those who build the same type of model hold similar views as to the procedures by which their models can be validated.

TABLE 3.1.--Classification of Computer Simulation Experiments.

	Discrete	Continuous	Positive	Normative	Behavioral	Physical
Cohen	X		X		X	X
Bonini	X		X		X	
Clarkson	X		X		X	
Balderston and Hoggatt		X		X		X
Forrester	X			X		X
Amstutz	X		X			X
LREPS	X		X	X		X

Unifying Validation Concepts

From the study of these six models general concern is directed in varying degrees to two distinct types of validation--validation of the basic underlying processes of the model and validation of the data stream output of the model. Because the basic design and assumptions used in any model are certain to differ from those used in any other model, design validation procedures must of necessity be tailored to the particular model under consideration. This type of validation is probably best carried out by interactions between the model builders and those who are familiar with the real system being modeled both during and after construction of the model. After completion of the model, the Turing test can be used to increase confidence in the validity of the basic design. This type of model validity will be called design validity; validity of the output data stream will be called output validity.

This study will not consider design validity to any great extent for two reasons. First, as indicated, design validity is a concept specific to the particular model at hand; and second, if the model satisfies the requirements of output validity, it is not unreasonable to assume that the basic processes of the real system must have been modeled reasonably accurately. Friedman adds weight to the decision not to consider design validity. He believes that the validity of a theory is not based on the realism of its assumptions (complete "realism" is unattainable), but on the accuracy of its predictions.

Design validity is the point at which many normative model builders (in particular Forrester) stop. They argue that a normative model is not built to represent the actual system, but to represent the system the way it should be. Missing from this argument is a rational method of moving from the actual state to the desired state. A functional normative model might well be one which first models the actual system (at which point output validity testing can be carried out) and then the desired corrections are made from this basis.

The Cohen and Bonini models, and even the more recent Amstutz model, after a rather thorough description of validity testing, use subjective and basic statistical tests for validity. It is reasonable to conjecture that in general validation of currently built simulation models is not carried out at a much, if any, higher level of sophistication.

Balderston and Hoggatt's original analysis for validity is also rather limited and basic, although Hoggatt's later analysis is the most sophisticated of those employed in the models discussed.

Data produced from a strictly behavioral model such as Clarkson's is very limited. His analysis is quite adequate for the purpose of his model.

Two main points arise from this examination of some of the most well known simulation models. The first point is that regardless of the type of simulation used or the aims of the analyst, much of the activity that has to be carried out in order to validate the model is the same. The second point is the obvious need for the use of more extensive and more reliable techniques in the validation process.

CHAPTER III--FOOTNOTES

[1]K. J. Cohen, Computer Models of the Shoe, Leather, Hide Sequence (Englewood Cliffs, N. J.: Prentice-Hall, 1960), p. 60.

[2]Ibid., pp. 62-63.

[3]Ibid., p. 63.

[4]R. M. Cyert, E. A. Feigenbaum, and J. G. March, "Models of a Behavioral Theory of the Firm," Behavioral Science, Vol. 4, No. 2 (April, 1959), pp. 81-95.

[5]C. P. Bonini, Simulation of Information and Decision Systems in the Firm (Englewood Cliffs, N. J.: Prentice-Hall, 1962), p. 52.

[6]G. P. E. Clarkson, Portfolio Selection: A Simulation of Trust Investment (Englewood Cliffs, N. J.: Prentice-Hall, 1962), p. 55.

[7]A. M. Turing, "Can a Machine Think?" The World of Mathematics, ed. by J. R. Newman (New York: Simon and Schuster, 1956), pp. 2099-2123.

[8]F. E. Balderston and A. C. Hoggatt, Simulation of Market Processes (Berkeley, California: Institute of Business and Economic Research, 1962), p. 33.

[9]A. C. Hoggatt, "Statistical Techniques for the Computer Analysis of Simulation Models," Appendix in Studies in a Simulated Market. L. E. Preston and N. R. Collins (Berkeley, California: Institute of Business and Economic Research, 1966), pp. 92-122.

[10]G. E. P. Box and K. B. Wilson, "On the Experimental Attainment of Optimum Conditions," Journal of the Royal Statistical Society, B, XIII (1951), pp. 1-45.

[11]Hoggatt, p. 94.

[12]J. W. Forrester, Industrial Dynamics (Cambridge, Mass.: The M.I.T. Press, 1961), p. 13.

[13]*Ibid*., p. 115.

[14]*Ibid*., p. 115.

[15]*Ibid*., p. 118.

[16]A. E. Amstutz, *Computer Simulation of Competitive Market Response* (Cambridge, Mass.: The M.I.T. Press, 1967), p. 18.

[17]*Ibid*., p. 369.

CHAPTER IV

THE MODEL

Introduction

When large amounts of money and manpower have been applied to a project over an extended period of time, there is a natural reluctance (maybe not explicitly stated or felt) to subject the finished model to scrutiny, the result of which may indicate the worthlessness of the expenditures. Because our industrial sponsor did not discourage critical examination of the completed model, this dissertation is a formal analysis of the model's validity. Rather than narrow the focus to the validity of one specific model, validation of simulation models as a class will be examined with particular reference to this one model. A description of the long-range environmental planning simulator for a physical distribution system (LREPS) follows.

The Systems Approach

During the post-war period there has been an increasing use of quantitative analysis (usually discussed as operations research or management science methods) of industrial problems in order to supply an added dimension to the decision making process. Use of these techniques in physical distribution has on the whole been applied to isolated segments of the entire system.[1] Only recently has the firm's fixed facility network, transport capability, inventory allocations, communications, and unitization (material handling, packaging, containerization) procedures been conceptualized as an integrated physical distribution system.[2] Suboptimization can occur without an orientation toward an integrated system. For example, suppose a corporation is organized into four functional areas: purchasing, finance, manufacturing, and sales. The responsibility for physical distribution activities is allocated as follows: inbound materials under purchasing, branch plant shipments and order processing under finance, traffic and shipping under manufacturing, and inventory control and public warehousing under sales. If planning is not carried on from the point of view of the corporation as a system, suboptimization might occur if purchasing determined the quantity of raw materials required solely on the basis of price per unit. This would probably mean large inbound shipments and non-optimal raw material inventory due to high storage costs. Many other situations can occur where the optimal action for a particular corporate functional area is suboptimal for the company as a whole. Recognition of the possibility of this type of suboptimization has led to the establishment of integrated physical distribution systems by many corporations.

The argument for integration using the systems concept could be extended. Why not integrate the functions of the firm? Why not integrate firms into a model of the economy? Given the capacity limitations of the present generation of computing machinery, the trade-off exists between cost benefits from the "systems effect" and loss of ability to represent the system components accurately in the required detail. At the desired level of detail a great deal of effort had to be expended in order to ensure that the size of LREPS did not exceed the capacity of the available computing machinery. Integration beyond the level of the physical distribution system would have required a lower level of model refinement. But the systems concept is a vital development which will be extended with future technological advances.

Model Structure

The actual physical distribution system is modeled in terms of the general structure given in Figures 4.1 and 4.2. The five basic components of an integrated physical distribution system (the fixed facility network, transport capability, inventory allocation, communication, and unitization) are evaluated at three stages in the channel structure. These three stages are:

1. The manufacturing control center (MCC) which produces a partial product line and distributes

PHYSICAL DISTRIBUTION SYSTEM

MANUFACTURING CONTROL CENTERS (MCC)
MULTI-LOCATION
EACH PRODUCES LESS THAN FULL LINE
EACH PRODUCT IS PRODUCED AT MORE THAN ONE MCC

REPLENISHMENT CENTERS (RC)
MULTI-LOCATION
EACH STOCKS ALL PRODUCTS MANUFACTURED AT MCC

DISTRIBUTION CENTERS (PDC) (RDC)
MULTI-LOCATION
FULL LINE - PRIMARY DC (PDC)
FULL OR PARTIAL LINE - REMOTE DC (RDC)

CONSOLIDATED SHIPPING POINT (CSP)

TRANSPORTATION
COMMON CARRIER - TRUCK, RAIL, AIR

INVENTORY
STOCKS AT RC, PDC, RDC

COMMUNICATIONS
COMPUTER, TELETYPE, MAIL, TELEPHONE

UNITIZATION
AUTOMATED OR MANUAL

PRODUCT PROFILE

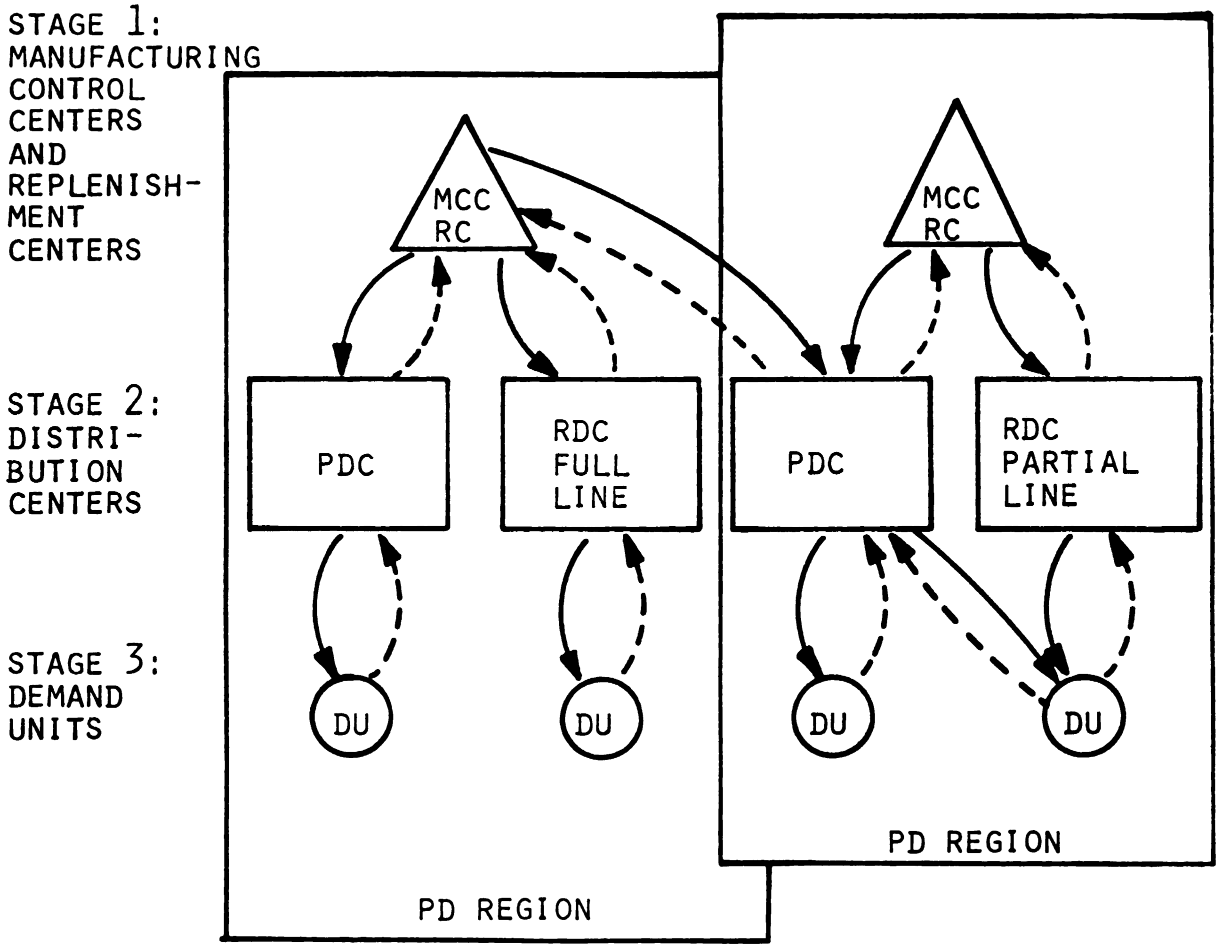
MULTI-PRODUCT LINE
KEY PRODUCT GROUPS FOR EACH CUSTOMER CLASS OF TRADE

MARKET PROFILE
MULTI-CUSTOMER CLASSES OF TRADE
TOTAL U.S. MARKET

COMPETITIVE PROFILE
MULTI-COMPETITORS

Figure 4.1.--General Description of Firm-Distribution Audit.[1]

[1]D. J. Bowersox, *et al.*, *Dynamic Simulation of Physical Distribution Systems*, Monograph (East Lansing, Michigan: Division of Research, Michigan State University, Forthcoming).



------INFORMATION FLOW ——————PRODUCT FLOW

REGION..THE REGION IS DEFINED BY THE ASSIGNMENT OF RDCS AND DUS TO A PDC.
MCC.....EACH MANUFACTURING CENTER PRODUCES A PARTIAL LINE.
RC......REPLENISHMENT CENTERS STOCK ONLY PRODUCTS MANUFACTURED AT COINCIDENT MCC.
RDC.....REMOTE DISTRIBUTION CENTER, FULL OR PARTIAL LINE.
PDC.....PRIMARY DISTRIBUTION CENTER, EACH PDC IS FULL LINE AND SUPPLIES ALL PRODUCTS TO DUS ASSIGNED TO THE PDC REGION; PRODUCT CATEGORIES NOT STOCKED AT THE PARTIAL LINE RDCS IN THE REGION ARE ALSO SHIPPED BY THE PDC.
DU......THE DEMAND UNIT CONSISTS OF ZIP SECTIONAL CENTER(S).
CSP.....CONSOLIDATED SHIPPING POINT.

Figure 4.2.--Stages of the Physical Distribution Network.[1]

[1]D. J. Bowersox, et al., Dynamic Simulation of Physical Distribution Systems, Monograph (East Lansing, Michigan: Division of Research, Michigan State University, Forthcoming).

these products from the adjoining replenishment center (RC).

2. The distribution center (DC) which provides a product selection at a location from which customer service requirements can be satisfied.

3. The demand unit (DU) which is an individual customer's demand or the agglomeration of several customers' demands.

The items manufactured at the MCC move to the customer through the distribution centers. Four different types of distribution center exist at the DC stage. Primary distribution centers (PDC) handle a full line of the firm's products and have the potential to serve all the demand units in a defined region of the total market area. Remote distribution centers full line (RDC-F) also handle all of the firm's products, but service only a pre-assigned subset of the DU's within the PDC market region. A remote distribution center which handles only a fraction of the firm's total product line is called a remote distribution center partial line (RDC-P). The last type of DC is the consolidated shipping point (CSP) which is an RDC-P which handles no products, but functions as a point at which the demand of several DU's is agglomerated and served from a PDC. The PDC's are capable of serving the same demand units as an RDC-P, but cannot serve the demand units affiliated with an RDC-F.

This model structure presents the physical distribution system at an integrated level, a level which allows the accumulation of information pertinent to the particular project in progress, but which also allows the same model (with minor modification) to be used in a wide range of other applications.

Consideration of the physical distribution system as an integrated unit offers management financial advantages. Can the operations research techniques used to analyze the elements of the system be extended to the system in its entirety? Usually not. The interaction between the elements of the system normally introduces a degree of complexity such that analytical procedures cannot be used. Fortunately numerical procedures exist which provide a method for studying this class of larger, more complex, problems. Such a numerical procedure is simulation. Simulation as a tool is less accurate and more costly than an analytical technique, but it is feasible.

As a design specification of the project was for a ten year time horizon, the model must be dynamic--dynamic because information is required of the system at all points along the time horizon, not just the end. The effect of a decision at time n is dependent upon the timing and nature of the decisions made prior to time n. LREPS has the facility to change over time both the endogenous variables, using internal feedback mechanisms and the exogenous variables, which represent the system's environment.

A dynamic simulation model is desired which will analyze the cost and service trade-offs between the elements or subsystems of the physical distribution system caused by any given sequence of decisions made over a long-range planning horizon.

The two main aspects which set the model apart from previous studies are the consideration of both spatial and temporal dimensions of the physical distribution system in one model and the concept of flexibility. The description of the model subsystems to follow will indicate the method of including both spatial and temporal considerations. Due to the stochastic nature of the system being modeled, several acceptable outcomes are possible from a given managerial decision. The flexibility of one particular outcome is the degree to which it is representative of the whole range of acceptable outcomes.

Subsystem Detail

The model[3] is constructed in three main parts: The Data Support Subsystem, the four subsystems which comprise the actual operating model (the Demand and Environment Subsystem, the Operations Subsystem, the Measurement Subsystem, and the Monitor and Control Subsystem), and the Report Generator Subsystem. This structure is shown in Figure 4.3.

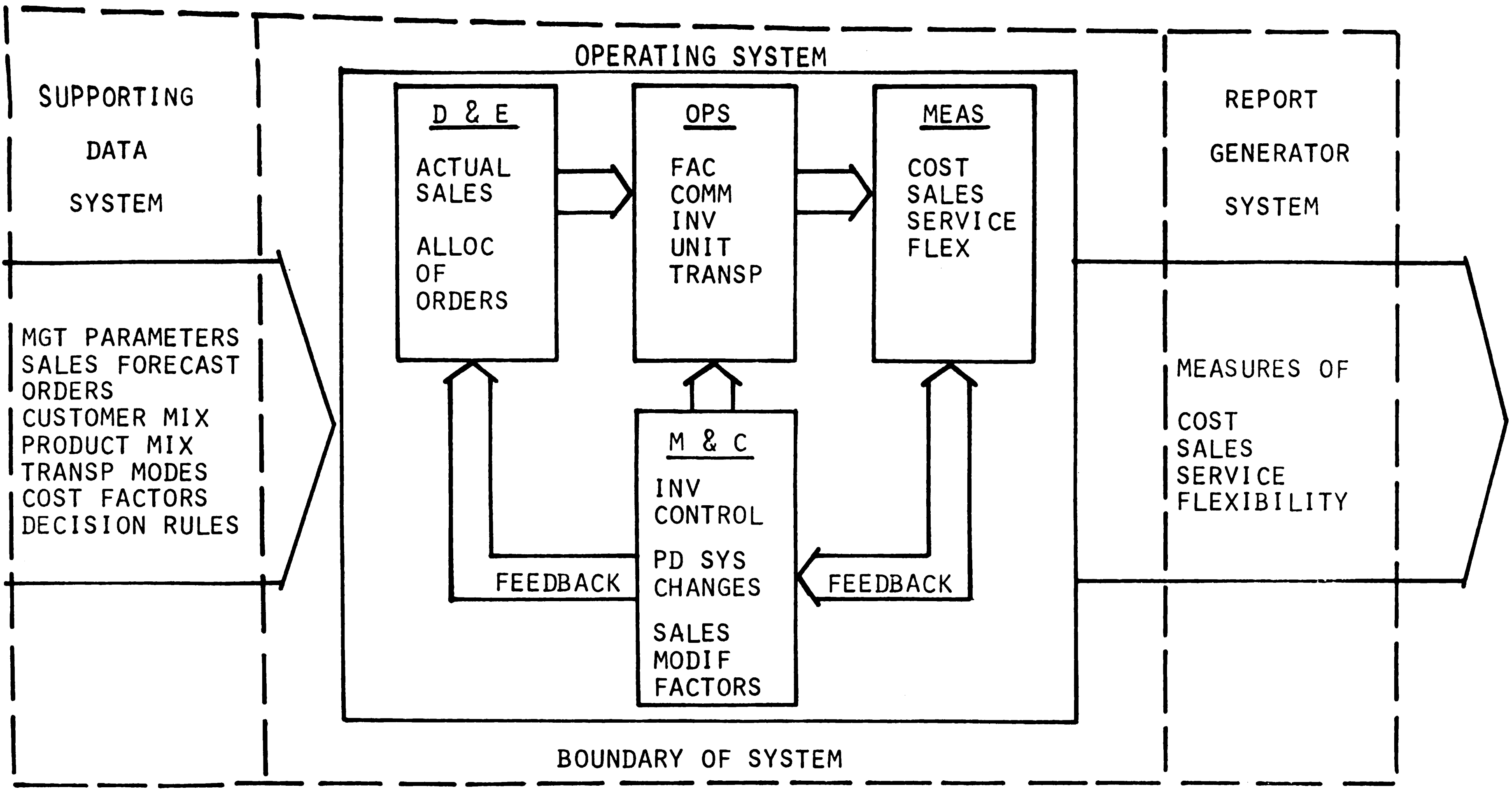


Figure 4.3.--LREPS Systems Model Concept.[1]

[1]D. J. Bowersox, et al., Dynamic Simulation of Physical Distribution Systems, Monograph (East Lansing, Michigan: Division of Research, Michigan State University, Forthcoming).

The Data Support Subsystem generates the input tape for the model. Contained on this tape are the constant exogenous variables for a particular experiment using the model and also the amount and timing over the ten year planning horizon of changes in controllable variables. The controllable variables are order characteristics, product mix, new products, customer mix, facility network, inventory policy, transportation, communications and unitization.

The second main segment of the model contains a mathematical representation (difference equations) of demand generation and allocation, the driver of the model, and the five elements of the physical distribution system: transportation, inventory control, facility location, unitization, and communications.

The Demand and Environment Subsystem subdivides the national sales forecast to the individual demand units, generates actual customer orders by product, allocates these orders to the demand units, and assigns a distribution center to service each demand unit. To avoid dealing with individual customers, demand was summarized by Zip Sectional Center. The product orders representing this demand were drawn in blocks at random from the order matrix until the demand unit's daily sales forecast[4] was satisfied. Blocks on the order matrix contain orders for a stratified sample of fifty products, or about 12% of the total product line. These orders can be constructed to be representative of historical conditions, or "pseudo orders" can be generated. Testing new product lines, changing demand patterns, or observing the dynamics of alternative inventory policies is possible by generating "pseudo orders" with the desired characteristics. Finally, demand units are assigned to distribution centers according to one of these decisions rules: minimum distance, minimum transit time, minimum transportation cost or a heuristic combination of these three factors.

The Operations Subsystem uses the information supplied by Demand and Environment and processes the product and information flows through the physical distribution system. Orders arrive each day at the distribution centers from the demand units. If inventory on hand is sufficient to meet this demand, the order is prepared and shipment is made, but if inventory on hand is not sufficient, a backorder is created, and at the time indicated by the inventory policy in use, an order is sent to the replenishment center. This transmittal time for the order, together with order processing and preparation time, the delay to the next scheduled shipping time, and the transit time to the distribution center, make up the reorder cycle. The average customer order cycle time, a measure of the system's service capability, can then be calculated as the total of customer order transmittal time, customer order processing and preparation time, the mean reorder cycle time, and the customer transit time. One of three inventory policies trigger the reorder cycle--a daily reorder point system, an optional replenishment system or a hybrid combination of these two. Communication policies can be tested by varying the distribution from which the transmittal time is selected. An order system based on mail, for example, would be represented by a distribution of order transmittal times with a larger mean and variance than would an order system using a teletype.

The Measurement Subsystem develops cost, service, and flexibility measures of the activity levels of the Operations Subsystem. Fixed facility investment cost, transportation cost, communications cost, average inventory carrying cost, reorder cost, and throughput or unitization cost per distribution center are summed to the total cost associated with the physical distribution system. The annual fixed facility investment cost is obtained by depreciating the dollar investment for the facility over its functional life span. The dollar investment is assumed to be constant for a given size and type of facility. To determine transportation costs, both inbound from the replenishment center to the distribution center and outbound from the distribution center to the demand units, the appropriate freight rate for the distance is multiplied by the weight. The freight rates were determined by regression analysis in order to account for such factors as freight class, weight breaks, regional differences, negotiated rates and average shipment size. The number of orders and lines processed are used to determine communication costs for each network link and each facility size, again by regression analysis. Inventory costs (average carrying cost and reorder cost) are determined for a sample product category and then extrapolated up by the appropriate sample to product line ratio. Average throughput costs per unit of volume moved through distribution centers of each size and type have been calculated. Throughput cost for the distribution center is then volume times the appropriate cost per unit.

Also calculated in the Measurement Subsystem are such service characteristics as the number of stockouts, total order cycle time, and the percentage of demand satisfied within a specified number of days' transit time.

