APPROPRIATE SUMMARY STATISTICS FOR EVALUATING THE HISTORICAL FIT OF SYSTEM DYNAMICS MODELS

John D. Sterman
Assistant Professor
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

February 1983

D-3393

ABSTRACT

System Dynamics modelers are often faulted for their reluctance to employ formal measures of goodness-of-fit when assessing the historical behavior of models. As a result, the validity of system dynamics models is often questioned even when the model's correspondence to historical behavior is quite good. This paper argues that the failure to present formal analysis of historical behavior creates an impression of sloppiness and unprofessionalism. After reviewing the concept of validity in simulation modeling, the paper proposes a simple set of summary statistics appropriate for system dynamics models (the root-mean-square error and Theil inequality statistics). The statistics allow the error due to individual behavior modes to be analyzed, do not require the use of formal parameter estimation procedures, and can be conveniently computed. A large model of the U.S. economy is used to illustrate the use of the statistics.

Introduction

System Dynamics modelers are often faulted for their reluctance to employ formal measures of goodness-of-fit when assessing the historical behavior of models. As a result, the validity of system dynamics models is often questioned even when their correspondence to historical behavior is quite good. This paper argues that the failure to present formal analysis of historical behavior creates an impression of sloppiness and unprofessionalism. After reviewing the theory of validity in system dynamics, the paper proposes a simple set of summary statistics appropriate for system dynamics models. The statistics allow the error due to individual behavior modes to be analyzed, do not require the use of formal parameter estimation procedures, and can be conveniently computed.

The "Validation" of System Dynamics Models

Debate over the concept of "validity" in system dynamics has as long a history as the field itself.[2] Discussions of validation in system dynamics have stressed three basic points:

1. There can be no absolute test of validity,
2. There can be no objective tests of validity,
3. There can be no single test of validity.

From the first, system dynamicists have rejected the notion that the validity of models can be established absolutely. Rather, as Forrester emphatically states,

> The validity (or significance) of a model should be judged by its suitability for a particular purpose. A model is sound and defendable if it accomplishes what is expected of it. ... validity, as an abstract concept divorced from purpose, has no useful meaning (3).

A model intended for short-term prediction must be evaluated by different procedures than models designed for long-term policy analysis, exploration of possible future behavior modes, or theory testing, a view widely shared by other modelers and social scientists.[4]

Rejected also is the notion that, even given a clear purpose, there can be objective criteria for validity. Forrester correctly states that

> Any "objective" model-validation procedure rests eventually at some lower level on a judgment or faith that either the procedure or its goals are acceptable without objective proof (5).

For example, one of the most common tests of the significance of parameter estimates in regressions is the t-statistic.[6] The t-statistic is used to test, within some level of significance (typically 5%) the hypothesis that the estimated parameter is equal to zero (the analyst usually hopes to be able to reject the hypothesis, establishing a significant nonzero value for the estimated parameter). Econometrics texts teach that "one may interpret a significant t-statistic ... as evidence tending to

validate the model ... [and] an insignificant statistic would lead toward invalidation of the model."[7] However, the t-test in the context of the standard single-equation least-squares repression model rests on several assumptions ("maintained hypotheses") that are not verifiable or go unquestioned in practice, including

> ... perfect specification of the model being estimated (including zero-mean, normally-distributed noise inputs in each equation) and perfect measurement of all variables (8).

Because the maintained hypotheses are not verified, the t-test is not a test of validity: an insignificant value may indicate one of the maintained hypotheses is violated rather than an insignificant relationship, a result that has been demonstrated through synthetic data experiments.[9] To treat the test as an indicator of validity, then, is necessarily to make a subjective judgment that the maintained hypotheses are in fact true. It may be objected that the t-test and standard linear model are simplistic and unrepresentative of actual practice. Econometricians have developed many powerful procedures that allow maintained hypotheses to be tested, including tests of model specification.[10] However, useful though they may be, such tests necessarily invoke other maintained hypotheses, shifting the locus of the inevitable a priori but never eliminating it.

