IR 83-2672

Selected Assessment Strategies Applied to Short-Term Energy Models

U.S. DEPARTMENT OF COMMERCE
National Bureau of Standards
National Engineering Laboratory
Center for Applied Mathematics
Operations Research Division
Washington, DC 20234

March 1983

Sponsored by:
nergy Information Administration
.S. Department of Energy
/ashington, DC 20461

NBSIR 83-2672

SELECTED ASSESSMENT STRATEGIES APPLIED TO SHORT-TERM ENERGY MODELS

Patsy B. Saunders, Editor

U.S. DEPARTMENT OF COMMERCE
National Bureau of Standards
National Engineering Laboratory
Center for Applied Mathematics
Operations Research Division
Washington, DC 20234

March 1983

Sponsored by:
Energy Information Administration
U.S. Department of Energy
Washington, DC 20461

U.S. DEPARTMENT OF COMMERCE, Malcolm Baldrige, *Secretary*
NATIONAL BUREAU OF STANDARDS, Ernest Ambler, *Director*

PREFACE

This report describes the fiscal year 1980 activities of the Center for Applied Mathematics at the National Bureau of Standards (NBS) to develop, extend, and refine procedures for assessing analysis systems (models) utilized by the Energy Information Administration (EIA) of the Department of Energy (DoE). These activities have as their goal the development of methods for determining the degree of confidence in a system's results and the circumstances under which such a system may be used to represent current and/or anticipated energy market conditions and the consequences of alternative policy scenarios.

The initial phase of these activities focused on the DoE Midterm Oil and Gas Supply Model (MOGSM) as a test vehicle for idea development and experimentation. That effort produced a series of reports which describe assessment activities related to documentation, data, mathematical structure, forecasting techniques, sensitivity, confidence, and portability. The second phase aimed at further development and refinement of these assessment procedures. The substitution of the Short Term Integrated Forecasting System (STIFS) for MOGSM as a sample subject for continued development of an assessment methodology was intended to increase the likelihood of the general applicability of the resulting methods and guidelines, or alternatively, to cast some light on the degree to which an assessment methodology is limited by being model or model-type specific.

The authors would like to acknowledge the significant contributions made to this study by Dr. Bert W. Rust of the Scientific Computing Division, Center for Applied Mathematics. His review of the document and suggestions for improving the treatment of certain topics in the paper were most useful.

ABSTRACT

This report is one in a series focusing on the evaluation of complex mathematical models. The basic approach pursued in this document is patterned after an earlier analysis of the Department of Energy's Midterm Oil and Gas Supply Model (MOGSM). Several extensions of the earlier methodology are presented which assist the analyst in defining the degree to which certain evaluation activities are model dependent. The Department of Energy's Short Term Integrated Forecasting System (STIFS) was used as a vehicle for exercising the revised methodology. The technical content of the report is divided into three parts, reflecting three basic issues of model form, sensitivity and forecast performance. The first issue addressed relates to the structure of STIFS. It includes not only the mathematical assumptions implicit in the model but also data and software considerations. The approach to the second issue focuses on the measurement of climatological uncertainties and uses as its basis a Monte-Carlo experiment. The final issue deals with several techniques for evaluating the predictive performance of a model. Both classical statistical methods and an information theoretic approach are used to illustrate how such an analysis would be carried out in practice.

Keywords: Assessment; documentation; energy; information theory; mathematical models; senstivity analysis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 The Short Term Integrated Forecasting System (STIFS)[1]

STIFS is a historical data base and a system of computer programs which simulates the network of national energy supplies, conversion processes, and demands. Its purpose is to produce automated monthly forecasts of integrated energy supply/demand balances, including stock changes over the short term (e.g., 12-18 months). The forecast energy types correspond to those for which historical data are published in EIA's Monthly Energy Review (MER). These types include motor gasoline, distillate fuel oil, residual fuel oil, jet fuel, natural gas, coal, and electricity. STIFS explicitly recognizes that the demands for all fuels and the production patterns for them are governed by a complex set of interrelations. These complex interrelations are generally not recognized in a stand-alone, fuel-specific model.

STIFS monthly forecasts may be used to predict shortages of each energy type before they develop. In addition to the need for a conceptual short term forecasting framework there also exists the need to be able to specify alternative scenarios to such a model, and to evaluate both the comparative impacts that these scenarios would have and the consequences of policy decisions that could be made in response to the forecasts made by the system.

[1] This section draws heavily on the report by Collins, et al. (Collins, Dwight E., Mary L. Barcella and Michael L. Shaw, Short Term Integrated Forecasting System (STIFS): Methodology and Model Descriptions, Logistics Management Institute, Washington, D.C., November 1979).

The system's integration function is carried out by an energy balancing model. This model monitors primary energy production through energy conversion from primary to usable form, and delivery to final consumption, balancing energy flows in the process. Like most other national and regional energy modeling systems, STIFS is based on the premise that the nation's energy system has the characteristics of a network. In addition, STIFS has been designed to ensure that all energy flows are properly identified and accounted for. It consists primarily of two processes. First, a set of computations known as a "closing routine" operates in conjunction with the historical data base to balance historical energy supplies and demands, isolating data discrepancies in the process. Second, more than 100 network flow variables are forecast by means of certain statistical and econometric procedures, and another closing routine balances forecast energy accounts, isolating implied shortages or surpluses by energy type.

In the first part of the system, the historical data base is used as an input to the closing routine and as the basis for many of the STIFS forecasts. The data base provides data on all significant elements within the nation's energy network on a monthly basis back to June 1, 1977. In some cases where greater volume of data is needed for purposes of providing a statistically valid data base the data extends further back into time. The historical closing routine is a computerized double-entry energy accounting system denominated in quadrillion Btus which balances energy supplies and demands.

The second part of the system consists of the means to forecast more than 100 different variables and a closing routine to balance the forecast energy accounts. The forecasting system represents future energy flows under a variety of scenarios, simulating energy supply/demand responses to such circumstances as alternative availabilities of petroleum imports, a coal strike, droughts, etc. While the present forecast closing routine is oriented primarily to petroleum imports scenarios, future development of the system may incorporate several versions of the forecast closing routine for a wider range of scenario types to be modeled. The items to be "closed" will be those which are generally used to manage perturbations in the nation's energy supply such as imports, changes in stock levels, and changes in refinery utilization.

1.2 Outline of the Evaluation Effort

The evaluation effort was shaped to a certain extent by the character of information sources concerning STIFS that were available to the project staff. These information sources centered on the following four areas:

(1) The nominal documentation of STIFS in draft form during the current phase of the project. While no attempt was made at NBS to conduct a formal assessment of these documents as was done for the Midterm Oil and Gas Supply Model (MOGSM) documentation, the drafts did provide one important source of information on STIFS.

(2) Exchanges of information which occurred at a March 1980 meeting at DoE of the Short-Term Forecasting Technical Consulting Panel. This meeting was attended by model developers, users, and assessors from DoE, NBS, academe, and private consulting firms.

(3) Consultation with DoE staff in the Short-Term Analysis Division. This group answered questions regarding the current version of STIFS and provided a "frozen" version of the integrating portion of the STIFS computer code complete with input data and the job control language (JCL) needed to run the code on the IBM computer at DoE.

(4) The computer code itself. Though computer code should never be considered a substitute for good documentation, reading and understanding code is in fact the only certain way to learn what a model does, as opposed to what it is supposed to do. For this assessment effort the STIFS code was not read in detail; the objective in reading the code was to verify our understanding of the model and to be able to make minor changes to obtain relevant output information pursuant to sensitivity experiments.

Unfortunately, much that was expected of STIFS as an assessment target proved to be disappointing. At the March 1980 panel meeting we learned that many, if not all, of the satellite models described in the documentation had been discarded for several reasons, most notable of which was a lack of data needed to support the models. This left the assessment effort in the position of having documentation for a discarded analysis system and no documentation for the more relevant, in-use system which was to be the assessment target. As a result of this dilemma, it was necessary to rely on documentation in which unspecified large portions were subject to replacement, on verbal information provided by DoE staff in the Short-Term Analysis Division, and on reading the computer code in a frozen version of STIFS provided for this effort. The frozen version of the code was not retrieved from archive status. Rather, it was created in response to an NBS request for access to the STIFS code. It is not known if this version of the code and data remains available at DoE for future use. Finally, it was learned that the STIFS code and data are subject to frequent undocumented modification and revision and that earlier versions are not archived. This makes it virtually impossible to obtain a copy of the model and data which had been used in making a particular short-term forecast. Furthermore, at the time of this study, the available track record of model use was limited to the February 1980 Short-Term Energy Outlook, with too few forecasts to permit meaningful applications of the procedures for comparing real world outcomes with model forecasts. Although STIFS did not provide the assessment opportunity that was expected, it was still possible to utilize the available system (primarily the integrating portion of STIFS) for limited assessment activities.

1.3 Scope and Approach

The goal of this study was to increase the likelihood of developing an evaluation methodology which was generally applicable. Consequently, the approach taken in this study significantly extended the methodology used in the evaluation of MOGSM. It is believed that although the evaluation of STIFS was flawed in several areas, the information presented in this report will highlight the degree to which various aspects of an assessment methodology are limited by being model or model-type dependent. The basic format of the study consists of a description of the structure of STIFS, an assessment of climatological uncertainties via Monte-Carlo techniques, and an analysis of several techniques which can be used to evaluate a model's predictive performance.

Specifically, Chapter 2 deals with operation of the computer model. Because STIFS is somewhat provisional and unformed relative to MOGSM, it serves as a case study of the feasibility of effecting "third party" assessment of a modeling system early in the development process, when model construction may have outrun documentation. The specific motivation was a desire to perform a set of sensitivity analyses (described below), but a wider significance of the findings concerns the feasibility of "emergency" operation of a new and possibly incomplete model, for decision making purposes. The section is primarily an account of what the evaluators know about STIFS and how this information was acquired. Chapter 2 also includes a brief investigation of the algebraic structure of STIFS. This investigation was undertaken to see if the structure of the model could be exploited in procedures for assessing the sensitivity of the model's outputs to perturbations of one or more of its inputs. Preliminary documentation describes STIFS as a system for the integration and reconciliation of energy flows among disparate sectors in the economy and various energy production sources. This suggests that if the system is linear, it is equivalent to a standard set of Input-Output equations, and thus possibly amenable to the use of "canned" linear algebra computing packages for streamlining senstivity calculations.

Chapter 3 describes sensitivity analysis experiments with STIFS using a Monte-Carlo technique successfully employed last year for investigation of the dependence of MOGSM supply forecasts on uncertainty in the resource base (undiscovered reserves). In contrast to the cost/supply orientation of MOGSM, STIFS bases forecasts on an explicit representation of the demand for energy by sectors of the economy in terms of activities; e.g., industrial production, home heating, commercial lighting, driven partially by "externalities" or variables arising outside of the economy. One such externality is climatic conditions measured by heating degree days and cooling degree days (cumulative annual departures from a standard base temperature). These are intuitively appealing candidates as predictor variables for energy demand. According to the sensitivity experiments, however, these variables do not affect energy consumption at the national level of aggregation of the model.

Because STIFS time horizon is sufficiently small, its operation is amenable to direct measurement of forecast error. Chapter 4 is devoted to the exploration of methods for forecast evaluation with emphasis on recently developed methods of error measurement based on Information Theory, a collection of ideas and procedures that arose about 30 years ago from the study of communication networks. The section occupies a major portion of the report. It encompasses an expository survey of the development of forecast evaluation methods and an analytic comparison of the information theoretic techniques with conventional measures of forecast accuracy. This comparison is accomplished in turn through the application of the various methods to the estimation of forecast error in two widely used model types: (1) single equation linear regression of one variable on another, and (2) linear econometric models (vector regression involving supply-demand relationships based on economic theory). The data sets used in these investigations are all input or intermediate series from STIFS. The model system itself is not invoked in the analysis. (Thus, the work described in the section was not constrained by the absence of valid model documentation).

The final section of the report contains findings, conclusions, and recommendations concerning directions for subsequent research. Some of these findings and conclusions are as follows:

(1) Third party assessment early in model development is of limited value. Early assessment is worthwhile if the assessors can be part of the development team.

(2) Assessors at any stage must be prepared to read computer code.

(3) Weather variations apparently do not affect annual energy consumption at the national level in the continental United States.

(4) STIFS is not yet sufficiently formalized to be operated independently to furnish forecasts for policy analysis.

(5) A definitive forecast-quality measure continues to be elusive. Information-theoretic methods yield results consistent with accepted traditional tests, supplemented by other useful information, at a slightly higher computational cost. Promising directions for continued research can be identified.

2. SOME COMMENTS ON THE STRUCTURE OF STIFS

by

Javier Bernal, Robert E. Chapman and Lambert S. Joel

Center for Applied Mathematics
National Bureau of Standards
Washington, D.C. 20234

This section is concerned with the structure of STIFS and the effect it exerts on the model evaluation process. In order to present the subject matter in a more coherent manner as well as promote a smooth transition to the latter parts of the report, this section begins with a brief discussion of the "ideal" model evaluation process. Several of these issues, especially with respect to the implied linearity of STIFS, are addressed in more detail in section 2.2 The structure of the data bases and the STIFS software are discussed in section 2.3. Some comments relating to the STIFS documentation and source code are also advanced.

2.1 An Overview of the Model Evaluation Process[1]

The evaluation of mathematical models is a complicated subject requiring both resourcefulness and ingenuity on the part of the evaluator. To a certain extent, the type of model and its intended use will govern how much time and effort should be expended in carrying out an assessment. Although some

[1] This section draws heavily on the paper by Saul Gass, Evaluation of Complex Models, University of Maryland, MS/S 76-002, May 1976.

disagreement exists among researchers as to what constitutes a comprehensive evaluation, most would agree that three topics occupy a central role in any model evaluation activity. These topics are:

(1) appropriateness;
(2) validation; and
(3) documentation.

In the discussion which follows, each of the terms will be defined. A description of how one would assess whether or not a particular model adequately covered the issue under consideration will also be presented.

Appropriateness

Many modelers believe that the purpose of building models is either to gain insight into how a particular process operates or to predict the future behavior of the system. In either case, the logical first step is to hypothesize a theoretical model of the process. In cases where the process closely follows an accounting framework as for STIFS, the theoretical model may be quite simple. Complexity is introduced when simplifying assumptions must be made regarding what is important or unimportant or what is controlled or uncontrolled. These decisions should be documented and carefully assessed because they govern the functional relationships which form the basis of an operational model. The appropriateness issue thus focuses on the structure of the abstract model which establishes the foundation upon which the rest of the modeling effort is based. If the abstract model is revealed to be seriously flawed, then the entire exercise of constructing and validating an operational model is likely to be of little use.

Validation

The validation of a complex model aims at demonstrating that the model bears a close resemblance to the system being modeled. The validation process is, in reality, three separate tasks: (1) technical validity; (2) operational validity; and, (3) dynamic validity.

Technical Validity

Technical validity requires the identification of all model assumptions, including those dealing with data requirements and sources. It is especially important to document any divergences from perceived reality. Sections 2.2 and 2.3 focus on several aspects of technical validity discussed below.

As a first step, all stated and implied assumptions, all decision variables, and any hypothesized relationships between variables should be identified. This step attempts to shed light on the correspondence between the model and the real world phenomena it attempts to explain. Three types of assumptions may be readily defined. First, the mathematical assumptions implicit in any large model include its functional form (e.g., linearity) and the continuity of its relationships. A second type, content assumptions, define all model terms and variables. They should also define the scope and limitations of the model. The final type, causal assumptions, are concerned with the assumed or hypothesized relationships between terms and variables (e.g., the relationship between STIFS and its satellite models). Since these assumptions define the direction of flow in a model, they are capable of producing significant divergences from the real world phenomena.

Ideally, one would like to buid a model which would produce true conclusions (predictions) whenever all of the assumptions were true. This property is known as _modus ponens_ among logicians. Since most assumptions have their roots in an empirical argument, however, statements enter in a probabilistic rather than an absolute sense. Unfortunately, the strength of _modus ponens_ does not apply to probabilistic statements, so that some other criteria must be used to reveal whether or not a model has a logical flaw. At one level, the axioms of Aristotle can be used to critique a model (e.g., the model contains one or more circular arguments). At a second level, logical formalism can be used to assess the adequacy of the translation of the model form into a numerical process that produces solutions. This involves:

(i) determining if the mathematical or numerical calculations are correct and accurate,

(ii) analyzing if the logical flow of data and intermediate results are correct and consistent with our causal assumptions; and,

(iii) ensuring that variables and relationships have not been omitted.

A concept which may be called the "principle of model economy" focuses on data validity. For example, an elegant model based on poor quality data may be somewhat less useful than a simple model with slightly better quality data.

In many cases, models use both raw data and transformed data in producing solutions. Since raw data are the basis for transformed data, their validity is essential. Basically, the problem is one of assessing if the raw data are true in terms of accuracy, impartiality, and representativeness. If any of these characteristics are violated and the nature of the violation can be isolated and bounded, then a transformation may produce the desired data set. Data transformations can also have a significant impact on the validity of data. For example, aggregating time series data can force a system which is, in reality, recursive to become simultaneous. Similarly, the pooling of cross-sectional data may be inappropriate if two or more distinct populations are involved.

Assessing the predictive performance of complex models is the subject of a rapidly growing literature. The techniques used range from graphical analysis to the tabulation of sample statistics to the use of information theory. This topic from both the viewpoint of technical validity and operational validity is the subject of chapter 4. All techniques, whether highly qualitative or quantitative, seek to analyze the relationship between actual outcomes and predicted outcomes generated by the model. Predicted outcomes are of two kinds:

(i) the values of the parameters used in the internal computational process of the model, and

(ii) the values of decision variables.

On the first level, one would review the statistical or related techniques used to obtain estimates for the relevant parameters. The main thrust here would be to demonstrate that the deviation of the estimated parameter value is as small as possible. The second level of analysis is more complicated, because it attempts to both document any divergences of actual and predicted values and identify what portions of the model appear to have caused them. These error decomposition techniques are especially difficult in models with a large number of interrelated equations.

Operational Validity

Operational validity attempts to assess the importance of any errors or divergences encountered when the model's technical validity was reviewed. It is concerned with whether or not the model can produce bad answers for proper ranges of parameter values; i.e., the model should be robust in that a user would find it difficult to make the model yield (in terms of the decision maker) an ostensibly wrong answer.

Sensitivity analysis is related to, but distinct from, robustness. This technique seeks to vary systematically the values of the model parameters to determine how much the solution changes. The literature on sensitivity analysis has rapidly grown as more and more decision makers wish to know the answers to "what-if" questions. The techniques for performing sensitivity analyses range from Monte-Carlo experiments to highly structured parametric programming. The mechanics of performiong a Monte-Carlo experiment are discussed in detail in chapter 3; exploiting the structure of the model to perform a systematic sensitivity analysis is the subject of section 2.2.

The last aspect of operational validity and the most difficult is implementation validity. Implementation validity is concerned with the extent to which the real world system being modeled will respond in a manner indicated by the recommended solution. This task is difficult because if a decision maker knew how the system would respond to a given change in a parameter or decision variable, there would be considerably less need for a model. Consequently, even after the model is built and shown to predict historical occurrences well, it may not produce a good or even useful prediction of some future event. One possiblity is that under some conditions the system does not appear to respond as the model predicts, whereas under others, it does. If one could determine under what conditions the system does not respond as predicted by the model, then some of the techniques used in assessing the accuracy of predictions could be used to attempt to isolate the cause. Several additional comments relating forecast performance to a "policy decomposition" are explored in chapter 4.

Dynamic Validity

Dynamic validity is concerned with determining how the model will be maintained during its life cycle so it will continue to be an acceptable representation of the real system. The two aspects associated with dynamic validity are updating and review. Considering the inchoative condition of STIFS it was not possible to address this issue. The basic concepts are defined to illustrate how one would perform such an analysis on a completed modeling system.

In updating, the evaluator needs to be satisfied that the model developers have established a procedure by which information is collected and analyzed to determine if and when model parameters or model structure need to be changed. It is also important that a process exists by which such changes can be incorporated into the model and disseminated to users.

A regular schedule for reviewing the success or failure of the model during its life cycle is also necessary. These reviews should be carried out regularly and should focus on documenting any systematic divergences between the solution predicted and the actual outcomes. The implications and means of accomplishing any proposed model changes should also be described. Periodically an extensive review should be conducted which produces a new evaluation of model validity that yields either a formal pronouncement that the model is still valid or that it should be substantially reworked or scrapped.

Documentation

From a model user's point of view, documentation (the written description of the model) is essential if the model is to be usable, useful, and used. Since the abstract model is a mathematical representation (whereas the operational model appears as a tableau format, a numerical representation, or a computer code), it is necessary to verify that there exists a unique relationship between the abstract (mathematical) model and the operational model. It is important to point out that a third party assessment should be sufficient to verify this relationship. This would spare each user from carrying out the in-depth analysis that this task requires. An issue which each model builder should address is whether or not the computer code which describes a complex model is machine independent. Portability requires not only that the program can be run on a variety of machines, but that it produces the intended results. This step again should be a part of a third party assessment, although the provision of test cases which could be run and compared to a known solution would facilitate a move from one system to another. The final point in the evaluation of documentation is user friendliness. Complex models often involve subtle techniques, which in the absence of a buffer between the user and the model could cause frustration and lead to a highly inefficient use of the model. Documentation and an executive code should serve to shield the user from unnecessary detail without withholding any information which is essential for confident use of the model. One way in which the model developers can test the user friendliness of their model is, again, through the use of a third party assessment.

A second, and for this case perhaps more useful exercise, is to rely on a group of "casual users." These people might be expected to have some familiarity with the problem under study but would be less likely to analyze the model as deeply as its developers or a third party assessment team. Casual users are thus likely to represent better the intended user than a team of model builders or assessors. Time and funding constraints did not permit an evaluation of STIFS by a "casual" user.

2.2 Mathematical Concepts in STIFS

If STIFs is essentially linear and can easily be configured (e.g., in the form $Ax = b$, where A is a matrix of conversion coefficients, b is a vector of energy sector demands or requirements, and x is a vector whose components are fuel quantities distinguished additionally by provenience), then its linearity could be exploited in procedures for assessing the effect on the model's outputs of perturbations of one or more of its inputs. In particular, if STIFS is equivalent to a standard set of input-output equations, it may be amenable to the use of "canned" linear algebra packages for streamlining sensitivity calculations. (These packages may also be useful in assessing robustness.)

STIFS is essentially a collection of "procedures" for projecting regional supply and consumption of energy, and an integrating process which derives import requirements and stock changes from the outputs of the procedures.

There is a preponderance of evidence which suggests that significant portions of STIFS may in fact be nonlinear. The source of the nonlinearities stem from many of the forecast procedures which involve nonlinear functions of (basically exogenous) variables that are believed to determine the quantities of interest (supply and consumption). The likelihood of nonlinearities is also increased because some of the procedures are not specified in any system documentation, while others have been modified frequently over the course of operation of the system (in what follows this point will be discussed further).

Finally it is doubtful that the benefits of linearity could be realized at least from a practical point of view in any event. The logistics of configuring the sheaf of computer estimation procedures in a uniform matricial structure would be formidable even if all the functional relations were linear.

The integrating process itself is linear since it consists merely of "accounting" relations in the form of a balance sheet. From the point of view of facilitating sensitivity analysis, however, its mathematical structure is unfortunately _too_ well behaved, because it decomposes into independent sub-blocks. Thus, it affords little opportunity to examine the behavior of large collections of endogenous variables through feedback from variations in small subsets of singleton exogenous (or endogenous) variables.

From the previous discussion it can be asserted that most of STIFS is a simple accounting framework. There are several statistical techniques which are rather complex, however. These for the most part are located in subroutines that contribute to the internal forecasting procedures.

Descriptions of the internal forecasting procedures employed in the model were provided in the STIFS documentation. By reading the computer code of the modules involved in these procedures, however, we were able to learn some additional details about these procedures. These details seem to be otherwise undocumented and are as follows:

1. Internal forecasts in STIFS are generated primarily through a multiple linear regression that follows the computational design for least square estimation.

2. A STIFS variable is internally forecasted only if its historic data series contains data for at least 12 consecutive months. Otherwise, the variable is either forecasted externally and fed into STIFS through the scenario data base, or receives its value through calculations in subroutines FUTURE, URISHR, STKMOD, STKERR, and REFYLD. The nature of these calculations will depend on the function of the given subroutine.[1]

[1] See Table 2.1 for a brief description of the subroutines referenced in this chapter. More detailed descriptions are found in the STIFS documentation.