The Monitor and Control Subsystem provides an alternative to specifying all changes in controllable variables in the Data Support Subsystem prior to the actual running of the model. In Monitor and Control, desired and actual levels of cost, service, and flexibility are compared at specified stages over the time horizon, and modifications are made automatically to the physical distribution system on the basis of the size of the variance. The modification might take the form of an expansion, addition or deletion of physical facilities for future periods or it might be an alteration of the sales forecasts for future periods.

The final main segment of the model is the Report Generator Subsystem which organizes the output data of the model into management reports.

Validation

Effort to validate a computer simulation model can be directed in two ways--to validate the design or method of construction of the model and to validate the output of the model. As indicated in Chapter III, too much emphasis has been placed on design validity in the past. This dissertation will concentrate on methods to establish the output validity of computer simulation models in general, and in particular the LREPS model.

Given this emphasis, it is still important to recognize the need to test for design validity during the process of constructing the model and as an initial procedure upon its completion. This testing involves checking the functioning of the model and its components for reasonableness. Do the values of the endogenous variables fall within acceptable limits? This procedure is sometimes known as determining the model's face validity, that is, determining the extent to which the assumptions of the model agree with known facts and also the internal consistency or "deductive veracity" of the model. In other words, the model must "make sense." Table 4.1 contains the face validity analysis for LREPS. A comparison of simulated versus actual data for an information category is designated "within limits" if the variance is less than 5%.

The third output validation procedure proposed is to examine the sensitivity of the major assumption employed by the model. To the extent of the analysis of data streams before and after a change in these assumptions, this is output validity. But the determination of the particular assumptions to be examined is a problem of design validity.

Gross malfunctions of a particular model can be discovered by analysis for face validity or design validity. Once the model has satisfied these criteria, the more general and sophisticated procedures for establishing output validity can be applied. These methods as applied to the LREPS model are now briefly discussed (the next three chapters take up each of the methods in greater detail).

Data streams for several endogenous variables need to be generated by the model over an extended time period. This is so the stability or viability of the model over the long run can be established. Do the data streams examined show persistent behavior over this time interval?

TABLE 4.1.--LREPS Face Validity.

Information Category	Simulated Versus Actual	PD Stages
Cust Sales	Within Limits	DU, DC and Domestic
Cust Dollar Sales/Order	Within Limits	DC and Domestic
Cust Wt Sales/Order	Within Limits	DC and Domestic
Line Items per Order	Within Limits	DC and Domestic
Cust Serv--		
NOCT-Avg	Within Limits	DC and Domestic
NOCT-Std Dev	No Data Avail.	DC and Domestic
T4-Avg	Within Limits	DC and Domestic
T4-Std Dev	No Data Avail.	DC and Domestic
Dollar-Preps	No Data Avail.	DC only
Order Preps	Within Limits	DC only
DC-MCC Reorders	Within Limits	DC only
DC Stockouts	No Data Avail.	DC only
DC Avg IOH	Within Limits	DC only
Cust ship Accums	Difficult to Compare Because of Small Sample Averages in Cust Order Blocks	
MCC Ship Accums	Within Limits	MCC only
Total Product Demand	Within Limits	Domestic only
Total PD Cost--	Within Limits	DC and Domestic
Facilities	Within Limits	DC and Domestic
Transportation		
Inbound	Within Limits	DC and Domestic
Outbound	Within Limits	DC and Domestic
Inventory	Within Limits	DC and Domestic
Communications	Within Limits	DC and Domestic
Throughput	Within Limits	DC and Domestic
Cum Wt Indicies	Within Limits	DU, DC and Regional

The second validation procedure requires a measure of the extent to which the model is an accurate representation of the real system. Time paths of selected endogenous variables, which are representative of the physical distribution system's behavior, will be generated by the model over a past time period. Statistical analysis of this data with actual historical data over the same time period will provide the required measure.

Two critical building blocks in the model are the use of a stratified sample of fifty products to represent the total product line and the method of generating demand unit orders. The model should be constructed so that reasonable changes in these two procedures do not have a significant effect on the model output. To carry out this third validation procedure, analysis of selected endogenous data streams before and after the change will be required. An example of such a change is the alteration of the composition or size of the stratified sample.

As indicated, the methods used, and the results obtained, with these three types of analysis will be examined in later chapters.

CHAPTER IV--FOOTNOTES

[1]For example:

Transportation--

W. H. Hausman and P. Gilmour, "A Multi-Period Truck Delivery Problem," Transportation Research, Vol. 1, No. 4 (December, 1967), pp. 349-357.

Warehousing--

A. A. Kuehn and M. J. Hamburger, "A Heuristic Program for Locating Warehouses," Management Science, Vol. 9, No. 11 (July, 1963), pp. 643-666.

Inventory--

A. F. Veinott, "The Status of Mathematical Inventory Theory," Management Science, Vol. 12, No. 11 (July, 1966), pp. 745-777. (This article includes an extensive bibliography.)

[2]D. J. Bowersox, E. W. Smykay, and B. H. LaLonde, Physical Distribution Management (New York: The Macmillan Company, 1968), Chapter 5.

[3]A more detailed description of the model can be obtained from the monograph "Development of a Dynamic Simulation Model for Planning Physical Distribution Systems: Formulation of the Conceptual Approach and Research Design" which is in process at the Graduate School of Business Administration, Michigan State University.

[4]The daily sales forecast for the demand unit is a function of population, retail sales, personal income and effective buying power associated with the Zip Sectional Center.

CHAPTER V

STABILITY OF THE MODEL

Introduction

The first aspect of validity to be subjected to detailed analysis is long-term stability. Stability is the ability of the model to generate endogenous data streams which show persistent behavior over the long run. Over this time period the data streams will exhibit convergence properties or the rate of change of each endogenous variable being examined will be proportional to or acceptable to the rate of change in all other endogenous variables. The ten-year planning horizon of LREPS is considered "long-term."

This type of analysis follows naturally the establishment of the model's face validity. While face validity is a statement of the model's reasonableness over the short run (preliminary runs of any model are usually not for the entire planning horizon), the analysis of this chapter is a statement of the model's reasonableness over the long run.

Endogenous data streams of sales weights for the three products are examined. This analysis is carried out in two ways. The first way is to study the time series or data stream and then make statements as to the reasonableness of its variability over the time horizon. Spectral analysis is used for this purpose. The second type of analysis is to lag the original time series by k units and then compare this lagged time series with the original set of observations. This comparison should indicate a reasonable correspondence between the two data streams. Given this particular analysis a 10 unit lag was selected, as a large proportion of the variance of the time series could be expected to occur over a two week period.

Graphical Analysis

Gross instability of the endogenous data stream under consideration is indicated rather clearly when the data is graphed. But it must be pointed out that the amount of variability contained in the data can appear to increase or decline with a contraction or expansion of the range of the ordinate. Figure 5.1 is the graph of sales weight for each of the three products over a ten-year period (Product 1 is plotted with "+'s," Product 2 with octagons, and Product 3 with triangles). No inordinate amount of fluctuation is observable from this graph.

Parameters of the data streams are of relatively little value because of the averaging effect over a large number of observations and also because a comparison of two different data streams is not being made. Recognizing this fact, the means, variances, skewness, and kurtosis of

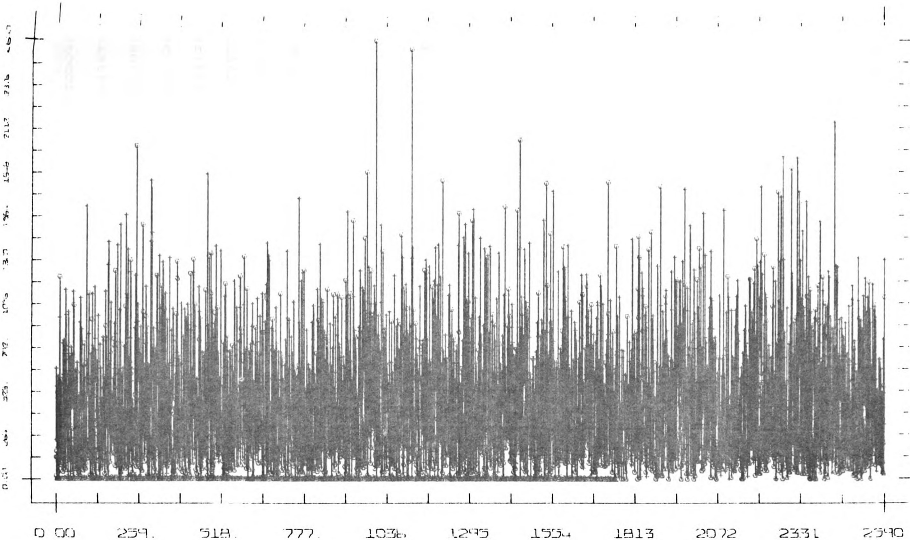


Figure 5.1.--Sales Weight--Three Products for 10 Years.

the three data streams are given in Table 5.1. The means and variances are of limited value as absolute quantities. A normal value for kurtosis is 3, and the symmetry of a symmetrical distribution is 1. The distribution for Product 1 is remarkably symmetric. The distributions of the other two products are nonsymmetric and leptokurtic ("humped" to a degree greater than normal).

TABLE 5.1.--Means, Variances, Skewness, and Kurtosis.

	Sales Weight		
	Product 1	Product 2	Product 3
Mean	530.95	326.61	1.78
Variance	100162.81	89553.83	13.79
Skewness	1.00	2.05	2.81
Kurtosis	1.23	6.37	9.40

Correlation

The amount of correlation between a time series and the same time series with observations lagged by k units is of interest. This can be shown by the coefficient of determination (r^2) which expresses the percentage of the total variation in the original variable which is "explained" by the regression line of this variable on the lagged variable. Also conveying the same type of information is the autocorrelation of a time series at time t and at time (t + k). The autocorrelation of order k is given by

$$e_k = \frac{\mathrm{COV}(\mu_t, \mu_{t+k})}{\sqrt{(\mathrm{VAR}\ \mu_t)(\mathrm{VAR}\ \mu_{t+k})}} .$$

The first task is to examine the coefficient of correlation (r). The range of this coefficient is from -1 to +1, or from perfect negative correlation to perfect positive correlation. The values of r for original data on lagged data are given in Table 5.2 as well as the results of the null hypothesis that r is significantly different from zero. In order to accept the null hypothesis with 95% confidence, r must be greater than 0.197.[1] The hypothesis is rejected for Product 3. This product is a slow mover, and so the variation in sales weight between a given time and a time two weeks later could be considerable (for example a positive sales weight against no sale or zero sales weight). So this result appears reasonable.

TABLE 5.2.--Test of Correlation Coefficients.

	Sales Weight	
	r	Ho
Product 1	0.6162	Accept
Product 2	0.5421	Accept
Product 3	0.1075	Reject

The values of the coefficient of determination are given in Table 5.3. A moderate amount of the total

TABLE 5.3.--Coefficients of Determination.

	Sales Weight
Product 1	0.3797
Product 2	0.2939
Product 3	0.0116

variation in the original data for Products 2 and 3 is explained by the lagged data--enough to suggest the absence of instability over two-week periods.

Usually the presence of autocorrelation is a burden to the analyst of time series. But for the present purpose, autocorrelation indicates an inherent relationship between observations in the time series at point n and those at point (n + k). The existence of such a relationship limits the susceptibility of the time series to excessive fluctuation. The autocorrelations of order (k = 10) for the three data streams are listed in Table 5.4.

Theil's Inequality Coefficient

The quality of predicted results, given the availability of the actual outcomes, is measured by Theil's inequality coefficient. If the coefficient is zero, the forecasts are perfect; and if the coefficient has a value of one, it means that the forecasting method has generated results no better than those obtained by no-change

TABLE 5.4.--Autocorrelation.

	Sales Weight
Product 1	0.0044
Product 2	-0.0199
Product 3	0.0394

extrapolation. The inequality coefficient has no finite upper bound.

Forecasting outcomes to be equal to those which occurred two weeks previously is not good forecasting technique, and the results of this test are not expected to be good. But if the inequality coefficient has a value close to one, it means that the variation occurring in the time series over a two-week period is minimal and also that movement within the series is gradual. The coefficients given in Table 5.5 show that this is indeed so. As expected, the covariance proportion accounts for all of the disparity between forecast and actual (Table 5.6).

TABLE 5.5.--Test of Predictive Quality.

	Sales Weight
Product 1	0.7222
Product 2	0.9609
Product 3	1.1886

TABLE 5.6.--Inequality Proportions.

	Sales Weight		
	Bias	Variance	Covariance
Product 1	0.0000	0.0000	1.0000
Product 2	0.0000	0.0000	1.0000
Product 3	0.0000	0.0000	1.0000

Spectral Analysis

The techniques discussed up to this point in the chapter have been applied to analyze the relationship of the observations in a time series at point t with observations in the same time series at point (t + k). The other form of testing for long-term stability is to inspect the variability contained in the original data stream. Examination of the power spectrum of this data stream allows the determination of the extent to which particular frequency bands contribute to the total variance. If the graph of the logarithm of the power spectrum does not violate Granger and Hatanaka's[2] simulataneous confidence interval at some specified confidence level, then the original time series can be said to exhibit stability for that time period.

Figure 5.2 is a graph of the logarithm of the power spectrum of 2590 observations of the sales weight for Product 1 against 120 frequency levels. Figures 5.3 and 5.4 are similarly graphs for Products 2 and 3 respectively.

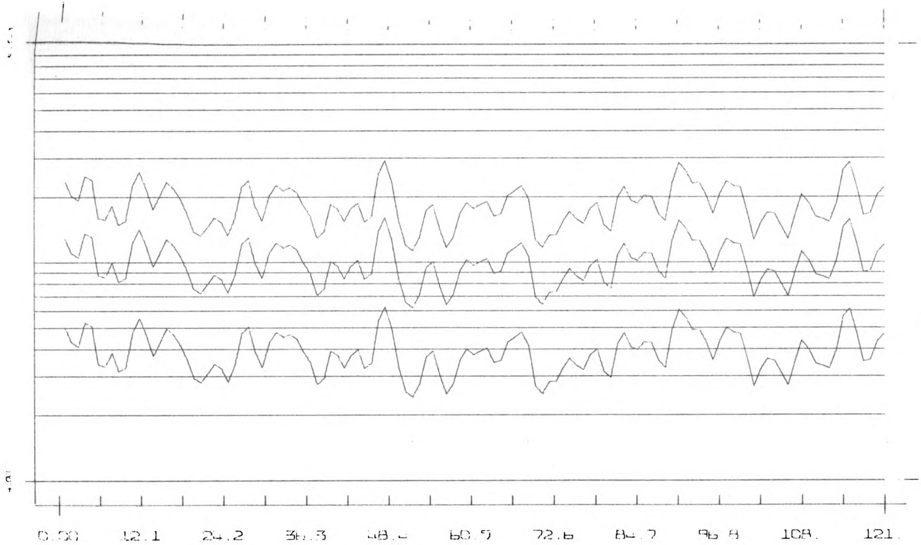


Figure 5.2.--Estimated Power Spectrum--Sales Weight for Product 1.

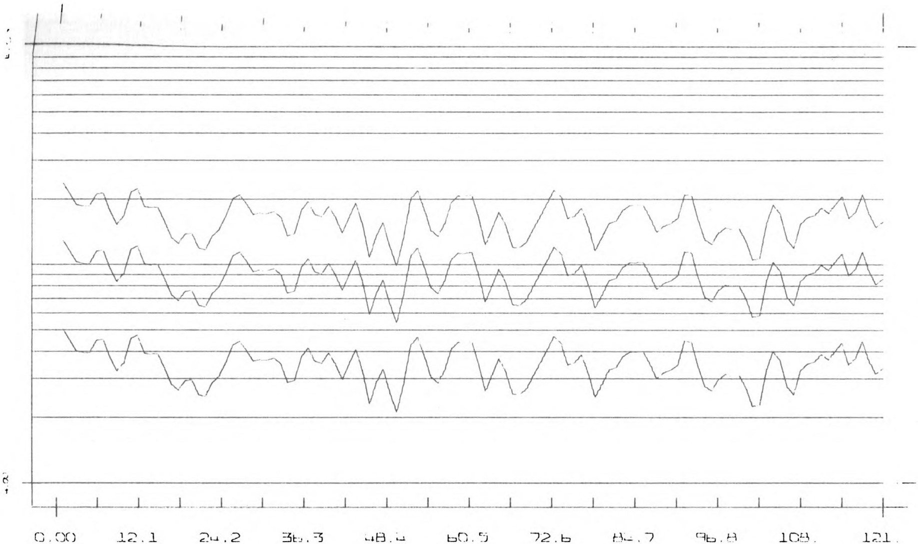


Figure 5.3.--Estimated Power Spectrum--Sales Weight for Product 2.

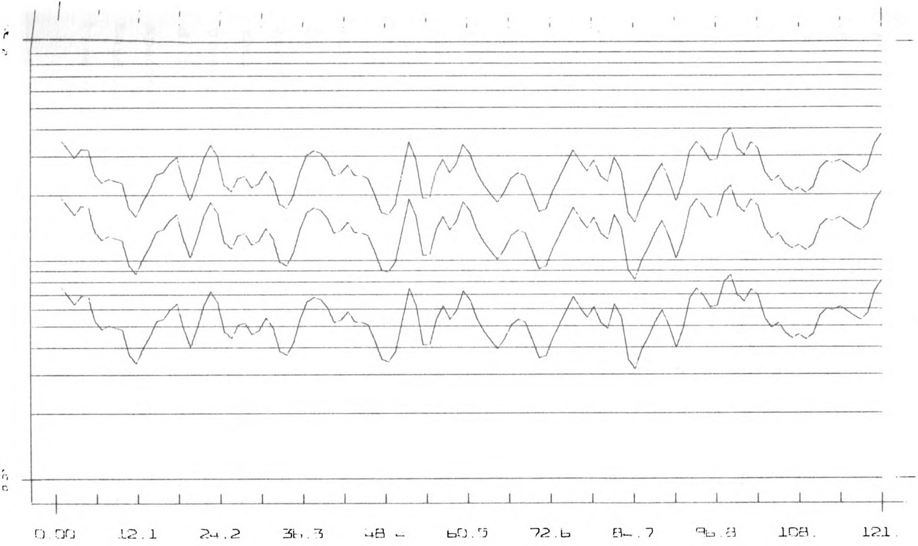


Figure 5.4.--Estimated Power Spectrum--Sales Weight for Product 3.

Stability is indicated in all three cases by the fact that a smooth curve could easily be drawn between the confidence limits. Another method of analysis is possible if the confidence intervals are constructed, not from the basis of the power spectrum itself, but from a smooth line of best fit for the power spectrum. In this case the power spectrum will violate the confidence intervals if long-term stability does not exist.

Stability of the Model

The analysis of a solitary data stream is more difficult than the analysis of the differences and similarities between two or more data streams. Fewer statistical techniques can be used, and even some which have been used generate information of dubious value.

Two main avenues are followed in the analysis of this chapter. The first is to examine the relationship between observations within the same time series separated by a particular time increment. If this relationship is strong (the series is relatively highly autocorrelated), then the possibility of the series' being unstable is greatly diminished. The other avenue is to examine several different frequency components of the time series (using spectral analysis) and establish that no one frequency band contributes in excess to the overall variance of the time series.

This analysis considered time series which are 2600 observations long. Detailed statistical analysis of even a few variables of this length consumes rather large amounts of computer time. As with all types of analysis, a point is reached where the value of additional information does not justify the costs involved in obtaining it. This makes the selection of the variables to study an important decision. Sales weight for a high volume product, a medium volume product, and a low volume product were selected as the variables to study because it was felt that these variables will reflect in general the total model operation.

The results of this chapter must be interpreted to conclude that the model does generate persistent endogenous behavior and is stable over the long run.

CHAPTER V--FOOTNOTES

[1]J. Riggs, Production Systems: Planning, Analysis and Control (New York: John Wiley and Sons, Inc., 1970), p. 70.

[2]C. W. J. Granger and M. Hatanaka, Spectral Analysis of Economic Time Series (Princeton, N.J.: Princeton University Press, 1964), p. 62.

CHAPTER VI

THE MODEL'S PREDICTIVE ABILITY

Introduction

The second major validation task is to compare the output of the simulation model for some past time period with the actual historical data that was recorded for that time period. This type of analysis comes most readily to mind when considering validation. Accountants, for example, place a great deal of emphasis on the analysis of the difference between actual figures and expected or forecast figures.

Several methods of comparing simulated endogenous data streams with the actual data streams are presented in this chapter. While the results of these statistical tests are given here, detailed evaluation is contained in Chapter VIII.

The results of this type of validation testing are dependent upon the quality and length of the actual data streams. The quality of the data is a function of the organization's accounting system and information transmission capability. Because of the random component deliberately included in a computer simulation model, the actual data stream must be of sufficient length for variables to approach the distributions and parameters modeled. Data collected on a weekly basis will be more likely to average out the vagaries of the accounting and information systems than data collected on a daily basis. It would seem reasonable, then, that a shorter data stream of weekly data would provide statistical information of similar quality to a longer data stream of daily data. The same case can obviously be made for information collected on a monthly basis against information collected weekly.

The industrial sponsor of the LREPS project was able to supply actual historical data for three products from the stratified sample. Dollar sales, sales weight, and inventory on hand for these three products was supplied for one region on a daily basis for a period of 103 days. Information for a longer period and on a weekly basis was requested, but was not available. The premonition was that the quality and length of these data streams were unacceptable. If the tests of this chapter are not satisfied, the next task must be to continue to accumulate more extensive historical information and conduct the tests again. As the industrial sponsor cannot obtain the required data prior to the first day of the 103 day's information on hand, the rerunning of the tests of this chapter would have to be delayed for the several months required for data accumulation.

Discussion of the techniques used will include terms as defined in Chapter II. Not all of these terms are defined again in this chapter.

Graphical Analysis

A graph showing the time paths of the simulated data stream and the corresponding actual data stream enables the analyst to make a very gross qualitative appraisal of the model's predictive ability. From the available data, nine such graphs could be constructed: actual against simulated dollar sales for the three products; actual against simulated sales weights for the three products; actual against simulated inventory of each of the three products on hand at the distribution center. Because of the lack of a reasonable degree of correspondence between any of the simulated and actual time series, only the graph of daily dollar sales (simulated plotted with octagons, and actual plotted with triangles) for Product 1 is reproduced (Figure 6.1).