Validation is an inherently social process. It depends on the cultural context and background of the model builders and model users. It depends on whether one is an "observer" (e.g., an academic researcher) or an "operator," (e.g., a decisionmaker who must act without waiting for more data or further analysis).[11] Churchman goes so far as to argue that the process is entirely social:

> ... a point of view, or a model, is realistic to the extent that it can be adequately interpreted, understood, and accepted by other points of view. (12)

Recognizing the ultimately subjective nature of all "objective" tests means one can never validate a model in the sense of establishing its truth. Rather, the notion of objective validity has been replaced by the confidence the model builders and users place in the model and its conclusions:

> No model has ever been or ever will be thoroughly validated. ... "Useful," "illuminating," "convincing," or "inspiring confidence" are more apt descriptors applying to models than "valid" (13).

Emphasizing the process of building confidence in a model means there can be no single test or measure of validity. No responsible model builder or user would ever be satisfied with a single test. Confidence must be developed through a process of testing and evaluation along many dimensions, a point emphasized by many.[14]

A wide variety of tests have been developed to aid the diagnosis of errors and to assist the confidence-building process in system dynamics. The tests, summarized in Table 1, include tests of the structure, parameters, behavior, and policy recommendations of the model.

The Role of Historical Data in the Confidence-building Process

A corollary of the three principles outlined above is that the single most common measure of validity in the social sciences, the historical fit of a model, is a weak test that contributes little if anything to confidence. Analysis of the historical fit of a model is a part of the Behavior Reproduction test (Table 1). But the Behavior Reproduction test is more than comparing the correspondence of simulated and actual data on a point-by-point basis. The test usually focuses on the character of the simulated data: does it exhibit the same modes, phase relationships, relative amplitudes, and variability as the real data?[15]

Point-by-point or event-oriented comparisons to historical data have been minimized in system dynamics for several reasons. The behavior of any real system is the result of both the systematic forces relevant to a particular model and purpose and the peculiarities of historical circumstance: the randomness or noise, that is, the aspects of behavior that are not relevant for

the purpose of the study. The historical behavior of a social system, then, can be viewed as analagous to a particular simulation of a model with stochastic elements. The randomness represents those aspects of decisionmaking that are weakly coupled to the system of interest and have not been modeled. Forrester has shown that point-prediction of social systems beyond at most one-quarter of their natural period is impossible in principle even when one has a perfectly specified and estimated model, knows the nature of the noise or error terms, and lacks only the precise values of the noise, assumptions which can never be met in real life and only poorly approximated.[16] At the same time, one can always fit any set of data to any degree of precision required. Phelps Brown puts it even more strongly:

> The case for validating assumptions by testing their implications really rests on the possibility of controlled experiment, but that possibility is generally denied the economist ....Where, as so often, the fluctuations of different time series respond in common to the pulse of the economy, it is fatally easy to get a good fit, and get it for quite a number of different equations....running regressions between time series is only likely to deceive. (17)

Because historical fit is a weak test, system dynamicists have tended to ignore or minimize the comparison of the behavior of their models to historical data, preferring to focus the confidence-building process on the stronger tests outlined in Table 1. When historical fit is considered, it is usually presented in a highly informal manner. Typically, the modeler

presents a graph of the historical behavior alongside the simulated version of the same data and asks the reader to judge whether the degree of fit is "close enough" (the so-called "Mistaken Identity Test").[18]

The failure of system dynamicists to treat historical fit more rigorously is unfortunate. Although reproducing historical behavior is only one of a large number of tests and activities required to build confidence in a model, it is nonetheless an extremely important one. Failure to satisfy a client or reviewer that a model's historical fit is satisfactory is often sufficient grounds to dismiss the model and its conclusions. Passing the historical behavior test, while far from sufficient, is a necessary step in the confidence-building process. Arnold Zellner's response to Forrester's description of the use of information in system dynamics modeling is perhaps typical of the attitude of econometricians and other quantitative social scientists:

> One difference between Forrester's approach and those of others, however, is that Forrester apparently does not make explicit use of formal statistical inference techniques....I do believe that it would be worthwhile for Forrester to consider incorporating...appropriate and relevant statistical techniques in his approach (19).

More often than not, the historical fit of a system dynamics model is sufficient for its purpose. The problem arises from the

informal way in which goodness-of-fit is demonstrated. The Mistaken Identity test is considered naive and unprofessional, and visual comparison alone is viewed as "sloppy" by economists and other social scientists reared in more quantitative and statistical methods. Like it or not, system dynamics models are reviewed and evaluated by persons who expect a formal measure of goodness-of-fit, and who are reluctant to place confidence in a model unless its historical performance is appraised with some summary statistics with which they are familiar. System dynamicists, who emphasize the social nature of the confidence-building process, should be the first to employ formal measures of goodness-of-fit when the purpose of their models is to communicate results to social scientists with quantitative biases.