TABLE 2.1 REFERENCED SUBROUTINES FROM THE STIFS LIBRARY

Subroutine Name	Description
RDFILE	RDFILE reads and controls information flow from the STIFS data base to MAIN; reads information into a data packet which is then transferred to the appropriate FORTRAN variable.
STKMOD	STKMOD adjusts stock level limits near the beginning forecast period to prohibit large jumps over these periods due to unrealistic forecasts of the first period's stock targets relative to very recent stock levels.
STKERR	STKERR tests whether stock level values computed on a trial basis are within a user specified range; if not it resets the stock level to fall just within this range.
REFYLD	REFYLD forecasts refinery sector yield vectors as a function of historical yield vectors.
URISHR	URISHR calculates the share factors used to determine the contributions of electricity generation by residual fuel oil, distillate fuel oil, crude oil, and natural gas; the shares are determined parametrically as a function of twelve seasonal dummy variables, mean centered heating and cooling degree days, time, and the existence/nonexistence of a coal strike.

Source: Collins, Barcella and Shaw, Short Term Integrated Forecasting System Methodology and Model Descriptions.

3. Most internal regressions utilize variables representing specified months as independent variables. The names of those variables are EJAN, EFEB, EMAR, EAPR, EMAY, EJUN, EJUL, EAUG, ESEP, EOCT, and ENOV. As their names indicate, they represent the months from January to November, respectively. A variable named EDEC, representating December, also exists, but it is never used in any of the internal regressions. EJAN, identified with January, is defined for a given one-month period as follows:

$$\text{EJAN (period)} = \begin{array}{l} 1 \text{ if period is in January} \\ \\ 0 \text{ if period is not in January} \end{array}$$

The other "month" variables are defined in a similar fashion.

Subroutines PAST and FUTURE perform the historical and forecast closures as described in the STIFS documentation. The available documentation does not provide any apparent reason for the selection of the tolerances used in the cross-checks of either closure. Also, through an analysis of the computer code of subroutines PAST and FUTURE we concluded that STIFS does not discriminate between opposite results of the cross-checks. It always executes the same instruction immediately after a given cross-check, even when the corresponding tolerance has been violated.

2.3 Data Bases and Software Description

The STIFS model version, as of May 1980, was "certified" to run properly by the DoE staff, and thus designated frozen to be utilized in our assessment effort. Since the execution of this version can be achieved at any time by following an existing fixed procedure, the programmer is not required to be familiar with the STIFS software. However, more information about the software is necessary to determine its range of application.

The STIFS Methodology and Model Description report provides a satisfactory discussion of the model. The evidence presented in the previously mentioned document was reinforced by reading much of the computer code. However, as the structure and functions of STIFS became clear, it became evident that the model did not perform the expected regressions involving the historic data series for heating degree days and cooling degree days. Besides being employed by subroutine URISHR to calculate the share factors used in determining the contribution to electricity generation by residual fuel oil, distillate fuel oil, crude oil, and natural gas, these data series are used in regressions that create forecasted data corresponding to the domestic demand for distillate oil and residual oil, and the total generation of electricity. These regressions are performed in satellite models (models not in STIFS) and the forecasted data are placed in the scenario data base to be read by STIFS. Thus, any variation in the historic data for heating degree days and cooling degree days requires the availability of these models. This discovery provided a rationale for sensitivity analysis described in chapter 3 of the report. As part of this sensitivity analysis, we sought to determine which variables in STIFS are affected by variations in the historic data series for heating degree days and cooling degree days (e.g., variables representing electricity generation by residual oil, net U.S. imports of crude oil, and share factors are of this type). Petroleum products are related to heating degree days and cooling degree days by equations appearing in subroutines FUTURE and URISHR. In order to determine the names of these variables, we analyzed the sequential data set created by execution of the STIFS model; this data set is made up of two-dimensional FORTRAN arrays containing all alphanumeric and numeric data of energy and econometric variables obtained from data bases and internal computations. Accordingly, the "data packet" structure of a given energy or econometric variable is defined as the corresponding array in this data set. Through the WYLBUR editing system we compared data sets created by separate runs of the frozen version of STIFS under different sets of historic data for heating degree days and cooling degree days. By identifying "data packet" structures in whch discrepancies occur, we were able to identify variables affected by modifications in the historical data series for heating degree days and cooling degree days. As the names of these variables became available, we could then proceed to analyze their data series as they appear in the "data packet" structures.

A sample data file member for motor gasoline production is illustrated in the STIFS documentation. Through a close examination of the computer code of subroutine RDFILE, additional information on the structure and format of the data bases came to light. This information, apparently omitted from the supplied documentation, was considered essential for carrying out the sensitivity analysis in which modifications were to be made to the historical data series of variables EHDD (heating degree days), ECDD (cooling degree days), and to the scenario data series of variables XDDS (domestic demand for distillate oil excluding electric utilities), XDRS (domestic demand for residual oil excluding electric utilities), and XGEL (total generation of electricity).

As a first step, all of the READ statements and the corresponding FORMAT statements for subroutine RDFILE were located and analyzed. This facilitated the definition of the type of input data which STIFS expects, and the format (byte make up of a record) for these data. This, together with samples of data series appearing in the available DoE computer listings, revealed the structure (organization of records) in which the input data must exist.

The mechanics of the sensitivity analysis were carried out through the use of a FORTRAN program that generates the desired data series for heating degree days, total generation of electricity, etc., and places these values in sequential data sets with the proper structure and format for STIFs to read.

The DoE staff in the Short-Term Analysis Division provided the necessary job control language to execute the frozen version of the STIFS model on the DoE IBM computer. This includes giving the computer sufficient information about the input/output requirements of STIFS.

Under normal circumstances, during the execution of STIFS, the historical and scenario data bases are read from data sets that exist on disk. However, as part of the sensitivity analysis, the STIFS model was to be executed a large number of times and each time variations would be made to the data series of variables EHDD, ECDD, XDDS, XDRS, and XGEL. Thus it seemed appropriate to modify the above job control language and STIFS, so that when executed, the model would not read these data from data sets on disk but through input stream data sets. To accomplish this two READ statements were added to the STIFS software; the required control statements to the job control language were also added. This gave STIFS the capability to read the above data in-stream.

Apparently, it is not documented that the STIFS model must read the historic data series of all of the econometric variables (EHDD, ECDD, EJAN, ETIME, etc.) before it reads those of STIFS variables (energy variables). This is because a STIFS variable can be forecasted internally in terms of econometric variables immediately after its historic data series is read. Since subroutine RDFILE does not automatically read all of the historic data series in the required order (it will read them in any order), they must be arranged properly before they are read. This is taken care of through the job control language.

After considerable effort we were able to execute STIFS in accordance with our objectives. Since we were unfamiliar with the model, we relied on the available documentation and computer code of STIFS as our principal sources of information. Because of difficulties in recognizing the relevance of our search for certain pieces of information, the DoE staff's role as a source of information was limited. On a number of points, crucial to the sensitivity analysis, the documentation seemed insufficient. These points included the execution of the program, the format and structue of the data bases, and the mathematical concepts in endogeneous and exogenous subroutines of STIFS. In most cases the necessary information was obtained by reading every line of computer code in specific sections of the software. Otherwise, the DoE staff was the only other source of information.

Our scrutiny of the computer programs uncovered two errors. In one of the subroutines (URISHR), the key punching transposition of a comma and a neighboring character resulted in the mislabeling of a variable. This typographical error apparently did not affect the model results materially, at least for the sample scenarios; outputs before and after correction did not differ at the level of precision defined by the computer output formats. In the subroutine FUTURE, direct imports of electrical power (XMEL) are counted twice in calculating total demand for electricity (XGEL). Ths discrepancy was not discovered prior to execution of the sensitivity experiments. Because the magnitude of this component is never as large as one percent (typically 3/4 of 1 percent) of the total demand, we judged that the effect of the error would not have been perceptible. The experiments were therefore not repeated.

3. AN ASSESSMENT OF CLIMATOLOGICAL UNCERTAINTIES USING MONTE-CARLO ANALYSIS

By

Carl M. Harris
Carl M. Harris and Associates
1625 I. Street, N.W.
Washington, D.C. 20006

The backbone of the effort described in this section was a Monte-Carlo analysis of the effects of climatological uncertainty on STIFS demand forecasts. (This should be distinguished from seasonal changes which come about with near certainty such as automobile usage.) In particular, it was aimed at establishing a more meaningful way than currently used by EIA of assessing the impact, if any, on annual (as opposed to seasonal) weather variations on energy supply and demand.

3.1 Design Considerations

Broadly, the overall range of uncertainty in STIFS's results is generated from the joint probability law governing the endogenous parameters (call them $\underset{\sim}{\beta}$) and exogenous input data elements (say $\underset{\sim}{X}$), together represented by the vector $(\underset{\sim}{\beta}, \underset{\sim}{X})$. This experiment explores the effect of the underlying stochastic character of part of the $\underset{\sim}{X}$ set, namely, the weather as represented by heating and cooling degree days.[1] More precisely, we worked with the subset of input

[1] Degree days are a well-accepted measure of space heating and cooling requirements, with one degree day representing one degree of difference from a base temperature of (below) 65° F for heating and (above) 68°F for cooling.

data corresponding to the (in fact, random) heating and cooling degree days data input into the model and processed by STIFS through its forecast horizon. We have specified the number of heating degree days per month during the *first* and *fourth* quarters of each forecast year to be independent, normally distributed, random variables; we do likewise for the cooling degree days in the *second* and *third* quarters (see Kaczmarek, 1970, for example). Motivated by STIFS's own specialized scenarios, the means of these *12* normal distributions are set at the most recent 30-year empirical estimates. Note, for example, that the current EIA approach to model sensitivity forms a "harsh winter" scenario by setting each month at a heating-degree-day level equal to its mean plus one standard deviation.

The properties of the resulting probability distribution on STIFS output (call it $F(\underset{\sim}{y})$) cannot be obtained by any closed-form analytic operations because of STIFS's complexity. To obtain such a distribution--one measure of uncertainty in results--we turned to Monte-Carlo methods.

In our experiment then: (1) values of the weather components of the vector $\underset{\sim}{X}$ are drawn randomly according to their distribution (here equal to the product of the marginals, in light of independence) according to a specific sampling plan; and (2) values of the (vector) dependent variable $\underset{\sim}{Y}$ are obtained by running STIFS tests with $\underset{\sim}{X}$ in the input set. When this procedure is carried out n times, it generates a sample of size n for $\underset{\sim}{Y}$. This sample permits a detailed analysis of measures of central tendency, dispersion, and the distribution function itself. The generation of a complex Monte-Carlo sample is rather expensive because of the size of the subject model. Therefore, this study employs a modified sampling procedure (as in Harris and Hirshfeld, 1980) which reduces the cost and time required for the experiment, without sacrificing any statistical precision. This approach uses a multidimensional version of the classical Latin Square--called the Latin Hypercube--as described by McKay *et al* (1979). Results presented by those authors indicate an approximate reduction in sampling of 75 percent versus conventional methods, at constant precision.

The Monte-Carlo experiment thus comprised two main steps:

1) Generating *50* vectors of estimates of heating and cooling degree days, where every vector is made up of elements, each of which is an estimate of a month's weather for each of the forecast horizon's *24* months. (*We call such vectors weather realization vectors.*) The vectors are randomly constructed according to a Latin Hypercube design described in the following section. To generate the weather vectors, we begin from the assumption of *normal* distributions for the degree days data series. With the means and variances of these normal cumulative distribution functions (CDFs), we can easily derive random normal deviates for the degree days. The actual parameters used are indicated in Table 3.1.

TABLE 3.1

PARAMETERS OF NORMAL
DEGREE-DAY DISTRIBUTIONS

Month	Heating	Cooling	Mean	Standard Deviation
Jan	X		947.0500	47.3525
Feb	X		786.9100	39.3455
Mar	X		665.4200	33.2710
Apr		X	27.6644	1.9365
May		X	92.4340	6.4704
Jun		X	207.3000	14.5110
Jul		X	311.2600	21.7882
Aug		X	282.3200	19.7624
Sep		X	150.5600	10.5392
Oct	X		258.5200	12.9260
Nov	X		555.0000	27.7500
Dec	X		860.9300	43.0465

NOTE: A letter X has been placed next to each month in position to indicate whether heating or cooling degree days are randomly varied for that month in our experiment.

2) Running one STIFS test for each of the 50 weather realization vectors, with all other model inputs unchanged, to obtain the corresponding complete energy forecasts. The end result of this effort is a short-term prediction revised (from the base case) for each of the weather realization vectors.

The results of the total set of runs can be summarized via a number of empirical statistical measures. The major ones we offer are the means and some key percentage points. The experiment was run as of January 1980 for the usual eight quarters of forecast period (though we display only six quarters of results for purposes of comparison).

3.2 Implementation Considerations

In accordance with the foregoing reasoning, the procedure for Step 1 of our experiment, generating the set of (random) weather realization vectors, is as follows:

(i) Divide the range of W_k, the random weather variable for each month k, into 50 intervals, j, of equal probability (0.02) (j = 1, 2,...,50; k = 1, 2,...,24).

(ii) Let the probabilistic midpoint of each interval be W_{kj}, corresponding to the j^{th} odd percentage point r_j of the standard unit normal.

(iii) Generate the order in which the 50 percentage points are to be used in each of 50 STIFS runs by creating a sequence of 24 (for each month) unique random permutations of the integers 1 to 50.

(iv) Form the required $\{W_k\}$ vector for the first run by taking the leading percentage point from each of the 24 random permutations. Continue to do this by similarly matching the i^{th} elements of each permutation to the i^{th} weather realization vector, until all 50 are formed.

Each of these vectors, then, is a probabilistic estimate of the weather as represented by heating and cooling degree days. The experimental procedure described here is, of course, not restricted to 50 or any other arbitrary number of STIFS tests. We specified 50 runs for our experiment, because it appeared to offer a reasonable trade-off between statistical precision and resource expenditures.

Step 2 in the experiment is the conversion of heating and cooling degree days into residual and distillate fuel demands, and the determination of the effect of such weather patterns on electricity generation. Heating and cooling degree days relate to these three via (so-called) satellite linear models month-by-month.

Model Flow

Ordinarily, heating and cooling degree days enter the overall STIFS framework in the so-called base case at monthly levels equal to the mean number computed for each month from the most recent 30 years (the standard period for defining weather "normals" in the U.S. Weather Service). These heating and cooling degree figures impact the forecasts in basically four places (though many other variables are ultimately affected through the usual material balance relationships):

- Residual fuel demand (excluding utilities);
- Distillate fuel demand (excluding utilities);
- Total electricity generated; and
- Primary source distribution of electricity.

For non-utility residual fuel demand, a satellite model was created for estimating such demand (in Millions of Barrels/day). At the time of this study, this model was a trigonometric function of month, and linear in national income, an industrial production index for the subject month, heating degree days (HDD) for the months of the first and fourth quarter, and a monthly real price estimate of the fuel. Such equations have typically been recalibrated for each quarterly forecast using approximately five years of data.

In the form of the equation used for the current study, the coefficient of the heating degree days is .421235 x 10^{-3} million barrels per day, per heating degree day. Thus, for example, a 100 degree-day increase from a month's normal total leads to a predicted increase in demand of .0421235 MMBbl/day, or 42,000 Bbl/day.

For non-utility distillate fuel demand, another exogenous model was formed. It is a trigonometric function of month; and linear in monthly heating degree days for first and fourth quarter months, the retail price of distillate for first and fourth quarter months, and real disposable personal income. The parameters of this equation are also reestimated at each quarterly forecast point. The coefficient of the heating degree days was .00162219 million barrels per day, per heating degree day, approximately four percent of the HDD coefficient in the equation for non-utility residual fuel demand.

The total electricity generation forecast for a given month is now a totally linear model, but its form was quite different at the time of our study. At that point, the demand for electricity generation, XGEL, was modeled on a per capita basis. The average (XGEL/N) was then specified to be a semilogarithmic function of cooling degree days (CDD) and heating degree days (HDD), but linear in the real price of electricity (PEL), real disposable income per capita (YD72/N), and previous levels of electricity generation, expressed as electricity generation per capita 12 months earlier (see August 1980 OUTLOOK). Detailed data analysis of an earlier EIA procedure (see February 1980 OUTLOOK) indicated that the null hypothesis of no serial correlation could not be accepted and thus that the model's error term may well be serially correlated. Thus a first-order serially-correlated error term was specified, and the model written as

$$\begin{aligned}\ln\,(XGEL/N) = CONSTANT &+ B1\ CDD + B2\ HDD + B3\ \ln\,(PEL)\\ &+ B4\ \ln(YD72/N) + B5\ \ln(XGEL(-12)/N(-12))\\ &+ ERROR,\end{aligned}$$

where ERROR is the first-order serially-correlated disturbance term.

The model's parameters were estimated using the Cochrane-Orcutt first-order serial-correlation regression technique, on monthly data from January 1975 to December 1979, representing 60 observations. The lag term used a data series that began 12 months earlier. The data used are more completely defined in Table 3.2.

It is interesting to note how the problem is presently perceived (see the May 1982 Outlook). Now total demand is forecast by a model fully linear in HDD and CDD, PEL, YD72, and the demand one-month lagged. The estimation uses data going back to January 1977. Observe that the partial semi-log structure has been dropped, that calculations are no longer per capita, and that the lag has

Table 3.2 DATA DEFINITIONS AND SOURCES FOR ELECTRICITY DEMAND MODEL

Variable	Definition	Source
CDD	Cooling degree days. National-1976 population weighted.	National Oceanic and Atmospheric Administration (NOAA), U.S. Dept. of Commerce.
HDD	Heating degree days. National-1976 population weighted.	NOAA
PEL	Real price of electricity to industrial consumer.	1979; U.S. Dept. of Energy FERC Form 5, "Monthly Statement of Electric Operating Revenue and Income." 1975-1978: Edison Electric Institute Statistical Yearbook 1978. PGNP deflater from Bureau of Economic Analysis.
YD72	Real personal disposable income in millions of 1972 dollars.	1979; Bureau of Economic Analysis, U.S. Dept. of Commerce.
XCEL(-12)	Generation of Electricity in billions of kilowatt hours, lagged 12 months.	U.S. Dept. of Energy FERC Form 4, "Monthly Power Plant Report."

Table 3.3 ELECTRICITY DEMAND MODEL ESTIMATES

Variable	Parameter Symbol	Parameter Estimate	Standard Error
Constant	BO	-12,768.	14,247.9
Cooling Degree Days	B1	29.9	2.0
Heating Degree Days	B2	126.2	6.2
Price	B3	-10,347.	6,951.5
Income	B4	920.	13.9
Log	B5	0.51	0.04

Summary Statistics

Adjusted R-Square = 0.94
Durbin-Watson Statistic = 2.20

been reduced to one month from twelve. Table 3.3 presents the current coefficient estimates and standard errors as offered in the May 1982 Outlook, as well as some summary fit statistics.

The final of the four models provides an estimate for the percentage contribution to total electricity by all major (fossil-fuel) primary sources. This model is endogenous to STIFS and allows the final estimation of total residual and distillate fuel when taken together with the other three models.

Electrical generation by coal-fired and nuclear generating plants is estimated using current and planned capacity additions and historical data on operating rates and heat content. Hydropower, geothermal, and other generation possibilities are estimated using data on historical seasonal patterns and trends.

Since the fractional share for each of four--residual fuel oil, distillate fuel oil, crude oil, and natural gas--must be between zero and one, an ordinary linear model is unacceptable since the forecasts might violate the constraint. As is not unusual in these situations (see Pindyck and Rubenfeld, 1976), a multinomial logit model was selected.

The binary-choice logit model presupposes the form

$$S = [1 + \exp(a + \sum_{j=1}^{n} b_j x_j)]^{-1} \qquad (1)$$

for the share S of the first (of two) fossil fuels where the $\{x_j\}$ are the exogenous factors affecting the share of fuel 1. A linear form can be established easily from Equation (1) by a little bit of algebra:

$$(1-S)/S = \exp(a + \sum_{j=1}^{n} b_j x_j)$$

Thus,

$$\ln [(1-S)/S] = a + \sum_{j=1}^{n} b_j x_j \quad , \qquad (2)$$

where we recognize that the left-hand side of the above is the logarithm of the ratio of the share of the second fuel to the first.

Instead of the binary-choice approach, STIFS uses a four-choice model, which is written in a form similar to that of Equation (2) for each share S_i ($i = 2, 3, 4$) with S_1 set as the norming variable (here residual fuel oil):

$$\begin{aligned}
\ln(S_2/S_1) &= a_2 + \sum_{j=1}^{n} b_{2j} x_j && \text{(distillate)} \\
\ln(S_3/S_1) &= a_3 = \sum_{j=1}^{n} b_{3j} x_j && \text{(crude)} \qquad (3) \\
\ln(S_4/S_1) &= a_4 + \sum_{j=1}^{n} b_{4j} x_j && \text{(natural gas)}
\end{aligned}$$

$$S_1 = 1 - S_2 - S_3 - S_4 \quad .$$

All the parameters can be estimated by means of either nonlinear least squares or, after appropriate transformations, by ordinary least squares. However, neither of these would use all of the information effectively. There are two problems: (1) the errors need not be consistent as required in the ordinary linear model context; and (2) the cross-equation correlation ought to be acccounted for directly in the estimation. The first problem is generally handled by estimating the variances of each of the independent variables and then weighing the least square errors accordingly. The second issue is approached by the application of a form of generalized least-squares regression due to Zellner (1962) to account for the correlation among the error terms associated with each equation. By this procedure, we are thus able to make use of a limited number of data points.

Monthly data on oil (steam and non-steam), natural gas (steam and non-steam), and combined cycle generation were obtained from the Federal Energy Regulatory Commission (FERC) Form 4 currently collected by EIA. Since data are not specificially collected for residual fuel oil, distillate fuel oil and crude oil generation, the approach used by EIA was to apportion the oil steam and oil non-steam into these components using the percentages derived from a detailed analysis of the literature.

The independent variables used in the estimation consisted of the following:

- Two variables traversing a sine and cosine curve, respectively, to capture the regular seasonal pattern of shifting proportions between months. Thus, for example, residual-fired generation on average is most prominent in December (0.57) when natural gas-fired generation is smallest (0.36).

- Heating degree days and cooling degree days.

- Coal strike indicator variables accounting for strikes which may occur (as in December 1974 and December 1977 through March 1978). To incorporate the potential impact of a coal strike on shifting the relative proportions of fossil fuel generation (i.e., an increase of distillate consumption relative to residual), a variable was defined to equal one for the strike period and zero otherwise.

- Qualitative hydroelectric generation shortage variable. A variable was inserted in each of the estimated relationships to account for hydroelectric generation reduction as in 1977 arising from the deficiency in the rainfall in the Pacific Northwest. It was defined to equal one for January 1977 through October 1977 and zero for the remainder of the sample period. Its statistical significance reflects the changing relative share during the drought period of, say, natural gas consumption to residual consumption.

- Fuel prices. The possibility of interfuel substitution in the fossil fuel generation of electrical energy exists. To capture this potential effectively, the ratio of each fuel price to the price of coal was introduced (i.e., the price reflecting the average delivered price of each fuel to electric utilities in a given month). Prices of distillate and crude were not available.

The results of the estimation procedure as of the February 1980 Outlook are presented in Table 3.4. All of the estimated coefficients were found to be significantly different from zero at the 5 percent level with the exception of three: hydro shortages on distillate, coal strike on natural gas, and gas/coal price ratio on gas.

One major result of the model is that variations in weather provide for significant alterations in relative shares. An increase in winter severity, measured by an increase in the number of heating degree days, results in a large relative reduction (compared to the base case) in the natural gas share as that fuel is diverted to high priority consumers. The differential between the actual and the desired level of generation is made by residual fuel oil and distillate fuel oil fired generation. During the cooling season, however, the results lead to a relative increase in distillate fired generation.

TABLE 3.4

FUEL SHARE FORECASTING EQUATIONS

Equation/ Variable	log (Distillate/Resid.)	log (Crude/Resid.)	log (Natural Gas/Resid.)
T1	-0.207	0.000448	-0.0959
T2	-0.347	0.000746	-0.0783
C	-3.02	-5.26	-0.462
HDD	0.000966	-0.00000205	-0.000630
CDD	0.000754	-0.0000014	-0.000537
PRESID/PCOAL	0.204	-0.000443	0.329
PGAS/PCOAL	-0.221	0.000439	-0-
COALSTK	0.149	-0.000318	-0-
HYDSHO	-0-	-0.0000210	-0.143
R^2	0.738	.999	0.945

T1, T2 denote sine and cosine, respectively; C denotes a constant term; HDD, and CDD denote heating and cooling degree days, respectively; PRESID, PCOAL, PGAS denote the prices of residual, coal, and natural gas; COALSTK denotes the coal strike variable; and HYDSHO denotes the hydrogeneration shortage.

SOURCE: Short-Term Energy Outlook, February 1980.