From these graphs the number, timing, and direction of turning points, amplitude of fluctuations for corresponding time segments, average amplitude over the entire series, simultineity of turning points, average values, probability distributions, variation about the mean, and exact matching can be determined. This was not done because later tests will perform similar comparisons in a more sophisticated

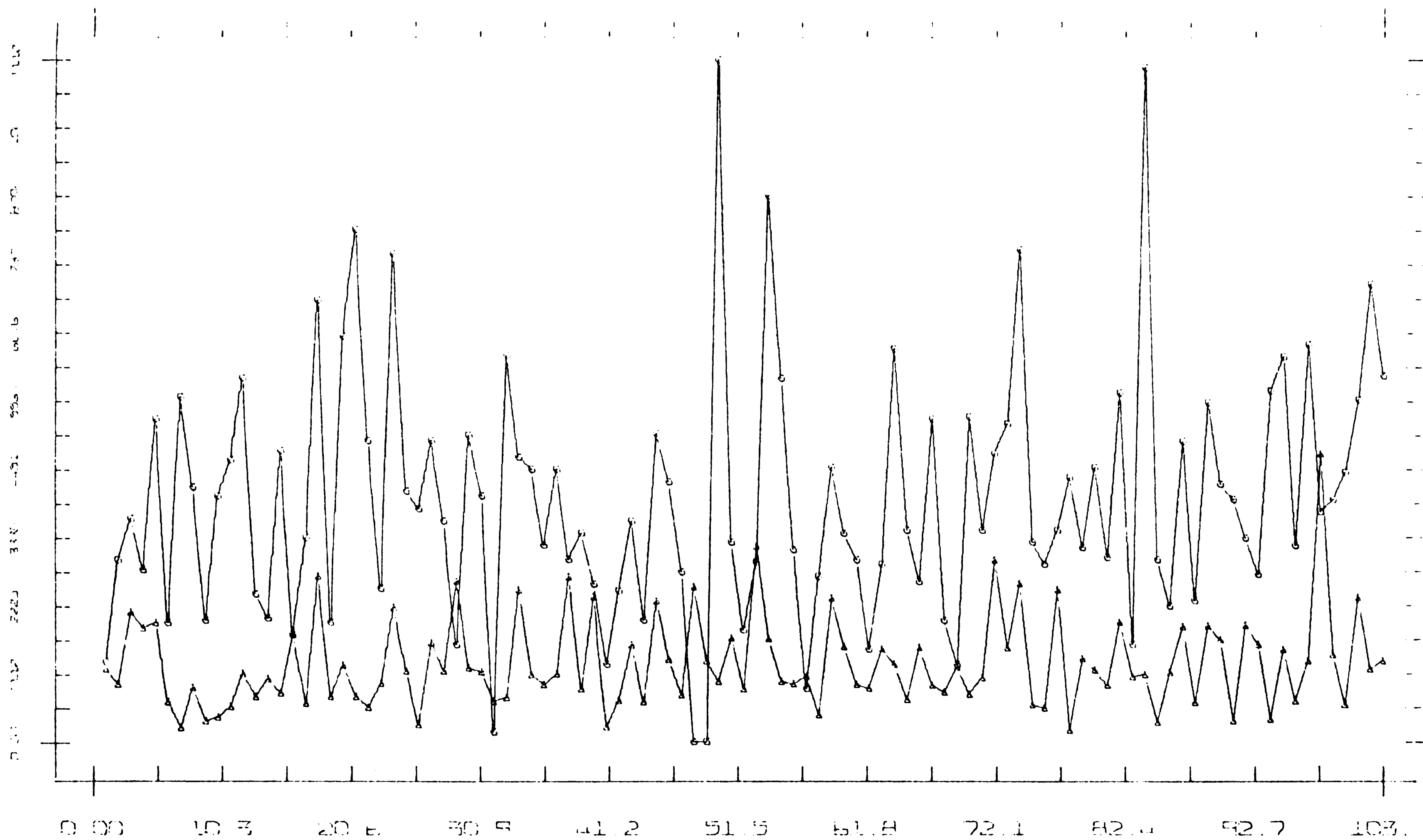


Figure 6.1.--Simulated and Actual Dollar Sales--Product 1.

manner, although the means and variances of the data streams are given in Table 6.1 and their skewness and kurtosis in Table 6.2.

It should be noted from Table 6.1 that the simulated inventory on hand for Product 3 is maintained at zero units. Product 3 is a slow mover, and on the infrequent occasions when this product is demanded, it is placed on back order. The information of this table shows large discrepancies between actual and simulated means and variances for all products over the three variables.

Skewness is a measure of the departure of a distribution from symmetry. This measure would take on the value zero if the distribution was symmetrical. Most of the time series considered are not very symmetrical (Table 6.2). Kurtosis is a measure of the "hump" of a single humped distribution. This measure centers on the value 3, platykurtic distributions having a kurtosis value less than 3, and leptokurtic distributions having values greater than 3. While this measure is of little value for the study at hand, most of the time series considered are platykurtic.

Analysis of Variance

A one-way analysis of variance is conducted to test the null hypothesis that the mean of the simulated data stream is not significantly (at a 95% significance

TABLE 6.1.--Means and Variances.

	Dollar Sales		Sales Weight		Inventory On Hand	
	Simulated	Actual	Simulated	Actual	Simulated	Actual
Product 1						
Mean	1290.63	3979.62	502.79	683.40	663.90	1965.40
Variance	574413.22	4232224.25	87148.79	124791.74	45505.18	705742.18
Product 2						
Mean	246.18	175.72	366.60	233.46	43.72	695.94
Variance	40539.37	25554.49	89897.11	45131.67	1925.80	46173.80
Product 3						
Mean	4.33	7.50	1.86	23.01	0.00	140.56
Variance	75.54	33.71	13.44	327.70	0.00	2924.34

TABLE 6.2.--Skewness and Kurtosis.

	Dollar Sales		Sales Weight		Inventory On Hand	
	Simulated	Actual	Simulated	Actual	Simulated	Actual
Product 1						
Skewness	1.34	0.92	1.34	0.92	-0.02	1.63
Kurtosis	2.84	1.53	2.84	1.53	-0.97	3.09
Product 2						
Skewness	1.47	2.11	1.47	2.11	1.41	0.33
Kurtosis	2.06	7.27	2.06	7.27	2.09	-0.91
Product 3						
Skewness	2.70	0.68	2.62	0.68	0.00	0.33
Kurtosis	8.04	0.25	7.55	0.15	0.00	-0.79

level) different from the mean of the actual data stream. Table 6.3 contains the results of this analysis.

The decision to reject the null hypothesis is made if the calculated F value (MSp/MSe) is greater than the tabled F value for the appropriate degrees of freedom. If the null hypothesis is rejected, the means at this level of confidence are significantly different.

The model indicated that inventory on hand for Product 3 should be maintained at a zero level so an F value could not be calculated. In all cases tested the null hypothesis was accepted at the 95% confidence level.

Multiple Comparison

Multiple comparison is a technique which can be used to test if a particular statistic from a simulation is significantly different from the same statistic in the control. The control in this case is the actual historical data, and the statistic to be tested is the mean. This analysis should confirm the results obtained using analysis of variance.

If the absolute difference between the mean of the simulated data stream and the mean of the actual data is greater than an appropriate Dunnett statistic multiple of the square root of twice the mean square error over the number of variables, then the hypothesis that the means are equal must be rejected. The appropriate Dunnett

TABLE 6.3.--Test of Means.

	Dollar Sales		Sales Weight		Inventory on Hand	
	F	Ho	F	Ho	F	Ho
Product 1	1.21	Accept	1.21	Accept	2.74	Accept
Product 2	0.58	Accept	0.58	Accept	1.66	Accept
Product 3	1.61	Accept	1.75	Accept		

statistic is indexed by the desired confidence level (95%), the number of "plans" to be compared (k=2), and the degrees of freedom for the mean square error term. The results of Table 6.4 show that in all cases this hypothesis was rejected.

The F Test

Similarity of simulated to actual mean values has been evaluated using analysis of variance and multiple comparison. The F distribution is to be used to test if a significant difference exists between the variances. It should be noted that other methods, such as multiple comparison, could be used. The F Test was selected because of the relatively small sample size.

The ratio of the actual variance to the simulated variance is distributed as F. With a knowledge of the number of degrees of freedom contained in each variance calculation and the significance level desired (95%), the correct F value can be found. If the tabled value of F is less than the F statistic, then the hypothesis that the two variances are equal at this significance level is rejected.

The number of degrees of freedom in both the numerator and denominator of the ratio of the variances is 102, and the F value at the 95% confidence level is 1.37. The hypothesis that the variances are equal will

TABLE 6.4.--Multiple Comparison Test of Means.

	Dollar Sales			Sales Weight			Inventory on Hand		
	A	B	Ho	A	B	Ho	A	B	Ho
Product 1	2688.99	169.00	Reject	180.61	65.82	Reject	1301.50	35.67	Reject
Product 2	70.46	49.74	Reject	133.14	74.08	Reject	652.22	9.14	Reject
Product 3	3.17	1.95	Reject	21.15	0.82	Reject			

$A = |\overline{X}_j - \overline{X}_c|$

$B = d\sqrt{MSe/n}$

only be accepted if both the ratio of actual to simulated and the ratio of simulated to actual variances are less than 1.37. The information in Table 6.5 shows that in no case is this true, and so the null hypothesis must be rejected every time.

Correlation

The coefficient of determination expresses the percentage of the total variation in one variable which is "explained" by the regression line of this variable on another variable. Taking the square root of the coefficient of determination gives the coefficient of correlation r. The range of r is -1 to +1 or perfect negative correlation to perfect positive correlation. For there to be some degree of correlation between two variables, r must be shown to be significantly different from zero. Tables are available[1] which show the value which r must be greater than, at a particular confidence level, to be considered different from zero. At a 95% confidence level this value is r=0.197. Analysis of the r values is contained in Table 6.6. The null hypothesis is that the value of r is significantly different from zero.

As stated previously, the value of the coefficient of determination is the proportion of the sum of the squared deviations from the regression line accounted for by the independent variable. The values of r^2 are given in Table 6.7.

TABLE 6.5.--Test of Variances.

		Dollar Sales		Sales Weight		Inventory on Hand	
		Ratio	Ho	Ratio	Ho	Ratio	Ho
Product 1	A	0.14	Accept	1.59	Reject	2.24	Reject
	B	7.37	Reject	0.63	Accept	0.45	Accept
Product 2	A	0.70	Accept	1.99	Reject	0.04	Accept
	B	1.43	Reject	0.50	Accept	24.38	Reject
Product 3	A	0.06	Accept	0.04	Accept		
	B	15.51	Reject	23.98	Reject		

$$A = \frac{\text{Variance of Simulated Series}}{\text{Variance of Actual Series}}$$

$$B = \frac{\text{Variance of Actual Series}}{\text{Variance of Simulated Series}}$$

TABLE 6.6.--Test of Correlation Coefficients.

	Dollar Sales		Sales Weight		Inventory on Hand	
	r	Ho	r	Ho	r	Ho
Product 1	0.1138	Reject	0.1138	Reject	-0.3513	Reject
Product 2	0.0002	Reject	0.0005	Reject	-0.0583	Reject
Product 3	0.1811	Reject	0.1689	Reject		

Table 6.7.--Coefficients of Determination.

	Dollar Sales	Sales Weight	Inventory on Hand
Product 1	0.0130	0.0129	0.1234
Product 2	0.0000	0.0000	0.0034
Product 3	0.0328	0.0285	

Regression Analysis

If perfect correlation existed between the values of the simulated endogenous data streams and the actual data, then the regression line of either of these two variables on the other would be a straight line passing through the origin with a slope of one. Another test of the degree of correlation between these two variables is to determine if the regression line of actual on simulated has an intercept significantly different from zero and a slope significantly different from one. The difference between the sum of the squared deviations between each actual and simulated datum and the sum of the squared deviations between the regression line and each simulated observation divided by the number of observations n all divided by the residual sum of squares divided by n-1 is distributed as F. If this value is greater than the tabled F value (F=3.97) indexed by the degrees of freedom and the confidence level, then the hypothesis that the intercept is not significantly different from zero and the slope is not significantly different from one is rejected.

The results of this test are given in Table 6.8 with the hypothesis being rejected in half the cases.

The Chi-Square Test

For the validity testing of simulated against actual data streams, the Chi-square test is not used in the

TABLE 6.8.--Test of Regression Lines.

	Dollar Sales		Sales Weight		Inventory on Hand	
	F	Ho	F	Ho	F	Ho
Product 1	43.88	Reject	3.87	Accept	1234.37	Reject
Product 2	0.46	Accept	0.43	Accept	2888.83	Reject
Product 3	0.49	Accept	34.59	Reject		

accustomed manner. Whichever is larger, the range of the actual data or the range of the simulated data, is divided into ten equal parts. The number of observations from the actual data which fall into each of these cells becomes the expected frequencies, and the number of simulation observations falling into each cell are the observed frequencies. Summing the squared differences between observed and expected frequencies divided by the expected frequency gives the Chi-square value. This value is compared with a tabled value given a confidence level and degrees of freedom, and if the calculated value is larger than the tabled value, then the hypothesis that there exists a significant correspondence between observed and expected frequencies is rejected.

With nine degrees of freedom and a 95% significance level, the appropriate value of Chi-square is 16.9. The values of Chi-square given in Table 6.9 are compared to the value 16.9, and if smaller, then the hypothesis that the actual and simulated frequencies show reasonable correspondence is not rejected.

Theil's Inequality Coefficient

Theil's Inequality Coefficient U measures the quality of predicted results against actual outcomes. The coefficient has a range from zero to infinity. If U=0, the forecasts are perfect, and U=1 indicates a prediction

TABLE 6.9.--Chi-Square Test.

	Dollar Sales		Sales Weight		Inventory on Hand	
	Chi-Square	Ho	Chi-Square	Ho	Chi-Square	Ho
Product 1	12.57	Accept	5.45	Accept	16.33	Accept
Product 2	3.85	Accept	2.48	Accept	37.71	Reject
Product 3	3.91	Accept	7.09	Accept	61.16	Reject

error equal to that obtained by extrapolation assuming no change.

From Table 6.10 it can be seen that when considering the simulation output as a forecast of the actual daily observations, the prediction is of rather poor quality. Table 6.11 shows that the disparity between forecast and actual is not consistently due to one particular inequality proportion, although the variance proportion is of less effect than the bias or covariance proportions.

Spectral Analysis

When considering the Fourier representation of a time series, the contribution that a particular frequency or frequencies make to the overall variance of the series is of interest. This type of analysis is possible because the frequency band $(\omega, \omega + d\omega)$ contributes $f_{(\omega)}\, d\omega$ to the total variance ($f_{(\omega)}$ is the power spectrum as defined in Chapter II). The number of frequency bands or lags m to consider should be less than $\frac{n}{3}$ (where n is the number of observations in the series), and if n is not large, m should be about $\frac{n}{5}$ or $\frac{n}{6}$.[2] For the n=103 of this analysis, m=20 was chosen.

Examination of the power spectra of the actual time series and the simulated time series will show which frequencies contribute the most to the total variance. If the frequencies were the same or close for both series,

TABLE 6.10.--Test of Predictive Quality.

	Dollar Sales	Sales Weight	Inventory on Hand
Product 1	0.7629	0.6105	0.7500
Product 2	1.1203	1.2365	0.9452
Product 3	1.0184	0.9328	0.9969

TABLE 6.11.--Inequality Proportions.

	Dollar Sales	Sales Weight	Inventory on Hand
Product 1			
Bias	0.6189	0.1477	0.6588
Variance	0.1445	0.0153	0.1528
Covariance	0.2366	0.8370	0.1884
Product 2			
Bias	0.0699	0.1161	0.8963
Variance	0.0242	0.0500	0.0616
Covariance	0.9059	0.8339	0.0421
Product 3			
Bias	0.0992	0.5838	0.8711
Variance	0.0825	0.2721	0.1289
Covariance	0.8183	0.1440	0.0000

similarity of the original series would be indicated. The log of the power spectrum is plotted against j in Figures 6.2 and 6.3 in order to construct Granger and Hatanaka's[3] simultaneous confidence bands (100-α)% for all j (α = confidence level). Notable "power" exists at frequencies where a smooth curve cannot be drawn easily between the confidence limits.

The shape of the power spectra of Figures 6.2 and 6.3 are quite different. The frequency band centered on the component with a period of about 2.67 days for the actual data shows a significant lack of contribution to the overall variance. For the simulated time series the frequency band centered on the component with a period of about 6.67 days provides significant positive contribution to total variance. No reasonable interpretation can be found for periods of 2.67 or 6.67 days. It is also noticable that the low-frequency range of the power spectra (within which the "long-run" components are concentrated) did not contribute to the extent that is normally found in economic time series. A detailed explanation of these rather poor results is contained in the final chapter.

A measure of the correlation between the frequency components of two series is given by

$$C_{(\omega)} = \frac{c^2(\omega) + q^2(\omega)}{f_x(\omega) + f_y(\omega)} ,$$

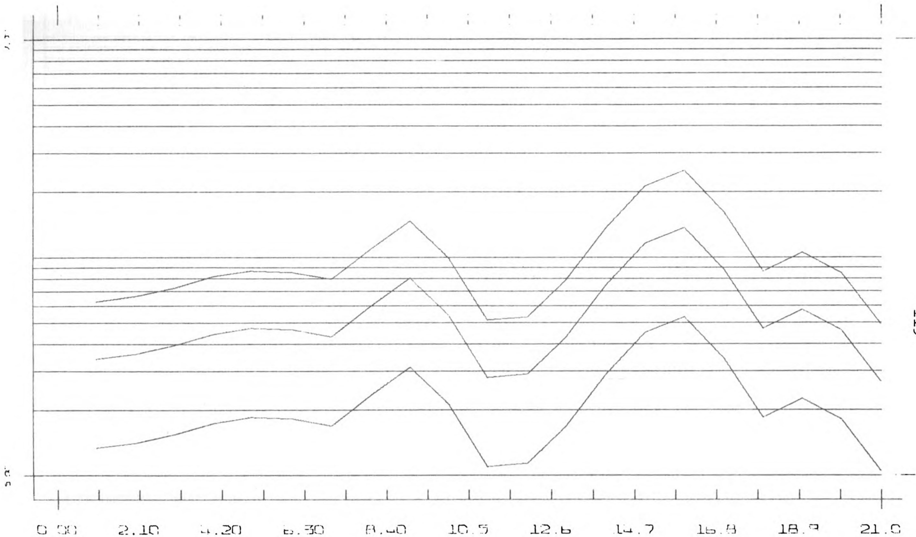


Figure 6.2.--Estimated Power Spectrum--Actual Dollar Sales for Product 1.

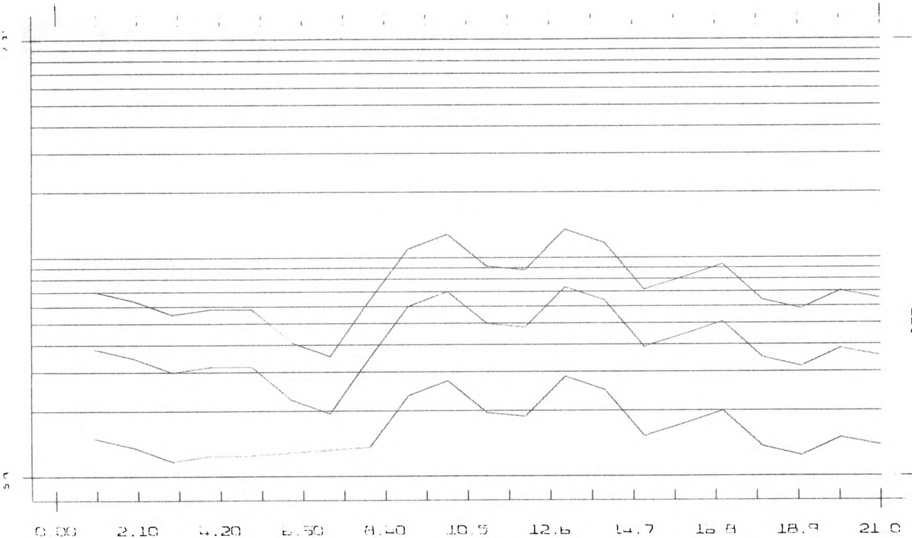


Figure 6.3.--Estimated Power Spectrum--Simulated Dollar Sales for Product 1.

where $c(\omega)$ is the co-spectrum, $q(\omega)$ is the quadrature spectrum, $f_x(\omega)$ the power spectrum of x, and $f_y(\omega)$ is the power spectrum of y. $C(\omega)$ is the coherence at ω. The range of $C(\omega)$ is from zero to one and its value can be interpreted as the square of the correlation coefficient.

The coherence of actual and simulated dollar sales for Product 1 is not great at any frequency although a stronger relationship does exist for frequencies of one month, one week and half a week (Figure 6.4). Tests established by Goodman[4] hypothesize that the true coherence at all frequencies in Figure 6.4 is zero.

A relationship may exist between one time series at point n and another at point (n+k). A measure of the phase difference between the frequency components of two series is

$$\psi(\omega) = \mathrm{TAN}^{-1}\left(\frac{q(\omega)}{c(\omega)}\right) .$$

From the phase diagram of Figure 6.5 no such relationship appears. There is no trend in the phase diagram which would indicate a time lag, neither are there oscillations about a constant other than zero indicating an angle lag.

A final diagram which may indicate the nature of a relationship between two time series is the gain diagram. The gain $R_{xy}(\omega)$ is defined by $f_y(\omega)R^2_{xy}(\omega) = f_x(\omega)C(\omega)$.

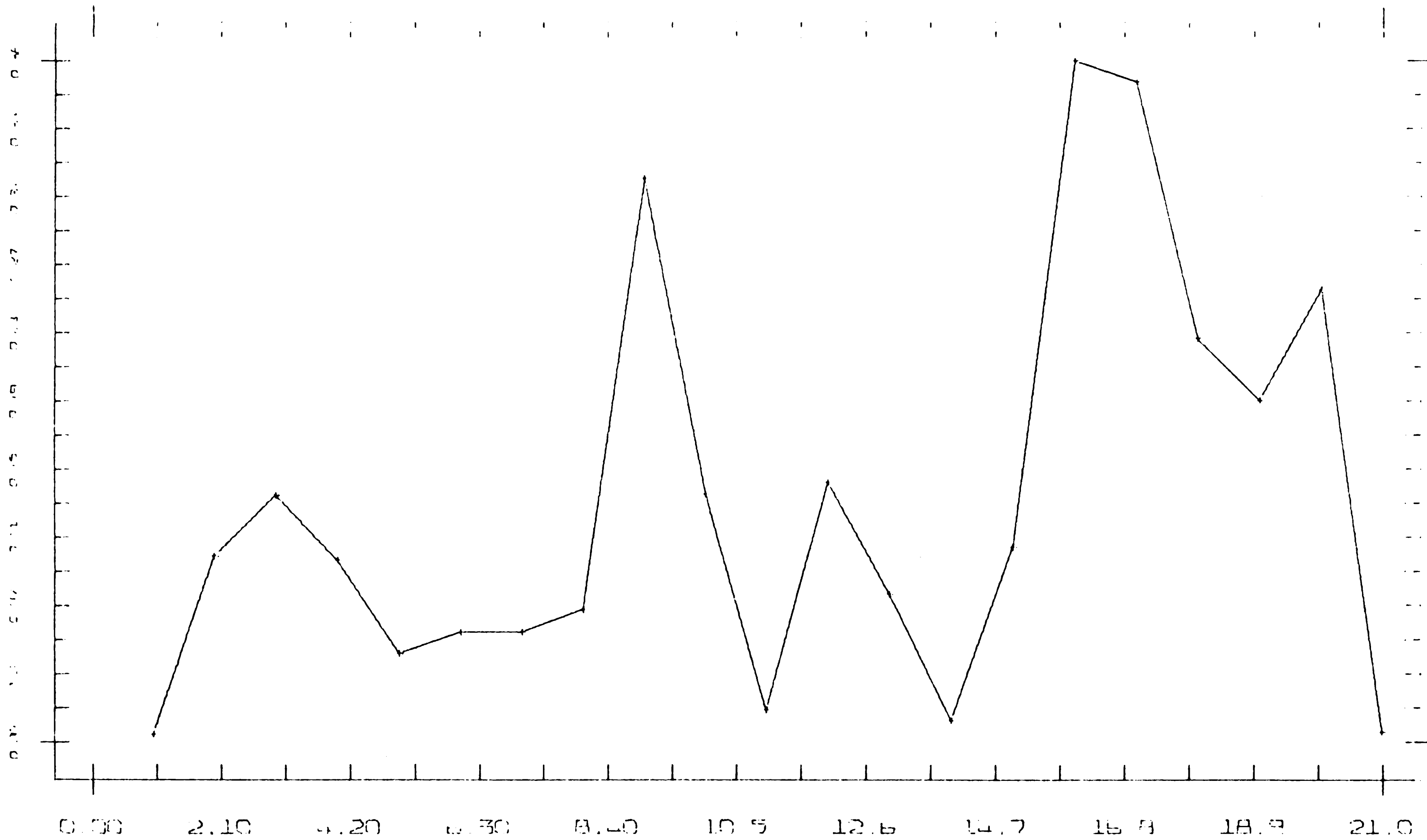


Figure 6.4.--Coherence of Actual and Simulated Dollar Sales--Product 1.