However, the use and interpretation of formal measures of goodness-of-fit must remain true to the purpose of system dynamics models and the confidence-building process. Historical fit is a necessary but far from sufficient test. Matching historical data must never become an end in itself, nor can the availability of numerical data be allowed to dictate the structure of a model. A good system dynamics model is expected to generate the historical behavior of the system endogenously, and without the extensive use of exogenous or dummy variables. Historical data should not be used to estimate the parameters of

a model directly; rather, parameters should be estimated from data "below the level of aggregation" of the model--that is, from interviews, engineering data, surveys, or other disaggregate studies that draw on descriptive knowledge of the system's structure rather than its aggregate behavior.[20]

Further, system dynamics models do not usually employ formal estimation procedures that guarantee a minimum sum-of-squared-errors over the range of available data, as in a regression.[21] As a result, the error between simulated and actual data may be larger than typically found in regression models. There may also be systematic bias between simulated and actual data. Yet precisely because exogenous and dummy variables are not used and the historical data are not used to derive the parameters that minimize some measure of error, larger errors than are typical in regression models do not necessarily compromise the validity of system dynamics models or imply lack of confidence in their results. In addition, system dynamics models are designed for a specific purpose and may deliberately exclude some of the modes of behavior present in the historical data. For example, a model of long-term economic growth may exclude the business cycle. The simulated GNP in such a model may not match the historical GNP, which fluctuates with the business cycle, on a point-by-point basis. The total error may be large even if the model matches the relevant growth mode extremely well.

For these reasons, the summary statistic most commonly used to evaluate goodness-of-fit in regression models, the coefficient of determination or R^2 (which measures the fraction of the total variation explained by the model), is inappropriate for system dynamics models.

Appropriate Summary Statistics for System Dynamics

To develop appropriate summary statistics to evaluate the historical fit of system dynamics models, it is useful to review the role of historical data in regression models such as econometric models based on time-series data. Often only the first portion of the available data is used to estimate the parameters of a model. Within the period of fit, the R^2, t-statistics, and other usual measures of goodness-of-fit and significance are applicable. The model is then simulated beyond the period of fit, to generate an *ex post* forecast. Simulating the model beyond the available data produces an *ex ante* forecast.[22]

The purpose of an *ex post* forecast is precisely the same as the purpose of analyzing the historical behavior of system dynamics models: to build confidence in the model. An *ex post* forecast provides a test of the model's ability to replicate the behavior of the real system that is independent of the process by which the structure and parameters of the model were chosen.

(Using the entire set of available data to estimate a model is a much weaker test, if passed, even if the resulting behavior is a closer match because the data in that case are directly used to find the structure and parameters that best match the data.)

Because system dynamics models typically do not employ the aggregate historical data in developing the structure or estimating the parameters, the behavior of the model over the entire range of available data may be analyzed as an ex post forecast, and summary statistics designed to measure forecast error are thus the appropriate measures of fit.[23]

The measurement and interpretation of forecast error has been studied extensively by statisticians and econometricians. One of the most common measures of forecast error is the mean-square-error (MSE), defined as

$$\frac{1}{n}\sum_{t=1}^{n}(S_t-A_t)^2$$

where

n = Number of observations ($t = 1, \ldots n$)

S_t = Simulated value at time t

A_t = Actual value at time t.

The MSE error has the advantage that large errors are weighted more heavily than small ones, and that errors of opposite sign do not cancel each other out. Often the square root of the mean-square error is taken, yielding the root-mean-square (RMS) error. The RMS error provides a measure of error with the same units as the variable under consideration.

It is often more convenient to compute a normalized measure of error. A common and easily interpreted dimensionless measure is the root-mean-square percent error (RMSPE),

$$\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left[\frac{(S_t-A_t)}{A_t}\right]^2}.$$

Other normalizations are possible; the choice of an appropriate measure depends on the purpose of the error analysis and the nature of the data.[24]

Error Decomposition

In addition to the size of the total error, it is important to know the sources of error. Failure to fit the data may be caused by a poor model or by a large degree of randomness in the historical data. The total error may be large if a mode of behavior in the real system is deliberately excluded as irrelevant to the purpose of the model. While there is ultimately no substitute for plotting the simulated and actual data side-by-side,

several statistical methods are available to decompose the total error into systematic and random portions.