3.3 Summary of Findings

The results of our work are set forth in the following:

- Tables 3.5 and 3.6 show experimental 25th, 32nd, 50th (median), 68th, and 75th percentiles, plus the mean of non-distillate demand (total distillate in 3.6) for each of the six quarters beginning January 1, 1980 (in Quads/quarter). The numbers are contrasted in each case to the base case result.

- Tables 3.7 and 3.8 show the simulation 25th, 32nd, 50th, (median), 68th, and 75th percentiles, plus the mean of non-utility residual fuel oil demand and utility demand, respectively, for each of the six quarters beginning January 1, 1980.

- Table 3.9 shows the experimental interquartile ranges and medians for total electricity generation induced by weather and the relative contributions made by the three liquid fossil fuels.

- Table 3.10 shows the effect of weather on net imports of crude.

The most important aspects of these results can be summarized as follows. When Tables 3.5 through 3.10 are combined, we find that the simulated interquartile ranges provide much tighter estimates for the range of variability resulting from the adverse and favorable weather scenarios than that indicated by DoE in the February 1980 OUTLOOK. The comparisons are presented in Table 3.11.

From Table 3.9, we see that the net effect of the weather variability on total electricity generation is really quite small compared to the base--a maximum interquartile range of less than 0.5 percent. For the imports, Table 3.10, the maximum interquartile range is also fairly small, reaching 1.3 percent.

TABLE 3.5

DISTILLATE FUEL OIL DEMAND
(EXCLUDING ELECTRIC UTILITIES)
BASE CASE VS. EXPERIMENTAL OUTPUT
(Quads per Quarter)

	1980				1981	
Quarters	1st	2nd	3rd	4th	1st	2nd
Demand in 50 States (Base Case)	1.936	1.441	1.310	1.673	1.897	1.474
Monte-Carlo Results:						
25th Percentile	1.923 (-.6%)[a]	X	X	1.662 (-.6%)	1.883 (-.7%)	X
32nd Percentile	1.927 (-.5%)	X	X	1.664 (-.6%)	1.887 (-.5%)	X
Median	1.936 (-.5%)	X	X	1.674	1.895	X
Mean	1.936	X	X	1.673	1.897	X
68th Percentile	1.942 (.3%)	X	X	1.682 (.5%)	1.902 (.3%)	X
75th Percentile	1.947 (.5%)	X	X	1.683 (.6%)	1.907 (.5%)	X

NOTE: The X columns are those in which there was no random variation since only changes in heating degree days affect distillate use.

[a] Values within parentheses are the percentage deviations of the results of the Monte-Carlo experiment from those of the base case.

TABLE 3.6

TOTAL DISTILLATE FUEL OIL DEMAND
BASE CASE VS. EXPERIMENTAL OUTPUT
(Quads per Quarter)

	1980				1981	
Quarters	1st	2nd	3rd	4th	1st	2nd
Base Case	2.050	1.537	1.440	1.769	1.999	1.569
Monte-Carlo Results:						
25th Percentile	2.034 (-.8%)	1.537	1.439 (-.07%)	1.754 (-.8%)	1.983 (-.8%)	1.568 (-.06%)
32nd Percentile	2.038 (-.6%)	1.537	1.440	1.757 (-.7%)	1.988 (-.6%)	1.568 (-.06%)
Median	2.049 (-.05%)	1.537	1.440	1.777 (-.2%)	1.998 (-.05%)	1.569
Mean	2.050	1.537	1.440	1.700	1.999	1.569
68th Percentile	2.057 (.3%)	1.538 (.07%)	1.441 (.07%)	1.779 (.7%)	2.006 (.4%)	1.569
75th Percentile	2.062 (.6%)	1.538 (.07%)	1.441 (.07%)	1.781 (.7%)	2.011 (.6%)	1.569

TABLE 3.7

RESIDUAL FUEL OIL DEMAND (EXCLUDING ELECTRIC UTILITIES) BASE CASE VS. EXPERIMENTAL OUTPUT

(Quads per Quarter)

	1980				1981	
Quarters	1st	2nd	3rd	4th	1st	2nd
Demand in 50 States (Base Case)	0.906	0.670	0.562	0.689	0.867	0.676
Monte-Carlo Results:						
25th Percentile	0.902 (-.4%)	X	X	0.685	0.863	X
32nd Percentile	0.903 (-.3%)	X	X	0.686 (-.4%)	0.864 (-.3%)	X
Median	0.905	X	X	0.689	0.866	X
Mean	0.906	X	X	0.689	0.867	X
68th Percentile		X	X			X
75th Percentile	0.909 (.3%)	X	X	0.691 (1.0%)	0.870 (.4%)	X

NOTE: The X columns are those in which there was zero random variation since only changes in heating degree days affect distillate use.

TABLE 3.8

TOTAL RESIDUAL FUEL OIL DEMAND
BASE CASE VS. EXPERIMENTAL OUTPUT
(Quads per Quarter)

	1980				1981	
Quarters	1st	2nd	3rd	4th	1st	2nd
Base Case	1.708	1.361	1.402	1.371	1.554	1.255
Monte-Carlo Results:						
25th Percentile	1.694 (-.8%)	1.360 (-.07%)	1.402	1.358 (-.9%)	1.540 (-.9%)	1.254 (-.08%
32nd Percentile	1.698 (-.6%)	1.360 (-.07%)	1.402	1.360 (-.8%)	1.545 (-.6%)	1.255
Median	1.707 (-.06%)	1.360 (-.07%)	1.402	1.374 (.2%)	1.553 (.06%)	1.255
Mean	1.708	1.361	1.402	1.371	1.554	1.255
68th Percentile	1.714 (.4%)	1.361	1.403 (.07%)	1.380 (.7%)	1.560 (.4%)	1.256 (.08%)
75th Percentile	1.718 (.6%)	1.361	1.403 (.07%)	1.385 (1%)	1.563 (.6%)	1.256 (.08%)

TABLE 3.9

TOTAL ELECTRICITY GENERATION: WEATHER SENSITIVITY ANALYSIS
(Billions of Kilowatt Hours per Quarter)

		1980				1981	
Quarters		1st	2nd	3rd	4th	1st	2nd
Base Case Total for 50 States		566.90	555.73	628.78	568.92	582.70	571.22
Primary Source							
	Favorable	7.62	5.68	7.52	5.92	6.94	5.58
Distillate	Median	7.79	5.71	7.59	6.17	7.11	5.61
	Adverse	7.98	5.73	7.66	6.26	7.29	5.63
	Favorable	75.85	57.30	74.70	66.44	67.20	49.02
Residual	Median	76.77	57.33	74.76	67.61	68.15	49.09
	Adverse	77.56	57.37	74.81	68.45	68.78	49.13
	Favorable	0.38	0.29	0.34	0.32	0.34	0.25
Crude Oil	Median	0.39	0.29	0.35	0.32	0.34	0.26
	Adverse	0.39	0.29	0.35	0.33	0.36	0.26
Total	Favorable	565.61	555.72	628.77	567.76	581.41	571.21
	Adverse	567.81	555.73	628.80	569.93	583.61	571.22

NOTE: The favorable and adverse scenarios correspond to the 25th and 75th percentiles, respectively, of the Monte-Carlo experiment.

TABLE 3.10

NET U.S. IMPORTS OF CRUDE OIL

(MBbl per day)

		1980			1981
Quarters	2nd	3rd	4th	1st	2nd
Base Case	5,193	5,764	4,497	5,095	5,502
Monte Carlo Results:					
25th Percentile	5,191	5,762	5,559 (-.7%)	5,054 (-.8%)	5,499
32nd Percentile	5,192	5,763	5,565 (-.6%)	5,067 (-.5%)	5,501
Median	5,193	5,764	5,606	5,093	5,502
Mean	5,193	5,764	5,597	5,095	5,502
68th Percentile	5,193	5,765	5,622 (.4%)	5,111 (.3%)	5,504
75th Percentile	5,193	5,765	5,631 (.6%)	5,123 (.5%)	5,505

TABLE 3.11

COMPARISON OF EXPERIMENTAL RANGES
WITH
DOE RANGES

(MBbl/day)

	1980				1981
Quarters	1st	2nd	3rd	4th	1st
Distillate:					
Experimental Range					
Nonutility	0.044	-0-	-0-	0.038	0.044
Utility	0.011	0.002	0.004	0.010	0.011
TOTAL	0.055	0.002	0.004	0.048	0.055
DOE Range					
TOTAL	0.64	0.07	0.08	0.47	0.69
Residual:					
Experimental Range					
Nonutility	0.000	-0-	-0-	0.000	0.000
Utility	0.035	0.001	0.002	0.041	0.032
TOTAL	0.035	0.001	0.002	0.041	0.032
DOE Range					
TOTAL	0.85	0.36	0.68	0.58	0.81

4. ASSESSING THE FORECAST ACCURACY OF SHORT-TERM ENERGY MODELS: BASIC PRINCIPLES AND EXTENSIONS[a]

by

Robert E. Chapman
Center for Applied Mathematics
National Bureau of Standards
Washington, D.C. 20234

and

Anthony E. Bopp
James Madison University
Harrisonburg, Virginia 22807

Generating forecasts, quantitative estimates about the likelihood of future events taking place based on past and current information, is one of the major reasons for constructing models. Although there are numerous types of models which could be applied in any given situation, the techniques outlined in this chapter can be illustrated through reference to two alternative model specifications designed to forecast the real price of motor gasoline. They are: (1) an econometric model which is a reduced-form equation from an inventory-price adjustment model used to capture the effects of supply and demand factors; and (2) a simple regression model which links the price of crude oil to motor gasoline prices.

[a] The authors wish to acknowledge the influence that P.A.V.B. Swamy of the Division of Research and Statistics at the Federal Reserve Board has exerted on the treatment of several theoretical issues discussed in this section.

In its purest sense, the forecasting process involves an extropolation beyond the period over which a model was estimated. Consequently, as the time horizon (i.e., the length) of the forecast is extended, one would expect the quality of the information provided by the forecast to deteriorate. The purpose of this chapter is to present a set of methods for measuring the performance of a model's forecast. A discussion of techniques for measuring how rapidly the model's performance deteriorates over time and to what extent this undesirable attribute can be mitigated through the formulation of a composite forecast will also be included. This chapter in its focus on forecast performance or accuracy assessment will not address issues such as the quality of the data underlying the model, the mathematical structure of the model, or its dynamic stability, all of which could, in principle, affect the accuracy of a model's forecasts. Readers interested in these and other evaluation topics are referred to the article by Dhrymes *et al.*[1]

In most applications two types of forecasts are found to be useful. Point forecasts predict a single number in each forecast period, while interval forecasts indicate in each forecast period an interval in which the realized value will hopefully lie. The focus of this chapter will be on point forecasts. It is also useful to distinguish between two classes of forecasts. These two classes are known as *ex post* and *ex ante* and are illustrated with respect to time in Figure 4.1. Three dates are shown on Figure 4.1. These

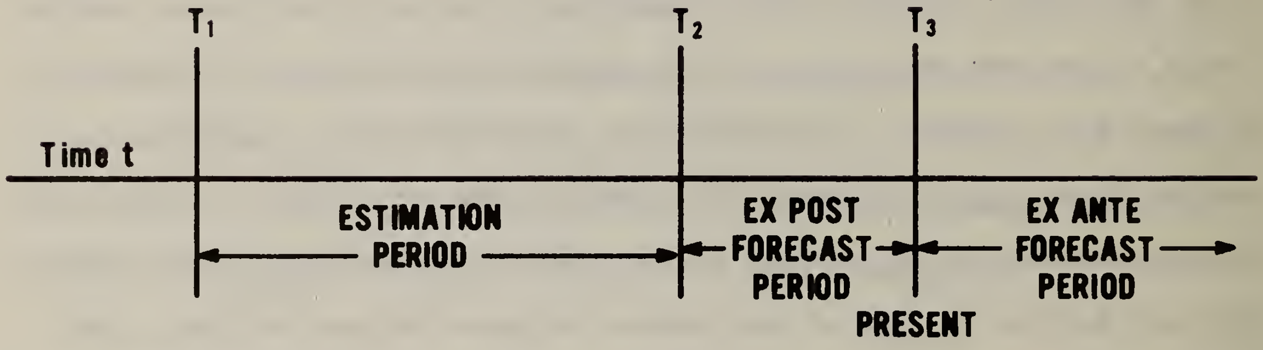
[1] Dhrymes, Phoebus J., E. Phillip Howrey, Saul H. Hymans, Jan Kmenta, Edward E. Leamer, Richard E. Quandt, James B. Ramsey, Harold T. Shapiro, and Victor Zarnowitz, "Criteria for Evaluation of Econometric Models," *Annals of Economic and Social Measurement*, Vol. 1, No. 3 (1972), pp. 291-324.

dates refer to the estimation period (between T_1 and T_2, where T_1 is the earliest date for which the complete set of data required to make an estimate is used), the *ex post* forecast period (between T_2 and T_3, where T_3 is the present time period), and the *ex ante* forecast period (beyond T_3). In an *ex post* forecast, the forecast period is such that observations on both endogenous[1] variables and exogenous[2] explanatory variables are known with certainty. Thus, *ex post* forecasts can be checked against existing data and provide a means of evaluating a forecasting model. An *ex ante* forecast predicts values of the dependent (endogenous) variables beyond the present using explanatory variables which (depending on the nature of the data and the length of the lags associated with the explanatory variables) may or may not be known with certainty. The emphasis of this chapter is on a modified version of an *ex post* forecast. This approach was taken since it permits the accuracy measures to be computed in a straightforward and unambiguous manner. The term modified *ex post* is used because in generating our forecasts, every effort was made to simulate the steps that an analyst located at time period T_2 in Figure 4.1 would have undergone in making an *ex ante* forecast of motor gasoline prices for the periods between T_2 and T_3. Thus the values of the exogenous variables used in the modified *ex post* forecast correspond to values which would have been based on predictions (i.e.,they are made *as if* the analyst were located at time period T_2). This point serves to highlight a difference between conditional and unconditional forecasts. In an

[1] Endogeneous variables are variables whose values are determined within the model's framework.

[2] Exogeneous variables are variables whose values are determined outside the model's framework.

FIGURE 4.1 THREE PHASES IN THE EVALUATION OF AN ECONOMIC FORECAST[a]



[a] Source: Pindyck, Robert S. and Daniel L. Rubinfeld. Econometric Models and Economic Forecasts, (New York: McGraw-Hill Book Company, 1976), p.157.

unconditional forecast the values of all the explanatory variables in the forecasting equation are known with certainty. Therefore, any pure ex post forecast is an unconditional forecast. However, depending on the lag structure of the explanatory variables, ex ante forecasts may also be unconditional. Since the values of the exogenous variables used in making the modified ex post forecast between time periods T_2 and T_3 were based on predictions we have a means for comparing the quality of the conditional forecast to the unconditional forecast. This approach is particularly desirable because in a real forecasting situation the analyst does not know the values of all explanatory variables with certainty. In addition, a measure of conditional forecast performance is potentially more useful in formulating a composite forecast because it is a better indication of how well the models would have done in a real forecasting situation.

4.1 The Classical Approach

The classical approach discussed in this section seeks to identify the "best" forecast. The mechanics of this approach may be carried out in a variety of ways. In all cases, the criteria used in determining the best forecast is to select the model, or combination of models, which produces the forecast error having the minimum variance. As is shown in Pindyck and Rubinfeld,[1] the error associated with a forecasting procedure can come from a combination of four distinct sources. First, the random nature of the additive error process guarantees that forecasts will deviate from true values, even if the model is specified correctly and its parameter values are known with certainty.

[1] Pindyck, Robert S. and Daniel L. Rubinfeld, Econometric Models and Economic Forecasts (New York: McGraw-Hill Book Company, 1976).

Second, the process of estimating the regression parameters introduces error because estimated parameter values are random variables which may deviate from the true parameter values. For any given sample, the estimated parameters are unlikely to equal the true values of the underlying parameters, even though they will (if they are unbiased) equal those parameters on the average. Third, in the case of a conditional forecast, errors are introduced when calculated guesses are made for the values of the explanatory variables during the forecast period. Finally, errors may be introduced because the model specification may not be an accurate representation of the true underlying process.

4.1.1 Qualitative Methods

As the title of this subsection suggests, the methods which will be discussed will stress the qualitative aspects of the forecast evaluation process. The term qualitative should not be taken in a pejorative sense because although the measures are somewhat subjective, they may be extremely useful to an experienced analyst (as is pointed out in the article by Dhrymes, et al.[1]). The concept of qualitative methods of forecast evaluation is quite broad, so that some selectivity was necessary to focus on a particular subset of these methods which also provides a transition into the classical quantitative and information theoretic measures of the next sections. More precisely, the

[1] Phoebus J. Dhrymes, et al., op cit.

emphasis of this subsection is on three graphical approaches for measuring goodness-of-fit. Each approach makes use of eight years of data for the econometric model[1] during a subset of its estimation period (January 1972 through December 1975) and a modified _ex post_ forecast covering January 1976 through December 1979.

The first graphical approach is the simplest; it consists of a plot of the realized and predicted values for the real price of motor gasoline versus time in months. The second approach, a plot of residuals,[2] introduces the key concept of error decomposition by permitting one to check the model for systematic biases.[3] The third approach, the Theil prediction-realization diagram, further extends the concept of error decomposition by permitting one to develop measures of the bias of the forecast and, loosely speaking, its efficiency.[4]

The first graphical approach is illustrated in Figures 4.2 and 4.3. Both figures provide comparisons of the predicted and realized values for a

[1] A complete description of the econometric model is given in section 4.3.1.

[2] A residual is the difference between the realized value and the predicted value. Residuals thus represent errors in the forecast.

[3] A forecast is unbiased if the expected (average) value of the predictor is equal to the realized value. All forecasts for which this statement is not true are referred to as biased.

[4] The term efficiency as it is used here refers to a case where a forecaster does _not_ systematically underestimate (overestimate) high values _and_ overestimate (underestimate) low values.

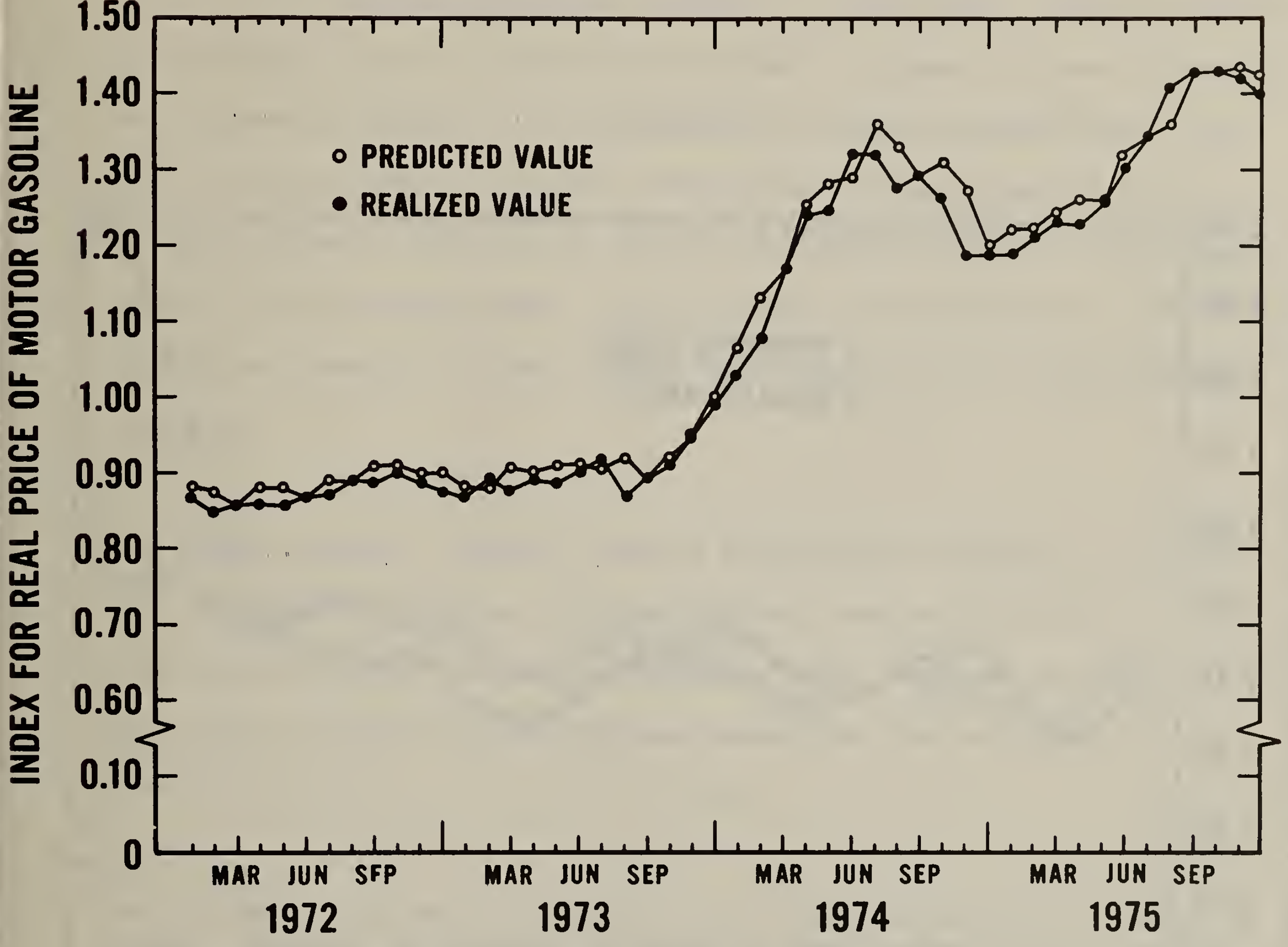
computed index[1] of the real price of motor gasoline during a four year period. Figure 4.2 treats the period January 1972 through December 1975 which is a subset of the period over which the model was estimated. Figure 4.3 treats the period January 1976 through December 1979 which is a modified ex post forecast. In both figures, the computed index is displayed along the vertical axis and the month and year along the horizontal axis.

Turning first to Figure 4.2, we see, as one would expect during the estimation period, that the model "tracks" the process quite well. During the periods of fairly stable prices (January 1972 through September 1973), the two series show close agreement, although the model "appears" to miss turning points by one month (compare the realizations of the price series in the neighborhood of March 1973 to the predicted series). It is reassuring to note that the model tracks the explosive growth in motor gasoline prices during the embargo period, although it still tends to miss turning points by one month. The model also captures well the upward trend in gasoline prices in 1975.

An examination of Figure 4.3 reveals some significant differences. First, the seasonal effects tend to be more pronounced now than before in the realizations of the price series, peaking in July and August and bottoming out in February and March. Second, during the last three quarters of 1979

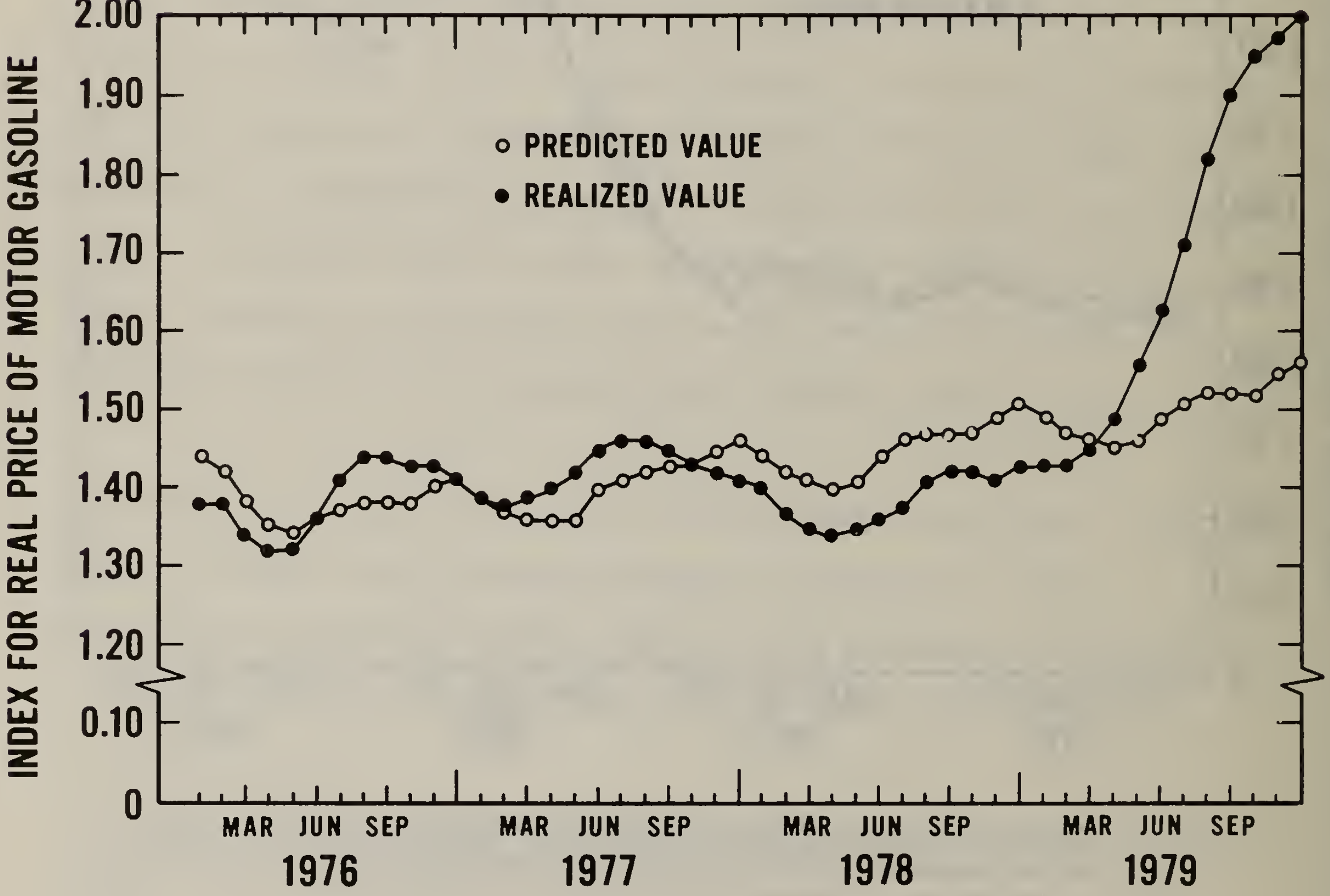
[1] The real price index for motor gasoline in period t is computed by dividing the wholesale price index for motor gasoline (not seasonally adjusted) in period t by the wholesale price index (not seasonally adjusted) for that period.