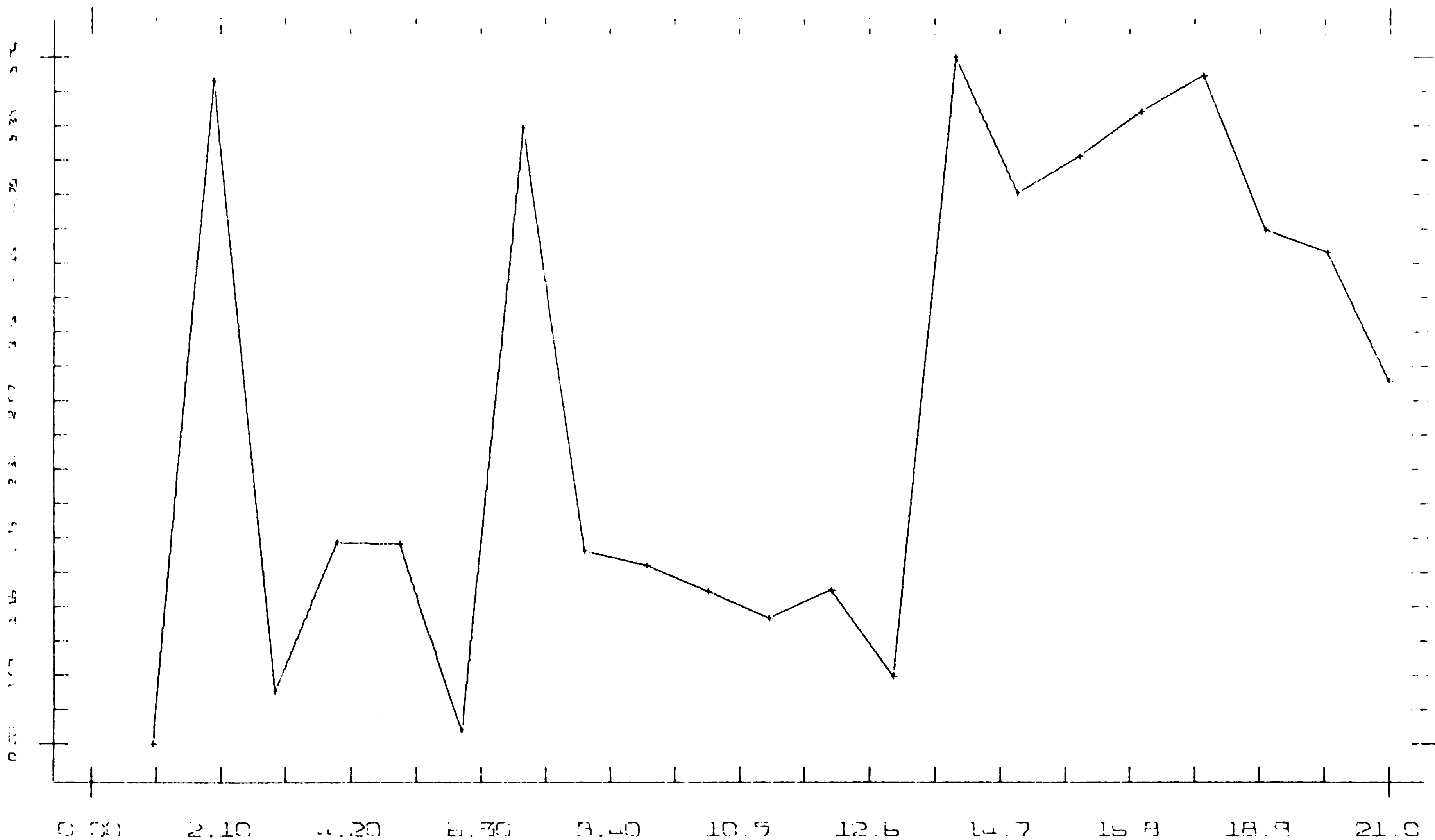


Figure 6.5.--Phase of Actual and Simulated Dollar Sales--Product 1.

Gain can be considered as the regression coefficient of process $\{X_t\}$ on process $\{Y_t\}$ at frequency ω (Figure 6.6).

The results of the other eight comparisons of actual time series with simulated time series were of comparable quality to those presented for daily dollar sales of Product 1 and so they are not reproduced here.

Factor Analysis

Cohen and Cyert[5] suggest comparison of the factor loadings of simulated results with the factor loadings of actual results as a method of appraising the quality of the simulated output.

A factor analysis of the nine actual data streams (dollar sales, sales weight, and inventory on hand, each for three products) produced most meaningful factor loadings with three factors. This was also the case with the nine simulated data streams. It is now of interest to determine the extent to which the three actual factors and the three simulated factors differ in ability to describe the actual and simulated data respectively. Table 6.12 is the similarity matrix for these three factor pairs. Each element in the matrix has a range of values from -1 to 1, significant correspondence between the factors occurring only for values of 0.78868 or greater. The best factor pairings are: actual 1 with simulated 2, actual 2 with simulated 3, and actual 3 with simulated 1. Only the second pairing is significant.

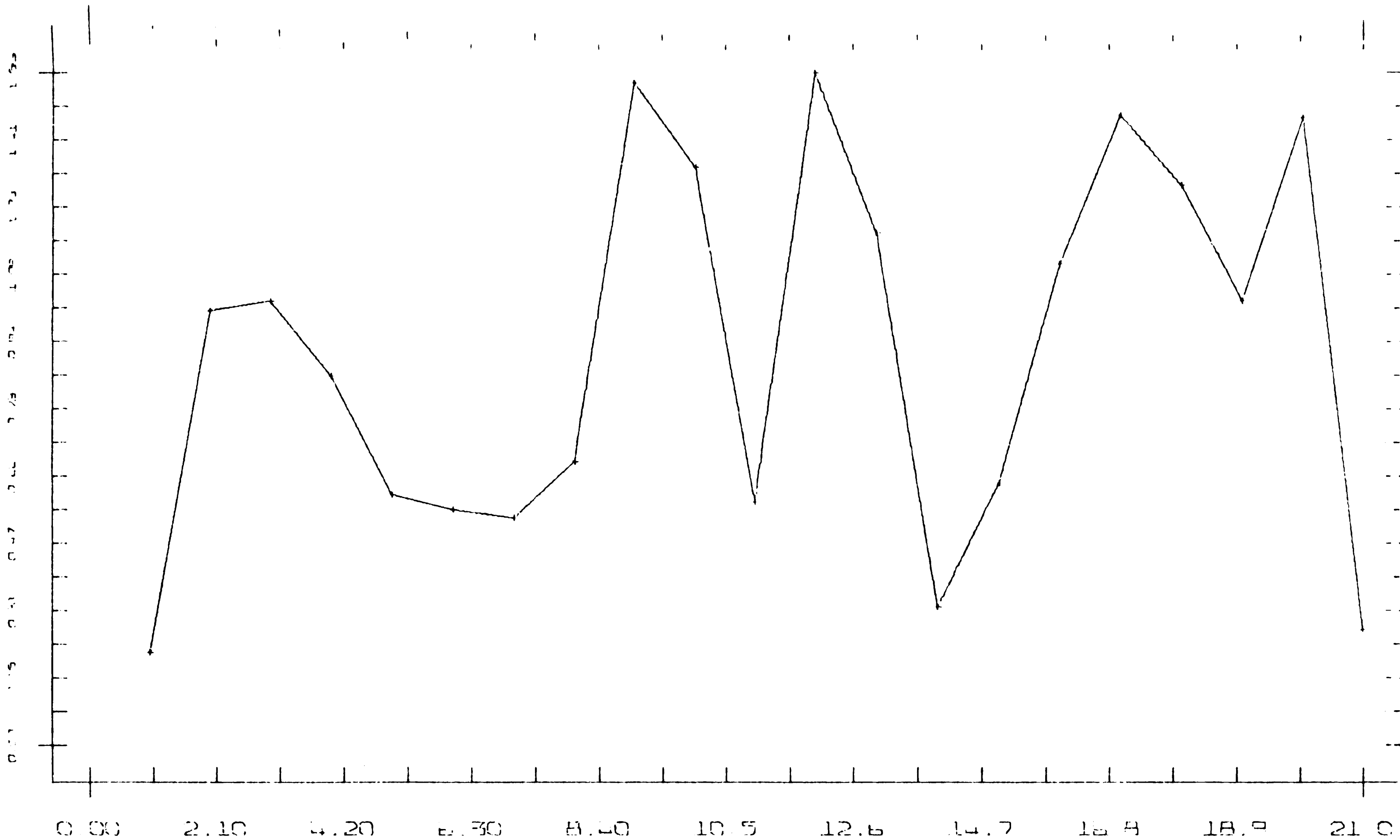


Figure 6.6.--Gain of Actual and Simulated Dollar Sales--Product 1.

TABLE 6.12.--Similarity Matrix for Factor Loadings.

		Factor Loading for Simulated Data		
		1	2	3
Factor Loading for Actual Data	1	0.6235	0.6619	0.0222
	2	-0.3452	0.3771	0.8431
	3	0.0874	-0.0012	-0.0373

The Model's Predictive Ability

The ability of the LREPS model to predict the behavior of the actual system has not been established. The results presented in this chapter are poor and at times contradictory. But neither has any major defect in the model been established. The only conclusion to be drawn is that the validity of the model's predictive capability has not been established. In order to do this, these same tests must be repeated with a larger number of observations collected at a longer time increment.

CHAPTER VI--FOOTNOTES

[1]J. Riggs, Production Systems: Planning, Analysis and Control (New York: John Wiley & Sons, Inc., 1970), p. 70.

[2]C. W. J. Granger and M. Hatanaka, Spectral Analysis of Economic Time Series (Princeton, N.J.: Princeton University Press, 1964), p. 61.

[3]*Ibid.*, p. 62.

[4]N. R. Goodman, Scientific Paper No. 10 (New York, N.Y.: New York University, Engineering Statistics Laboratory, 1957).

[5]K. J. Cohen and R. M. Cyert, "Computer Models in Dynamic Economics," The Quarterly Journal of Economics, Vol. LXXV, No. 1 (February, 1961), pp. 112-127.

CHAPTER VII

SENSITIVITY OF THE MODEL'S MAJOR ASSUMPTIONS

Introduction

The third and final part of the validation procedure is to determine the degree to which the characteristics of the endogenous data streams change when the form of one of the model's major assumptions is altered. Assumptions are usually made to simplify the complexity of real situations and so make the modeling process easier. Indeed, model construction may not be possible in many situations without incorporating rather stringent assumptions. But it is undesirable to have the model output dependent on the nature of the assumptions embodied in the model. It seems reasonable that the endogenous data streams of a valid computer simulation model will not change significantly even with rather severe changes to the assumptions which are incorporated into the model. This chapter describes the analysis performed in order to test this statement for the LREPS model.

The LREPS model contains two major assumptions. The first concerns the way in which demand from the consumer level is generated, and the second concerns the selection of products from the total product line over which this demand will be allocated. Both of these assumptions are required because a firm of reasonable magnitude producing consumer products can expect to handle hundreds of thousands of orders for hundreds of different products during the course of a year. The dilemma created is: too much detail cannot be handled by available computing machinery; too much aggregation of this detail will reduce the model's ability to test the effects of such changes as the introduction of new products, different inventory policies or different demand patterns. Solution of the dilemma comes with the introduction of assumptions.

A stratified sample of 50 products from the total product line was selected.[1,2] The products in the sample must be representative of the entire product line so that the information generated on the basis of the sample can be extrapolated to the level of the total corporate operation. The sample products were selected on the basis that a product be representative of the company's inventory and movement costs. Products were classified into four categories on the basis of annual dollar sales, with the first category containing "high-movers" and any products management might want to give special consideration.

Rather than attempt to account for each of several hundred thousand individual orders, a random selection is made of a year's invoices. A particular number of individual orders for the sample products is summarized into a block. The number of orders so summarized is called the blocking factor. These blocks are then combined into an order file or order matrix from which a block of orders is randomly drawn to generate the demand for each time period.[3]

This chapter investigates the effect of four changes in the assumptions for LREPS product analysis and order generation. The normal blocking factor for order generation is 10--blocking factors of 5 and of 20 are considered. A stratified sample of 50 products is used, these products divided into four categories--a new sample of 50 products is generated, and the effect of using only 3 product categories is investigated.

So the net result is the analysis of the control endogenous data stream (the output of the model in its unmodified condition) with the endogenous data streams resulting when each of the four proposed changes is put into effect. To simplify the presentation of the results of the statistical tests used, these five situations will be designated plans viz.:

Plan A The control--no change in model structure

Plan B Blocking factor of 5 used in order generation

Plan C Blocking factor of 20 used in order generation

Plan D 3 categories used for sample products

Plan E New product sample used

Graphical Analysis

An approximate idea as to the degree of change occurring in the model's output data streams with a change in assumptions can be obtained by examination of the graph of these data streams before and after the change. Six endogenous data streams of the unmodified model are obtained: dollar sales for each of the three products for a two-year period and sales weights for the three products over the same two years. Comparison of each of these six Plan A's with each of the other four plans gives a net result of 30 data streams or 24 one-on-one comparisons of Plan A with another plan. An exceedingly large volume of data is recorded if the results of all tests for all products for both variables are included. In this section and the Spectral Analysis section only the results for dollar sales of Product 1 are presented and even then the amount of data included is considerable. The results not included do not add any new dimension to the analysis which might justify their inclusion.

Figure 7.1 shows the dollar sales of Product 1 Plan A (the control) against Plan B. Figures 7.2 to 7.4 are the graphs of Plan A and Plan C, Plan A and Plan D, and Plan A and Plan E. The high degree of intermeshing of each of the pairs of data streams indicates no radical change in results for any of the four plans tested.

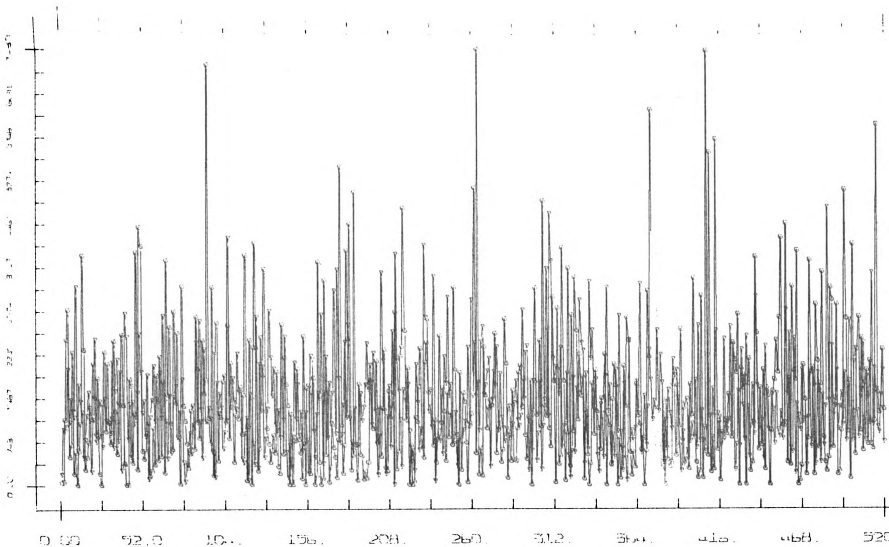


Figure 7.1.--Plan A and Plan B--Dollar Sales for Product 1.

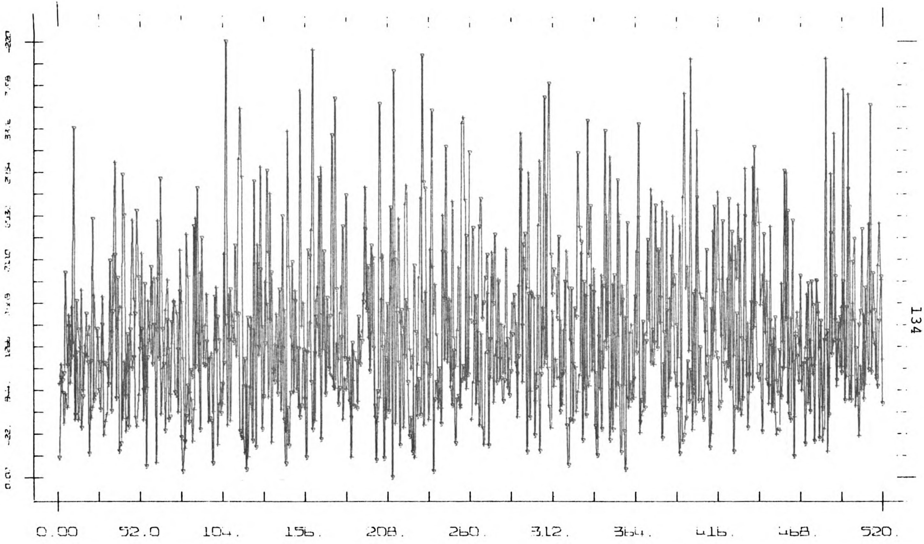


Figure 7.2.--Plan A and Plan C--Dollar Sales for Product 1.

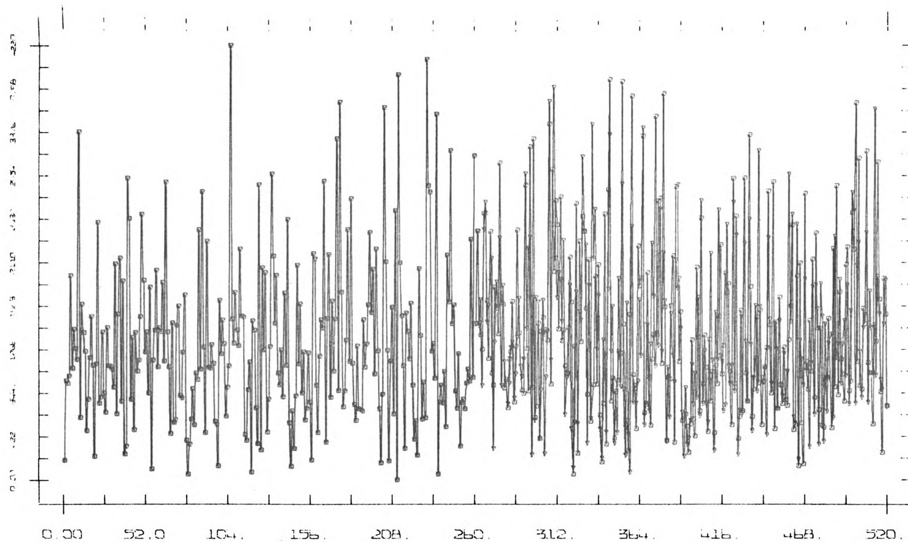


Figure 7.3.--Plan A and Plan D--Dollar Sales for Product 1.

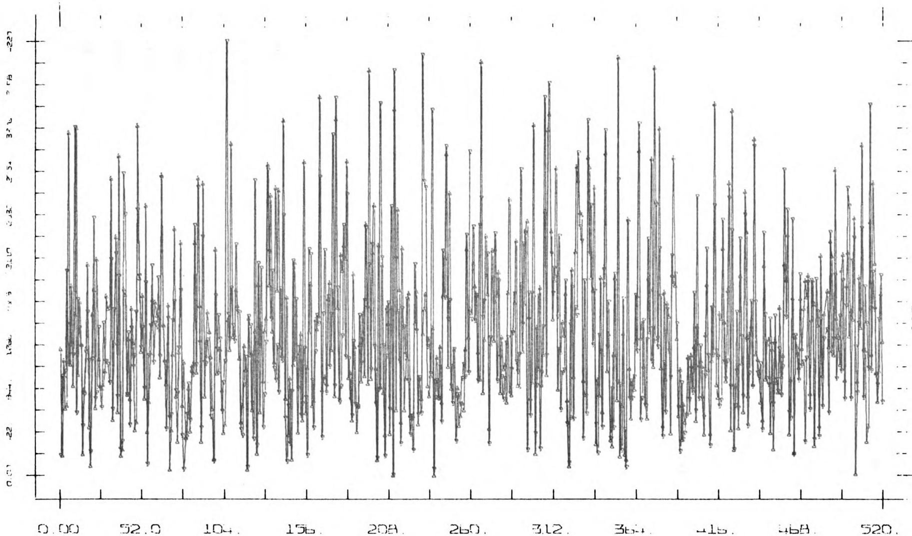


Figure 7.4.--Plan A and Plan E--Dollar Sales for Product 1.

Added information can be obtained by a detailed analysis of the graphs as outlined in the preceeding chapter. But again this will not be done, as later tests will provide similar information by more reliable methods. The means (Table 7.1), variances (Table 7.2), skewness (Table 7.3), and kurtosis (Table 7.4) are given. These parameters show no remarkable change between the control and any of the other four plans.

Analysis of Variance

Analysis of variance is used to test for any difference between the mean of the control (Plan A) and the means of the other four plans. The null hypothesis that at the 95% confidence level no difference exists between the control mean and the other means is examined in Table 7.5. The null hypothesis is rejected if the calculated value of F (MSp/MSe) is greater than the tabled value of F for the appropriate degrees of freedom.

The null hypothesis is accepted when Plan A is compared with Plans B or C, but is rejected when Plan A is compared with Plans D or E. Remember that Plans B and C involve changes in the order generation process, while Plans D and E involve alterations to the product sampling procedure. Given a change in the method of order generation, a particular product should still be contained in the average order to the same extent. But when the number

TABLE 7.1.--Means.

	Plan A	Plan B	Plan C	Plan D	Plan E
Dollar Sales					
Product 1	1361.74	1414.25	1428.58	1403.69	1409.05
Product 2	217.58	252.76	229.54	219.80	216.79
Product 3	4.15	2.48	5.36	3.74	6.95
Sales Weight					
Product 1	530.49	550.94	556.52	546.82	548.93
Product 2	324.03	376.39	341.85	326.28	704.63
Product 3	1.78	1.06	2.32	1.61	9.35

Table 7.2.--Variances.

	Plan A	Plan B	Plan C	Plan D	Plan E
Dollar Sales					
Product 1	643436.02	1525642.69	527771.21	741116.72	703880.74
Product 2	38024.76	75415.02	24135.74	38729.53	41502.32
Product 3	76.44	61.35	64.59	64.50	199.01
Sales Weight					
Product 1	97648.49	231520.58	80086.61	112471.44	106821.38
Product 2	84331.09	167221.95	53520.93	85902.55	438492.12
Product 3	13.61	11.05	11.56	11.55	361.61

TABLE 7.3.--Skewness.

	Plan A	Plan B	Plan C	Plan D	Plan E
Dollar Sales					
Product 1	1.04	1.64	0.98	1.10	1.10
Product 2	1.73	1.70	1.12	1.87	1.80
Product 3	2.76	4.63	2.19	2.91	2.54
Sales Weight					
Product 1	1.03	1.64	0.98	1.10	1.10
Product 2	1.73	1.70	1.12	1.87	1.80
Product 3	2.70	4.58	2.13	2.85	2.53

TABLE 7.4.--Kurtosis.

	Plan A	Plan B	Plan C	Plan D	Plan E
Dollar Sales					
Product 1	1.36	3.73	1.10	1.23	1.37
Product 2	3.89	3.11	1.80	5.05	3.86
Product 3	8.69	31.92	6.51	10.02	7.51
Sales Weight					
Product 1	1.36	3.73	1.10	1.23	1.37
Product 2	3.89	3.11	1.80	5.06	3.86
Product 3	8.42	30.93	6.29	9.74	7.42

TABLE 7.5.--Test of Means.

	Plan A - Plan B		Plan A - Plan C		Plan A - Plan D		Plan A - Plan E	
	F	Ho	F	Ho	F	Ho	F	Ho
Dollar Sales								
Product 1	1.36	Accept	1.18	Accept	134.81	Reject	79.71	Reject
Product 2	0.53	Accept	0.27	Accept	127.36	Reject	84.47	Reject
Product 3	0.44	Accept	0.47	Accept	287.32	Reject	89.12	Reject
Sales Weight								
Product 1	1.44	Accept	1.18	Accept	131.85	Reject	79.19	Reject
Product 2	0.51	Accept	0.28	Accept	138.47	Reject	89.00	Reject
Product 3	0.92	Accept	0.46	Accept	218.26	Reject	72.44	Reject

of product categories or the particular products included in the sample are changed, the extent of a particular product's presence in the average order might well vary.