One elegant decomposition of the mean-square-error is provided by the Theil inequality statistics. The Theil statistics are derived from the following decomposition of the MSE:[25]

$$\frac{1}{n}\sum_{t=1}^{n}(S_t-A_t)^2 = (\bar{S}-\bar{A})^2 + (s_S-s_A)^2 + 2(1-r)s_S s_A$$

where $\bar{S}$ and $\bar{A}$ are the means of S and A

$\frac{1}{n}\sum S_t$ and $\frac{1}{n}\sum A_t$, respectively;

s_S and s_A equal the standard deviations of S and A

$\sqrt{\frac{1}{n}\sum(S_t-\bar{S})^2}$ and $\sqrt{\frac{1}{n}\sum(A_t-\bar{A})^2}$, respectively;

and finally r equals the correlation coefficient between simulated and actual data

$$\frac{\frac{1}{n}\sum(S_t-\bar{S})(A_t-\bar{A})}{s_S s_A}.$$

The term $(\bar{S}-\bar{A})^2$ measures the bias between simulated and actual series. The term $(s_S-s_A)^2$ is the component of the MSE due to a difference in the variances of the simulated and actual series, and measures the degree of unequal variation between the two series. Finally, the term $2(1-r)s_S s_A$ is the component of the

error due to incomplete covariation between the two series, and measures the degree to which the changes in the simulated series fail to match the changes in the actual series on a point-by-point basis.

By dividing each of the components of the error by the total mean-square-error, the "inequality proportions" are derived:

$$U^M = \frac{(\bar{S}-\bar{A})^2}{\frac{1}{n}\sum(S_t-A_t)^2}$$

$$U^S = \frac{(s_S-s_A)^2}{\frac{1}{n}\sum(S_t-A_t)^2}$$

$$U^C = \frac{2(1-r)s_S s_A}{\frac{1}{n}\sum(S_t-A_t)^2}.$$

Of course, $U^M + U^S + U^C = 1$, so U^M, U^S, and U^C reflect the fraction of the mean-square-error due to bias, unequal variance, and unequal covariance, respectively.

Interpretation of the Inequality Statistics

To see how the inequality statistics apply, consider each term in turn. Bias, indicated by a large U^M and small U^S and U^C, can be thought of as a translation of one series by a constant amount at all points in time (Figure 1a). A large bias (indi-

cated by both a large MSE and a large U^M) reveals a systematic difference between the model and reality. Errors due to bias are potentially serious, possibly indicating specification or parameter errors. Alternatively, bias may be due to acceptable simplifying assumptions which do not compromise the model.

Error due to unequal variance may also be systematic. Consider two cases: suppose unequal variation (U^S) dominates the error, with U^M and U^C small. Then the two series match on average and are highly correlated, but the magnitude of the variation in the two around their common mean differs. One variable is a "stretched out" version of the other. In Figure 1b, U^S is large because the magnitude of the trend in the two variables is different. Such a case reveals a systematic difference between simulated and actual series and directs attention to the assumptions of the model, much as bias does. Systematic error is also the verdict in Figure 1c, in which the magnitude of a cyclical mode in one variable is underestimated by the other, though the phasing is correct. Such a case would direct attention to the factors controlling the amplitude and damping of the cyclical mode.[26]

Alternatively, if U^S is large, but both series have the same mean (U^M=0) and if at least one variable is nearly constant, U^C will be small because the standard deviation s_S or s_A will be

small. In such a case (Figure 1d) the error would reflect random noise or a cyclic mode in one of the series not present in the other. The interpretation of such a situation depends on the purpose of the model. If the model is designed to investigate the cyclic mode, the failure of the model to generate the cycle would clearly be a systematic error. But if the purpose of the model is analysis of long-run behavior that abstracts from the short-term cycle, failure to represent the cycle is unimportant. The cycle becomes unsystematic noise relative to the model purpose.