FIGURE 4.2 COMPARISON OF PREDICTED AND REALIZED VALUES OF THE COMPUTED INDEX FOR THE REAL PRICE OF MOTOR GASOLINE[a] DURING THE LAST FOUR YEARS OF THE ESTIMATION PERIOD.



[a] Wholesale price index of motor gasoline (not seasonally adjusted) divided by the wholesale price index (not seasonally adjusted).

FIGURE 4.3 COMPARISON OF PREDICTED AND REALIZED VALUES OF THE COMPUTED INDEX FOR THE REAL PRICE OF MOTOR GASOLINE DURING THE FOUR YEAR FORECAST PERIOD: JANUARY 1976 THROUGH DECEMBER 1979.



gasoline prices increased explosively mirroring the explosive growth in crude oil prices. Since the likeihood of the events producing this surge in crude oil prices being foreseen in December 1975 is nil, one would not expect the model to perform well during that period.[1]

When both time series are examined together, it can be seen that the predicted series tends to lag the realized price series. Although this same phenomenon was observed in Figure 4.2 its consequences are more pronounced here. As a result of these lags, if the model overestimated (underestimated) in the previous period it will tend to overestimate (underestimate) in the subsequent period. Thus one would expect to see a pattern in the residuals if they were plotted as a function of time which serves to motivate the second graphical approach.

The second approach, residual plots, is illustrated in Figures 4.4 and 4.5. As in the previous figures, the month and year associated with each observation is plotted along the horizontal axis. Along the vertical axis of each figure is measured the difference between the realized value and the predicted value (i.e., the forecast error).

[1] Note that the real price of crude oil is an exogenous variable in this model. However, the realized value was not used because we are attempting to simulate the forecasting process as if the analyst were situated at December 1975 on the time line (see Figure 4.1) and wished to perform an ex ante forecast.

FIGURE 4.4 RESIDUAL PLOT OF THE COMPUTED INDEX FOR THE REAL PRICE OF MOTOR GASOLINE DURING THE LAST FOUR YEARS OF THE ESTIMATION PERIOD.

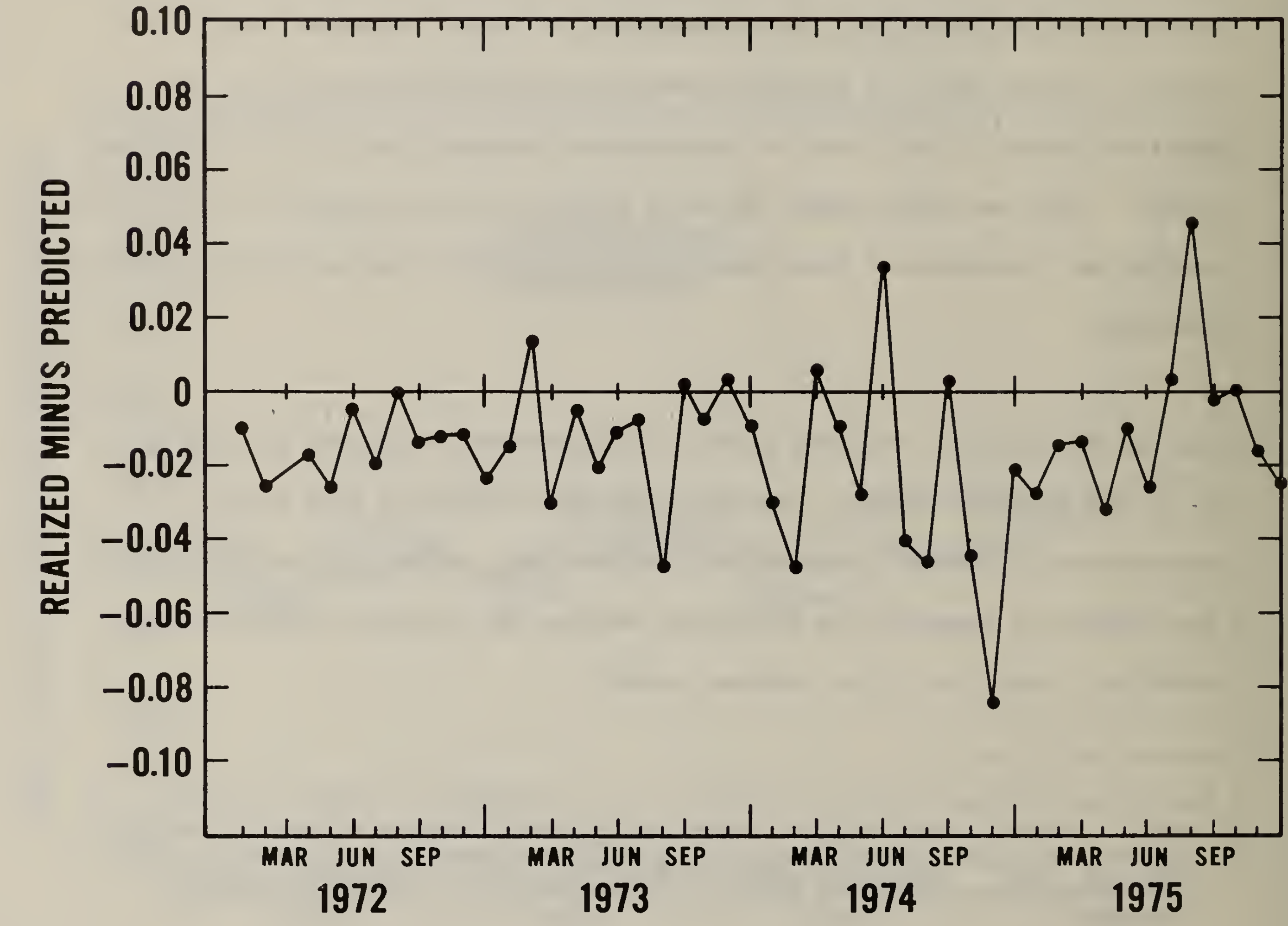
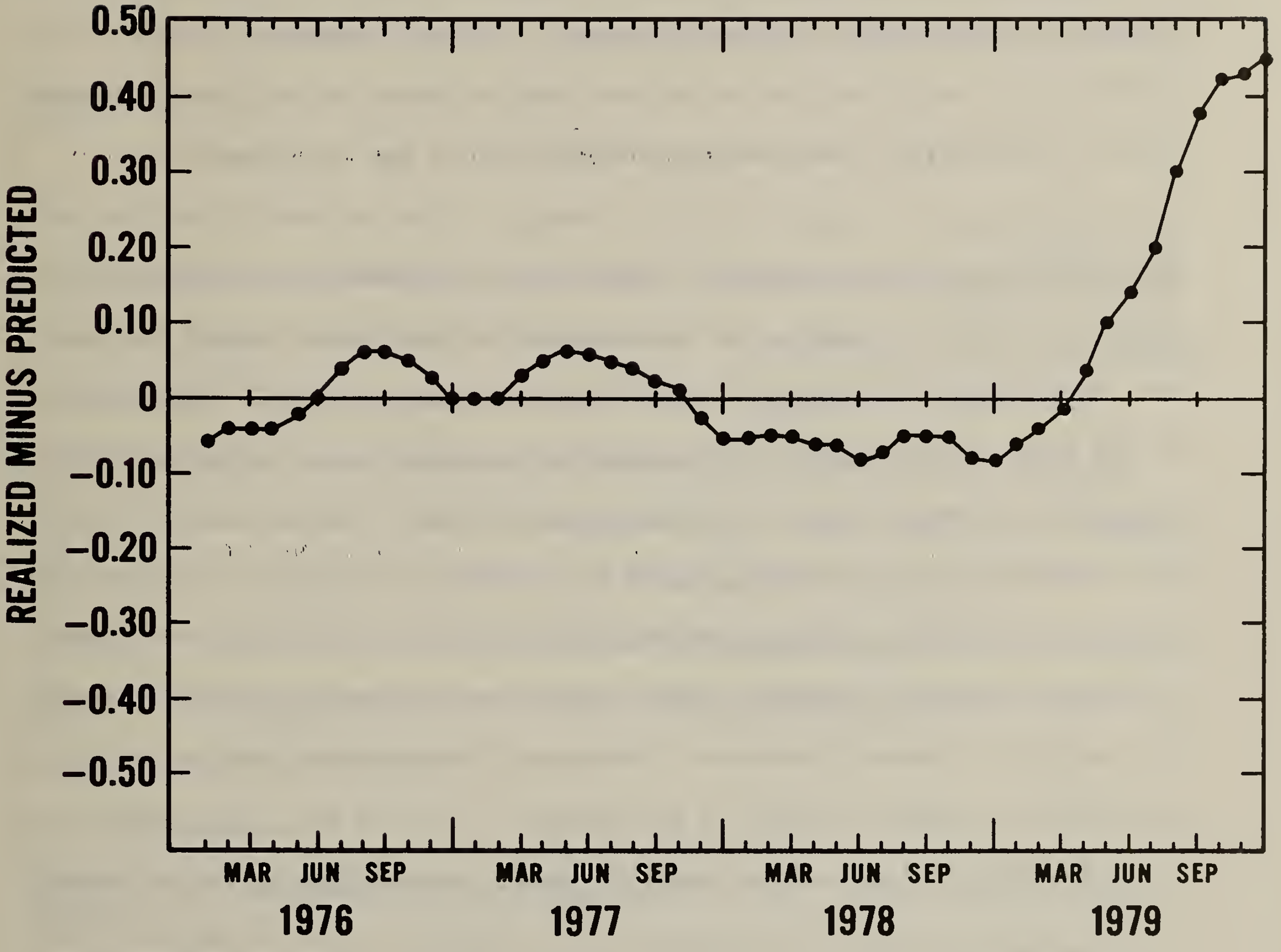


FIGURE 4.5 RESIDUAL PLOT OF THE COMPUTED INDEX FOR THE REAL PRICE OF MOTOR GASOLINE DURING THE FOUR YEAR FORECAST PERIOD: JANUARY 1976 THROUGH DECEMBER 1979.



Turning now to each figure, a careful examination of Figure 4.4 reveals a rather jagged pattern in the residuals. From a statistical viewpoint, this pattern is desirable because it appears to be random. That is, the residuals are likely to be generated by a "white noise" process, which is a necessary requirement for the validity of several tests performed in the model estimation and diagnostic checking phases. A closer examination of the residual plot would lead one to believe that the model overestimates slightly (0.02 on the average) and that heteroscedastricity may be present.[1]

When Figure 4.4 is contrasted with Figure 4.5, an interesting observation results. First, the pattern of the residuals is not at all jagged, in fact the changes are quite smooth. Second, as was observed earlier (Figure 4.2), if the model overpredicted (underpredicted) in the previous period it tends to overpredict (underpredict) in the subsequent period. Both of these observations can be explained in part by noting that Figure 4.4 is a plot of residuals from the estimation period, whereas Figure 4.5 is a plot of the residuals during the forecast period. Since the estimation process minimizes the sum of the squared deviations (residuals), one seeks to construct a model which does not have a pattern in its residuals. (If it did, this information could be incorporated into the model to reduce the variance of the estimation

[1] Heteroscedasticity implies that the variance of the residuals changes over time. From the plot, one might hypothesize that the variance of the residuals is smaller during the pre-embargo period than for periods after and including the embargo.

error.) On the other hand, during the forecasting process it is not possible to place constraints on the residuals because, a priori, one does not know what the realized value will be. Thus, unless there is a clear cyclical pattern in the residuals (e.g., the model systematically underpredicts in the driving season and overpredicts in the heating system), one can say very little about the model's forecast performance except through resort to a more rigorous test.

One qualitative measure which attempts to fill the void just mentioned is the Theil prediction-realization diagram. This procedure is developed in detail in one of Theil's books[1] and is discussed as an analytical tool in the studies by Mincer and Zarnowitz[2] and Zarnowitz.[3] The basic concept behind the prediction-realization diagram is to construct a scatter diagram relating predictions and realizations. In a probabilistic sense, this scatter diagram can provide a rough measure of both the likelihood of a particular prediction-realization pair occurring jointly as well as a particular value of a realization occurring conditional upon some specified predicted value (e.g., the average predicted value, $\overline{P}$).

[1] Henri Theil, Applied Economic Forecasting (New York: Rand McNally and Co., 1966.

[2] Jacob Mincer and Victor Zarnowitz, "The Evaluation of Economic Forecasts," in Economic Forecasts and Expectations, Jacob Mincer (ed.), National Bureau of Economic Research, New York 1969.

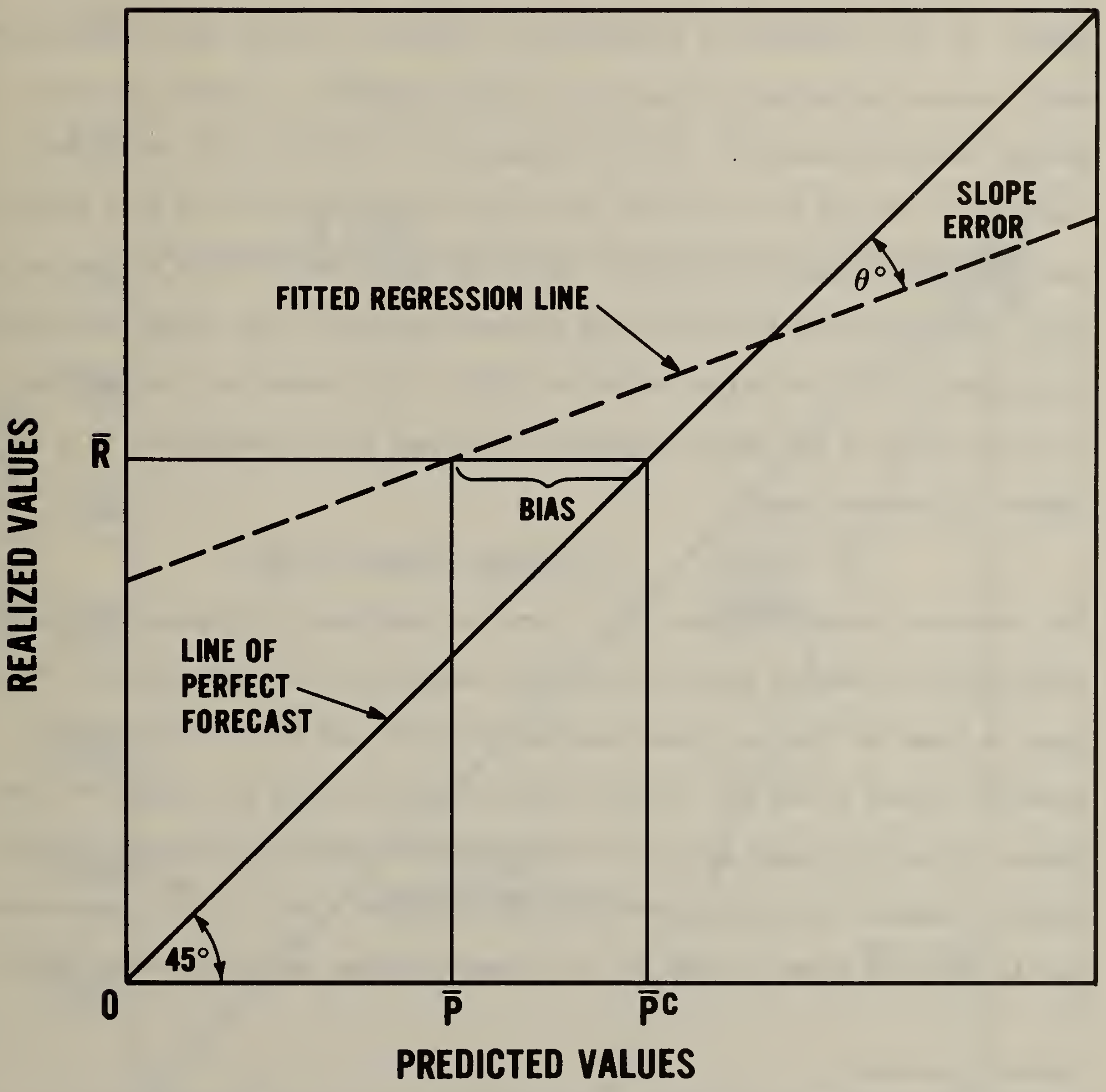
[3] Victor Zarnowitz, An Appraisal of Short-Term Economic Forecasts, National Bureau of Economic Research, New York 1967.

Figure 4.6 provides an example of how a prediction-realization diagram can be used to decompose forecast errors into two systematic components. In the figure, the realized values are measured along the vertical axis whereas the predicted values are measured along the horizontal axis. The 45 degree line which divides the diagram is referred to as the line of perfect forecast, since along this line the predicted value is equal to the realized value. Once all of the predictions and realizations have been tabulated, it is possible to calculate the average realized value, $\overline{R}$, and the average predicted value, $\overline{P}$. The difference between the two, $\overline{R} - \overline{P}$, is the model's bias and is so labeled on Figure 4.6. Thus, based on this computation, it is possible to correct for bias by moving $\overline{P}$ rightward along the horizontal axis to the point $\overline{P}^c$, the corrected average. Unfortunately, this correction ignores another source of forecast error. This error, which is also systematic, involves an overprediction (underprediction) of the realized value when that value is high. This problem may be addressed by fitting the following regression line

$$R(t) = \alpha + \beta P(t) + v(t)$$

where R(t) is the realized value in period t, P(t) the predicted value and v(t) is a "white noise" process error term. It is important to point out that P(t) should be used as the explanatory variable in the regression because it is available before R(t). In the case of a perfect forecast, all observations lie on the line of perfect forecast which implies that α is equal to zero and β is equal to one. The difference between the estimate of β, $\hat{\beta}$, and

FIGURE 4.6 EVALUATING A FORECAST WITH THE THEIL PREDICTION-REALIZATION DIAGRAM[a].



[a]Source: Mincer, Jacob and Victor Zarnowitz. Op Cit., p. 7.

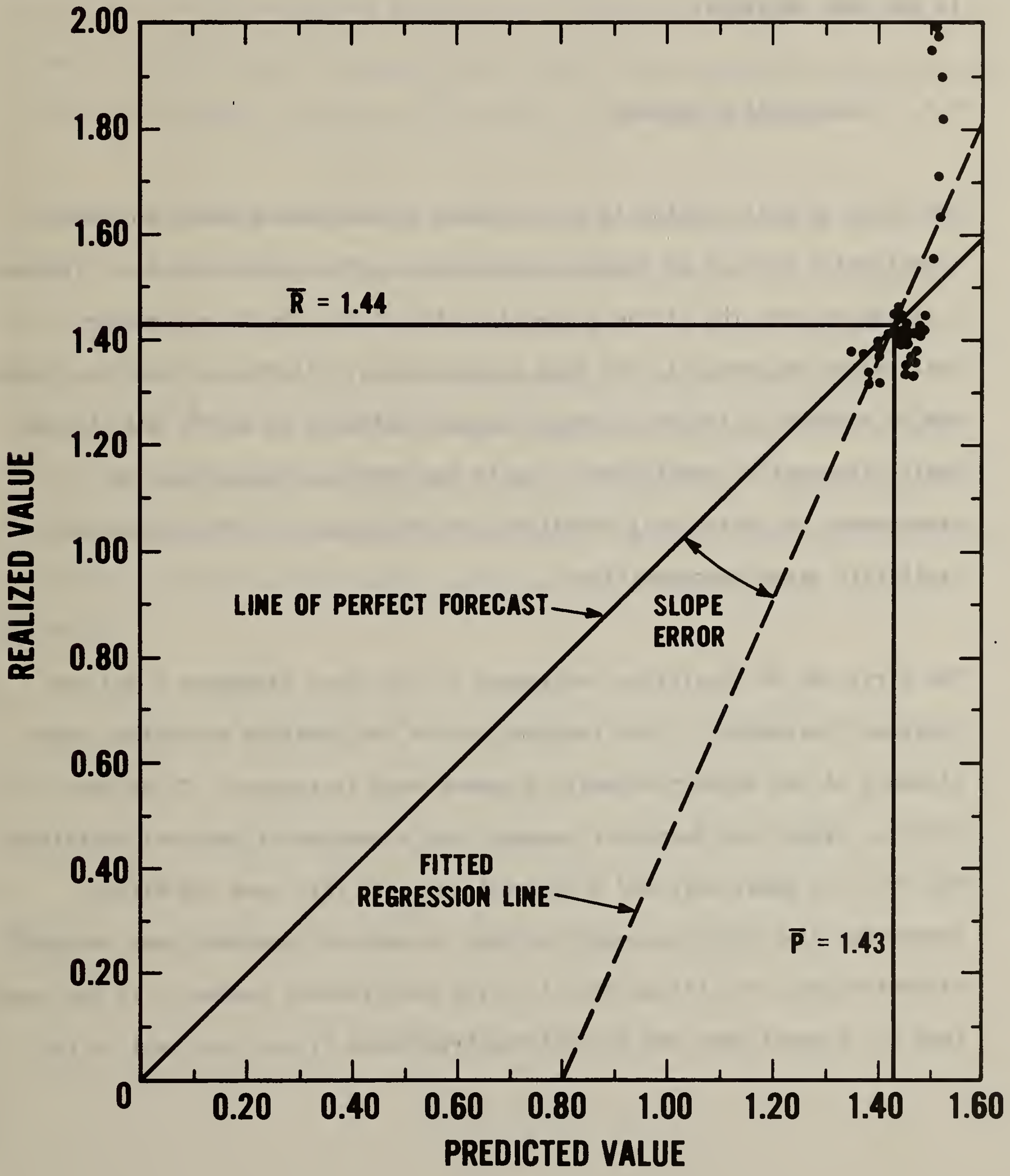
one can be used to compute the slope error θ shown on Figure 4.6.[1]

The Theil prediction-realization diagram thus permits the analyst to "correct" the forecasts by first translating the fitted regression line from the point $(\overline{R}, \overline{P})$ to the point $(\overline{R}, \overline{P}^c)$ and then rotating the line by θ degrees. An example of this approach is illustrated in Figure 4.7 where data from the four year forecast period are plotted on a scatter diagram. In this case, the average realized value, $\overline{R}$, and the average predicted value, $\overline{P}$, are almost equal ($\overline{R} = 1.44$ and $\overline{P} = 1.43$) but the fitted regression line is much steeper than the line of perfect forecast. Note the eight observations on the extreme right of the diagram and which form a nearly vertical line; these observations correspond to the last eight months of 1979. It is important to recognize that the slope of the fitted regression line may be very sensitive to a small number of observations.

The previous discussion should also serve to highlight the danger that merely correcting the model's forecasts without recognizing the importance of events, such as those of the last three quarters of 1979, may be an inappropriate solution, since it may not be the forecasts which should be changed but the model. Since the model purports to explain the underlying process, if that process changes, one should first determine if the model is still appropriate and if not take steps to make it so. These problems which adversely affect

[1] Through a property of ordinary least squares the fitted regression line will pass through the point $(\overline{R}, \overline{P})$.