Multiple Comparison

To test for significant difference in a particular statistic between the control and alternative plans, multiple comparison is used. Again multiple comparison is used to confirm the analysis of variance testing of the mean values.

The absolute difference between the mean of the control and the mean of the particular alternative plan under consideration must be less than a specified amount. Otherwise the null hypothesis that no significant difference exists between the means cannot be accepted. This specified amount is an appropriate Dunnett statistic multiple of the square root of twice the mean square error divided by the number of variables involved. The correct Dunnett statistic is found with a knowledge of the desired confidence level (95%), the number of plans (2), and the degrees of freedom for the mean square error.

Table 7.6 contains the results of this analysis. The analysis of variance testing is confirmed only to a moderate degree. General acceptance of the null hypothesis is shown for all plans for Products 1 and 2, while general rejection of the null hypothesis is shown for Product 3.

TABLE 7.6.--Multiple Comparison Test of Means.

	Plan A - Plan B			Plan A - Plan C		
	A	B	Ho	A	B	Ho
Dollar Sales						
Product 1	52.51	125.48	Accept	66.84	73.88	Accept
Product 2	35.18	28.06	Reject	11.96	15.91	Accept
Product 3	1.67	0.80	Reject	5.36	1.21	Reject
Sales Weight						
Product 1	20.45	48.86	Accept	26.03	28.78	Accept
Product 2	52.36	41.73	Reject	17.82	23.63	Accept
Product 3	0.72	0.34	Reject	0.54	0.35	Reject

	Plan A - Plan D			Plan A - Plan E		
	A	B	Ho	A	B	Ho
Dollar Sales						
Product 1	41.95	54.92	Accept	47.31	61.80	Accept
Product 2	1.52	11.67	Accept	0.79	13.72	Accept
Product 3	0.41	0.46	Accept	2.80	1.11	Reject
Sales Weight						
Product 1	16.33	21.54	Accept	18.44	24.11	Accept
Product 2	2.25	19.57	Accept	2.25	19.57	Accept
Product 3	0.17	0.20	Accept	7.57	1.49	Reject

$A = |\bar{X}_j - \bar{X}_c|$

$B = d\sqrt{MSe/n}$

The F Test

While analysis of variance and multiple comparison have been used to test means, the F distribution is used to test for significant differences between variances.

The ratio of the variance of the control (Plan A) to the variance of one of the other plans is distributed as F. This F value, if greater than the appropriate tabled value of F, will cause the null hypothesis that the two variances are equal to be rejected. The correct tabled value of F is selected with knowledge of the desired significance level (95%) and the degrees of freedom of each of the variances (519). The tabled F value of 1.11 is used for the results of Table 7.7.

For Products 1 and 2 the null hypothesis is accepted for all plans except Plan C. Plan C uses the large blocking factor which provides an individual product with a greater probability of being included in the block, and therefore decreases the variability (and variance) for the product. Generally, the null hypothesis is rejected for Product 3. This product is a slow mover, and so it occurrs in an order with a great degree of irregularity, forcing the variance to be relatively large and unpredictable.

Correlation

The square of the correlation coefficient r is the coefficient of determination which expresses the amount of

TABLE 7.7.--Test of Variances.

	Plan A - Plan B		Plan A - Plan C		Plan A - Plan D		Plan A - Plan E	
	F	Ho	F	Ho	F	Ho	F	Ho
Dollar Sales								
Product 1	0.42	Accept	1.22	Reject	0.87	Accept	0.91	Accept*
Product 2	0.50	Accept	1.58	Reject	0.98	Accept*	0.92	Accept*
Product 3	1.25	Reject	1.18	Reject	1.19	Reject	0.38	Accept
Sales Weight								
Product 1	0.42	Accept	1.22	Reject	0.87	Accept	0.91	Accept*
Product 2	0.50	Accept	1.58	Reject	0.98	Accept*	0.19	Accept
Product 3	1.23	Reject	1.18	Reject	1.18	Reject	0.04	Accept

$$F = \frac{\text{Control (Plan A)}}{\text{Alternative plan}}$$

*Only these alternatives are acceptable if the inverse of F is also constraining.

the total variation contained in one variable which is "explained" by the regression line of this variable on another variable (Table 7.9).

The range of the correlation coefficient is from -1 to +1. If a relationship exists between two variables, the most basic test is to show that the correlation coefficient is significantly different from zero. Given a particular confidence level, the calculated value of r must be larger than a tabled r value[4] in order to accept the null hypothesis that the value of r is significantly different from zero. At a 95% confidence level this tabled value of r is 0.197. Table 7.8 gives the results of such testing for the correlation coefficient of the control plan and each of the alternative plans. While the correlation coefficients are not significantly different with changes in order generation (Plans B and C), they are significantly different with changes in the product sample characteristics (Plans D and E). This is confirmed by the values of the coefficients of determination.

Regression Analysis

A regression line passing through the origin with a slope of one indicates perfect correlation between the dependent and independent variable(s). The regression lines of the control values (Plan A) against the values

TABLE 7.8.--Test of Correlation Coefficients.

	Plan A - Plan B		Plan A - Plan C		Plan A - Plan D		Plan A - Plan E	
	r	Ho	r	Ho	r	Ho	r	Ho
<u>Dollar Sales</u>								
Product 1	0.0220	Reject	0.0359	Reject	0.8125	Accept	0.7117	Accept
Product 2	0.0460	Reject	0.0042	Reject	0.8266	Accept	0.7562	Accept
Product 3	-0.0183	Reject	0.0109	Reject	0.8499	Accept	0.6783	Accept
<u>Sales Weight</u>								
Product 1	0.0220	Reject	0.0359	Reject	0.8126	Accept	0.7117	Accept
Product 2	0.0462	Reject	0.0041	Reject	0.8267	Accept	0.7562	Accept
Product 3	-0.0173	Reject	0.0061	Reject	0.8507	Accept	0.6827	Accept

TABLE 7.9.--Coefficients of Determination.

	Plan A - Plan B	Plan A - Plan C	Plan A - Plan D	Plan A - Plan E
Dollar Sales				
Product 1	0.0005	0.0013	0.6602	0.5006
Product 2	0.0021	0.0000	0.6832	0.5718
Product 3	0.0003	0.0001	0.7224	0.4601
Sales Weight				
Product 1	0.0005	0.0013	0.6603	0.5006
Product 2	0.0021	0.0000	0.6835	0.5719
Product 3	0.0003	0.0000	0.7236	0.4661

of each of the other plans is constructed and the slopes and intercepts are tested to determine if they are significantly different from one and zero respectively (Table 7.10). The sum of the squared deviations between each observation of the control (Plan A) and another plan and the sum of the squared deviations between the regression line and the control observations are calculated. The difference between these two sums is divided by the number of observations n and the result divided by the residual sum of squares over n-1. The net result of this calculation is distributed as F. In order to reject the null hypothesis that the intercept is not significantly different from zero and the slope is not significantly different from one, this F value must be greater than a tabled value of F indexed by degrees of freedom and confidence level. The tabled F value is 3.00 for this testing.

Table 7.10 shows that this hypothesis is accepted in all but one case which involved Product 3.

The Chi-Square Test

The Chi-square test is used to compare the control (Plan A) with the other plans. The larger of the range of the control observations and the range of observations of the other plan being considered is divided into ten equal parts. The number of observations from the control

TABLE 7.10.--Test of Regression Lines.

	Plan A - Plan B		Plan A - Plan C		Plan A - Plan D		Plan A - Plan E	
	F	Ho	F	Ho	F	Ho	F	Ho
Dollar Sales								
Product 1	0.2345	Accept	0.6145	Accept	0.4357	Accept	0.4727	Accept
Product 2	0.2927	Accept	0.7921	Accept	0.4736	Accept	0.4929	Accept
Product 3	0.6848	Accept	0.6034	Accept	0.4646	Accept	0.4456	Accept
Sales Weight								
Product 1	0.2347	Accept	0.6145	Accept	0.4298	Accept	0.4712	Accept
Product 2	0.2918	Accept	0.7911	Accept	0.3708	Accept	-0.6042	Accept
Product 3	0.6484	Accept	0.6025	Accept	0.4464	Accept	-6.2863	Reject

which fall into each of these cells is considered the expected frequency, while the number of observations from the alternative plan which fall into each of the cells is the observed frequency. Then the Chi-square value is the sum of the squared difference between observed and expected frequencies divided by the sum of the expected frequencies. Given a confidence level and degrees of freedom, a tabled value of Chi-square is compared with this calculated Chi-square value. If the calculated value is larger than the tabled value, then the hypothesis that there exists a significant correspondence between the observed and expected frequencies is rejected.

For a 95% significance level and nine degrees of freedom, the Chi-square value is 16.9. Table 7.11 shows that in all cases a reasonable correspondence does exist between the control frequencies and the frequencies of the other plans.

Theil's Inequality Coefficient

The quality of a prediction when compared to the actual outcome is measured by Theil's Inequality Coefficient U. If U is equal to its lower limit of zero, then the forecasts have been perfect, while a value of 1 indicates a forecasting method no better than no-change extrapolation. U has no finite upper bound.

Table 7.12 shows the results of trying to predict the control (Plan A) from the values generated by one of

TABLE 7.11.--Chi-Square Test.

	Plan A - Plan B		Plan A - Plan C		Plan A - Plan D		Plan A - Plan E	
	Chi-Square	Ho	Chi-Square	Ho	Chi-Square	Ho	Chi-Square	Ho
<u>Dollar Sales</u>								
Product 1	1.39	Accept	0.74	Accept	0.62	Accept	0.67	Accept
Product 2	1.39	Accept	1.09	Accept	0.95	Accept	0.84	Accept
Product 3	0.69	Accept	0.61	Accept	0.48	Accept	0.95	Accept
<u>Sales Weight</u>								
Product 1	1.17	Accept	1.35	Accept	1.34	Accept	1.34	Accept
Product 2	0.66	Accept	0.66	Accept	0.64	Accept	0.63	Accept
Product 3	0.63	Accept	0.59	Accept	0.52	Accept	0.99	Accept

TABLE 7.12.--Test of Predictive Quality.

	Plan A - Plan B	Plan A - Plan C	Plan A - Plan D	Plan A - Plan E
Dollar Sales				
Product 1	0.7768	0.6642	0.3120	0.3815
Product 2	0.8868	0.8974	0.3914	0.4682
Product 3	1.4299	1.1949	0.5134	0.6753
Sales Weight				
Product 1	0.7765	0.6639	0.3119	0.3813
Product 2	0.8870	0.8978	0.3913	0.6346
Product 3	1.3913	1.1469	0.4948	0.8571

the alternative plans. All comparisons for Products 1 and 2 generate coefficients well below 1, with the values for Plans D and E being below those of Plans B and C. The coefficient for Product 3 also indicates better predictions using Plans D or E, but the coefficient values are higher in every comparison than for Products 1 and 2.

The covariance proportion consistently accounts for most of the disparity between the actual and forecast results (Table 7.13).

Spectral Analysis

Spectral Analysis is used to analyze the relationship between the control (Plan A) and alternative plans in the same manner as described in Chapter VI. The log of the power spectrum is plotted and around it are constructed simultaneous confidence bands for all frequencies. The power spectra of the two plans under examination should show similar characteristics; notable "power" should exist at similar frequencies. The correlation between frequency components of the two series is given in the coherence diagram. Relationship between different frequencies of the two series is shown in the phase diagram. Finally, the gain diagram is the graph of the equivalent of the regression coefficient of one process on the other at all frequencies.

Again this procedure is presented only for dollar sales of Product 1, as no significant additional information

TABLE 7.13.--Inequality Proportions.

		Plan A - Plan B	Plan A - Plan C	Plan A - Plan D	Plan A - Plan E
Dollar Sales					
Product 1	Bias	0.0013	0.0039	0.0067	0.0057
	Variance	0.0881	0.0050	0.0131	0.0035
	Covariance	0.9106	0.9910	0.9803	0.9908
Product 2	Bias	0.0113	0.0023	0.0002	0.0000
	Variance	0.0578	0.0253	0.0002	0.0039
	Covariance	0.9310	0.9724	0.9996	0.9961
Product 3	Bias	0.0196	0.0104	0.0078	0.0673
	Variance	0.0058	0.0035	0.0233	0.2482
	Covariance	0.9746	0.9861	0.9689	0.6846
Sales Weight					
Product 1	Bias	0.0013	0.0039	0.0067	0.0057
	Variance	0.0881	0.0051	0.0131	0.0035
	Covariance	0.9106	0.9910	0.9803	0.9908
Product 2	Bias	0.0113	0.0023	0.0002	0.3844
	Variance	0.0577	0.0253	0.0002	0.3668
	Covariance	0.9310	0.9724	0.9996	0.2488
Product 3	Bias	0.0204	0.0113	0.0074	0.1701
	Variance	0.0052	0.0033	0.0220	0.6976
	Covariance	0.9744	0.9854	0.9706	0.1322

is provided by the analysis of the control and four plans for the other five variables.

Observation of Figures 7.5, 7.6, 7.7, 7.8, and 7.9 shows that in all cases (Plans A, B, C, D, and E respectively) particular frequency bands contribute more to the overall variance than might reasonably be expected. These frequency levels occur at the equivalent of 20 days, 8 days, 4.5 days, 3.5 days, and just under 3 days. The 20 day (four week) and 4.5 day (almost one week) periodicity could well be expected, but an explanation for other three frequency bands is difficult to find. But anyway the main result to be obtained from the power spectra is that the frequency bands supplying notable power are the same for all plans. This implies that a significant difference does not exist between the original data streams.

Figures 7.10 to 7.13 show the coherence of each alternative plan with plan A, while Figures 7.14 to 7.17 give the phase and Figures 7.18 to 7.21, the gain of all possible comparisons with the control.

The coherence diagrams show that the correlation per pair of frequency components is stronger when the control is compared with the alternatives of a new product sample (Plan E) and three product categories (Plan D) than with the alternatives of changed blocking factors (Plans B and C).

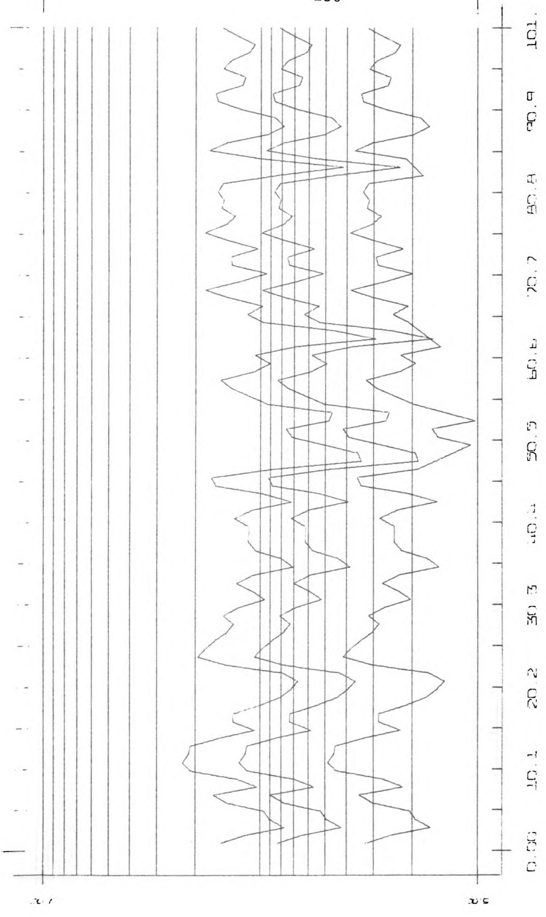


Figure 7.5.--Estimated Power Spectrum of Plan A--Dollar Sales for Product 1.

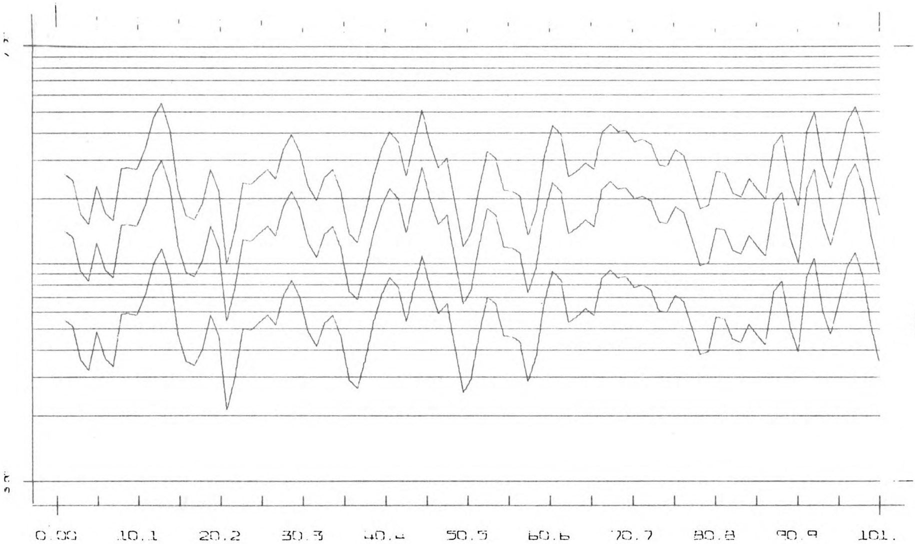


Figure 7.6.--Estimated Power Spectrum of Plan B--Dollar Sales for Product 1.

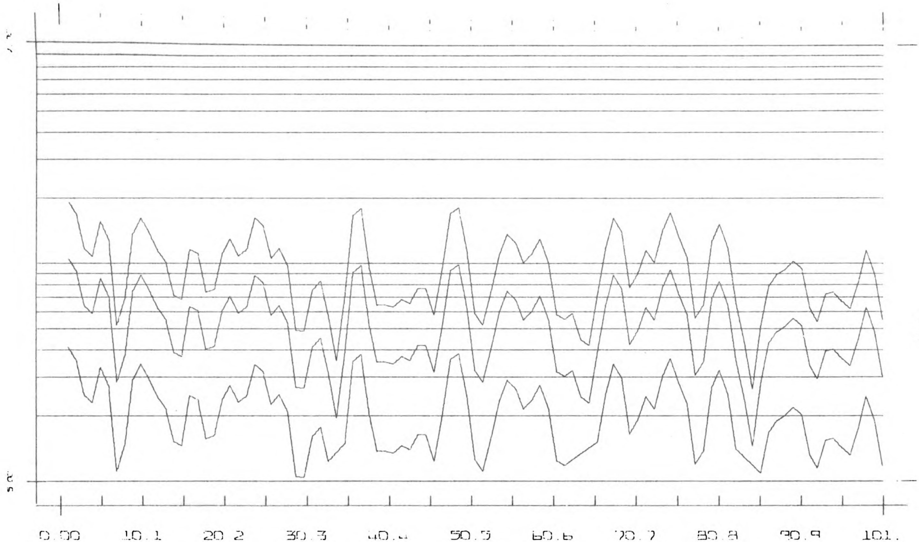


Figure 7.7.--Estimated Power Spectrum of Plan C--Dollar Sales for Product 1.

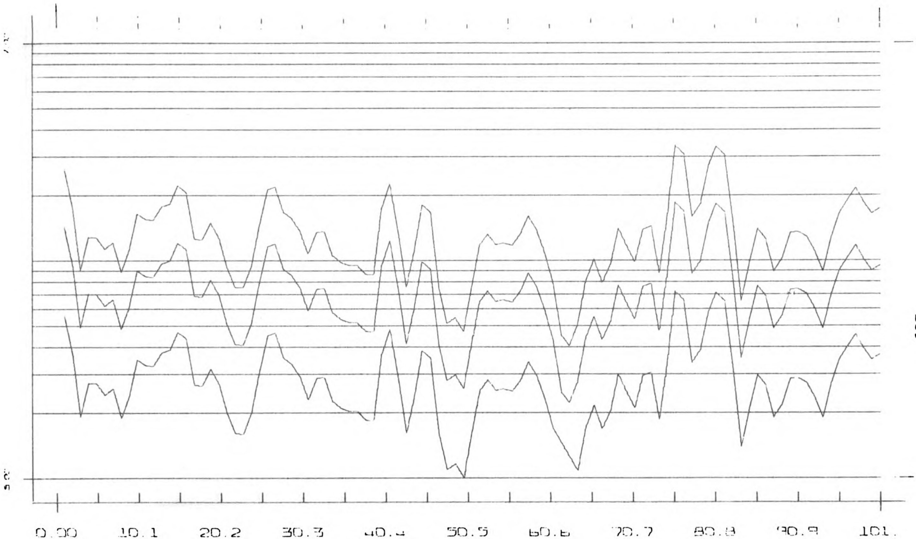


Figure 7.8.--Estimated Power Spectrum of Plan D--Dollar Sales for Product 1.

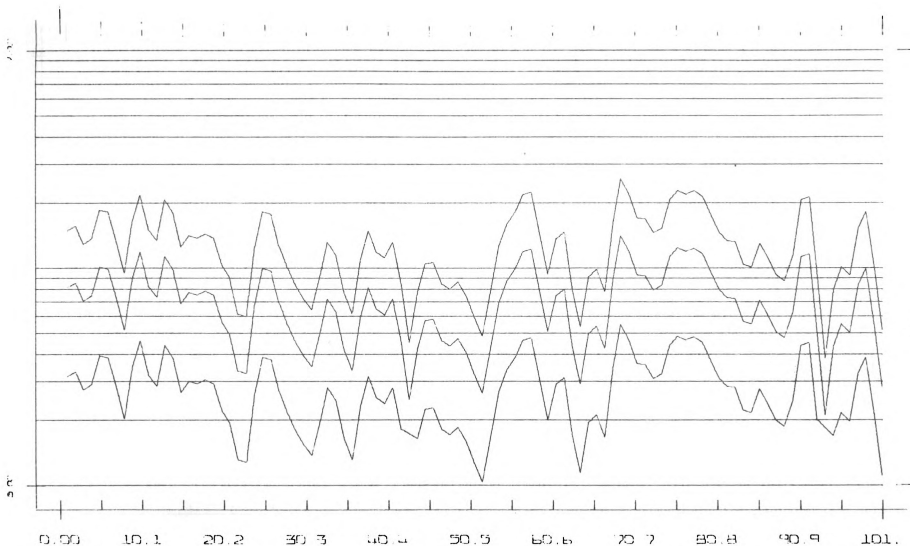


Figure 7.9.--Estimated Power Spectrum of Plan E--Dollar Sales for Product 1.