If the majority of the error is concentrated in unequal covariation U^C, while U^M and U^S are small, it indicates that the point-by-point values of the simulated and actual series do not match even though the model captures the average value and dominant trends in the actual data well. Such a case might indicate a fairly constant phase shift or translation in time of a cyclical mode otherwise reproduced well (Figure 1e). More likely, a large U^C indicates one of the variables has a large random component or contains cyclical modes not present in the other series. In particular, a large U^C may be due to noise or cyclical modes in the historical data not captured by the model. A large U^C indicates the majority of the error is unsystematic with respect to the purpose of the model, and the model should not be faulted for failing to match the random component of the data.[27]

Unsystematic error may also show up in U^S. Suppose the actual series has a trend as well as cyclic modes or noise (Figure 1f). If there is no bias, $U^M=0$. The MSE will be divided between U^S and U^C: by virtue of the cycles or noise, the two series will have slightly different variances and will be imperfectly correlated even if the model matches the trend in the data. The distribution of the MSE between U^S and U^C will depend on the magnitude of the noise relative to that of the trend. Even though $U^S>0$ here, the error is unsystematic and does not compromise the model.

In terms of building confidence in the ability of a model to endogenously generate the behavior of the system, the error should be small and unsystematic, that is, concentrated in U^C or U^S. Large total errors need not compromise the model's utility if they are due to excluded modes or noise in the historical data. Conversely, large biases or unequal trend errors should lead to questions about the assumptions of the model. As in all statistical tests, the choice of significance or tolerance levels depends on the purpose of the model and the characteristics of the data.

An Illustration

The mean-square-error and inequality statistics have been used to evaluate the historical fit of a large system dynamics

model of energy-economy interactions.[28] The purpose of the model is to investigate the effects of resource depletion and rising energy prices on economic growth over the long term (the simulations run from 1950 to 2050). The model focuses on long-run growth and explicitly excludes the business cycle. The model is a dynamic general disequilibrium representation of the U.S. economy and energy sector, including OPEC. Table 2 summarizes the major endogenous and exogenous variables. Typical of system dynamics models, the model boundary is quite wide. All the major economic and energy aggregates are generated endogenously. In contrast, there are but three exogenous variables. Of these, population and the index of technological progress are specified at ten-year intervals, and linear interpolation is used in intervening years. Only the historical OPEC price is represented annually (and it is generated endogenously after 1982). Therefore the behavior of the model and its ability to replicate historical data, to capture trends and turning points, is predominantly the result of the interaction of the endogenous variables. Starting the model in 1950 provides roughly thirty years of simulated data to compare to the actual behavior of the economy.

Table 3 summarizes the error analysis for eleven variables. The RMS percent error provides a normalized measure of the magnitude of the error. The MSE error and inequality statistics

provide a measure of the total error and how it breaks down into bias, unequal variation, and unequal covariation components.[29]

The RMS percent errors are below ten percent with the exception of real private investment, the fraction of energy imported, and real energy prices. Five variables including real GNP, consumption, consumption as a fraction of GNP, and total energy consumption have RMS errors under 5 percent.

While the small total errors in most variables show the model adequately tracks the major variables, the several large errors might raise questions about the internal consistency of the model or the structure controlling those variables.

The error decomposition helps resolve such doubts. Consider real private investment. The RMS percent error is 11.7. But only 2% of the mean-square-error is due to bias, and unequal variation accounts for only 10% of the total. The vast majority of the error (nearly 90%) is due to unequal covariation, indicating that simulated investment tracks the underlying trend in actual investment almost perfectly, but diverges point-by-point. Plotting the two series and their residuals (Figure 2) shows that actual investment is the culprit as it fluctuates with the business cycle around the simulated values. Since the business cycle is explicitly excluded from the purpose of the model, the

large RMS percent error is of little concern and does not compromise the conclusions of the study.

The energy import fraction reveals the same pattern. Only 12% of the MSE is due to bias, and virtually none to unequal variation. Imports, as the most costly source of energy, are the most volatile component of energy consumption. Like investment, the actual import fraction fluctuates with the business cycle around the simulated value, again causing nearly 90% of the error to show up as unequal covariation. As shown in Figure 3, the model captures the rapid rise in imports that began around 1970 quite well, even though the point-by-point match is poor.