FIGURE 4.7 THEIL PREDICTION-REALIZATION DIAGRAM FOR THE FOUR YEAR FORECAST PERIOD: JANUARY 1976 THROUGH DECEMBER 1979.



the value of qualitative methods are, to some extent, mitigated through the use of the classical quantitative and information theoretic methods discussed in the next sections.

4.1.2 Quantitative Methods

The focus of this section is on four sets of statistics which provide a quantitative measure of forecast performance. The statistics are: (1) the α, β estimates from the fitted regression line in the Theil prediction-realization diagram; (2) the mean squared error; (3) the ratio of the residual sum of squares to the total sum of squares referred to as R^2, and (4) the Theil U-inequality coefficient. As in the previous subsection, the development of topics will highlight the advantages of techniques which facilitate error decomposition.

The first set of statistics, estimates of the slope parameter β and the intercept parameter α, were touched upon in the previous subsection where elements of the Mincer-Zarnowitz argument were introduced. To be more precise, Mincer and Zarnowitz contend that a measure of forecast efficiency was for α to equal zero and β to equal one. In this case the fitted regression line coincides with the line of perfect forecast; some residual variation about the fitted line is still anticipated, however. In the event that the α equal zero and β equal one hypothesis is not true, all of the variation can not be attributed to the additive nature of the error term and it becomes possible to decompose the forecast variance into three components. These components are: (1) a bias component measuring the difference between $\overline{R}$ and $\overline{P}$; (2) a slope component measuring the difference between the slope of the line of perfect forecast and the slope of the fitted regression line; and (3) a residual component measuring the effects of the random error.

In order to perform this decomposition, another statistic known as the mean squared error must be introduced. (The root mean squared error is the positive square root of the mean squared error.) The mean squared error can be calculated in either of two ways depending on whether one wishes to forecast the level of a variable or the changes in the level of a variable. Using the notation of the previous subsection, let R(t) be the realized value of the variable under consideration in period t and P(t) the predicted value. The actual relative change, r(t), and predicted relative change, p(t) are therefore

$$r(t) = (R(t) - R(t-1))/R(t-1)$$

and

$$p(t) = (P(t) - R(t-1))/R(t-1)$$

The mean squared error for levels, MSE(L), is thus

$$MSE(L) = \frac{1}{n} \sum_{t=1}^{n} (P(t) - R(t))^2$$

and for changes, MSE(C),

$$MSE(C) = \frac{1}{n-1} \sum_{t=2}^{n} (p(t) - r(t))^2 = \frac{1}{n-1} \sum_{t=2}^{n} ((P(t) - R(t))/R(t-1))^2$$

where n is the number of prediction realization pairs.

Both mean squared error measures may be broken into two major components, bias and variance of the prediction errors. The first of these components, referred to as the bias component, measures the extent to which the size of the mean squared error is due to a tendency to overestimate or underestimate a value of the forecast variable. The variance of the prediction errors, may be further decomposed into a regression component and a residual component.[1] It is this decomposition which is consistent with the Mincer-Zarnowitz argument.[2]

[1] Henri Theil, 1966, op. cit.

[2] An alternative decomposition consists of a variance component (i.e., the difference in the variance of the predictor series P(t) and the realized series R(t) and a covariance component due to imperfect covariance between the P(t) and R(t) series). The covariance component is the most dangerous source of forecast error because a predictor series is unlikely ever to be perfectly correlated with the realized series. Both the bias and variance components, however, could be reduced if additional information were incorporated into the forecasting process.

A measure which is commonly used during the estimation phase of model development is the R^2 statistic. This statistic measures the proportion of the variation of the dependent variable, Y, which is explained by the model. R^2 has sometimes been used as an informal measure of goodness of fit. As such, it is used in a limited sense to compare the validity of alternative model specifications. The R^2 measure varies from 1, a perfect fit, to 0, no variation is explained by the model. Consequently, if one used R^2 as an objective criterion, the model which produced the highest R^2 would be revealed as "best."

Unfortunately, there are several problems associated with the use of the R^2 statistic. In the first place, the basis for all statistical results proceed on the assumption that the model has been correctly specified. Consequently, the R^2 statistic does not provide the needed flexibility to differentiate among alternative specifications which include different sets of explanatory variables. Second, the R^2 statistic is designed to be used in the model estimation phase (between time period T_1 and T_2 in Fig. 4.1). It is important to note that during this time period the model builder constrains the estimates of the parameter values in the model so as to minimize the residual sum of squares. However, during the ex ante or modified ex post forecast, the model builder does not have the ability to minimize the sum of squares of the forecast residuals. The previous statement can be justified based on the observation that minimizing the residual sum of squares implies that the realized values are known, which is clearly not the case in the usual forecasting environment. As a result, the formula usually used to compute the R^2 statistic is based on the following formula which ignores a cross product term:

$$\sum_{i=1}^{n} (Y_i-\overline{Y})^2 = \sum_{i=1}^{n} (Y_i-\hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i-\overline{Y})^2$$

where

Y_i = the realized value in period i;

$\hat{Y}_i$ = the predicted value in period i;

$\overline{Y}$ = the average of all realized values;

$\sum_{i=1}^{n} (Y_i-\overline{Y})^2$ = variation in Y;

$\sum_{i=1}^{n} (Y_i-\hat{Y})^2$ = residual sum of squares; and

$\sum_{i=1}^{n} (\hat{Y}_i-\overline{Y})^2$ = regression sum of squares.

The cross product term not included in the above equation,

$$2 \sum_{i=1}^{n} (Y_i-\hat{Y}_i)(\hat{Y}_i-\overline{Y}),$$

is indeterminate in sign in the ex ante or modified ex post forecast because, in the absence of the constraints imposed during the model estimation process either,

$$\sum_{i=1}^{n} \hat{\varepsilon}_i \neq 0 \text{ where } \hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

or

$$\sum_{i=1}^{n} \hat{\varepsilon}_i \hat{\beta}_j X_{ji} \neq 0$$

for $j = 1, \ldots, k$, where the X_{ji}'S are explanatory variables and the $\hat{\beta}_j$'S are the estimated coefficients.

Since the R^2 statistic ignores a component of forecast variation, one would expect the rankings to differ from those produced by the root mean squared error in some cases.

The fourth quantitative measure of the accuracy of a model's forecasts is the Theil inequality coefficient.[1] This statistic is designed to measure the performance of a model in predicting changes for the variable under consideration. The U inequality is defined by the expression

$$U = + \{[\sum_{t=2}^{n} (p(t) - r(t))^2] / \sum_{t=2}^{n} r(t)^2\}^{1/2}$$

[1] Henri Theil, 1966, op. cit.

where p(t) and r(t) are the predicted relative change and realized relative change[1] for period t.

A brief examination of the U-inequality reveals that it assumes values between zero and infinity. The smaller the value of the inequality coefficient, however, the better is the forecasting performance of the model. In the case where p(t) is equal to r(t) for all t, then U is equal to zero and we obtain perfect forecasts with our model. If p(t) is equal to zero for all t, then we are using what is referred to as the "naive" or zero change prediction. This forecasting strategy produces a value of one for the inequality coefficient and serves as a baseline against which a model's performance can be measured. Therefore, if U is greater than one, the predictive power of the model is worse than the zero-change prediction strategy.

Since the inequality coefficient is just a monotone (non-decreasing) transformation of the mean squared error, it may be decomposed into either bias, regression and residual proportions or bias, variance and covariance proportions. Each of these proportions gives a measure of the proportion of the total inequality which can be attributed to a particular component.

[1] Theil uses percentage change rather than relative change in his development of the inequality coefficient. Due to cancellations, however, both methods of computation produce the same value of U.

4.1.3 A Critique of the Classical Approach

The purpose of this section is to present some recent criticisms of the classical methods discussed in section 4.1.1 and 4.1.2. The approach taken follows closely the presentation in a recent article by Granger and Newbold.[1] The basic tenet of their criticism is that "the standards which a set of economic forecasts have been required to meet are not sufficiently stringent, since the object of any evaluation exercise ought to be self-critical rather than self-laudatory."

This argument can best be introduced by returning to Figure 4.3. In this figure the level of both realized and predicted series is plotted. The word level in the previous sentence is underlined because Granger and Newbold believe short-term forecasts of levels present the problem in an overly flattering light. A more meaningful representation would be a plot of changes. This point may be underscored by noting that many typical time series of economic levels are nearly a random walk.[2] Hence, one may be easily convinced, based on graphical appearances, that the model is an excellent predictor of the level of the series. In particular, Theil's naive zero change prediction strategy usually appears quite impressive. On a more

[1] C. W. J. Granger and P. Newbold, "Some Comments on the Evaluation of Economic Forecasts," Applied Economics, Vol. 5 (1975), pp. 35-47.

[2] Random walk as used in the above context refers to the process governing the time path of an economic variable, where the time path moves in steps, each step being determined by chance in regard to magnitude.

fundamental level, Granger and Newbold indicate that the use of graphical plots of levels can lead one to accept one random walk as a predictor of another independent random walk. Granger and Newbold claim that graphical representations of changes would not suffer as seriously from such limitations.

Another issue brought up by Granger and Newbold concerns the analysis of errors. This issue can be clarified through reference to Figures 4.4 and 4.5. Granger and Newbold argue that greater emphasis should be placed on an analysis of one-step ahead errors. This is because statistical decomposition can be performed more meaningfully on one-step ahead forecast errors than for arbitrary n-step ahead forecasts. Their assertion is based on the claim that the errors of an n-step ahead forecast should have autocorrelations of order n or greater equal to zero. This claim helps to explain the "smoothness" of the forecast errors in figure 4.5 as compared to those in Figure 4.4.

Turning now to the quantitative measures, Mincer and Zarnowitz suggested regressing realized values on predicted values. According to Mincer and Zarnowotiz the predictor P(t) is called efficient if α and β do not differ significantly from zero and one, respectively. Granger and Newbold argue that the Mincer-Zarnowitz definition "hardly constitutes a definition of 'efficiency' according to any acceptable interpretation of the word." They argue that any measure that looks at only the relationship between the predictor and realized series and not the magnitude and behavior of the prediction errors will give a misleading impression about the accuracy of forecasts.

Granger and Newbold also discuss the two methods for decomposing the mean squared error. In many software packages it is common practice to report the decomposition of mean squared error into the bias, variance and covariance components. However, as Granger and Newbold argue, it is hard to give any meaningful interpretation to the variance and covariance components. They consider the case where the realized series is generated by the following first order autoregressive process

$$R(t) = \gamma R(t-1) + \varepsilon(t) \quad 0 \leq \gamma \leq 1$$

where $\varepsilon(t)$ is a zero mean white noise process. Consider the predictor $P(t) = \gamma R(t-1)$. Then in the limit in large samples the bias component is zero, the variance component is $(1-\gamma)/(1+\gamma)$ and the covariance component is $2\gamma/(1+\gamma)$. If one varies γ from zero to one, the variance and covariance components can take on any values subject to the constraints that they are bounded above by one and below by zero and that they sum to one. Thus interpretation of these quantities is impossible. The second decomposition into bias, regression and residual components is more meaningful. If we consider the previous example, for this decomposition both the bias and the regression components go to zero for the optimal predictor.

When Theil first defined the inequality coefficient[1] it was found to produce misleading results. The original definition was

$$U^\circ = \left\{ MSE(C) \;/\; \left(\frac{1}{n-1} \sum_{t=2}^{n} r(t)^2 + \frac{1}{n-1} \sum_{t=2}^{n} p(t)^2 \right) \right\}^{1/2}$$

This statistic lies between zero and one. Granger and Newbold provide the following example as an illustration of how this statistic can produce misleading results:

$$r(t) = \gamma r(t-1) + \varepsilon(t) \qquad 0 < \gamma < 1$$

with a predictor series of

$$p(t) = \delta r(t-1) \qquad 0 < \delta < 1 \quad .$$

Granger and Newbold proceed to show that in large samples $(U^\circ)^2$ tends to $1 - [2\delta(1+\gamma)/(1+\delta)^2]$ which is minimized for δ equal to one (which maximizes the variances of the predictor series) rather than for the optimal value of $\delta = \gamma$. (The corrected inequality coefficient does not suffer from the previous criticism and is minimized for the optimal value, γ.) Another problem with

[1] Henri Theil, Economic Forecasts and Policy, North-Holland Publishing Company, Amsterdam, 1961.

Theil's inequality coefficient focuses on its misuse. Although any statistic is open to abuse, the inequality coefficient is often misinterpreted because it is frequently output with software packages within which the analyst is forecasting levels and not changes. Since the level of an economic variable is often an order of magnitude larger than its changes, the calculated inequality coefficient may be low. Consider the following deterministic model

$$R(t) = \delta + .1\delta \cos(\pi t/2).$$

If we trace the model through one complete cycle, at time zero R(o) is equal to 1.1δ; at time one, R (1) is equal to δ; at time two, R(2) is equal to .9δ; and at time three, R(3) is equal to δ. Now if the analyst were charged with forecasting the level of the series and decided to use the observed value at the beginning of each cycle as the predictor, P(t), throughout the cycle, we would have

$$\hat{P}(t) = 1.1\delta \quad t = 1, 2, 3, 4, . . .$$

If we assume the process goes through n complete cycles, then the calculated value of the inequality would be approximately 0.214, which is much closer to zero than one. If on the other hand, the changes, $\hat{p}(t)$, were computed and used to calculate the inequality coefficient, a value of approximately 1.225 would result. Thus, in this case, the naive zero-change forecast strategy would have clearly been superior.

The previous discussion has aimed at highlighting some of the complexities involved in forecast evaluation. Although the criticism of the classical approach for failing to address these complexities poses formidable difficulties, the approach outlined in the next section addresses most of them. This occurs because the use of information theory preserves much of what is desirable in the classical approach but does not suffer from the more serious criticisms of Granger and Newbold with regard to the adequacy and efficacy of a particular model in a given forecasting situation. In addition, an information theoretic approach provides further support for the strategy of combining forecasts discussed by Granger and Newbold.

4.2 An Information Theoretic Approach to Accuracy Assessment

The purpose of this section is to demonstrate how concepts from the discipline of information theory can be applied to the problem of accuracy assessment. More precisely, the concept of entropy will be used to measure the information content of a model's forecast as well as how rapidly that content changes as a function of the length of the forecast period and the model specification chosen. Entropy is a term used widely in statistical mechanics and communication theory. The definition chosen here comes from communication theory[1] and refers to the variability--and thus degree of uncertainty--of the outcome of a particular process (or, in the vernacular of communication theory, signals which contain information).

[1] S. Goldman, Information Theory (New York: Prentice-Hall, Inc., 1953).

4.2.1 Some Basic Concepts

In order to promote a more complete understanding of the topics in information theory presented in the subsequent sections, some of the basic concepts will be developed on an intuitive level.[1] As a basic premise, we require that information be defined as a function of one argument--the probability that an event will take place before we have received a reliable message (market signal) that it did in fact take place (i.e., was realized). A second premise is that the information content of the message is a decreasing function of the probability of its occurrence. In principle, one could choose any decreasing function; however, it is customary to take the logarithm of the reciprocal of the probability of an event taking place. The rationale behind the selection of the logarithmic function focuses upon its additivity in the case of independent events.

Consider the case where we have a set of n events, $x_1, \ldots, x_n$, (e.g., n candidate motor gasoline prices for the next period) and that they define a complete system since one and only one of them will occur. Thus, if the probabilities are $P(x_1), \ldots, P(x_n)$, they should be non-negative and sum to one. Now if we receive a reliable message stating that a particular event, x_1, has occurred, the information content of the message, is a function of

[1] The presentation given here parallels that given in Henri Theil, Economics and Information Theory (Amsterdam: North-Holland Publishing Company, 1967).

the probability of occurrence, $h(P(x_i))$, which is equal to $-\log(P(x_i))$. However, before the message is received we do not know with certainty how large the information content will be since x_i is just one of the n events. Thus the information content of a message could be $h(P(x_1))$ if the first event is realized or $h(P(x_2))$, . . . , $h(P(x_n))$ for each of the second through n^{th} events. In order to get around .this problem, we are forced to rely on an expected measure of information content. Since event x_i has a probability $P(x_i)$, the message that x_i has occurred will be received with the same probability. The information content is therefore $h(P(x_i)))$ with probability $P(x_i)$. When one considers all events, the expected information content, now called entropy and denoted H(X), is thus

$$H(X) = \sum_{i=1}^{n} P(x_i)h(P(x_i)) = -\sum_{i=1}^{n} P(x_i)\log(P(x_i)) \quad .$$

In forecasting the price level of a key energy commodity, such as motor gasoline, it is useful to generate a scatter diagram (see Figure 4.2 in Section 4.1.1) as an aid in computing the variability of motor gasoline prices from a time series of realized values. Thus, entropy characterizes the unexpectedness or noise inherent in the stochastic process which governs the time series. Entropy can also indicate how difficult, in relative terms, it would be to forecast one variable (e.g., motor gasoline prices) in comparison to some other variable (e.g., residual fuel oil prices). Furthermore, the quantitative nature of the entropy measure regarding the variability of the process being modeled establishes a more realistic approach to the evaluation of a particular model's forecasts. For example, one would expect, *ceteris paribus*, that a commodity having a time series which was fairly stable (low entropy) would be easier to forecast than one which was highly variable (high entropy). Although one could assert the same observation from a visual examination of the data, such a measure would only be subjective. Through the use of information theory, however, a quantitative measure can be attached to the variability of different processes. This measure, if based on forecast evaluations, could also provide an indication of the information gain of a particular model over some alternative formulation.

Marginal entropy is the first of the three information theoretic terms which will be used in assessing forecast acccuracy. Marginal entropy is concerned with a time series of *realized* values. It shows the inherent variability of the stochastic process which underlies that time series. In practice it would be computed from a set of *historical* realizations of the time series under consideration. In order to compute the marginal entropy of a given time series, however, one must first select a probability distribution which characterizes the stochastic process which underlies that time series. The probability distribution chosen may be either discrete or continuous depending on the process being modeled. If the process is discrete, the marginal entropy, $H^D(X)$, is defined with respect to the probabilities associated with the occurrence of that event:

$$H^D(X) = -K \sum_{n=1}^{N} P(x_n) \log P(x_n)$$

where

K = a constant associated with the base[1] adopted (since all measures will be computed with respect to the same base, without loss of generality K can be taken equal to 1);

x_n = individual events (e.g., corresponding to the potential price levels within the N-period long time series); and

$P(x_n)$ = the probability of the event x_n taking place.

If the process is continuous, the analogue of the probability of a particular event occurring is given by the density function f(x), and the information content of a message by -log (f(x)). The marginal entropy, $H^C(X)$, is thus defined as

$$H^C(X) = -K\int_{-\infty}^{\infty} f(x) \log (f(x))dx$$

where f(x) = the probability density function;

K = a constant associated with the base adopted (as in the previous case K can be taken equal to 1).

[1] Bases and their respective units commonly used in communication theory and information theory are: base 2 referred to as bits; base e (the base of the natural logarithms) referred to as nits; and base 10 referred to as decibels.

For purposes of illustration, should the probability distribution which characterizes the process be revealed as normal with a mean of zero and a variance of σ^2, it can be shown that marginal entropy is equal to

$$H^c(X) = \frac{1}{2} \log (2\pi e\sigma^2)$$

where $\log (2\pi e\sigma^2)$ = the logarithm to base e of $(2\pi e\sigma^2)$; and

$e \doteq 2.7183$ (the base of the natural algorithms).

Marginal entropy can also be used to construct a measure of the relative difficulty of forecasting one time series vis-a-vis another. For example, suppose two time series, say prices for motor gasoline and residual fuel oil, were to be forecast. If the processes were normally distributed and the variances associated with the motor gasoline price series, $X_1(t)$, and residual fuel oil price series, $X_2(t)$, were $\sigma_1{}^2$ and $\sigma_2{}^2$ respectively,[1] then a quantitative measure of the relative difficulty of forecasting residual fuel oil prices, $X_2(t)$, with respect to (WRT) motor gasoline prices, $X_1(t)$, for the same time period would be

$$RD(X_2WRTX_1) = H_2{}^c(X_2) - H_1{}^c(X_1) \quad .$$

[1] In the discussion which follows it is assumed that the two series under consideration have been normalized. This step can be accomplished by either deflating (dividing) each realized value $X_i(t)$, i = 1, 2, by the mean of the respective process X_i, i = 1, 2, or more appropriately, deflating each value prevailing at a common point in time such as January 1980.

From the properties of natural logarithms it can be shown that for the normally distributed case the measure reduces to

$$RD(X_2WRTX_1) = \log(\sigma_2/\sigma_1)$$

which is an increasing function of σ_2, the standard deviation of the second series. It is important to point out that this measure is quantitative and as such can be used as a benchmark for comparing the impact of the relative difficulty of forecasting on the accuracy assessment of a model or set of model forecasts. If, for example, the relative difficulty measure were equal to five, then one would not expect the logarithm of the ratio of the standard deviations of the forecast errors to be less than five. If the logarithm of the ratio were less than five, it would be an indication that the model used to forecast residual fuel oil prices was relatively more accurate than the one used to forecast motor gasoline prices. This analysis of forecast errors provides a motivation for introducing another entropy calculation.

The formal process of forecasting implies the existence of a mathematical representation of the time series under study. This representation or "model" may take on many alternative forms depending upon the needs, capabilities, and objectives of the person(s) making the forecast(s). Given the existence of a model of the process, the concept of conditional entropy can be introduced fairly easily. The models presented in the next section produce a set of values, a time series, which contains estimates of the future values of the real price of motor gasoline.[1] The historical series of forecasts (ex ante values) y(t), or "modified ex post" values if a forecast history is not available to the evaluator, may then be compared to the historical series of realized (ex post) values, x(t), to determine the probability distribution of the prediction error. Similarly, one could compute a joint distribution[2] [P(x,y) if the processes are discrete or f(x,y) if the processes are continuous] and a conditional distribution[3] [P(x|y) or f(x|y)] to measure how much of the uncertainty of the original time series still remains after one has made use of the forecasts produced by the model. For a given set of realizations and predictions, the remaining uncertainty or conditional entropy is defined as

$$H^d\ (X|Y) \ = \ -\sum\sum P(x,y) \log P(x|y)$$

if the process is discrete and

[1] The forecasts presented in the next section are "modified ex post" forecasts. They are designed to simulate the forecasting strategy of an analyst making an ex ante forecast.

[2] A joint distribution establishes a relationship between two or more sets of events. This concept can be better understood by referring to the prediction-realization diagram in Figure 4.7 where both predictions and realizations are plotted on a scatter diagram. Now suppose one were interested in the event R_i and the event P_j from, the set of all possible R-P combinations. The joint distribution would then tell you the probability of both R_i and P_j taking place.

[3] A conditional distribution tells you the probability of a particular event, say R_i, taking place given that another specified event, say P_j, has taken place. Referring once again to Figure 4.7, the probability of a particular realization taking place, say R = 1.80, is given by examining a vertical slice of the figure above some specified prediction, say P = 1.50.

$$H^c(X|Y) = -\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) \log (f(x|y)) \, dydx$$

if the process is continuous.

Since conditional entropy provides a measure (based on the historical simulated performance of the model) for the remaining or residual uncertainty once a set of forecasts have been made, it provides a natural means through which forecasts can be ranked. For example, if one were interested in forecasting the price of motor gasoline at various times in the future (say 1, 3, 6, 12, and 24 months ahead), it is possible that several model specifications would be considered as appropriate. Such an approach is consistent with both Bayesian econometric theory[1,2] and previous empirical studies.[3,4] One set of specifications which is illustrated in the next section is an econometric model which captures key policy variables and a regression model driven

[1] In Bayesian estimation the parameter is looked upon as a random variable which has a prior distribution reflecting either the strength of one's belief about the possible values it can assume, or collateral information. The posterior distribution obtained by combining this prior with the likelihood function via Bayes theorem then shows how the prior beliefs are modified by observation, or more specifically, by information provided by actual data.

[2] P. J. Harrison and C. F. Stevens, "Bayesian Forecasting," Journal of the Royal Statistical Society, No. 3 (1976), pp. 205-228.

[3] Anthony E. Bopp and John A. Neri, "The Price of Gasoline: Forecasting Comparisons," The Quarterly Review of Economics and Business, Vol. 18, No. 4 (Winter 1978), pp. 23-33.

[4] C. W. J. Granger and P. Newbold, Forecasting Economic Time Series (New York: Academic Press, 1977).

by the exogeneously determined price of crude oil. (For a detailed discussion of these models the reader is referred to Section 4.3.) Associated with each model specification is a measure of residual uncertainty referred to as conditional entropy which may be calculated directly from the respective models' error terms. If the error terms are normally distributed and the models produce unbiased forecasts, then the conditional entropy of each model, i, is given as

$$H^c(X|Y_i) = \frac{1}{2} \log (2\pi e \sigma_i^2)$$

where i = 1, econometric; and 2, regression.