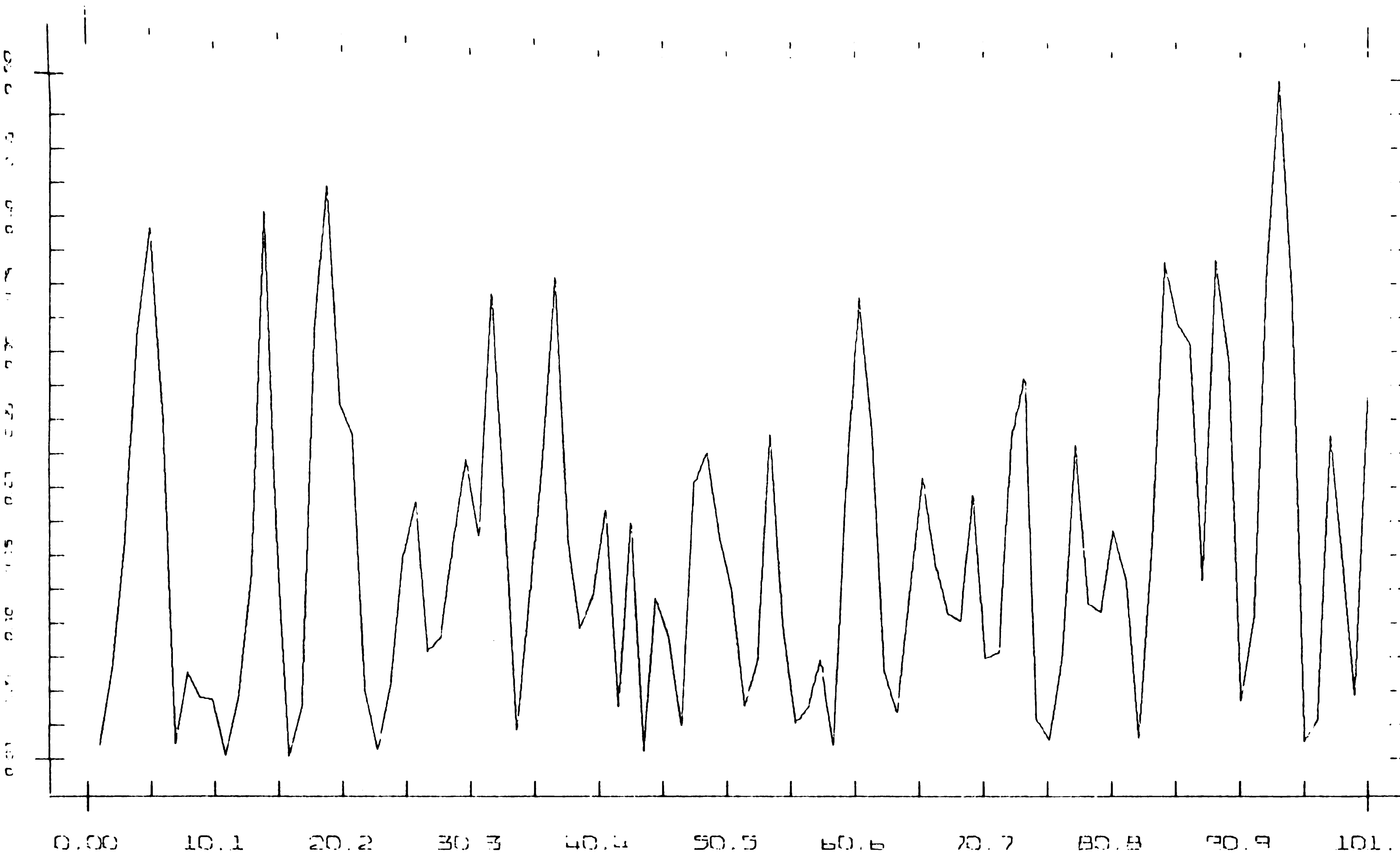


Figure 7.10.--Coherence of Plan A and Plan B--Dollar Sales for Product 1.

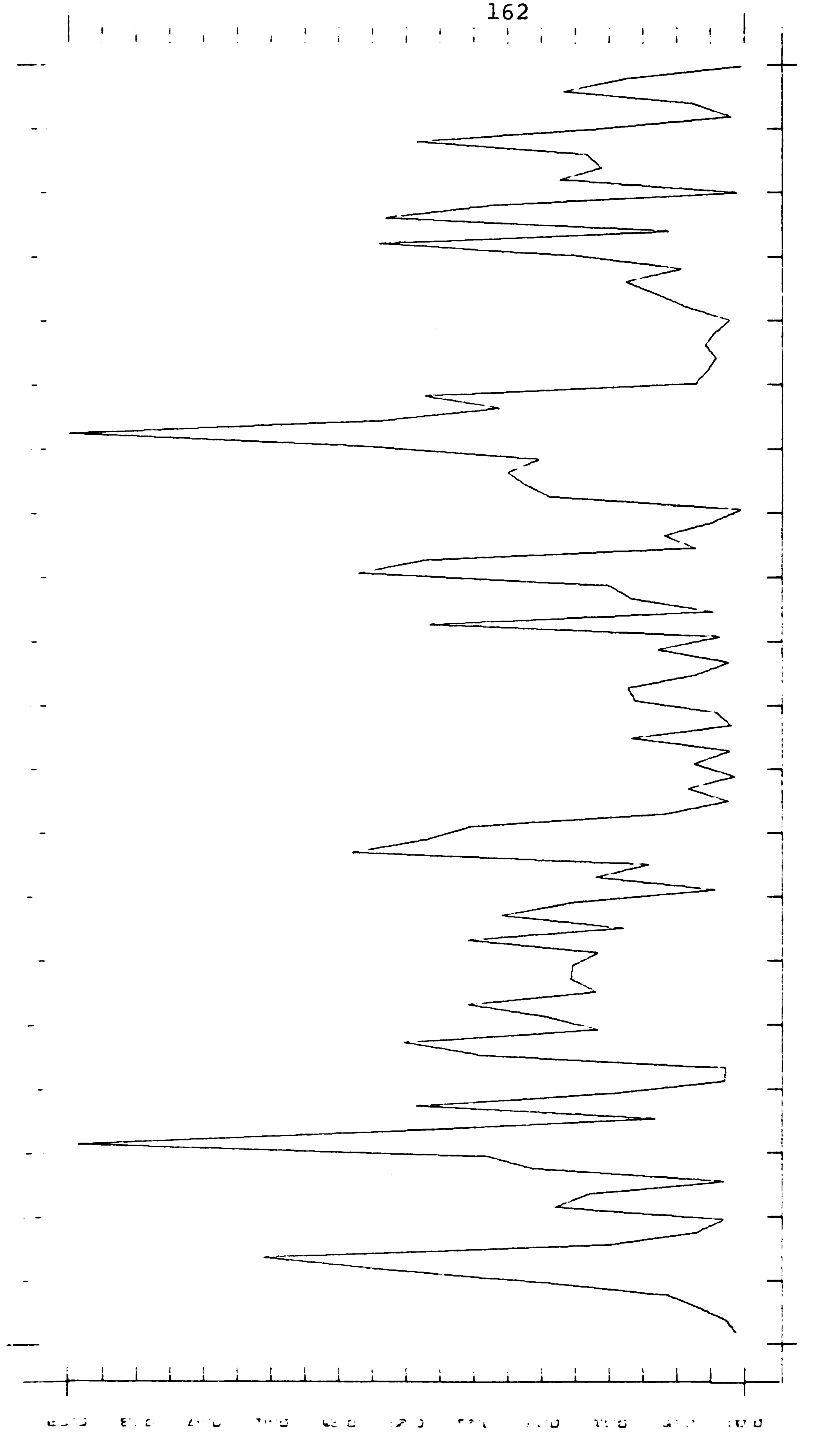


Figure 7.11.--Coherence of Plan A and Plan C--Dollar Sales for Product 1.

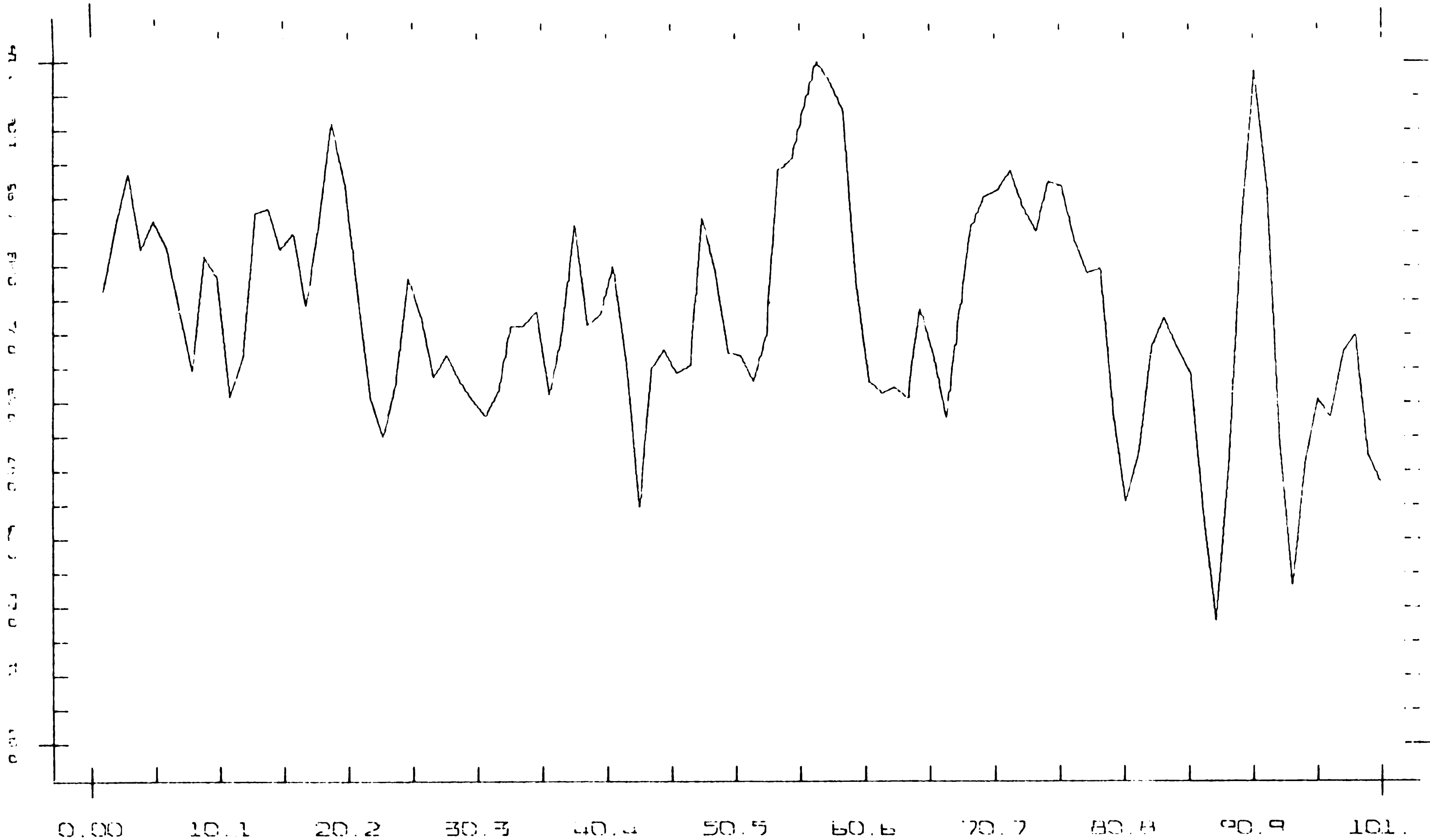


Figure 7.12.--Coherence of Plan A and Plan D--Dollar Sales for Product 1.

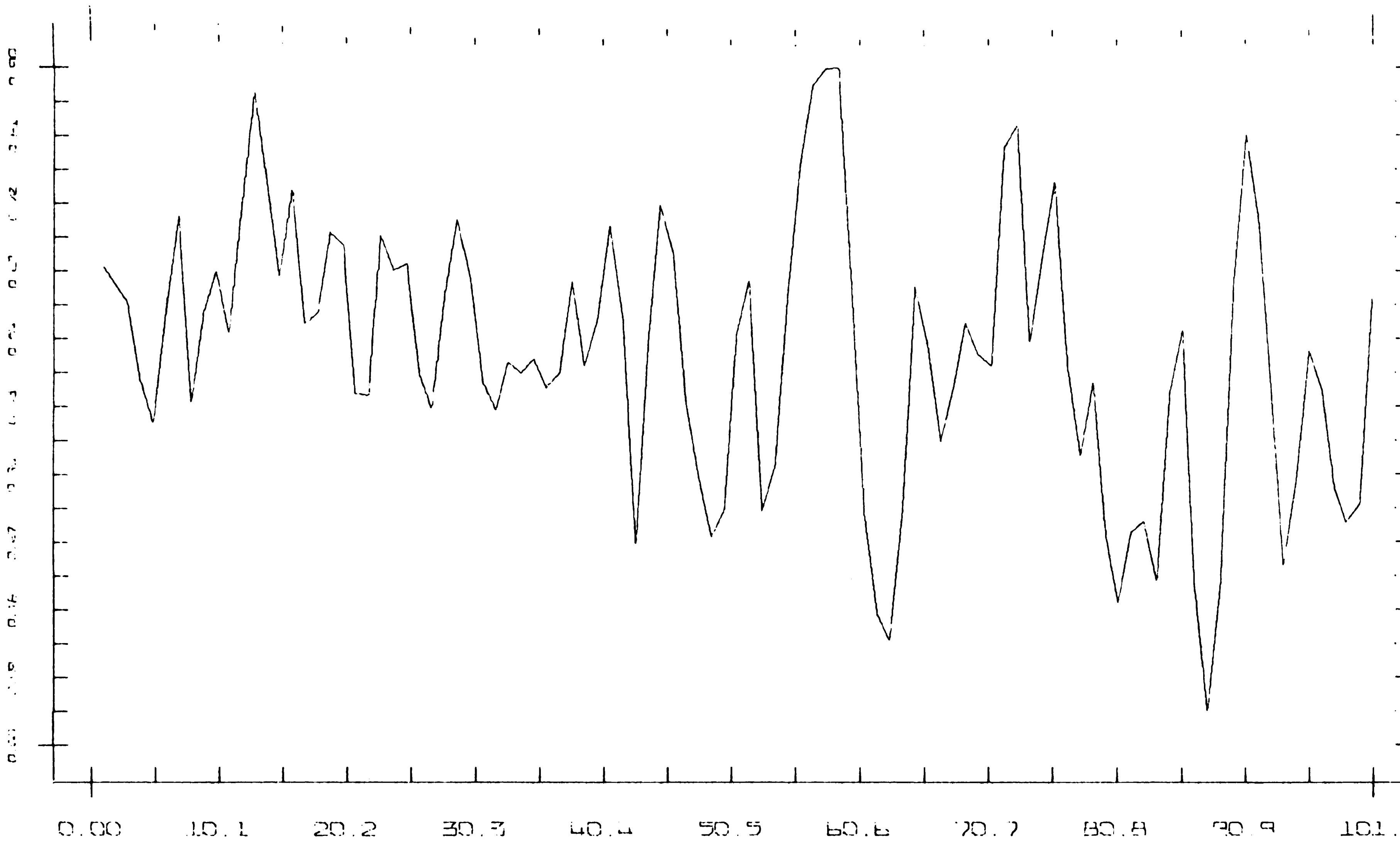


Figure 7.13.--Coherence of Plan A and Plan E--Dollar Sales for Product 1.

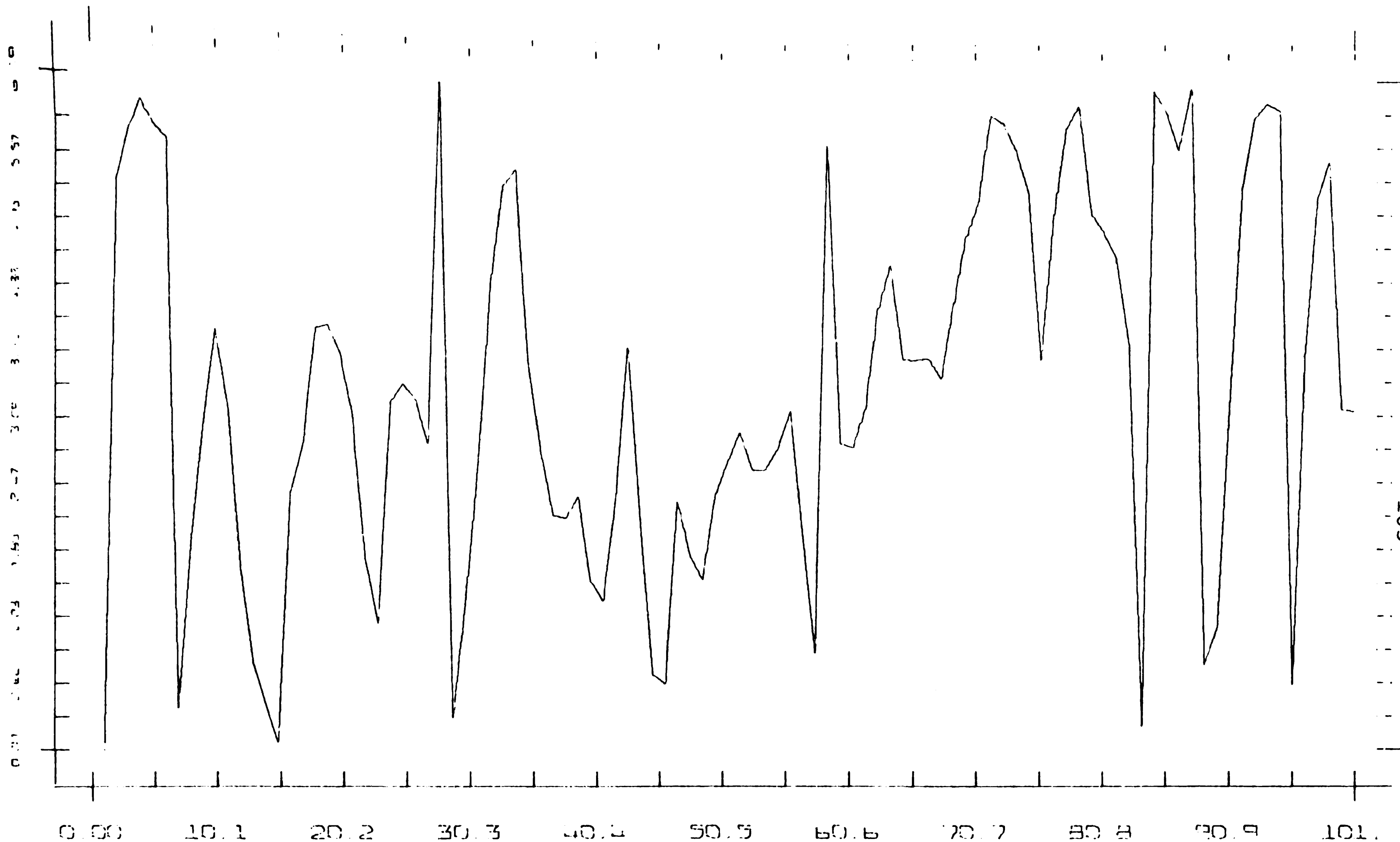


Figure 7.14.--Phase of Plan A and Plan B--Dollar Sales for Product 1.

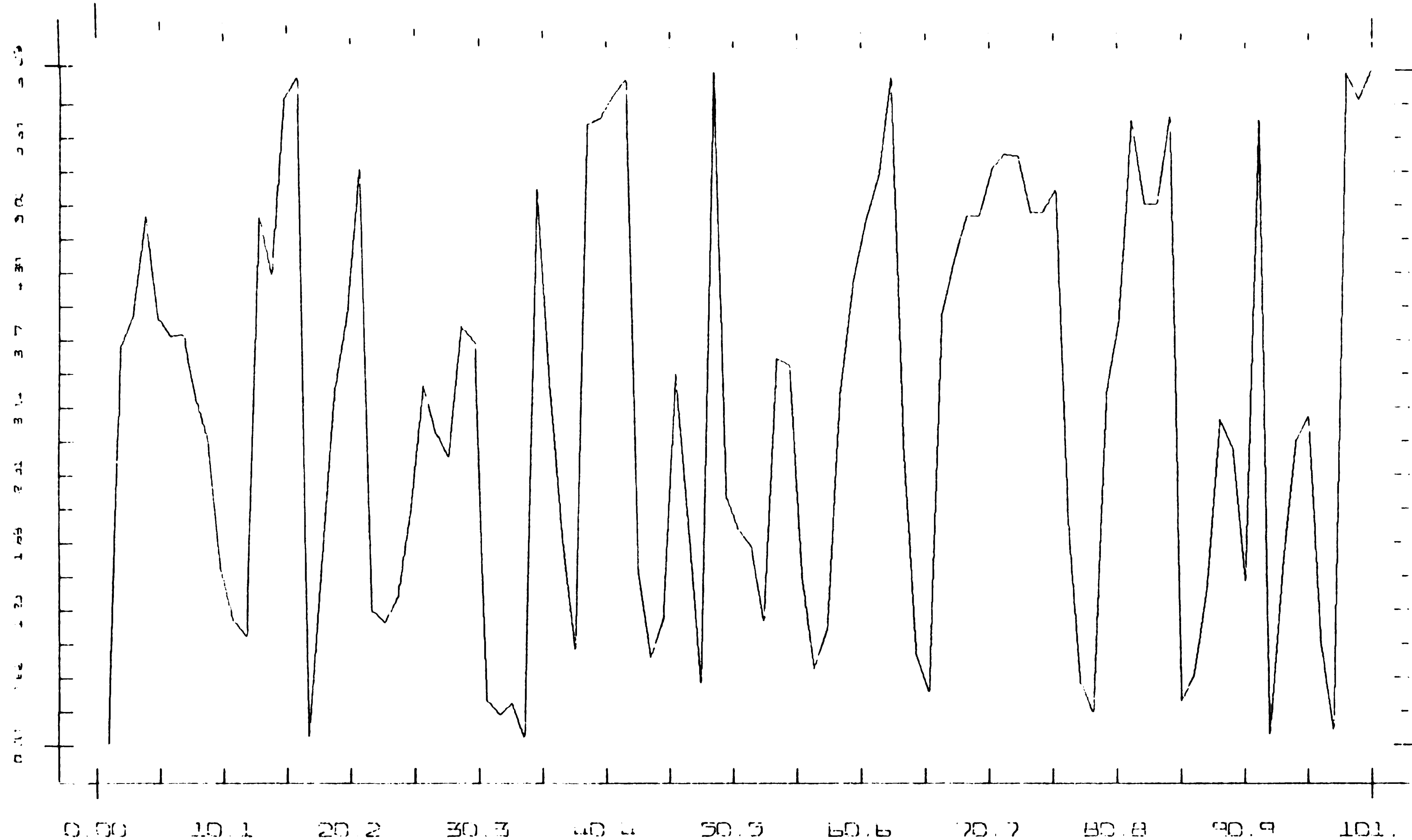


Figure 7.15.--Phase of Plan A and Plan C--Dollar Sales for Product 1.

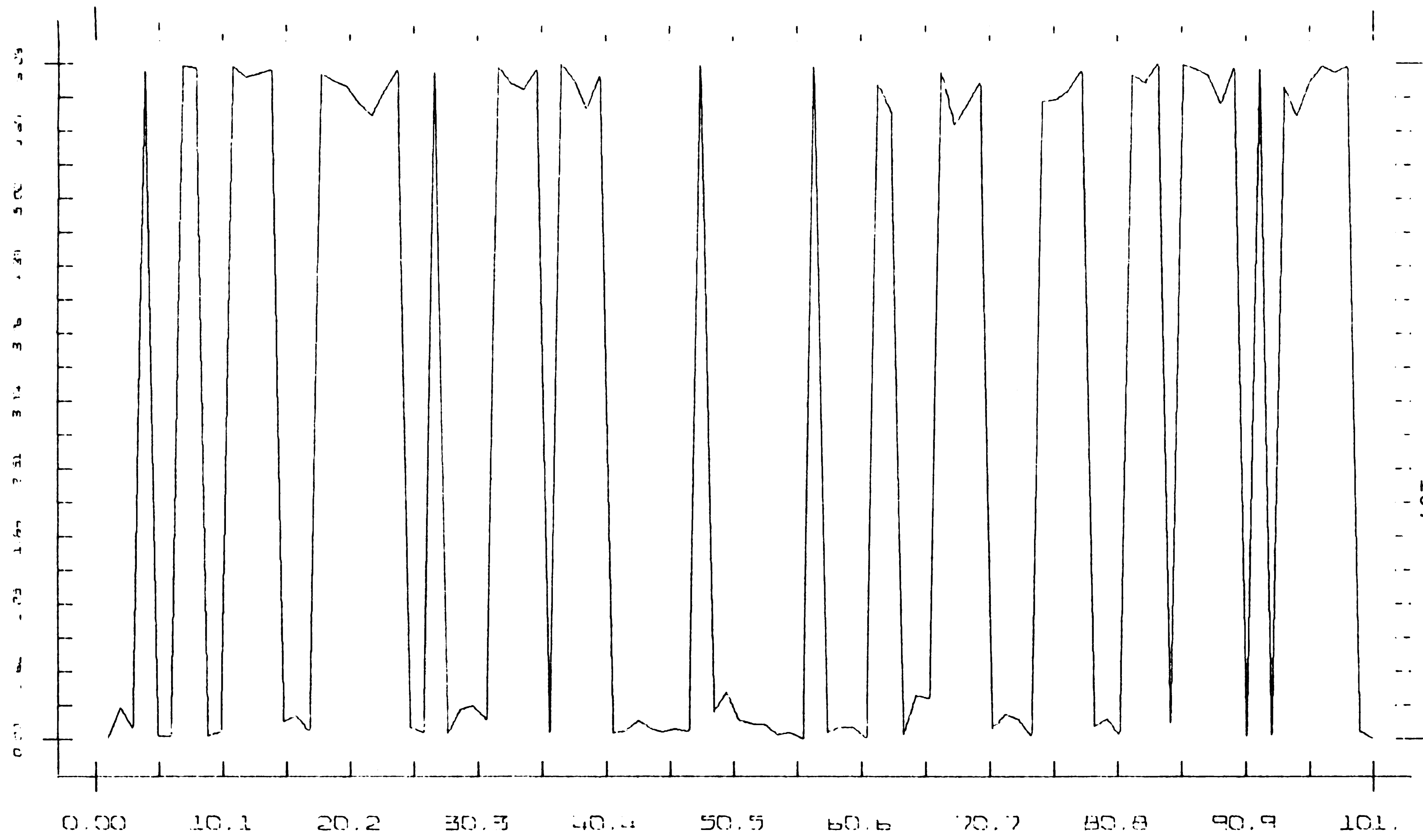


Figure 7.16.--Phase of Plan A and Plan D--Dollar Sales for Product 1.