The largest RMS percent error, 14%, shows up in the real energy price. Error decomposition shows the majority of the MSE to be due to bias (58%) with the rest due to unequal covariation. Plotting the two series (Figure 4) reveals the cause: The average real energy price fell over 30% between 1950 and 1970 before rising 130% in the next seven years. The model does not capture the full decline in real price, only some of which can be explained by the pressure of inexpensive imports before 1973. Several theories have been offered to explain the drop in real energy prices up to 1970: economies of scale associated with ever larger electric generation plants, higher than average technical progress in the energy sector, and discovery of less

costly resources. Economies of scale and technical progress could be represented by assuming technology in the energy sector improved faster than the average, but is excluded for simplicity and since such an assumption would be exogenous. Similarly, depletion is assumed to be strictly monotonic: in keeping with traditional resource theory, it is assumed the least expensive resources are exploited first. It would be an easy matter to "tune" the model to reproduce the decline in real energy price. If the purpose of the model were point-prediction or short-term forecasting, such tuning would be appropriate and would help build confidence in the utility of the model. But since the purpose is assessment of long-term trends and policy analysis, such tuning, relying as it would on exogenous variables and ad-hoc adjustments to parameters, would contribute nothing to the confidence-building process and might actually decrease confidence by obscuring the model's ability to endogenously capture the behavior of interest.

As a final illustration of error decomposition, consider the 9.7 RMS percent error in net energy consumption. Only 16% of the MSE is due to bias, but nearly two-thirds is due to unequal variance. Again, a simplifying assumption is responsible. Net energy consumption (gross primary consumption less conversion and distribution losses) is underestimated by the model during the 1950s and overestimated during the 1970s. Actual efficiency

dropped from 88% in 1950 to 70% in 1977, primarily due to the large conversion losses associated with electricity generation. For simplicity, the model assumes a constant average efficiency of 80%. Thus, as shown in Figure 5, simulated net consumption grows more rapidly than the actual value, resulting in the large error in variance. Since the model is not intended for forecasting but rather for policy analysis, the error in net energy consumption is of little concern as it will not affect the relative efficacy of policies.

Reviewing the other variables with RMS percent errors greater than 5% shows they all have small bias and unequal variance components. The bias fraction in the real wage (RMSPE=5.4) is just .10, while the bias in primary energy production (RMSPE=7.6) is just .14; the unequal variation terms are .23 and .26, respectively.

The error analysis shows the model reproduces historical behavior well. The small number of large errors are readily explained, with the help of error decomposition, as the result of modes of behavior outside the purpose of the model, noise in the historical data, or simplifying assumptions. In interpreting the statistical results, it must be stressed that the historical data were not used by a formal estimation procedure that guarantees the minimum sum-of-squared errors. Parameters were chosen on the

basis of disaggregate data, econometric estimation reported by others, and other managerial and engineering data. More important, no dummy variables and only three exogenous variables are used. The ability of the model to capture the trends and turning points in the historical data, over a three-decade span, is due to the interaction of the endogenous variables. To the author's knowledge, no other energy-economy model can make that claim.

Conclusion: Rigor Means Never Having to Say You're Sorry

Though historical fit is but one of many tests required to build confidence in a system dynamics model, and a weak one at that, it is nevertheless a necessary one. The process of building confidence in system dynamics models has been hampered by the reluctance of model builders to employ formal measures of goodness-of-fit, even though the models often fit the historical data quite well, generate the behavior endogenously, and pass a variety of structure and behavior adequacy tests. The statistics developed here provide a straightforward, easily interpreted method to lend rigor to the analysis of historical behavior. The root-mean-square percent error provides a simple way to gauge the magnitude of the total error between simulated and actual variables. The Theil inequality statistics are particularly well suited for system dynamics models because they allow the analyst to separate the fraction of the error due to excluded modes or

noise from the error due to systematic differences between the model and reality.

Other summary statistics may be more appropriate for some purposes and systems. For example, a model focusing on an oscillatory mode such as the business cycle will typically not be expected to reproduce the point-by-point behavior of the system because of the strong influence of noise on its exact trajectory. In such a case, establishing confidence in the model rests on the correspondence of the average period, amplitude, and phase relationships among the variables. Appropriate summary statistics might compare the means, variances, or spectral densities of the variables.