Consider the case where the models are used to make forecasts of the real price of motor gasoline in the next period. Now if for these one-step ahead forecasts we find $\sigma_1^2 < \sigma_2^2$, it would be reasonable to ask if the econometric model (i=1) produces <u>significantly</u> better forecasts (in terms of residual uncertainty) than the regression model. For this case, as well as for the general n-step ahead forecast, one would wish to test the null hypothesis

$$H^c{}_2(X|Y_2) - H_1{}^c (X|Y_1) > 0$$

versus the alternative hypothesis

$$H^c{}_2(X|Y_2) - H_1{}^c (X|Y_1) < 0.$$

Making use of the properties of logarithms once more, it can be seen that the test on the null hypothesis reduces to

$$\frac{1}{2} \log (\sigma_2{}^2/\sigma_1{}^2) > 0$$

versus the alternative hypothesis

$$\frac{1}{2} \log (\sigma_2{}^2/\sigma_1{}^2) < 0 \quad .$$

Guidelines for carrying out this "variance ratio" test for a set of one-step ahead forecasts are given in the text by Granger and Newbold. Unfortunately, due to the correlation structure of the errors for n-step ahead forecasts (n > 2) it is not possible to test the null hypothesis directly, although a simulation approach might offer a satisfactory solution. Under the assumption that one could satisfactorily perform the "variance ratio" test, it is possible that the ranking produced for a one-step ahead forecast may not hold for an n-step ahead forecast. Thus although a regression model may be revealed as best for a one-step ahead forecast, it may be revealed as worst on the average for a 24-step ahead forecast.

Transinformation, the third entropy calculation, is a measure of the information content of the model. On an intuitive level, it can be viewed as the information transferred by knowledge gained from the forecasts, y(t), to make the realizations, x(t), better defined. Transinformation, denoted as T(X,Y), may be defined as the "information" contained in the forecast. It results from a reduction in the uncertainty of the process due to the model. More succinctly, transinformation is equal to marginal entropy minus conditional entropy or

$$T(X,Y) = H(X) - H(X|Y) \quad .$$

Depending on how well the model "tracks" the commodity under consideration, transinformation can be positive, zero, or negative. An infinitely large positive value of T(X,Y) indicates a perfect fit; a positive value means that some information is being conveyed. A value of T(X,Y) less than or equal to zero would imply that the model's forecasts may be spurious; at best the model adds no information and in fact it may even be misleading.

As discussed earlier, conditional entropy can be used to rank models based on their forecasts. The application of the third measure, transinformation, provides a technical base for two additional refinements. These refinements relate to: (1) the maximum number of steps ahead which a model can forecast before its results may become spurious; and (2) the way in which the forecasts of a set of alternative model specifications can be optimally pooled.

The first issue is of crucial importance in all forecasting activities because as the forecast period is extended, the variance of the model's error term increases. At an intuitive level, the point in relative time at which transinformation becomes negative is attractive since it can be easily calculated and compared against the forecast horizon. Consequently, if the modeler wished to forecast 24 months ahead and transinformation turned negative six months ahead, then some real consideration should be given to the adequacy of the model. On a more subtle level, it is possible for a model to look "reasonable" as the forecast horizon is extended based on other accuracy measures, e.g., root mean square error, but still have a negative value of transinformation. This is quite serious because the model user has no _a priori_ test for determining how large the root mean square error can grow before use of model forecasts becomes dangerous. Transinformation fills that void while still preserving many of the desirable attributes of the root mean square error statistic.

The issue of optimally pooling the forecasts of several models has been addressed in a growing number of papers.[1,2,3] The usual approach is based on pooling forecasts to minimize the variance of the aggregated forecast. An optimal pooling based on the information theoretic approach can be defined as

[1] D. J. Reid, _A Comparative Stuidy of Time Series Prediction Techniques on Economic Data_, Ph.D. Thesis, Nottingham University, 1969.

[2] J. M. Bates and C. W. J. Granger, "The Combination of Forecasts," _Operational Research Quarterly_, Vol. 20, No. 4 (1969), pp. 451-468.

[3] C. W. J. Granger and P. Newbold, 1977, _op. cit_.

that pooling which maximizes the information content of the aggregated forecast. Now if the models are unbiased and their errors are distributed as a white noise process, it can be shown that a sufficient condition for maximizing the information content of the aggregated forecast is to pool the models so as to minimize their variance. Thus the technique for pooling developed by Reid, Bates and Granger, and Granger and Newbold, can be applied directly without loss of generality. Another desirable attribute of the information theoretic approach is that the information content of the model lends itself to a decomposition into a variety of components which are sensitive to changes in the forecast horizon.

4.2.2 Calculation Techniques

The physical task of calculating the three entropy measures requires more than a direct application of the definitions given in the previous section. In particular, certain decisions must be made regarding the structure of the forecasting process as well as the appropriateness of a particular probability distribution. Given a satisfactory resolution of these issues and presuming a continuous probability density function is chosen, one is also faced with the issue of choosing a discretized version of the continuous distribution which, although approximate, offers several conceptual advantages versus the "theoretically exact" continuous solution.

The actual forecasts of key energy commodities are not made in a vacuum. In practice there would not be a single forecast for a particular future date, but rather a sequence of forecasts for which the time between the future date (the forecast horizon) and the date at which the forecast is made is decreasing. For example, the initial forecast may attempt to predict the monthly change in the price of motor gasoline for a time 24 months in the future. Consider the following scenario which defines a 24-step ahead forecast: the analyst is located at December 1978, has estimated the model, and wishes to forecast the monthly change in the price of motor gasoline for December 1980. Periodically, and as additional information become available, new forecasts are made which include the December 1980 motor gasoline price figure. (In these cases the end points of the forecast would be later than December 1980.) If we denote each of these "stages" as s_h, h = 1 . . ., f, where s_f is the final stage where the value is realized, it then becomes possible to see how the uncertainty of the outcome changes with each stage. To do this, one would first obtain a forecast estimate for each stage, for a given month-year or quarter combination, and for a given variable. Thus if i were the index of the variable (e.g., motor gasoline, distillate, residual fuel oil, etc.), t were the index of the month-year combination, and h were the index of the stage, a given forecast estimate, X_{ith}, could be thought of as the estimated change from the previous month in the price level for the month-year combination denoted by t. From a policy viewpoint we are therefore concerned with the difference between the predicted and realized values of the changes in the price level, and the variance of that error, σ^2_{ith} .

Under certain plausible circumstance shown by Theil,[1] it is possible to decompose the above mentioned variance into a commodity component, A_i^2; a month-year component, B_t^2; and a stage component, C_h^2. That is, $\sigma_{ith}^2 = A_i^2 B_t^2 C_h^2$. Estimates for the A_i's, B_t's, and C_h's can be obtained by following an iterative process outlined by Theil. With regard to the variance decomposition presented above, it becomes possible to estimate the information gain (reduction in uncertainty) of going from a particular stage, h, to the next stage, h+1. (Similarly one could define the information gain of going from a particular stage, h, to some higher stage, $h+j < f$.) Recalling the formula for the conditonal entropy of the normal distribution presented in the previous section, it can be shown that the information gain of going to the next stage is

$$H^c(X_{itf}|X_{ith}) - H^c(X_{itf}|X_{ith+1}) = \log (C_{h+1}/C_h) \quad .$$

Through an analysis of the stage effect it is possible to compare alternative model specifications to see how rapidly each stage resolves the uncertainty inherent in the underlying process.

It was mentioned earlier that one of the advantages of transinformation was a capacity for decomposition into a variety of components. More precisely, the

[1] Henri Theil, 1966, op cit.

information gain associated with a stage transition of k steps, log $(C_{h+k}|C_h)$ $k \geq 1$, can be decomposed into three components. These components are due to: (1) refinements in the values of the variables used in the model, (2) more precise estimates of the true parameters of the model, and (3) the basic structure of the model. Prior to the computation of the information decomposition, however, one should first perform a structural change test. This approach is advisable because, as pointed out earlier, the model purports to explain the underlying process and should that process change it will be necessary to determine if the model is still appropriate. The structural change test is one way through which changes in the underlying process can be detected. In the presence of structural change, assessing whether or not the model is still appropriate is a complicated issue which goes beyond accuracy assessment.

The first component of the information content of the model relates to refinements in the values of the model's variables. These refinements are important even for simple regression models because as the forecast horizon is extended, the estimated values of the variable may diverge from those realized. An econometric model, on the other hand, may contain lagged endogenous variables whose values are determined exogenously (outside the model framework). Consequently, each exogenous variable must be used in making a forecast of the future value of the variable under consideration.

These forecasts, such as future values of gross natural product, national personal income, or the urban consumer price index, may be based on outputs from a large macroeconomic model, on extrapolations based on a time-series model of that variable, or some "naive" model. The value of the lagged endogenous variables are determined within the framework of the model. These forecasts thus represent a potential scenario regarding the future values of the endogenous and exogenous variables. Consequently, as more of the process unfolds, the scenario must be refined to reflect differences between expectations and realizations. In some cases the refinement may be insignificant and in others it may reflect a major change in policy, balance of payments, or other factor(s). The effects that these refinements have can be measured, on the average, by replacing the forecast values (or simulated forecasts of these values) for each variable with its realized values. As a result, the information content of the model should increase somewhat because it now reflects a more precise characterization of the process being modeled. The mechanics of this process would involve all stage transitions of k steps, $k > 1$. From the earlier discussion of the stage effect, it can be shown that the average information gain in going from stage one to stage h, $\hat{IG}(h)$, is defined by:

$$\hat{IG}(h) = \log(C_h/C_1) \quad .$$

Of the <u>total</u> average information gain, $\hat{IG}(h)$, the amount which can be attributed to refinements in the values of the variables in the model is denoted $\hat{IG}_1(h)$, where the subscript refers to the first of the three components mentioned earlier.

The second component of the information gain in going from the first to the h^{th} stage focuses on more precise estimates of a model's parameters. As more data become available, it is often advisable to reestimate the values of a model's parameters. This step is advisable because it may result in a reduction in the variance of the parameter estimate which in turn <u>should</u> produce a reduction in the variance of the predicted value. Since any reduction in the variance of the predicted value will reduce conditional entropy, some gain in information will result. As in the first decomposition, the mechanics would involve the use of the realized values of the endogenous and exogenous variables to reestimate the model(s) under consideration. Note that this step uses all the data used in computing the first decomposition measure so that it is a net addition to the information content of the model. The amount of the total average information gain accounted for by this component is denoted $\hat{IG}_2(h)$.

The third component of information gain is attributable to the basic structure of the model. It enters the calculation in the form of a residual. The inclusion of this component is required because refinements not covered under the first two components may be introduced into the model, the data base(s) used for a particular set of calculations may change, or the definition of a variable may change.

From the previous discussion we can thus assert that

$$\log(C_h/C_l) = \hat{IG}(h) = \hat{IG}_1(h) + \hat{IG}_2(h) + \hat{IG}_3(h)$$

which can be plotted as a function of h to obtain a "dynamic decomposition" of the information gain associated with a stage transition. Similarly, it is possible to form ratios $(\hat{IG}_j(h)/\hat{IG}(h)$, j = 1, 2, or 3) which show the relative sensitivity of the gain in information of moving from a lower to higher step due to an endogenous/exogenous variable component or a parameter value component. The ratios formed could thus provide a measure of model sensitivity as well as indicate conditions under which emphasis should be placed on models which offer greater policy flexibility.

For example, suppose one were to break the forecast period into two parts, one with an exogenous/policy variable (or weighted average of exogenous/policy variables) changing by at least ten percent versus one in which they changed by less than ten percent. Then, based on this policy decomposition, one would expect that a model with a low value of the first component of average information gain for the rapid change case (> 10 percent) would not be very useful if that policy scenario were pursued. Thus, if a policy maker has a particular scenario in mind, say a tax surcharge of 15 percent, then a model with a high value of the first component of average information gain for the rapid change case should be weighted more heavily than one with a low value for the desired n-step ahead forecast.

An issue which might be revealed through an analysis of the $\widehat{IG}_2(h)/\widehat{IG}(h)$ ratio would be the potential importance of a varying-parameter model. As Maddala[1] shows, if the estimated relationships are based on an optimization problem which involves some of the policy variables, then the economic agents, in determining their optimal activity level, would be taking these policy variables into account in their decisions. The variables would thus be entering the model not in an additive fashion but as determinants of the parameters in the model, which would imply that a varying-parameter model would be appropriate. Ratios of $\widehat{IG}_2(h)/\widehat{IG}(h)$ close to one would be a strong indication that variations in the parameter values were the source of the stage effect. In this case, a model whose parameters varied about some average value might be the best choice.[2]

A crucial step which must be taken prior to the application of any entropy measure is to determine which probability distribution is appropriate for the case at hand. This step is particularly important because the entropy measure is sensitive to the probability distribution chosen. For example, when the probability density function is continuous on $(-\infty, \infty)$ with a given variance, σ^2, it can be shown that the normal distribution has the greatest entropy. For a finite range and a given variance, the uniform distribution has the greatest entropy.[3]

[1] G. S. Maddala, Econometrics (New York: McGraw-Hill Book Company, 1977).

[2] P. A. V. B. Swamy and P. A. Tinsley, "Linear Prediction and Estimation Methods for Regression Models with Stationary Stochastic Coefficients," Journal of Econometrics, 12 (1980), pp. 103-142.

[3] S. Goldman, op cit.

The first step in selecting a distribution is to collect data on actual values and forecast values by stage for each variable. These data should then be grouped into intervals and plotted to construct a histogram. The histogram may then be compared to sample standard histograms. Based on these comparisons, a class of distributions which might fit the data can be hypothesized. The next step is to develop estimates (preferably maximum likelihood) for the parameters of the distribution under the assumption that the hypothesized distribution is correct. The final step is to perform a goodness-of-fit test on the data. If the fit is unacceptable, it will be necessary to hypothesize a new class of distributions and reestimate the parameters of the distribution. Several points along these lines are worth noting. First, there are at least two goodness-of-fit tests[1] which should be considered unless the analyst has strong _a priori_ beliefs about the distribution governing the data. These tests are referred to as omnibus tests because they are appropriate for any distribution; they are: (1) the Chi-square Test; and (2) the Kolmogorov-Smirnov Test. Second, it is possible to have either goodness-of-fit test accept more than one distribution as appropriate. This situation is rather common with the Chi-square Test. However, since the Chi-square Test is by far the easiest to apply, such a potential problem may not be deemed serious. Finally, it is important to recognize that the Kolmogorov-Smirnov Test is somewhat more robust (less

[1] L. Breiman, _Statistics: With a View Toward Applications_, Houghton Mifflin, Boston, 1973.

sensitive to the relaxation of the assumptions underlying the test) so if sample size is small the emphasis should be on the result of the Kolmogorov-Smirnov Test rather than the Chi-square Test.

In an information theoretic setting discrete distributions are often used to reflect qualitative characteristics whereas continuous distributions are used to refer to counts and levels which are of a more quantitative nature. Thus, given the nature of most economic variables, it is quite likely that a continuous distribution will be revealed as the one most appropriate for a given application. In using continuous distributions, however, one must recognize that not only the distribution but also the scale chosen affects the entropy calculation. In particular, both the marginal and conditional entropies would be affected. (This is why in computing the relative difficulty measure both series had first to be normalized.) The transinformation measure would not be affected, however, since it is a difference and the "scale" factor would be netted out. The same cancellation effect would occur for the calculation of the gain in information in going to the next stage.

The present discussion parallels that of several authors in the field of information theory who have advocated the use of discrete approximations to continuous distributions.[1] The arguments presented in the literature are

[1] J. Amorocho and B. Espildora, "Entropy in the Assessment of Uncertainty in Hydrologic Systems Behavior and in Mathematical Model Performance," International Symposium on Uncertainties in Hydrologic and Water Resource Systems, pp. 977-1008.

based on the assumption that the variables under study are bounded and that unbounded frequency distributions are fitted to the data primarily as a matter of convenience. If one accepts this assumption, interval widths, Δx, and the number of intervals, N, can then be defined such that the weighted summation over all intervals is bounded between some fraction, say 0.99 and 1.00. More succinctly, the probability that X is within the n^{th} interval, $n = 1, \ldots,$ N, is

$$P_n = P[x_n - (\Delta x/2) < X < x_n + (\Delta x/2)] = \int_{x_n-(\Delta x/2)}^{x_n+(\Delta x/2)} f(x)dx$$

that is,

$$P_n \doteq f(x_n)\Delta x$$

such that

$$0.99 < \sum_{n=1}^{N} f(x_n)\Delta x < 1.00 \quad .$$

The interval width chosen, Δx, could even be the same one used in building the histogram(s) to test the goodness-of-fit for the alternative probability distributions discussed earlier. Using this approach, the estimates for marginal and conditional entropy would now differ from the H(X) and H(X|Y) calculations presented earlier in section 4.2 by $-\log(\Delta x)$. (The transinformation measure, since it is a difference, is unaffected by the choice of interval width.) Suppose the increment Δx is initially chosen to reflect a variation of $\pm\alpha$ thousand barrels of motor gasoline, where α might reflect a very "tight" band about the consumption value. Now if, for all practical purposes, a forecast which was $\pm 10\alpha$ of the consumption value was satisfactory (possibly due to abnormally high stocks), the conditional entropy measure could be adjusted downward by $\log(10\alpha)$. After this adjustment has been made some values of the conditional entropy function may be less than zero. In these cases it is possible to assert that the model can predict the true consumption level with $\pm 10\alpha$. For cases where the conditional entropy measure lies above zero the model still contains residual uncertainty. However, through reference to transinformation it is still possible to define the values of the coefficients of correlation between historical forecasts and realizations for certain months, quarters, or other time periods, required to bring the predictions within $\pm 10\alpha$. Thus the discretized version of the continuous probability distribution gives both the analyst and the decisionmaker a handle on where a particular model might need refinement.

4.3 Applying the Entropy Measure to the Forecasts of a Short-Term Energy Model

The primary focus of this section is to illustrate the process of formulating and evaluating two alternative model configurations useful in forecasting the real price of motor gasoline.

4.3.1 Alternative Model Specification

In order to demonstrate the mechanics of a forecast evaluation program, the various measures discussed in Sections 4.1 and 4.2 were applied to two different model specifications. These specifications were: (1) an econometric model; and (2) a regression model. Each model specification was estimated and assessed based on monthly data on each explanatory variable compiled from January 1967 through December 1979. This period was selected both for data availability and because it highlights the basic difficulty in forecasting motor gasoline prices. This may be seen by noting that gasoline prices were relatively stable between 1960 and 1970, gradually increased between 1970 and early 1973, increased explosively between 1973 and the middle of 1974, gradually increased from the middle of 1974 through the middle of 1979, and then increased explosively.

Recalling the three phases in the evaluation of an economic forecast presented in Figure 4.1, each specification was first estimated in the period January 1967 through December 1975. A sequence of modified *ex post* forecasts[1] spanning the period January 1976 to December 1979 was then made for each

[1] Recall that a modified *ex post* forecast operates under the assumption that the analyst was situated at December 1975 on the time line (i.e., $T_2 = T_3 =$ December 1975 c.f. Figure 4.1) and was therefore forced to use forecasts for any future values of the explanatory variables. Consequently, the evaluation may be referred to as "conditional."

model. The difference between the realized values and predicted values were then computed for a one-year forecast horizon (i.e., January 1976 through December 1976), a two-year forecast horizon (January 1976 through December 1977), a three-year forecast horizon (January 1976 through December 1978), and a four-year forecast horizon (January 1976 through December 1979). Each specification was then reestimated using data from the period January 1967 through December 1976.[1] This approach was taken because forecasting models are constantly being upgraded and any critical evaluation of the forecasting process which ignored this issue was felt to be deficient. Furthermore, by augmenting the data base it is anticipated, based on theoretical considerations, that the variance of the parameter estimates will be reduced. This should translate into better forecast performance. Each model was then used to make a sequence of modified ex post forecasts covering the period January 1977 through December 1979. As in the previous case, the difference between the realized and the predicted values for one-, two-, and three-year forecast horizons were computed. A four-year forecast horizon was not attempted due to incomplete data for 1980. As a final step, each specification was reestimated using data from January 1967 through December 1977; three sequences of modified ex post forecasts for motor gasoline prices were then made for 1978 and 1979. In all, 18 ex post forecasts were made: two with a four-year horizon, four with a three-year horizon, six with a two-year horizon, and six with a one-year horizon. In order to avoid a great deal of repetition in the discussion which follows, the emphasis will

[1] In the context of Figure 4.1, T_3 is now December 1976.

be placed on the model specifications and coefficient estimates based on the estimation period January 1967 through December 1975. The description of the models follows closely that presented in the article by Bopp and Neri.[1]

An Econometric Model

The price of gasoline in this model is determined by factors that affect both the supply of and the demand for gasoline. The econometric specification presented here is a reduced-form equation from an inventory-price adjustment model used to capture the effects of supply and demand factors. Since gasoline stocks play an important role in the gasoline market, an inventory adjustment model seems appropriate. Following McCallum,[2] supply, demand, price adjustment, and inventory adjustment equations are first specified and then used to obtain the final reduced-form equation for estimation.

Dynamic demand and production equations and the market clearing equation are given in equations 1, 2, and 3, respectively. All terms used in these equations are defined in Table 4.1.

[1] Bopp, Anthony E. and John A. Neri, "The Price of Gasoline: Forecasting Comparisons," The Quarterly Review of Economics and Business, Vol. 18, No. 4 (Winter 1978), pp. 23-33.

[2] McCallum, B. T., "Competitive Price Adjustments: An Empirical Study," American Economic Review, Vol. 64 (March 1974), pp. 56-65.

Dynamic Demand Equation

$$CMG(t) = a_0 + a_1 \cdot RPG(t) + a_2 \cdot RY(t) + a_3 \cdot CMG(t-1) \qquad (1)$$

Production Equation

$$PMG(t) = b_0 + b_1 \cdot RPG(t) + b_2 \cdot RPC(t) + b_3 \cdot RPD(t) + b_4 \cdot PMG(t-1) \qquad (2)$$

Market Clearing Equation

$$CMG(t) = PMG(t) + SMG(t) - SMG(t-1) \qquad (3)$$

TABLE 4.1 VARIABLES USED IN SPECIFYING THE ECONOMETRIC MODEL

Variable	Definition
RPG	Real Price of Motor Gasoline
RPC	Real Price of Crude Oil
PMG	Production of Motor Gasoline
RY	Real Personal Income
SMG	Stocks of Motor Gasoline
CMG	Consumption of Motor Gasoline
RPD	Real Price of Distillate

Thus the quantity of gasoline consumed is assumed to depend upon its own price, income, and a consumption adjustment factor to reflect noninstantaneous stock and behavior adjustments. Production depends upon own-price, crude oil prices, distillate prices (a close substitute in production but not in consumption), and a production adjustment factor. Making use of McCallum's supply of storage model permits equations 1, 2, and 3 to be solved for the real price of motor gasoline in period t, RPG(t). The solution is given by equation 4.

Real Price of Motor Gasoline Equation

$$\begin{aligned} RPG(t) = {} & A + B \cdot RPC(t) + C \cdot PMG(t-1) + D \cdot RY(t) + E \cdot RPG(t-1) + \\ & F \cdot SMG(t-1) + G \cdot CMG(t-1) + H \cdot RPD(t) \end{aligned} \qquad (4)$$

A preliminary estimate of equation 4 using ordinary least squares indicated the presence of autocorrelated disturbances. A quasi first-difference form of equation 4 was therefore estimated by nonlinear least-squares to obtain asymptotically efficient estimates. The estimated parameters and supporting statistics are presented in Table 4.2.

The specification of equation 4 is rich in policy analysis capability. Not only can the impact of higher crude oil prices and changes in national income be assessed relative to their impact on gasoline prices, but even such seemingly unrelated events as natural gas curtailments can be evaluated relative to the gasoline market. Natural gas curtailments can be expected to push up distillate (heating fuel oil) prices which affect refiners' decisions about relative distillate/gasoline yield levels. It is this policy richness which makes econometric forecasts attractive and difficult--forecast values of exogenous variables are needed.

In order to be consistent, a larger model that would also forecast distillate prices should be constructed. One variable in the distillate price equation would be gasoline prices. The solution of the complete model would then simultaneously determine distillate and gasoline prices. In such a model the exogenous effect of cold weather or natural gas curtailments on gasoline prices could be traced. By looking at only one part of such a model--the gasoline price equation--such effects are captured only through exogenous distillate price changes.