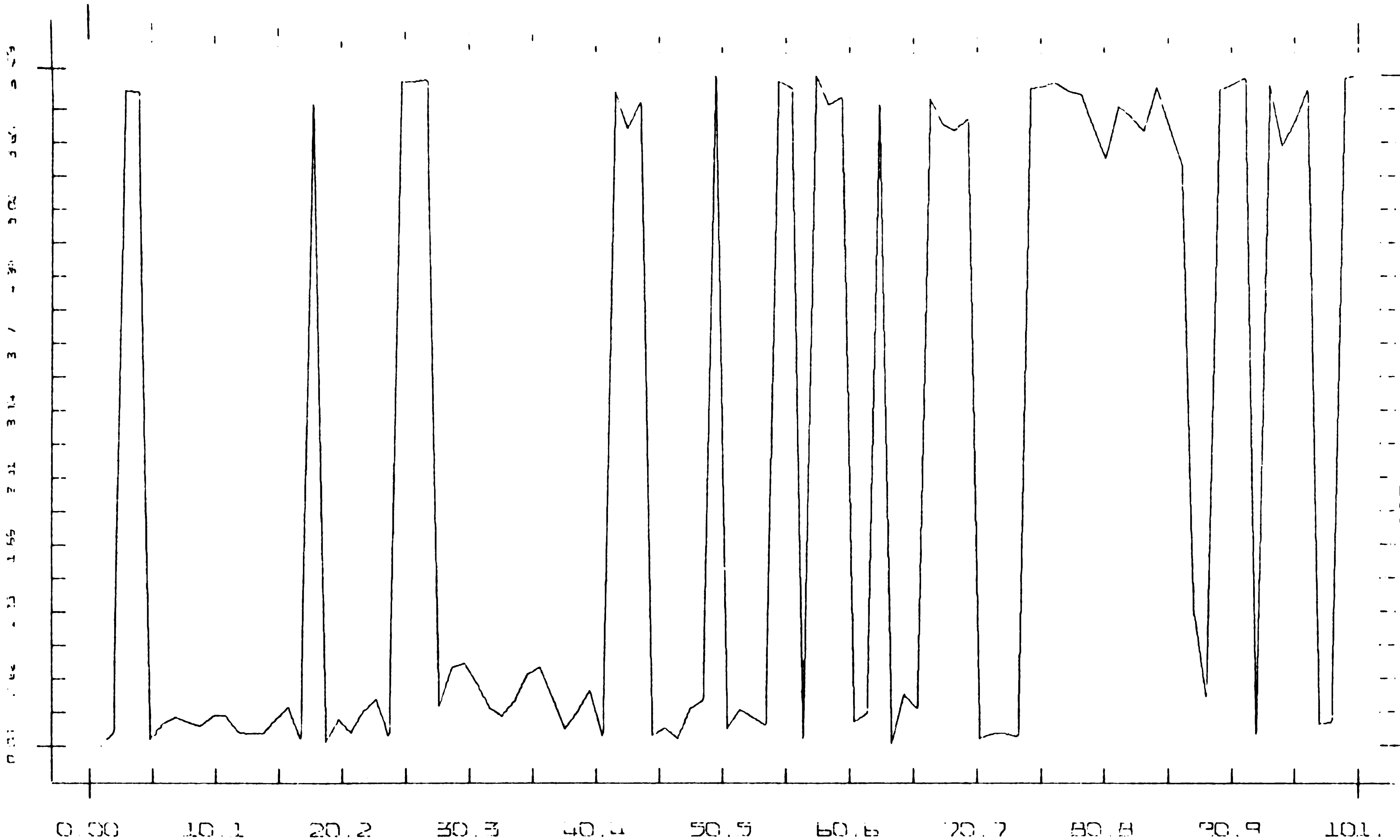


Figure 7.17.--Phase of Plan A and Plan E--Dollar Sales for Product 1.

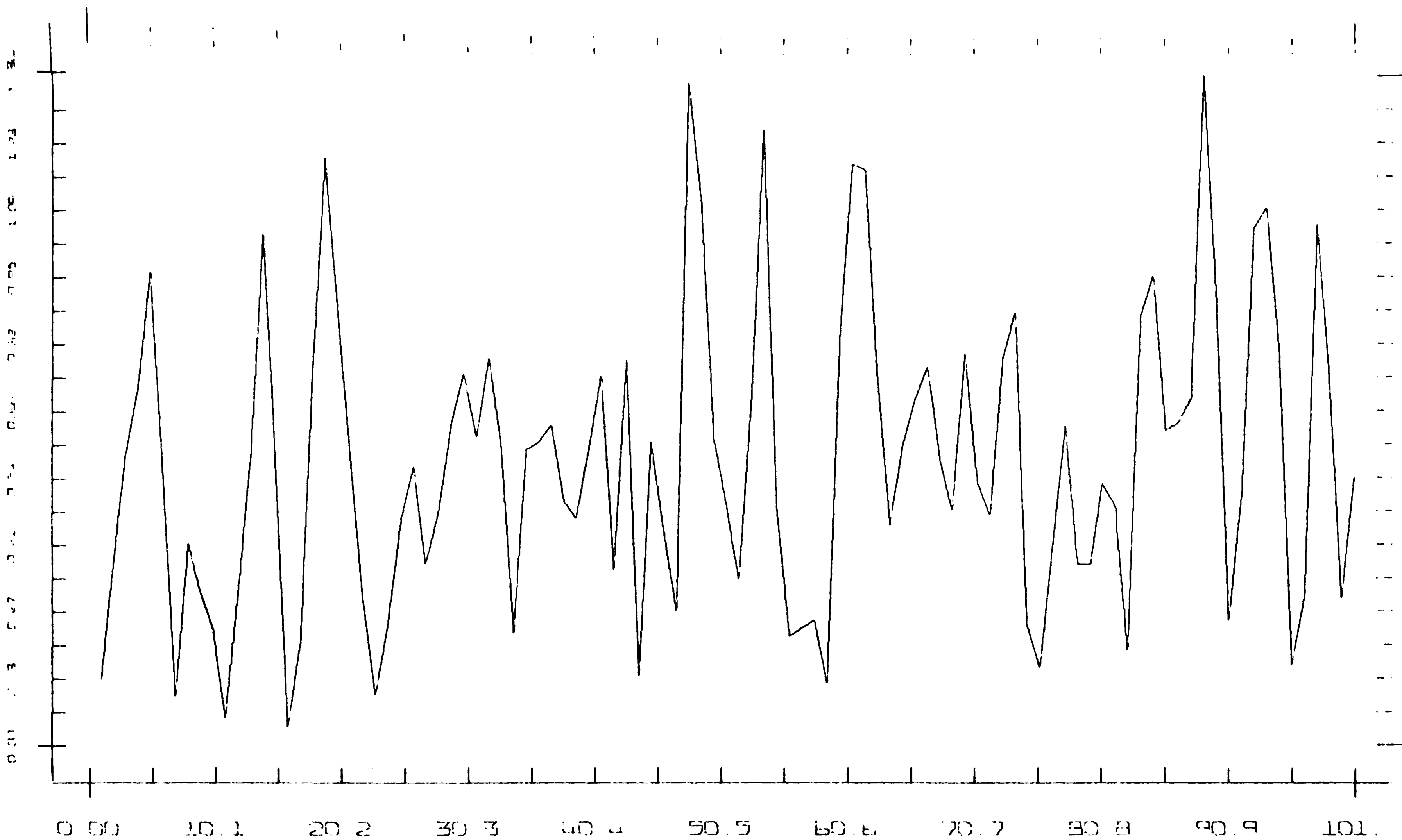


Figure 7.18.--Gain of Plan A and Plan B--Dollar Sales for Product 1.

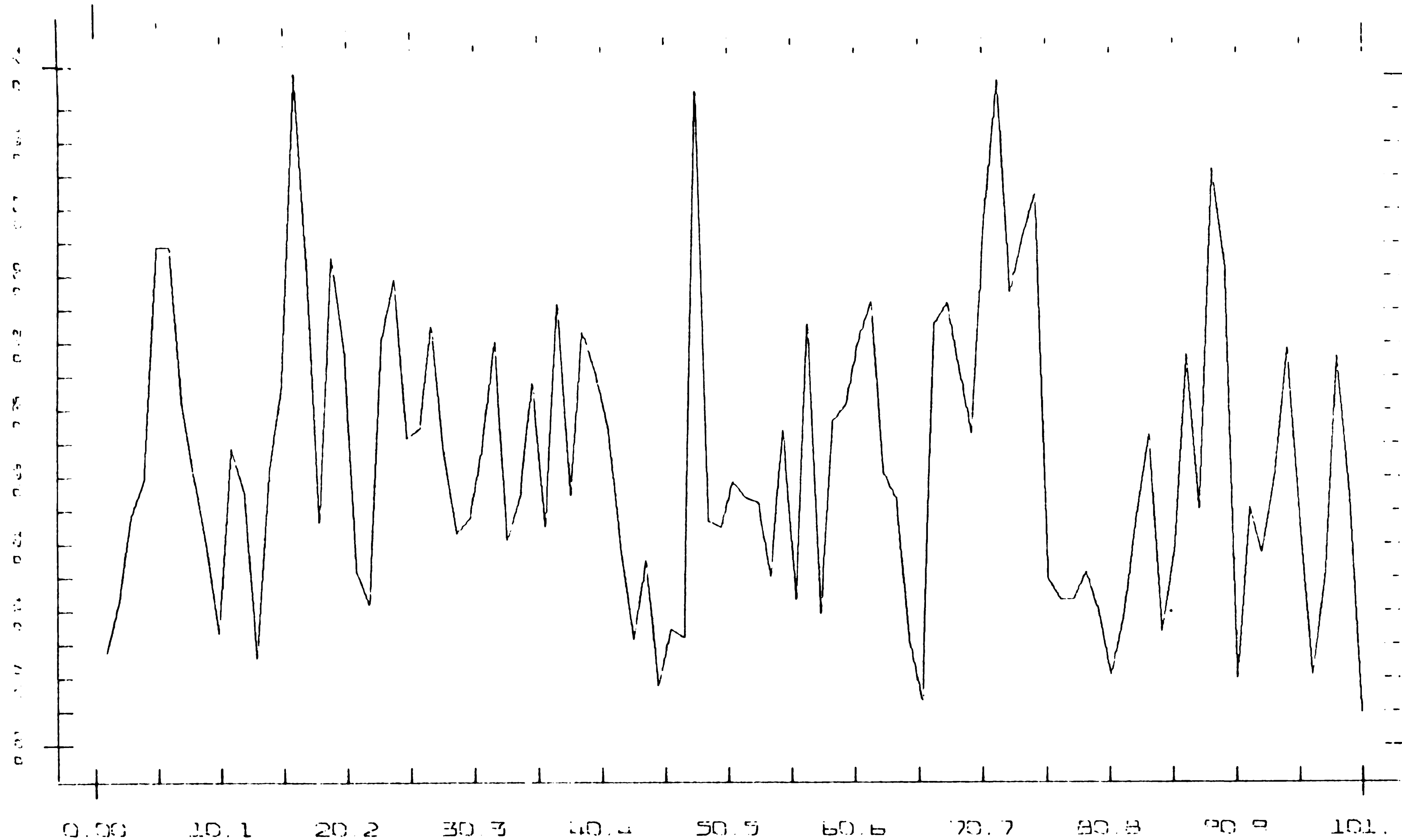


Figure 7.19.--Gain of Plan A and Plan C--Dollar Sales for Product 1.

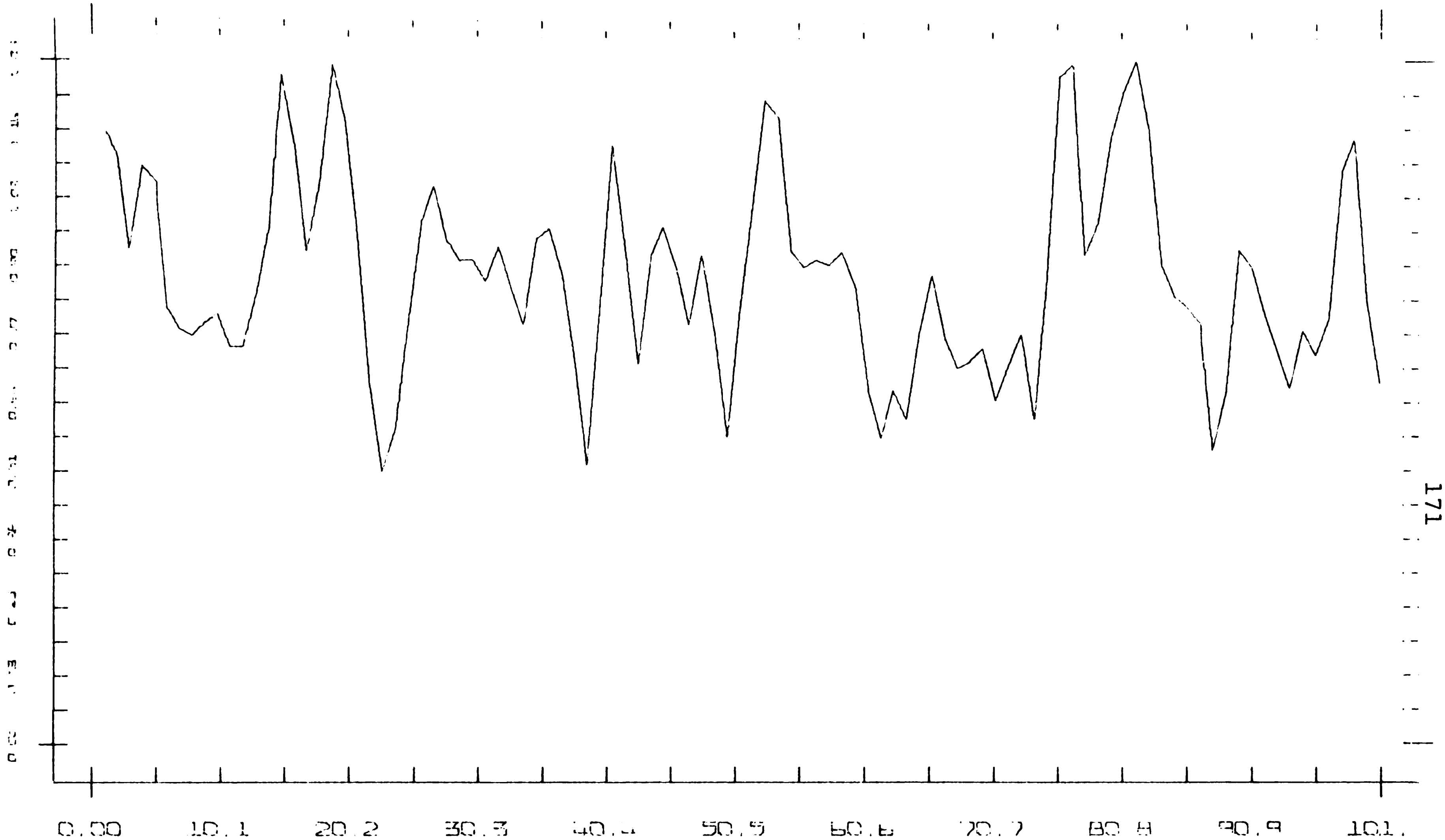


Figure 7.20.--Gain of Plan A and Plan D--Dollar Sales for Product 1.

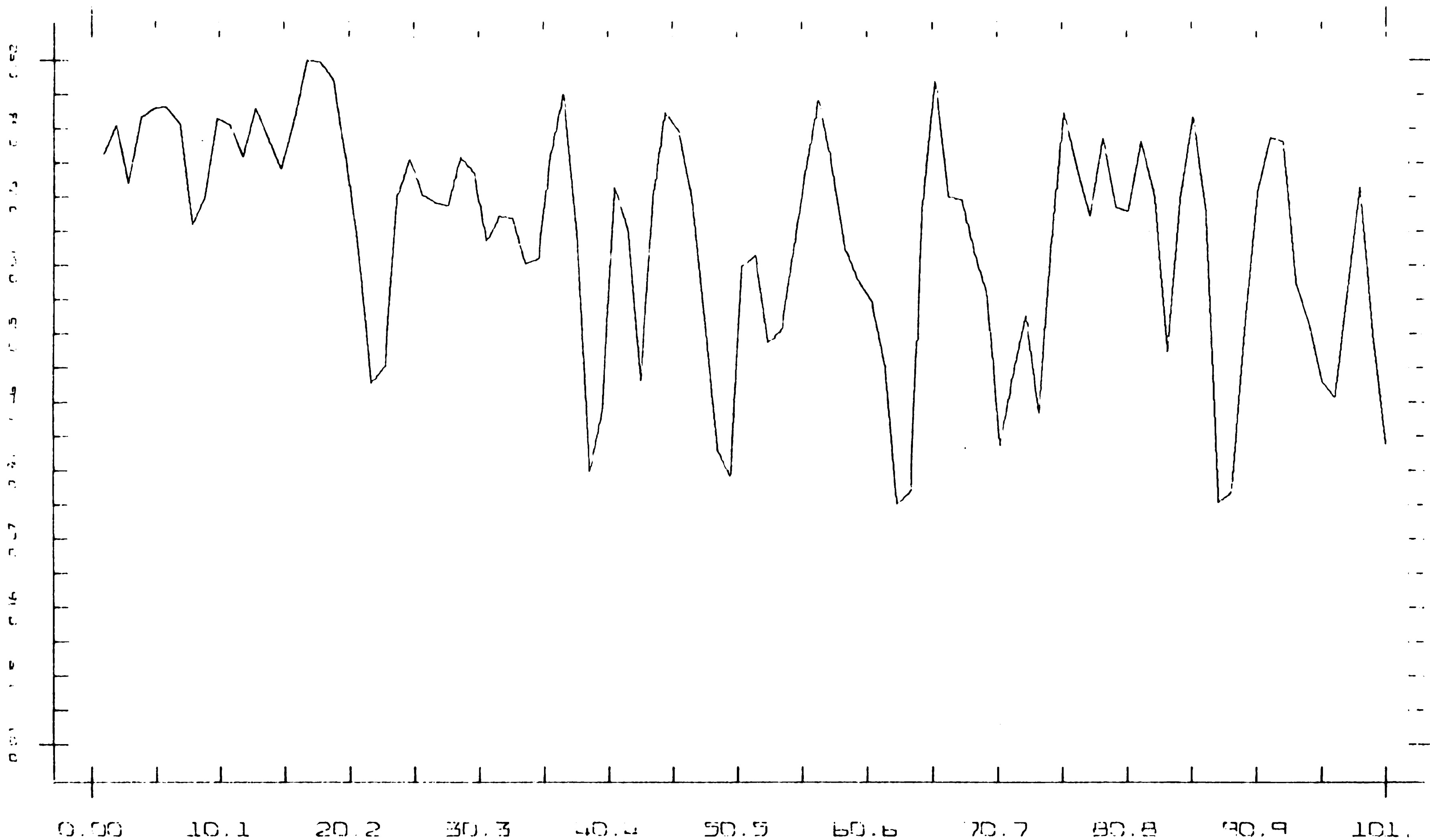


Figure 7.21.--Gain of Plan A and Plan E--Dollar Sales for Product 1.

The phase diagrams shows oscillations about a constant other than zero. This indicates the presence of a fixed angle lag rather than a fixed time lag (i.e. the lag is proportional to the inverse of the frequency which is the period of the component). Although this fact is of interest in the analysis of the time series, it is only germane to this study to the extent that this angle lag is present for all plans.

Values which can be interpreted as regression coefficients are given in the gain diagram. As with coherence, better results are obtained for Plans D and E than for Plans B and C.

Factor Analysis

The factor loadings of the control (Plan A) are compared to the factor loadings of each of the alternative plans.

A factor analysis of the six streams of data generated by Plan A (dollar sales and sales weight for each of the three products) produced factor loadings of most value with three factors. This was also true for each of the other four plans. These three factors in each case describe the data they represent, and the similarity of this descriptive power between plans indicates a similarity in the basic data.

Table 7.14 contains the similarity matrices for the factor loadings of all plans when compared with Plan A.

TABLE 7.14.--Similarity Matrices for Factor Loadings.

		Factor Loading for Plan A		
		1	2	3
Factor Loading for Plan B	1	0.0582	-0.9980	0.0129
	2	0.0300	-0.0171	0.9999
	3	0.9980	-0.0572	0.0284
Factor Loading for Plan C	1	0.0335	-0.9991	-0.0108
	2	0.0435	0.0056	0.9998
	3	0.9991	-0.0336	0.0434
Factor Loading for Plan D	1	0.9999	-0.0060	0.0293
	2	-0.0065	0.9998	0.0090
	3	0.0291	0.0041	1.0000
Factor Loading for Plan E	1	0.0041	0.9992	0.0346
	2	-1.0000	-0.0035	-0.0342
	3	0.0348	0.0186	0.9997

Each element in the matrix has a range from -1 to +1, significant correspondence between the factors occurring with a value of 0.78868 or above. From the table a significant one-to-one correspondence between factors is found for every comparison.

Sensitivity of the Model's Major Assumptions

Some of the early results of this chapter are contradictory. While analysis of variance indicated that the means of Plans D and E were significantly different from the mean of the control, multiple comparison provided results exactly opposite--accept the means of Plans D and E as being the same as the control mean. The F test of variances and the testing of the correlation coefficient rejected about half of the plans as being equal to the control. But all the remaining tests of the chapter accepted the alternative plans as being equal to the control plan.

While there was not 100% support from all analyses for the hypothesis that no significant difference exists between Plan A and Plans B, C, D, and E, neither was the hypothesis consistently rejected for any one plan (even when considering only those few tests which rejected the hypothesis for one or more plans).

Even prior to an evaluation of the relative merit of each form of analysis, the conclusion that the two major assumptions embodied in LREPS do not have a significant influence on the model's endogenous data streams can be accepted.

CHAPTER VII--FOOTNOTES

[1]A. H. Packer, "Simulation and Adaptive Forecasting as Applied to Inventory Control," Operations Research, Vol. 15 (July, 1967), pp. 660-679.

[2]O. K. Helferich, "Development of a Dynamic Simulation Model for Planning Physical Distribution Systems: Formulation of the Mathematical Model" (unpublished D.B.A. dissertation, Michigan State University, 1970), p. 98.

[3]*Ibid.*, p. 121.

[4]J. Riggs, *Production Systems: Planning, Analysis and Control* (New York: John Wiley & Sons, Inc., 1970), p. 70.

CHAPTER VIII

A GENERALIZED VALIDATION PROCEDURE

Introduction

Before the tests of the last three chapters can be evaluated as a generalized validation procedure, the results of the application of these tests for the LREPS model need to be more closely examined. The results obtained must be evaluated in light of the relative merit or value of the technique generating them. The merit of a technique is established from the number and severity of the assumptions of the technique.

The selection procedure for techniques to be used for each type of validity testing is given in the next section, and the following section is a discussion of the assumptions contained in the techniques which were selected. Two questions remain: is the LREPS model valid, and has a generalized validation procedure been developed? These two questions are answered in the final sections of the chapter.

Selection of Statistical Validation Techniques

Not all the statistical techniques presented in Chapter II were used for the validation procedures described in Chapters V, VI, and VII. A summary of those techniques which were used is given in Table 8.1. Four techniques were not used at all: sequential analysis, multiple ranking, the Kolmogorov-Smirnov test, and response surface analysis.

Sequential analysis provides a means of reducing computation if superfluous information is available. This technique was not considered because the primary difficulty in the analysis of the LREPS model was that caused by insufficient data. Use of this technique can save time and effort, but does not change the final results obtained.

Multiple ranking is a method to determine the "best" of several plans under consideration. To establish the validity of a model, the important task is to determine if significant differences exist between sets of data. The size of this difference is unimportant; the mere fact that it exists casts doubt on model validity. This technique could be used to advantage during model experimentation.

The Kolmogorov-Smirnov test establishes if a given sample is a sample from a particular distribution. This test could have been used to test the normality of data used in other techniques which assume normality. This was not done, as more powerful techniques not having this assumption were also used.

Response surface analysis is a technique which can be used to approximate the optimal value of a given function.

TABLE 8.1.--Use of Statistical Techniques.

	Stability (Chapter V)	Predictive Ability (Chapter VI)	Sensitivity of Major Assumptions (Chapter VII)
Graphical Analysis	X	X	X
Sequential Analysis			
Analysis of Variance		X	X
Multiple Comparison		X	X
Multiple Ranking			
The F Test		X	X
Correlation Analysis	X	X	X
Regression Analysis		X	X
The Kolmogorov-Smirnov Test			
The Chi-Square Test		X	X
Theil's Inequality Coefficient	X	X	X
Spectral Analysis	X	X	X
Factor Analysis		X	X
Response Surface Analysis			

Although this technique is recommended for testing simulation models by Naylor,[1] it was determined to be of relevance for design rather than the validity procedures developed in this dissertation.

All of the remaining techniques of Chapter II were used to test the model's predictive ability and the sensitivity of its assumptions. When establishing the long-term stability of the model, analysis is concentrated on a single endogenous data stream (all other analyses are comparisons between pairs of endogenous data streams). This limits the applicability of techniques for stability testing to the four listed in Table 8.1.

Comparative Value of Results

Because of the reasonably large number of techniques used and because the results obtained from these techniques were sometimes conflicting, the results of a technique need to be weighted by a measure of the technique's merit or value. This measure of value is established by the number of major assumptions which are contained in the technique.

The three most common assumptions in the techniques used are the assumed independence between individual observations, the assumed equality of variance, and the assumed normality of the variables under consideration.[2] Analysis of variance, the F test, and multiple comparison include all these three assumptions. The assumption of independence can be satisfied by the independence of the pseudorandom numbers generated.[3] This is not so for the type of testing carried out on the LREPS model. Inequality of variance for analysis of variance has little effect for a reasonable number of plans when the sample size is the same.[4] Departure from normality can have severe effects on inferences about variances, but little effect on inferences about means.[5]

The number of observations in a Chi-square test needs to be large (at least 50) in order for the excess of actual over expected frequencies to be normally distributed. Also the theoretical cell frequency must be an absolute minimum of 5 and a reasonable minimum of 10.[6]

Theil's Inequality Coefficient is always positive. Because it does not discriminate between the direction of forecast error, the coefficient might not be suitable for some applications.[7]

The main assumption of factor analysis is that the observed variables are linear functions of the factor variables. All observed variables must also be linearly related to one another.[8] This assumed relationship can be relaxed to monotonic, as a straight line can be assumed a good approximation to a monotonic function. While another assumption is that each observed variable must be normally distributed, considerable latitude from this assumption is often possible.

For correlation analysis the number of observations used must be reasonably large (even up to 100) or little reliability can be placed on the interpretation of the coefficient of correlation.[9]

The spectral analysis performed assumed the stochastic process under consideration to be covariance stationary.[10] That is, the second moment of the process is finite and a function only of reference time. If the process is not covariance stationary, the trend can be removed by filtering or transforming the time series. An effective method of performing this task is to apply a large term moving average to the data. The Tukey-Hanning estimate of the power spectrum was used for all analyses which allows very small leakage from one frequency band to another.[11] The effect of the covariance stationarity assumption is then minimized even if the data violates the assumption. But the most important fact about spectral analysis is that the technique does not assume independence of observations. This means that autocorrelated data (the form of the output of most simulation models) can be analyzed effectively.

A summary of these assumptions is shown in Table 8.2. Because the effect of an assumption can vary given the particular analysis, the important consideration is how many of these assumptions are violated for the analysis

TABLE 8.2.--Assumptions of Statistical Techniques.

	Independence	Equality of Variance	Normality	Number of Assumptions Violated
Graphical Analysis				4
Analysis of Variance	X	X	X	3
Multiple Comparison	X	X	X	3
The F Test	X	X	X	3
Correlation Analysis	(Number of observations must be large)			
	X			2
Regression Analysis	(Number of observations must be large)			
	X			2
The Chi-square Test	(Number of observations must be large)			
	(Theoretical cell frequency minimum of 10)			2
Theil's Inequality Coefficient	(Does not discriminate between signs)			1
Spectral Analysis	(Stationary time series)			0
Factor Analysis	(Observed variables must be linear functions of factor variables)			
	X		X	3

under consideration. So the last column of Table 8.2 indicates the number of assumptions for each technique which were violated by the nature of the output of the LREPS model. Attention is drawn to the fact that the worth of graphical analysis cannot be established in this manner. A factor equivalent to the "number of assumptions violated" is assigned by judgment.

Validity of LREPS

The statistical results generated by the three procedures for output validity testing are of major importance in establishing the validity of a simulation model. But two factors must be considered before conclusions are drawn from these results: the availability and adequacy of the historical data used to test the model's predictive ability, and the computational limitation on the number of endogenous data streams which can be analyzed by each of the three procedures.

Data collection is presently a major difficulty with any type of analysis. This will continue to be so until organizations implement management information systems which are designed from a basis of the data requirements of their planning and control tools (such as the LREPS model).

With a large simulation model many endogenous data streams are developed. Given the present generation of computing machinery, a computational limitation exists as to how many of these data streams can be statistically analyzed for validity. So it is important to select those data streams for analysis which will be representative of the behavior of the remaining data streams. This selection procedure and the critical examination of the nature of the data streams omitted from analysis must be made by the analyst working in close cooperation with the management of the client corporation. Graphical analysis can supply additional input to this face validity testing. Advancement in the computational capability of future generations of computing machinery may well cause the obsolescence of the use of face validity testing as a supplement to output validity testing--all endogenous data streams will be statistically analyzed for output validity. This is not to say that face validity testing will no longer be required. A coarse and inexpensive first estimate of the model's ability will always be provided by this type of testing.[12] These initial results often are of vital importance in the client's decision to provide adequate funds for the model's development and implementation.

Conclusions have been made in Chapters V, VI, and VII as to the ability of the output of the LREPS model to satisfy the validation procedures outlined in those chapters. Now it must be determined if the conclusions drawn will be affected if the results of each statistical test are not weighted equally, but are weighted inversely with the number of assumptions that are violated in using the technique. The significance placed on the results of a particular statistical test should vary with the quality of the results, and the quality can be determined by examining the number of assumptions violated in the process of applying the test.

Accounting for the quality of information generated in this manner allows greater confidence to be placed on the conclusions drawn in Chapters V and VII, as the tests showing positive results were those which violated the least assumptions. The LREPS model is stable over the long run. The model also satisfies the claims made concerning the generality of its structure--significant changes in specific major assumptions contained in the model did not result in significant changes in the output of the model.

While the results of these two validation procedures are positive, the ability of the model to duplicate actual historical data is still not established. But again it must be emphasized that the failure to establish the predictive ability of the model does not necessarily indicate any shortcoming in the structure of the model. A more accurate evaluation of this important aspect of the model can be made only when a longer stream of actual historical data, collected at a time increment greater than one day, is available. The results of the validation procedures of Chapter VI would undoubtedly be greatly improved for the LREPS model if two hundred observations of information collected weekly were used.

A Generalized Procedure

It is admitted that some degree of design validity testing (face validity testing) is required during the construction of any computer simulation model. To this extent a generalized validation procedure cannot be developed. But once the model passes this coarse testing, output validity can be established by a general procedure--a procedure composed of the three parts outlined in Chapters V, VI, and VII.

The procedure is general, but two inputs to the procedure are specific: the assumptions of the particular model under consideration to evaluate, and the assumptions of the statistical validation techniques which are violated in this particular situation. These two considerations are of similar significance to this generalized procedure, as is the need for specific endogenous data streams in any particular analysis.

Interpretation of the results of this procedure involves a reasonable judgmental factor even with the consideration of the violated assumptions. So it is of interest to consider the construction of a validity index for a given simulation model. Each statistical technique generates results a percentage of which are favorable to the proposition that the model is valid. This percentage is weighted by the inverse of the number of assumptions this technique has violated plus one. The result is summed for all techniques used in the validation procedure, and then this total is divided by the sum of the weights used. The result is an index of validity with a range from 0 to 1.

$$\text{Index of validity} = \frac{\sum_{i=1}^{n} \left\{ \left(\begin{array}{c}\text{Percentage of} \\ \text{favorable results}\end{array} \right) \left(\dfrac{1}{\begin{array}{c}\text{Number of assump-} \\ \text{tions violated +1}\end{array}} \right) \right\}}{\sum_{i=1}^{n} \left\{ \dfrac{1}{\begin{array}{c}\text{Number of assumptions} \\ \text{violated +1}\end{array}} \right\}}$$

where n = number of statistical techniques used.

Actually an index is determined for the long-term stability of the model I^s, another index calculated for the model's predictive ability I^{pa}, and a third index for the sensitivity of the model to its major assumptions I^a. The overall index of validity for the Model I is the mean of the three component indices. Table 8.3 shows these indices for the LREPS model.[13] Again the general results previously established are confirmed.

An analyst may not want to carry out all the tests of Chapters V, VI, and VII. He should select techniques for each of the three validation procedures starting with those that violate the least assumptions. If time and

TABLE 8.3.--Indices of Validity.

	Number of Assumptions Violated	Weight	Percentage of Favorable Results		
			Chap. V	Chap. VI	Chap. VII
Graphical Analysis	4	0.20	100.0	50.0	100.0
Analysis of Variance	3	0.25		100.0	50.0
Multiple Comparison	3	0.25		0.0	88.0
The F Test	3	0.25		50.0	75.0
Correlation Analysis	2	0.33	100.0	0.0	50.0
Regression Analysis	2	0.33		50.0	100.0
The Chi-square Test	2	0.33		88.0	100.0
Theil's Inequality Coefficient	1	0.50	50.0	50.0	75.0
Spectral Analysis	0	1.00	100.0	50.0	100.0
Factor Analysis	3	0.25		13.0	100.0
			$I^s=0.877$	$I^{pa}=0.466$	$I^a=0.867$
				$I=0.737$	

money permit, he can then move to techniques which violate more assumptions and provide information of poorer quality. With this selection procedure the value of the validity index may tend to vary inversely with the number of techniques used.[14]

The procedures detailed in this thesis provide a generalized validation procedure, and the validity index provides a basis for intra-model analysis and inter-model comparison.

CHAPTER VIII--FOOTNOTES

[1]T. H. Naylor, Computer Simulation Experiments with Models of Economic Systems (New York: John Wiley & Sons, Inc., 1971), pp. 172-175.

[2]T. H. Naylor, K. Wertz, and T. H. Wonnacott, "Methods of Analyzing Data from Computer Simulation Experiments," Communications of the ACM, Vol. 10 (November, 1967), p. 703.

[3]M. D. MacLaren and G. Marsaglia, "Uniform Random Number Generators," Journal of the ACM, Vol. 12 (1965), pp. 83-89.

[4]H. Scheffe, The Analysis of Variance (New York: John Wiley & Sons, Inc., 1959), p. 345.

[5]Ibid., Chapter 10.

[6]G. U. Yule and M. G. Kendall, An Introduction to the Theory of Statistics (London: Charles Griffin and Company Ltd., 1953), p. 469.

[7]H. Theil, Applied Economic Forecasting (Amsterdam: The North-Holland Publishing Co., 1966), p. 28.

[8]H. H. Harman, Modern Factor Analysis (Chicago: The University of Chicago Press, 1960), p. 380.

[9]Yule and Kendall, p. 231.

[10]E. Parzen, Stochastic Processes (San Francisco: Holden-Day, Inc., 1962), p. 70.

[11]C. W. J. Granger and M. Hatanaka, Spectral Analysis of Economic Time Series (Princeton, N.J.: Princeton University Press, 1964), p. 60.

[12]The results of face validity testing for the LREPS model are shown in Table 4.1.

[13]The indices were calculated omitting results pertaining to Product 3 because of this product's instability of demand.

[14]The truth of this statement can be established or rejected by sensitivity analysis. If the index does vary in this manner the appropriate corrective weighting system can also be determined.

BIBLIOGRAPHY

BIBLIOGRAPHY

Amstutz, A. E. Computer Simulation of Competitive Market Response. Cambridge, Massachusetts: The M.I.T. Press, 1967.

Balderston, F. E., and Hoggatt, A. C. Simulation of Market Processes. Berkeley, California: Institute of Business and Economic Research, 1962.

Bechhofer, R. E., and Blumenthal, S. "A Sequential Multiple-Decision Procedure for Selecting the Best One of Several Normal Populations with a Common Unknown Variance, II: Monte Carlo Sampling Results and New Computing Formulae." Biometrics, Vol. 18, March 1962.

Blackman, R. B., and Tukey, J. W. The Measurement of Power Spectra. New York: Dover Publications, Inc., 1958.

Bonini, C. P. Simulation of Information and Decision Systems in the Firm. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1962.

Bowersox, D. J., et al. Dynamic Simulation of Physical Distribution Systems. Monograph. East Lansing, Michigan: Division of Research, Michigan State University, Forthcoming.

Bowersox, D. J.; Smykay, E. W.; and LaLonde, B. H. Physical Distribution Management. New York: The Macmillan Company, 1968.

Box, G. E. P. "The Exploration and Exploitation of Response Surfaces: Some General Considerations and Examples." Biometrics, Vol. 10, 1954.

Box, G. E. P., and Wilson, K. B. "On the Experimental Attainment of Optimum Conditions." Journal of the Royal Statistical Society, Series B, Vol. XIII, 1951.

Buchan, J., and Koenigsberg, E. Scientific Inventory Management. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1963.

Carnap, R. "Testability and Meaning." Philosophy of Science. Vol. 3, No. 4, October, 1936.

Chu, K. Quantitative Methods for Business and Economic Analysis. Scranton, Pennsylvania: International Textbook Co., 1969.

Clarkson, G. P. E. Portfolio Selection: A Simulation of Trust Investment. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1962.

Cohen, K. J. Computer Models of the Shoe, Leather, Hide Sequence. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1960.

Cohen, K. J., and Cyert, R. M. "Computer Models in Dynamic Economics." The Quarterly Journal of Economics. Vol. LXXV, No. 1, February, 1961.

Conway, R. W. An Experimental Investigation of Priority Assignment in a Job Shop. Santa Monica, California: The Rand Corporation, RM-3789-PR, 1964.

Conway, R. W.; Johnson, B. M.; and Maxwell, W. L. "Some Problems of Digital Systems Simulation." Management Science. Vol. 6, October, 1959.

Cooper, G. R., and McGillem, D. C. Methods of Signal and System Analysis. New York: Holt, Rinehart and Winston, Inc., 1967.

Cyert, R. M. "A Description and Evaluation of Some Firm Simulations." Proceedings of the IBM Scientific Computing Symposium on Simulation Models and Gaming. White Plains, N.Y.: IBM, 1966.

Cyert, R. M.; Feigenbaum, E. A.; and March, J. G. "Models of a Behavioral Theory of the Firm." Behavioral Science. Vol. 4, No. 2, April, 1959.

Draper, N. R., and Smith, H. Applied Regression Analysis. New York: John Wiley & Sons, Inc., 1967.

Duncan, A. J. Quality Control and Industrial Statistics. Homewood, Illinois: Richard D. Irwin, Inc., 1965.

Dunnett, C. W. "A Multiple Comparison Procedure for Comparing Several Treatments with a Control." Journal of the American Statistical Association. Vol. 50, December, 1955.

Fishman, G. S. Digital Computer Simulation: Input-Output Analysis. Santa Monica, California: The Rand Corporation, RM-5540-PR, 1968.

Fishman, G. S. Digital Computer Simulation: The Allocation of Computer Time in Comparing Simulation Experiments. Santa Monica, California: The Rand Corporation, RM-5288-PR, 1967.

Fishman, G. S. Problems in the Statistical Analysis of Simulation Experiments: The Comparison of Means and the Length of Sample Records. Santa Monica, California: The Rand Corporation, RM-4880-PR, 1966.

Fishman, G. S., and Kiviat, P. J. Digital Computer Simulation: Statistical Considerations. Santa Monica, California: The Rand Corporation, RM-3281-PR, 1962.

Fishman, G. S., and Kiviat, P. J. Spectral Analysis of Time Series Generated by Simulation Models. Santa Monica, California: The Rand Corporation, RM-4393-PR, 1965.

Forrester, J. W. Industrial Dynamics. Cambridge, Mass.: The M.I.T. Press, 1961.

Goodman, N. R. Scientific Paper No. 10. New York: New York University Engineering Statistics Laboratory, 1957.

Granger, C. W. J., and Hatanaka, M. Spectral Analysis of Economic Time Series. Princeton, N.J.: Princeton University Press, 1964.

Harman, H. H. Modern Factor Analysis. Chicago: The University of Chicago Press, 1960.

Hausman, W. H., and Gilmour, P. "A Multi-Period Truck Delivery Problem." Transportation Research. Vol. 1, No. 4, December, 1967.

Helferich, O. K. "Development of a Dynamic Simulation Model for Planning Physical Distribution Systems: Formulation of the Mathematical Model." Unpublished D.B.A.dissertation, Michigan State University, 1970.

Hoggatt, A. C. "Statistical Techniques for the Computer Analysis of Simulation Models." Appendix in Studies in a Simulated Market. L. E. Preston and N. R. Collins. Berkeley, California: Institute of Business and Economic Research, 1966.

Holzinger, K. J., and Harman, H. H. Factor Analysis: A Synthesis of Factorial Methods. Chicago: The University of Chicago Press, 1941.

Jenkins, G. M., and Box, G. E. P. Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day, Inc., 1970.

Jenkins, G. M., and Watts, D. G. Spectral Analysis and its Applications. San Francisco: Holden-Day, Inc., 1968.

Karreman, H. F. Computer Programs for Spectral Analysis of Economic Time Series. Princeton, N.J.: Economic Research Program. Princeton University, Research Memorandum, No. 59, 1963.

Kraft, C. H., and Van Eeden, C. A Nonparametric Introduction to Statistics. New York: The McMillan Co., 1968.

Kuehn, A. A., and Hamburger, M. J. "A Heuristic Program for Locating Warehouses." Management Science. Vol. 9, No. 11, July, 1963.

MacLaren, M. D., and Marsaglia, G. "Uniform Random Number Generators." Journal of the ACM. Vol. 12, 1965.

Marien, E. J. "Development of a Dynamic Simulation Model for Planning Physical Distribution Systems: Formulation of the Computer Model." Unpublished Ph.D. dissertation, Michigan State University, 1970.

McMillan, C., and Gonzalez, R. F. Systems Analysis: A Computer Approach to Decision Models. Homewood, Illinois: Richard D. Irwin, Inc., 1968.

Naylor, T. H. Computer Simulation Experiments with Models of Economic Systems. New York: John Wiley & Sons, Inc., 1971.

Naylor, T. H., et al. Computer Simulation Techniques. New York: John Wiley & Sons, Inc., 1966.

Naylor, T. H.; Burdick, D. S.; and Sasser, W. E., Jr. "The Design of Computer Simulation Experiments." The Design of Computer Simulation Experiments. Edited by T. H. Naylor. Durham, N.C.: Duke University Press, 1969.

Naylor, T. H., and Finger, J. M. "Verification of Computer Simulation Models." Management Science, Vol. 14, October, 1967.

Naylor, T. H.; Wertz, K.; and Wonnacott, T. H. "Methods for Analyzing Data From Computer Simulation Experiments." Communications of the ACM. Vol. 10, November, 1967.

Naylor, T. H.; Wertz, K.; and Wonnacott, T. H. "Spectral Analysis of Data Generated by Simulation Experiments with Economic Models." Econometrica. Vol. 37, April, 1969.

Packer, A. H. "Simulation and Adaptive Forecasting as Applied to Inventory Control." Operations Research. Vol. 15, July, 1967.

Parzen, E. Stochastic Processes. San Francisco: Holden-Day, Inc., 1962.

Parzen, E., ed. Time Series Analysis Papers. San Francisco: Holden-Day, Inc., 1967.

Paulson, E. "Sequential Estimation and Closed Sequential Decision Procedures." The Annals of Mathematical Statistics. Vol. 35, September, 1964.

Popper, K. R. The Logic of Scientific Discovery. New York: Basic Books, 1959.

Riggs, J. Production Systems: Planning, Analysis and Control. New York: John Wiley & Sons, Inc., 1970.

Robinson, E. A. Multichannel Time Series with Digital Computer Programs. San Francisco: Holden-Day, Inc., 1967.

Rogers, R. T. "Development of a Dynamic Simulation Model for Planning Physical Distribution Systems: Experimental Design and Analysis of Results." Unpublished Ph.D. dissertation, Michigan State University, Forthcoming.

Sasser, W. E.; Burdick, D. S.; Graham, D. A.; and Naylor, T. H. "The Application of Sequential Sampling to Simulation: An Example Inventory Model." Communications of the ACM. Vol. 13, May, 1970.

Scheffe, H. The Analysis of Variance. New York: John Wiley & Sons, Inc., 1959.

Siegel, S. Nonparametric Statistics. New York: McGraw-Hill Book Company, 1956.

Theil, H. Applied Economic Forecasting. Amsterdam: The North-Holland Publishing Co., 1966.

Tocher, K. D. The Art of Simulation. London: The English Universities Press Ltd., 1963.

Tukey, J. W. "Discussion, Emphasizing the Connection Between the Analysis of Variance and Spectrum Analysis." Technometrics. Vol. 3, No. 2, May, 1961.

Tukey, J. W. "The Problem of Multiple Comparisons." Princeton, N.J.: Dittoed Manuscript. Princeton University, 1965.

Turing, A. M. "Can a Machine Think?" The World of Mathematics. Edited by J. R. Newman. New York: Simon and Schuster, 1956.

Van Horn, R. "Validation." The Design of Computer Simulation Experiments. Edited by T. H. Naylor. Durham, N.C.: Duke University Press, 1969.

Veinott, A. F. "The Status of Mathematical Inventory Theory." Management Science. Vol. 12, No. 11, July, 1966.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill Book Company, 1962.

Yaglom, A. M. An Introduction to the Theory of Stationary Random Functions. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1962.

Yule, G. U., and Kendall, M. G. An Introduction to the Theory of Statistics. London: Charles Griffin and Co., Ltd., 1950.