TABLE 4.2 SUPPORTING STATISTICS FOR THE ECONOMETRIC MODEL

Term	Coefficient	Estimate	t-statistic
Constant	A	0.157E-02	1.75
RPC(t)	B	0.144	2.06
PMG(t-1)	C	-0.142E-06	-1.29
RY(t)	D	0.923E-06	0.58
RPG(t-1)	E	1.070	11.06
SMG(t-1)	F	-0.396E-08	-1.96
CMG(t-1)	G	0.217E-06	0.10
RPD(t)	H	0.116	2.28

A Regression Model

A simple regression approach was also used to link gasoline prices to crude oil prices. This approach is presented here for two reasons. One is that other comparisons have used it as a naive version against which comparisons can be made.[1] It is used here to benchmark the forecasting power of the econometric technique. A second reason is that, especially in petroleum economics, it is tempting to link product prices directly to crude oil prices. A common linkage is to assume straight cost pass-throughs from crude oil to product prices. The straight cost pass-through assumption in conjunction with a response to own price in the previous period is used here. The specification estimated is given by equation 5 which is linear in logarithms.

$$\log(RPG(t)) = a + b\cdot\log(RPC(t)) + c\cdot HD(t) + d\cdot\log(RPG(t-1)) \qquad (5)$$

where

$\log(RPG(t))$ = the logarithm of the real price of motor gasoline in period t;

$\log(RPC(t))$ = the logarithm of the real price of crude oil in period t; and

$HD(t)$ = a dummy variable for the heating season (October, November, December, January, February, March, and April).

[1] Moore, G. H., "Forecasting Short-Term Economic Change," Journal of the American Statistical Association, Vol. 64 (March 1974), pp. 1-22.

Refiners adjust product yields and establish inventory priorities for the heating season (HD(t) = 1) and for the gasoline season (HD(t) = 0). The dummy variable thus captures the effects of changing the product mix and inventory policy on the gasoline price-crude oil relationship. Due to autocorrelated disturbances, a quasi-first differenced equation was estimated. The parameter estimates for equation 5 are given in Table 4.3.

TABLE 4.3 SUPPORTING STATISTICS FOR THE REGRESSION MODEL

Term	Coefficient	Estimate	t-statistic
Constant	a	12.797	0.38E-02
log(RPC(T))	b	0.408	5.25
HD(t)	c	-0.013	-2.53
log(RPG(t-1))	d	1.000	44.06

4.3.2 Comparison of Forecast Performance

The comparison of the forecasts produced by the two sets of models just described is aimed at illustrating the usefulness and the limitations of the various accuracy assessment measures discussed in Sections 4.1 and 4.2. It is important to point out that any conclusions regarding the superiority of a particular class of models based on the case studies presented in this section would be unwarranted. The purpose here is to illustrate an evaluation technique and not to critique each class of models. The models which are used in this section were deliberately kept simple. As a result of this simplicity, some desirable attributes had to be sacrificed. In particular, it was noted in Section 4.3.1 that a simultaneous equation formulation might be preferable to the single equation reduced-form model, since one could then explicitly model price interactions between motor gasoline, distillate, and other petroleum products.

The presentation in this section will proceed along the same lines as that in Sections 4.1 and 4.2. The first topic addressed will include a graphical analysis of 12 selected forecasts, six for levels and six for monthly charges. Four quantitative methods will then be discussed and checked for consistency in ranking the accuracy of these and other forecasts.

Three graphical methods--plots of predicted and realized values, plots of residuals, and the Theil prediction-realization diagram--were described in Section 4.1 as an aid to assessing the accuracy of a model's forecast. The presentation in this section will focus on the most common method, a simultaneous plot of both the predicted and realized values for the real price of motor gasoline series. In all of the figures which follow, the horizontal axis measures the time elapsed since the beginning of the forecast. Each forecast shown is either for a two-year period (24 months) or for a three-year period (36 months). The 24-month forecast begins January 1977 and runs through December 1978; the 36-month forecast begins in January 1976 and runs through December 1978.[1] Part A of each figure presents the 24-month forecast and Part B of each figure presents the 36-month forecast. The vertical axis of Figures 4.8 and 4.10 shows the real price index for motor gasoline. The vertical axis of plots appearing as Figures 4.9 and 4.11 are referred to as "changes" because they represent predicted and realized changes in the level of the series. In all figures the realized values (both levels and changes) are denoted by a set of shaded circles connected with a solid line and the

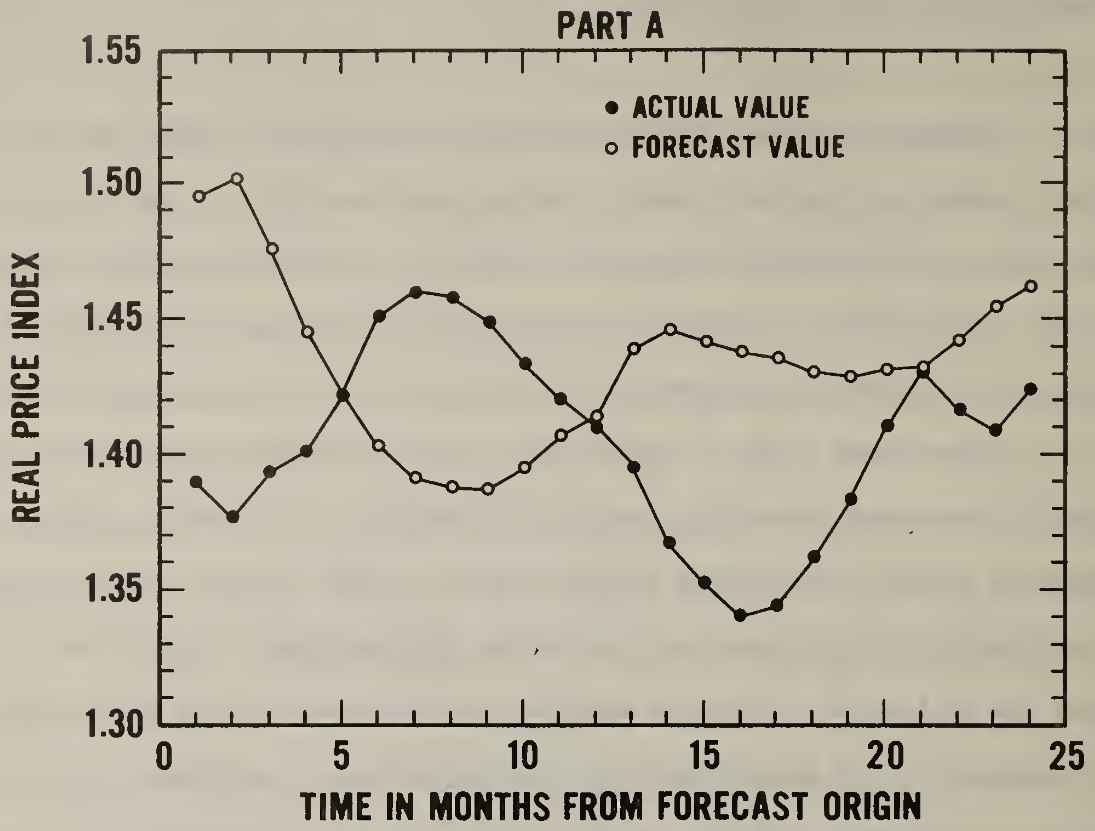
[1] The forecasts of monthly changes begin in February 1977 and 1976, respectively, and run through December 1978.

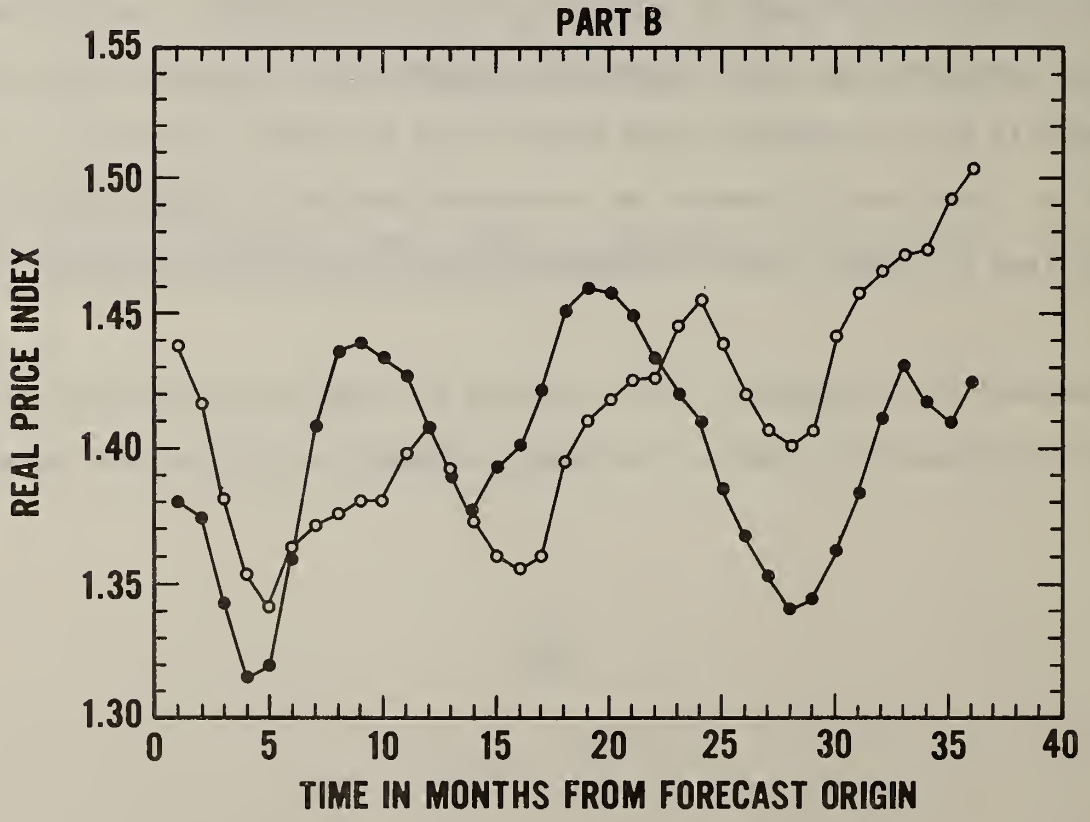
predicted values (both levels and changes) are denoted by a set of open circles connected by a solid line.

Figure 4.8 shows the values predicted using an econometric model and the realized values for the real price of motor gasoline index. Part A of the figure shows the 24-month forecast and reveals an interesting pattern--the predicted series peaks at month two bottoms out at approximately month eight, remains high for months 14 through 21 and then rises. An analysis of the realized series shows a dip in month two, a peak at approximately month seven and then a depressed period for months 14 through 18. Although a visual examination of the series would indicate that the fit is poor, the two series are correlated (in this case the two series are negatively correlated). Turning now to Part B, it can be seen that the forecast tracks better but is still somewhat out of phase with the realized values. The first decline in real prices is captured fairly well although the subsequent price rise is underestimated and the peak is missed by two to three months. As the forecast horizon increases, the phase relationship becomes more drawn out. The trough in month 14 is not predicted until month 16 and the peak in month 19 not until month 24. From month 24 onward the two series seem more in phase, however the predictions are biased (they persistently exceed the realized values).

An examination of Figure 4.9, Part A reveals the same lack of synchronization as seen in Figure 4.8, Part A. One again, although the fit is poor, we would

Figure 4.8 FORECAST VALUES OF MOTOR GAS PRICES USING AN ECONOMETRIC MODEL





expect the two series to be highly (negatively) correlated. A careful study of Part B, however, presents the model in a more flattering light. With the exception of the two segments centered around month 10 and month 22, the predicted monthly changes track the actual monthly changes very well.

The predictions based on the simple regression model are presented in Figures 4.10 and 4.11. Since the price of crude oil represents such an important component of the price of motor gasoline, one would expect that the regression model should forecast fairly well, at least over the short run. An examination of Part A of Figure 4.10 reveals this to be the case for the first ten months of the forecast period. Beyond month 10, however, the prediction produced by the model shows a fairly strong upward trend. In Part B of the figure it can readily be seen that the model does not perform well until the end of the third quarter of the forecast period. The model then tracks the process fairly well through the seventh quarter of the forecast period.

An analysis of the predicted and realized monthly changes in the real price of motor gasoline index reveals a much more choppy appearance than was seen for the econometric model over the same period. The fit for the first three quarters of the 24-month forecast period (Part A) appears reasonable as was the case for the levels forecast. The predicted changes over the 36-month forecast period do not appear to give an adequate representation of the process under study even in the short run. Part of the apparent inadequacy of

Figure 4.9 FORECAST MONTHLY CHANGES OF MOTOR GAS PRICES USING AN ECONOMETRIC MODEL

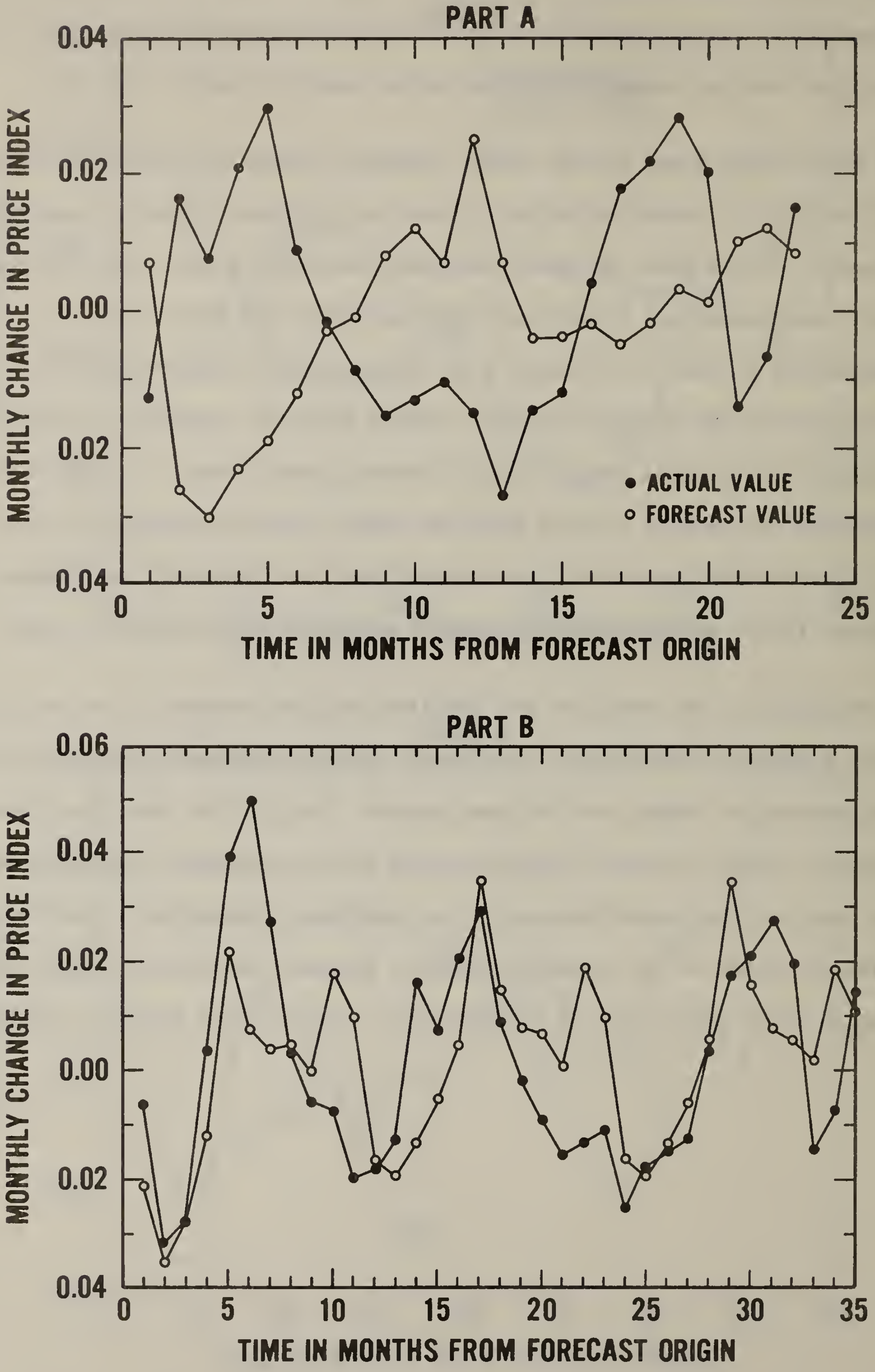


Figure 4.10 FORECAST VALUES OF MOTOR GAS PRICES USING A REGRESSION MODEL

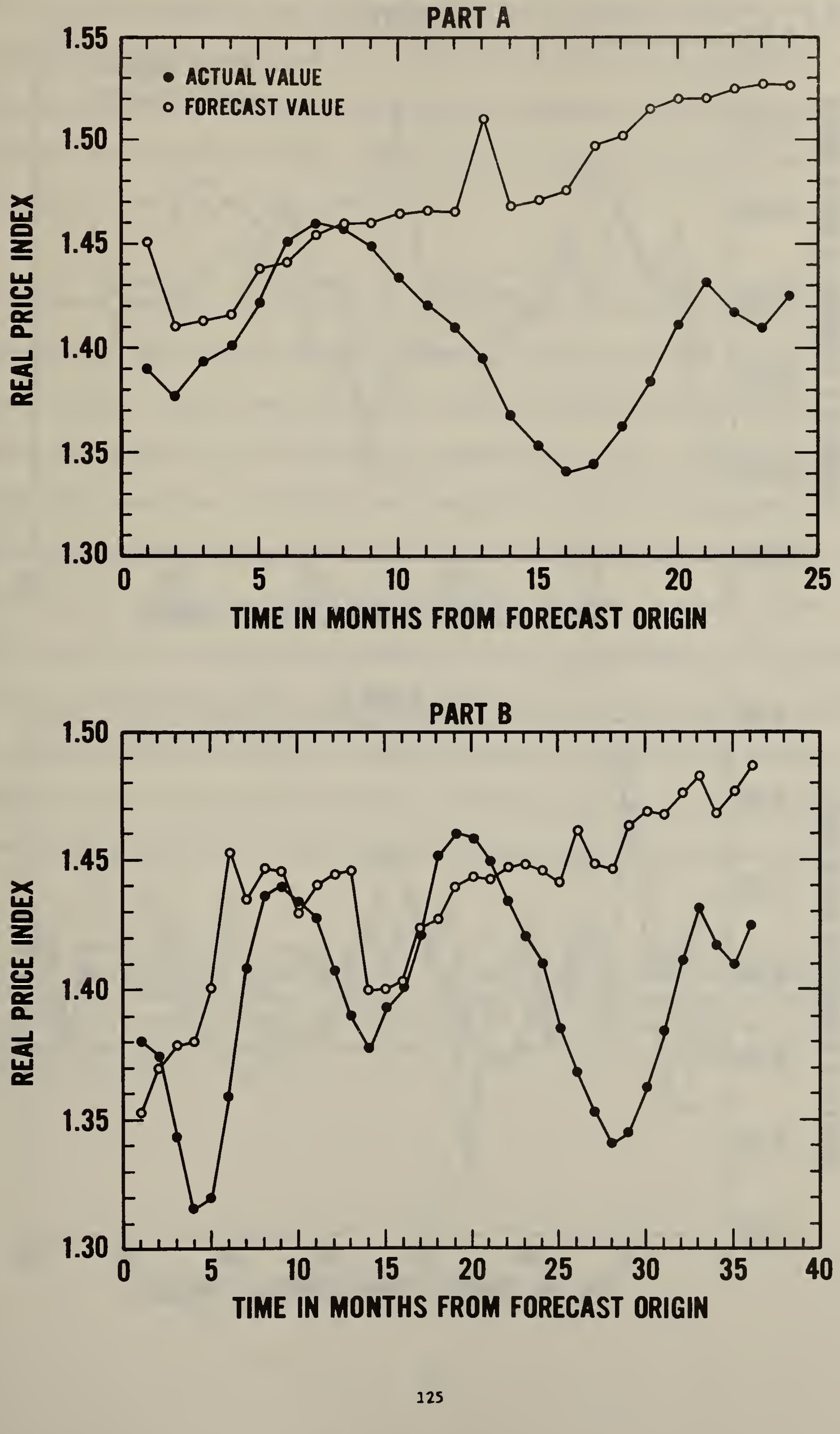
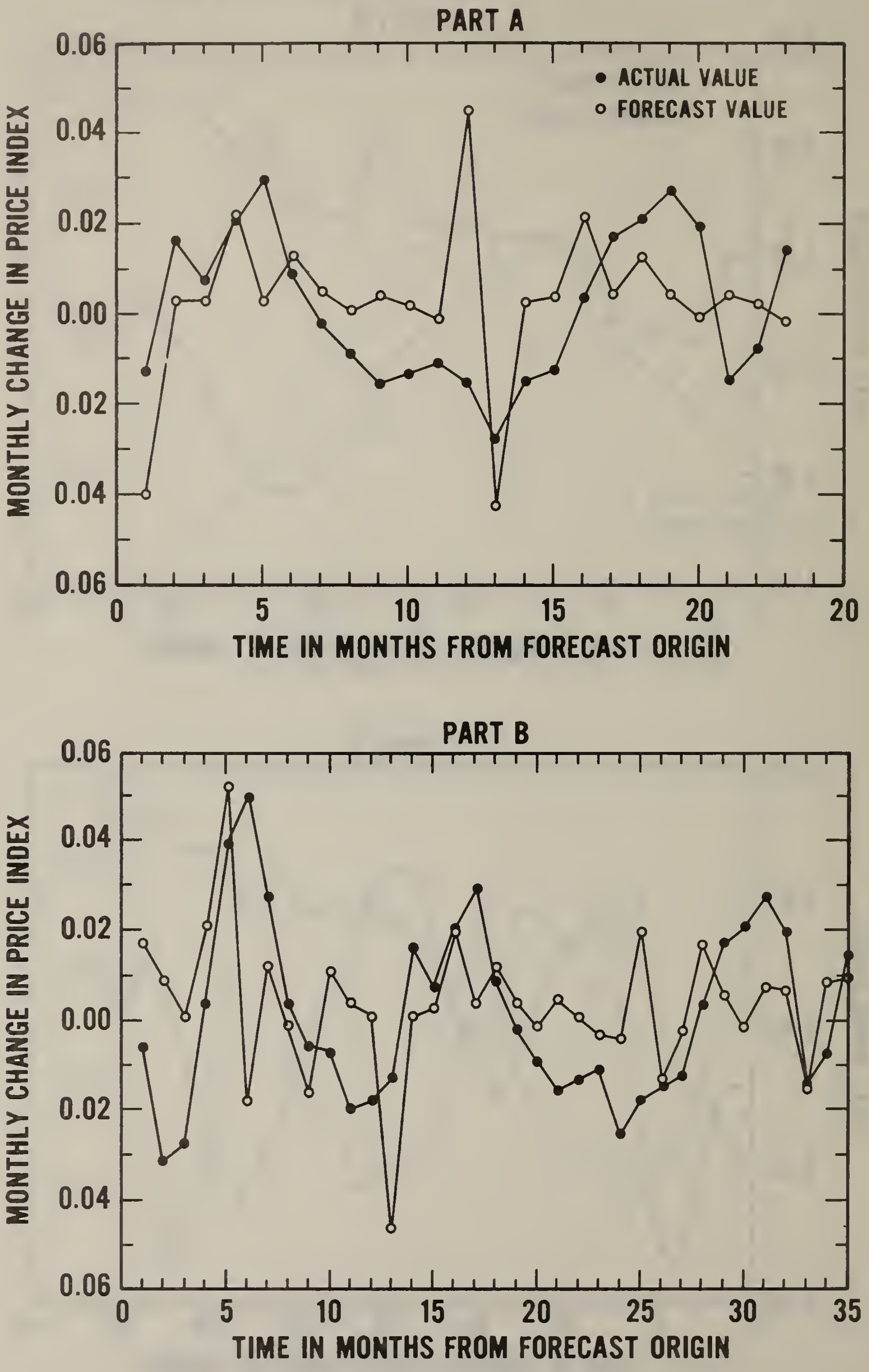


Figure 4.11 FORECAST MONTHLY CHANGES OF MOTOR GAS PRICES USING A REGRESSION MODEL



this model is undoubtedly due to the way in which the real price of crude oil was forecast. However, since these are modified *ex post* forecasts we must operate *as if* we did not know what the actual real price of crude oil was going to be. Thus the forecasts presented in Figures 4.10 and 4.11 are probably fairly similar to ones which would have been produced by an analyst in December 1976 and December 1977.

As discussed earlier, the graphical approach helps to reveal the strengths and weaknesses of a particular model; however it is not an end in itself. In order to gain a better indication of a model's predictive performance, we should also subject it to a battery of quantitative tests. Although each test described in Sections 4.1 and 4.2 was applied to each model for each forecast period, an attempt has been made to limit the discussion which follows to the salient parts of the quantitative analysis.[1] In order to simplify the identification of a particular forecast, or set of forecasts, it was necessary to adopt a numbering scheme. In the tables which follow, the first digit of the forecast number denotes the number of years in the original forecast period. The second and third digits of the forecast number denote the number of months in the subset of the original forecast period for which data were

[1] The following statistics were computed for each model and each forecast: (1) R^2; (2) root mean squared error; (3) Theil U inequality; (4) decompositions into bias, variance, and covariance and bias, regression, and residual variance; (5) standard deviation of the realized series during the forecast period; (6) marginal entropy; (7) conditional entropy,; and (8) transinformation.

analyzed. For example, the forecasts denoted as 312 contain the first 12 months of the 3-year forecast beginning in January 1977 and ending in December 1979. Each model specification is also coded; for example, all models preceded by an R are regression models. The coding for each model and each forecast are summarized in Table 4.4 for ready reference.

The rankings of the forecast performance for each model and each time period are presented in Table 4.5. Part A of the table presents the rankings for the models designed to forecast the levels of the real price of motor gasoline index. Part B of the table presents the rankings for the models designed to forecast monthly changes in the motor gasoline price index. Four basic statistical measures are presented and for three of these the models are ranked in order of performance measured against an objective criterion. The first column contains the ranking based on maximizing R^2. The second column contains the ranking based on minimizing the Theil U inequality. The title of the third column contains both the root mean squared error (RMSE) and transinformation (TI) because (for unbiased forecasts) both sets of statistics produce the same ranking (in this case, TI is a monotone transinformation of the RMSE). The fourth column indicates whether or not the model conveys useful information, i.e., whether transinformation is positive, +, or negative, −.

Even a cursory review of the data presented in Table 4.5 would reveal that the three major types of statistical measures of accuracy do not produce the same rankings. Statisticians have often pointed this out with regard to the R^2

TABLE 4.4 ABBREVIATIONS USED IN SUBSEQUENT TABLES

ABBREVIATION	DEFINITION	
RMSE	Root Mean Squared Error	
TI	Transinformation	
E	Econometric Model	
R	Regression Model	
212	First 12 months of the 2-year forecast:	1/78 - 12/78
312	First 12 months of the 3-year forecast:	1/77 - 12/77
412	First 12 months of the 4-year forecast:	1/76 - 12/76
224	Entire 2-year forecast:	1/78 - 12/79
324	First 24 months of the 3-year forecast:	1/77 - 12/78
424	First 24 months of the 4-year forecast:	1/76 - 12/77
336	Entire 3-year forecast:	1/77 - 12/79
436	First 36 months of the 4-year forecast:	1/76 - 12/78
448	Entire 4-year forecast:	1/76 - 12/79

TABLE 4.5

RANKINGS OF FORECAST PERFORMANCE FOR SELECTED STATISTICAL MEASURES

PART A

LEVELS	STATISTIC							
	R^2		THEIL U		RMSE&TI		POS.INFO	
FORECAST	E	R	E	R	E	R	E	R
212	1	2	2	1	2	1	–	+
312	1	2	2	1	2	1	–	+
412	2	1	1	2	2	1	+	+
224	2	1	1	2	2	1	+	+
324	1	2	1	2	2	1	–	–
424	2	1	2	1	2	1	+	+
336	2	1	2	1	2	1	+	+
436	2	1	1	2	2	1	–	–
448	1	2	1	2	1	2	+	+

PART B

CHANGES	STATISTIC							
	R^2		THEIL U		RMSE&TI		POS.INFO	
FORECAST	E	R	E	R	E	R	E	R
212	1	2	2	1	2	1	–	–
312	1	2	2	1	2	1	–	–
412	2	1	1	2	1	2	+	–
224	2	1	1	2	1	2	–	–
324	1	2	2	1	2	1	–	–
424	2	1	1	2	1	2	+	–
336	1	2	2	1	1	2	–	–
436	2	1	1	2	1	2	+	–
448	2	1	1	2	1	2	+	–

statistic and the root mean squared error statistic. However, it is also obvious that some major disagreement also exists between the root mean squared error and the Theil U inequality. For example, consider forecast 224 in Part A of Table 4.5 where, based on maximizing the R^2 statistic, the regression model is revealed as best (rank = 1), and the econometric model worst. If the criterion were the minimization of the Theil U inequality (recall that the U inequality is greater than or equal to zero with a value of zero indicating a perfect fit), then exactly the opposite ranking would occur. A somewhat similar reversal with respect to the root mean squared error occurs with forecast 324 when the criterion is maximizing the R^2 statistic. In this case the Theil U inequality also produces incorrect rankings. An examination of Part B reveals a rather shocking result, the R^2 and Theil U inequality criteria produce diametrically opposed rankings. This unfortunate state of affairs is further complicated by noting that the minimum root mean squared error or maximum transinformation criterion produce rankings which conflict with both of the other statistics. Given the desirable attributes of the root mean squared error, it is gratifying that the transinformation statistic produces the same ranking. One should not be too hasty and conclude that transinformation is just a variant of the root mean squared error. In particular, transinformation tells the analyst everything that the root mean squared error tells him. Transinformation also provides the analyst with a measure of how large the root mean squared error can get before use of the model may become misleading (i.e., the model's forecasts are spurious). An overview of the fourth column would reveal that six of the 18 forecasts for the level of the real price of motor gasoline index were, on the average, spurious. The previous statement does not imply that each and every forecast value within the given forecast period (e.g., 324) was spurious. As was shown in the discussion of the graphical results, some segments of the forecast periods may be surprisingly accurate. The previous statement reflects a basic difference between an assessment of the overall forecast versus an assessment of the various components of the forecasts. The approach taken in this section is similar to the empirical studies by Nelson[1] and Fromm and Klein[2] in that emphasis is placed on the performance of the model during the entire forecast period. An alternative approach would be to analyze the distributional properties of the residuals as a function of the model's lead time. Although the theoretical criteria upon which such an analysis would be based were outlined in section 4.2, the project staff determined that time and funding constraints could not justify the additional computational effort required for an empirical test.

[1] Nelson, C.R., "The Predictive Performance of the FRB-MIT-PENN Model of the U.S. Economy," American Economic Review, Vol. 62 (December 1972), pp.902-917.

[2] Fromm, G., and L. Klein, "The NBER/NSF Model Comparison Seminar: An Analysis of Results," Annals of Economic and Social Measurement, Vol 5 (1976), pp. 1-27.

By examining the fourth column of Part B one finds that 14 out of the 18 forecasts for monthly changes are, on the average, spurious. Thus, one can see why Granger and Newbold have claimed that evaluating a model based on its predictive performance for the level of an economic variable is perhaps giving it more credit than it is due.

In light of the difficulties concerning the interpretation of the various accuracy assessment statistics, it is useful to ask "how often should one expect the alternatives to the maximum transinformation (minimum root mean squared error) criterion to produce the same rankings?" Although we are undoubtedly concerned whether or not the criterion would produce the same ranking, especially if we are considering a combination of forecasts, we are probably more concerned whether or not the objective criterion can correctly identify the first ranked model. In order to calculate this probability it is necessary to determine how many times each criterion ranked the model first given that the maximum transinformation (minimum root mean squared error) criterion ranked it first. These conditional probabilities are given for each statistic for both levels and changes.[1] An examination of the probabilities in Table 4.6 reveals that the R^2 statistic outperforms the Theil U inequality

[1] All probabilities shown in Table 4.6 are based on data presented in Table 4.5. The probabilities are defined as the ratio of the number of times the R^2or U criterion (see Table 4.5) ranked the model first given that the maximum transinformation criterion ranked it first.

on levels but is outperformed on changes. One should interpret this result with some caution, however, since this is the U° rather than the U statistic. Theil has corrected many of the deficiencies of the U° statistic with the U statistic. Most software packages, however, continue to print out the U° rather than the U statistic and for that reason it is presented here. Since the corrected Theil U inequality is just a monotone transformation of the root mean squared error, minimizing it should produce the correct rankings. Recall that the Theil U inequality was designed for cases where changes and not levels were the object of the forecast, so that a value of U less than or equal to one would not necessarily imply that the model's forecasts were not spurious. It is also important to note that the corrected Theil U inequality makes no assumptions about the distribution governing either the process or the residuals so that meaningful measures of the information content of the forecast cannot necessarily be derived from it.

TABLE 4.6

COMPARISONS OF RANKINGS

Statistic	Classification	Probability that Rankings Agree Levels	 Changes
R^2	P (model ranked 1 given rank = 1)	0.667	0.556
U	P (model ranked 1 given rank = 1)	0.556	0.889

4.4 Concluding Remarks

The previous discussion has pointed out some of the weaknesses of the classical approach to accuracy assessment. These weaknesses were addressed both from a theoretical point of view in Section 4.1 and based on empirical considerations in Section 4.3. The major thrust of this chapter was to expose these weaknesses and demonstrate how they could be reduced or eliminated through resort to a new technique based on concepts from information theory. Although the advantages of the information theoretic approach have been established and some broad guidelines for implementing such an approach into a comprehensive forecast evaluation program have been given, much work remains to be done.

It is important to point out that the focus of the information theoretic approach presented in this chapter was based on an assumption that the models under consideration produced unbiased forecasts. Although this assumption is reasonable, it is probably overly optimistic, especially in cases where the values of one or more variables used in the model are rapidly changing. The mechanics of generalizing the information theoretic approach to measure the performance of biased forecasts is under study and is planned for inclusion in subsequent papers on the subject where some computational simplifications will also be outlined.

The analytics of performing the dynamic decomposition and the policy decomposition outlined in Section 4.2 should also be perfected so that these valuable tools can be readily accessible to decision makers. Two other areas where additional work is needed concern: (1) the sensitivity of the transinformation measure to the probabilistic structure of the process and its forecast errors; and (2) techniques for optimally combining the forecasts from competing models. A final area of concern involves the extension of the information theoretic measures from the single equation framework to simultaneous equation models of a dynamic system. At this stage it is unclear whether such an extension will permit the retention of a univariate measure or will entail the development of a "transinformation" vector or matrix. Since arguments can be made for both approaches, some analysis might be done to determine which approach offers the greatest advantages to both model builders, model users, and decision makers. P.A.V.B. Swamy has outlined, in a personal communication to the authors, a theoretical approach which stresses the importance of exact finite sample properties and Bayesian methods of estimation. The Swamy approach, which uses a quadratic loss function as its basis, first limits the class of candidate models to those which possess finite second-order moments. The axiomatic approach of DeGroot[1] is then used to insure that the maximization of expected utility is a valid criterion for

[1] DeGroot, Morris H., Optimal Statistical Decisions (New York: McGraw-Hill Book Company, 1970).

choosing among competing models. In a recent paper by Bernardo[1] and subsequent discussion by DeGroot[2], it has been shown that the use of the entropy measure is equivalent to considering a decision problem in which one must choose a density function, f, from the class of all densities on the parameter space, Θ, subject to the loss function

$$L(\theta, f)= -\log (f(\theta)).$$

Therefore one can interpret the techniques outlined in this chapter as a special case of Swamy's approach to forecast evaluation and model selection.

[1] Bernardo, Jose, "Reference Posterior Distribtuions for Bayesian Inference," Journal of the Royal Statistical Society, No. 2 (1979), pp. 113-128.

[2] DeGroot, Morris H., "Discussion of Professor Bernardo's Paper," Journal of the Royal Statistical Society, No. 2 (1979), pp. 135-136.

5. FINDINGS AND CONCLUSIONS

by
Lambert S. Joel
Center for Applied Mathematics
National Bureau of Standards
Washington, D. C. 20234

Our conclusions at the end of this second year of scrutiny of major energy models fall under two general headings, judgments on the model STIFS and judgments concerning assessment and model use methodology.

We will discuss briefly our findings concerning STIFS.

(1) STIFS is not in deliverable condition. Simply put, STIFS is not yet a model or a system of models, but a framework for a model system. Consequently, although STIFS can be pressed into service as a provisional policy analysis tool in the hands of its architects, while undergoing development, it cannot be operated for policy analysis by any independent user (no matter how broadly "independent" is interpreted) with or without existing documentation.

(2) While there is possibly some scope for selection of efficient off-the-shelf computational software for regression, time series analysis, and the like in the satellite component models, the integrating model, which we suspected to follow the (Leontieff) input-output structure and therefore to be amenable to efficiency improvement, is already in the simplest computational form possible.

(3) The superficially counterintuitive notion, produced by the developers in the process of coefficient estimation, that weather variations do not affect nationally aggregated annual energy consumption has been verified in our sensitivity analyses. The ostensible reason is that the size of the contiguous United States ($3x10^6mi^2$) permits fairly sizable regions to have severe seasonal weather while the nation is nominally subject to mild weather and vice versa.

In summary form, our conclusions on the wider issues of assessment are:

(1) Reading Code

Sooner or later, the assessment of a large-scale model will entail close examination of computer programs. There are two general reasons: outputs fail to conform to expectation in some way, or the model is to be operated in a manner not envisioned in the design of the system. In the first case, counterintuitive model outputs (or outputs differing from prespecified sample values) could result from a variety of causes, for instance erroneous conceptualization of the model by the developers (or the assessors), erroneous mathematization, defective computer programs, or interaction with an operating system or supporting data for the model. In any event, virtually no large-scale model is cast as a coherent "closed-form" mathematical function amenable to verification by means of, say, computation with a desk calculator. Therefore, difficulties must be resolved by reading computer code. The ease with which this process of reconciliation between the computer programs and the conceptual content of model can be accomplished is related to the quality of documentation and the extent of annotation (in the form of comment cards) in the programs. It also depends on the availability of the model developers for consultation. This suggests that during debugging of the computer programs by the analysts/programmers/developers of the model, which is a very close parallel to reconciliation by assessors, special effort should be made to refine the documentation so that difficulties encountered in debugging may be avoided in subsequent similar activities by assessors or users.

The second generic situation which can be expected to require reading code--non-standard operation--implies some modification of the model by assessors (or users). This could range from a change in the value of a "hard-wired" constant, that is, one which is not treated in the model as an input subject to change from run to run, to alterations in the structural form of the model. In our case, considerable computational time and analysis of outputs was saved in the sensitivity experiments, by changes in the procedure for initializing model runs so as to allow ganged sequences of model runs with "automatically" iterated scenario changes. Although we had cooperative assistance from EIA staff, the work would have been greatly drawn out if we hadn't studied the computer code.

(2) The Impracticability of Early Assessment

Technical assessment is an evaluation of validity and practicability of a model for its intended purpose. No model is ever constructed without technical assessment, because assessment for orientation is an inseparable aspect of analytical development. "Third-party" or "independent" assessment, however, is an administrative construct. Let us define it as evaluation of a model under specific assignment by someone other than the sponsor[1] or the developer(s) of the model. In the early stages of model formulation, third-party assessment is likely to be of limited value. Documentation will be nonexistent or at best extremely scanty and provisional, so that excessive effort will be required to assemble information that will enable proper assessment. (As model development approaches nominal completion, it is normally easier for an assessment effort to "get up to speed.") Moreover, model development is unlikely to be a process of unswerving progressive synthesis; when elements are discarded, time spent on these by assessors is time lost, by and large. If legal/ethical objections and contractual objections could be overcome, so that an assessment team would participate cooperatively in model development while maintaining an independent stance as technical evaluators, early assessment might yet be "cost effective."

[1] Conventionally this is done on a contractual basis. We will avoid ambiguity by designating a subordinate group within a sponsor institution and tasked with the responsibility for evaluating a model developed outside the group, as "not the sponsor."

(3) Monte Carlo Methods

Our exercise of portions of the model system and our experimental analysis of the relationship of climatological to energy-demand data, have reinforced[1] our confidence in the utility of Monte Carlo techniques for developing representative application scenarios as opposed to "best case" or "optimistic" vs. "worst case" or "pessimistic" scenario settings, which give, at best, conceptually possible extreme values that is, upper and lower bounds, with excessively small probability of actual occurrence.

(4) Information Theory and Forecast Accuracy

The determination of forecast quality for large-scale models remains a major unsolved problem related to, and rivaling the difficulty of, formulating the models and making the forecasts. The accuracy of a single short-term point forecast can, of course, be measured directly *ex post*, as the difference between a measured and a forecast value of some variable of interest, but such isolated differences do not tell us much about forecast quality. At the very least we would want some sizeable number, perhaps a time series, of observations and predictions before we would be willing to make statements of confidence in a forecast model. The richer the context of a forecast model,

[1] The Monte Carlo method was employed to good effect in the analysis of the MOGSM resource base in the first year of the study.

that is, the greater the number of significant components of the system under study, particularly those whose action or evolution is governed by chance insofar as we can tell, the less feasible is a faithfully representative model, and the less reliable are simple measurements of divergence as indicators of forecast quality.

One criterion for suitability of an algorithmic (that is, mechanically applicable) measure of forecast quality is that it should produce a ranking of a collection of forecasts consistent with a ranking based on, say, expert intuition. In chapter 4 methods of accuracy based on information theory have been demonstrated to be consistent with conventional measures based on various well-accepted statistical/econometric tests, and to furnish at slight additional computation costs, important additional information concerning the effect of imprecision in model parameters and the range of uncertainty of data inputs on the expected accuracy of forecasts. Technically, these results (consistency and superior power of discrimination) have been established only for single equation (scalar) models and unbiased forecasts; additional research is indicated to extend them to biased forecasts and to simultaneous equation (matrix) models.

REFERENCES

Amorocho, J., and B. Espildora, "Entropy in the Assessment of Uncertainty in Hydrologic Systems Behavior and in Mathematical Model Performance," International Symposium on Uncertainties in Hydrologic and Water Resource Systems, University of Arizona, 1972, pp. 977-1008.

Bates, J. M., and C. W. J. Granger, "The Combination of Forecasts," Operational Research Quarterly, Vol. 20, No. 4 (1969), pp. 451-468.

Bernardo, Jose, "Reference Posterior Distributions for Bayesian Inference," Journal of the Royal Statistical Society, No. 2 (1979), pp. 113-128.

Bopp, Anthony E., and John A. Neri, "The Price of Gasoline: Forecasting Comparisons," The Quarterly Review of Economics and Business, Vol. 18, No. 4 (Winter 1978), pp. 23-33.

Breiman, L., Statistics: With a View Toward Applications, (Boston: Houghton Mifflin, 1973).

Collins, Dwight E., Mary L. Burcella and Michael L. Shaw, Short Term Integrated Forecasting System (STIFS): Methodology and Model Description, Logistics Management Institute, Washington, D.C., November 1979.

DeGroot, Morris H., Optimal Statistical Decisions, (New York: McGraw-Hill Book Company, 1970).

DeGroot, Morris H., "Discussion of Professor Bernardo's Paper," Journal of Royal Statistical Society, No. 2 (1979), pp. 135-136.

Dhrymes, Phoebus J., E. Phillip Howrey, Saul H. Hymans, Jan Kmenta, Edward E. Leamer, Richard E. Quandt, James B. Ramsey, Harold T. Shapiro, and Victor Zarnowitz, "Criteria for Evaluation of Econometric Models," Annals of Economic and Social Measurements, Vol. 1, No. 3 (1972), pp. 291-324.

Fromm, G. and Klein, L., "The NBER/NSF Model Comparison Seminar: An Analysis of Results," Annals of Economic and Social Measurement, Vol. 5 (1979), pp 1-27.

Gass, Saul I., Evaluation of Complex Models, University of Maryland, MS/S 76-002, May 1976.

Goldman, S., Information Theory, (New York: Prentice-Hall, Inc., 1953).

Granger, C. W. J., and P. Newbold, "Some Comments on the Evaluation of Economic Forecasts," Applied Economics, VOl. 5 (1975), pp. 35-47.

Granger, C. W. J., and P. Newbold, Forecasting Economic Time Series, (New York: Academic Press, 1977).

Harris, Carl M., and D. S. Hirshfeld, "Sensitivity and Statistical Analysis of Models," in Validation and Assessment of Energy Models, Saul I. Gass (ed.), National Bureau of Standards, SP616, 1981, pp. 171-181.

Harrison, P. J., and C. F. Stevens, "Bayesian Forecasting," Journal of the Royal Statistical Society, No. 3 (1976), pp. 205-228.

Kaczmarek, Z., Statistical Methods in Hydrology and Meteorology, Publications in Transportation and Communication, Warsaw, 1970 (in Polish).

Leamer, Edward E., Specification Searches: Ad Hoc Inference with Nonexperimental Data, (New York: John Wiley and Sons, 1978).

Maddala, G. S., Econometrics, (New York: McGraw-Hill Book Company, 1977).

McCallum, B. T., "Competitive Price Adjustments: An Empirical Study," American Economic Review, Vol. 64 (March 1974), pp. 56-65.

McKay, M., W. J. Conover and R. J. Beckman, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code," Technometrics, Vol. 21 (1979), pp. 239-245.

Mincer, Jacob, and Victor Zarnowitz, "The Evaluation of Economic Forecasts," in Economic Forecasts and Expectations, Jacob Mincer (ed.), National Bureau of Economic Research, New York, 1969.

Moore, C. H., "Forecasting Short-Term Economic Change," Journal of the American Statistical Association, Vol. 64 (March 1974), pp. 1-22.

Nelson, C.R., "The Predictive Performance of the FRB-MIT-PENN Model of the U.S. Economy," American Economic Review, Vol. 62 (December 1972), pp. 902-917.

Pindyck, Robert S. and Daniel L. Rubinfeld, Econometric Models and Economic Forecasts, (New York: McGraw-Hill Book Company, 1976).

Reid, D. J., A Comparative Study of Time Series Prediction Techniques on Economic Data, unpublished Ph.D. Dissertation, Nottingham University, 1969.

Short-Term Energy Outlook: August 1980, Vol. II, Methodology and Analysis, Energy Information Administration, DOE/EIA-0202/4-2, 1980.

Swamy, P. A. V. B., and P. A. Tinsley, "Linear Prediction and Estimation Methods for Regression Models with Stationary Stochastic Coefficients," Journal of Econometrics, Vol. 12 (1980), pp. 103-142.

Theil, Henri, Economic Forecasts and Policy, (Amsterdam: North-Holland Publishing Company, 1961).

Theil, Henri, Applied Economic Forecasting, (Amsterdam: North-Holland Publishing Company, 1966).

Theil, Henri, Economics and Information Theory, (Amsterdam: North-Holland Publishing Company, 1967).

Zarnowitz, Victor, An Appraisal of Short-Term Economic Forecasts, National Bureau of Econoic Research, New York, 1967.

Zellner, A., "An Efficient Method for Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," The Journal of the American Statistical Association, Vol. 57 (1962), pp. 348-368.

NBS-114A (REV. 2-80)

U.S. DEPT. OF COMM. **BIBLIOGRAPHIC DATA SHEET** *(See instructions)*	1. PUBLICATION OR REPORT NO. NBSIR 83-2672	2. Performing Organ. Report No.	3. Publication Date April 1983

4. TITLE AND SUBTITLE

Selected Assessment Strategies Applied to Short-Term Energy Models

5. AUTHOR(S)

Patsy B. Saunders, Editor

6. PERFORMING ORGANIZATION *(If joint or other than NBS, see instructions)*

NATIONAL BUREAU OF STANDARDS
DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20234

7. Contract/Grant No.

8. Type of Report & Period Covered

9. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS *(Street, City, State, ZIP)*

Energy Information Administration
U.S. Department of Energy
Washington, D.C. 20461

10. SUPPLEMENTARY NOTES

□ Document describes a computer program; SF-185, FIPS Software Summary, is attached.

11. ABSTRACT *(A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here)*

This report is one in a series focusing on the evaluation of complex mathematical models. The basic approach pursued in this document is patterned after an earlier analysis of the Department of Energy's Midterm Oiland Gas Supply Model (MOGSM). Several extensions of the earlier methodology are presented which assist the analyst in defining the degree to which certain evaluation activities are model dependent. The Department of Energy's Short Term Integrated Forecasting System (STIFS) was used as a vehicle for exercising the revised methodology. The technical content of the report is divided into three parts, reflecting three basic issues of model form, sensitivity and forecast performance. The first issue addressed related to the structure of STIFS. It includes not only the mathematical assumptions implicit in the model but also data and software considerations. The approach to the second issue focuses on the measurement of climatological uncertainties and uses as its basis a Monte-Carlo experiment. The final issue deals with several techniques for evaluating the predictive performance of a model. Both classical statistical methods and an information theoretic approach are used to illustrate how such an analysis would be carried out in practice.

12. KEY WORDS *(Six to twelve entries; alphabetical order; capitalize only proper names; and separate key words by semicolons)*

Assessment; documentation; energy; information theory; mathematical models; sensitivity analysis

13. AVAILABILITY

☒ Unlimited
□ For Official Distribution. Do Not Release to NTIS
□ Order From Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.
☒ Order From National Technical Information Service (NTIS), Springfield, VA. 22161

14. NO. OF PRINTED PAGES

153

15. Price

$16.00