The statistics proposed here are not tests of validity, but are summary statistics: convenient, compact, ways to express the correspondence between a model's behavior and numerical data. The use of summary statistics (when numerical data exist) can help to establish confidence in system dynamics models without placing unwarranted emphasis on the point-by-point correspondence with historical data. But historical fit in itself must not be viewed as a test of validity. Building confidence in the structure of the model demands the analyst expose it to other, more severe tests. Such tests may include statistical tests where the important maintained hypotheses can be established, but

will rest primarily on the structure and behavior adequacy tests described in Table 1. The true test of a model is its ability to reproduce historical behavior endogenously, with structure and parameters that are consistent with descriptive knowledge of the system. These are strong requirements that few models in economics and social science can meet. Satisfying them is the process of building confidence in a model, and a well-built and carefully tested system dynamics model owes no apology to those who would judge validity by statistics alone.

APPENDIX

Computing the Summary Statistics with DYNAMO

The summary statistics presented above can be computed easily with DYNAMO using the following macros. In general, a simulation may start before and end after the period in which the historical comparison is to be made. Further, the model may compute the simulated values more frequently than the data are available. It is necessary to compute the summary statistics using sampled versions of simulated and actual data.

The Pick Function

```
MACRO PICK(ST,ET,PER)                                              1
    PICK   - PICK FUNCTION (DIMENSIONLESS) <2>
    ST     - START TIME FOR PICK FUNCTION (TIME UNIT)
    ET     - END TIME FOR PICK FUNCTION (TIME UNIT)
    PER    - PERIOD OF DATA FOR PICK FUNCTION (TIME UNITS)

PICK.K=PULSE(1,ST,PER)*(1-STEP(1,ET+DT))                          A,2
    PICK   - PICK FUNCTION (DIMENSIONLESS) <2>
    PULSE  - PULSE FUNCTION
    ST     - START TIME FOR PICK FUNCTION (TIME UNIT)
    PER    - PERIOD OF DATA FOR PICK FUNCTION
             (TIME UNITS)
    STEP   - STEP FUNCTION
    ET     - END TIME FOR PICK FUNCTION (TIME UNIT)
    DT     - TIME STEP FOR SIMULATION (TIME UNITS)

MEND                                                               3
```

The PICK function is used to sample a variable at a specified period PER over a specified interval (from ST to ET):

$$\text{PICK} = \begin{cases} 1 & \text{TIME=ST,ST+PER,ST+2PER,...,ET} \\ 0 & \text{otherwise.} \end{cases}$$

PICK has a value of zero before ST and after ET, and takes a value of 1 at intervals of PER between (and including) ST and ET.

Root-Mean-Square Percent Error Macro

```
MACRO RMSPE(HV,SV,ST,ET,PER,PE)                                    4
    RMSPE   - ROOT-MEAN-SQUARE PERCENT ERROR (%) <6>
    HV      - HISTORICAL VARIABLE (UNITS)
    SV      - SIMULATED VARIABLE (UNITS)
    ST      - START TIME FOR PICK FUNCTION (TIME UNIT)
    ET      - END TIME FOR PICK FUNCTION (TIME UNIT)
    PER     - PERIOD OF DATA FOR PICK FUNCTION (TIME UNITS)
    PE      - PERCENT ERROR BETWEEN SIMULATED AND ACTUAL
              VARIABLES (%) <5>
```

The RMSPE macro computes the root-mean-square percent error between simulated and historical variables and also the percent error at each moment. The RMSPE is computed every PER time units between ST and ET, inclusive.

```
PE.K=100*(SV.K-HV.K)/HV.K                                          A,5
    PE      - PERCENT ERROR BETWEEN SIMULATED AND ACTUAL
              VARIABLES (%) <5>
    SV      - SIMULATED VARIABLE (UNITS)
    HV      - HISTORICAL VARIABLE (UNITS)
```

The error between simulated and historical variables is computed as a percent of the historical value.

```
RMSPE.K=SQRT($SSPE.K/$N.K)                                         A,6
    RMSPE   - ROOT-MEAN-SQUARE PERCENT ERROR (%) <6>
    SQRT    - SQUARE ROOT
    $SSPE   - SUM OF SQUARED PERCENT ERRORS (% SQUARED)
              <9>
    $N      - NUMBER OF OBSERVATIONS (DIMENSIONLESS)
              <7,18,30>
```

The root-mean-square percent error is defined as the square root of the mean of the squared percent errors.

```
$N.K=$N.J+(DT/DT)*$IN.J                                            L,7
$N=1E-20                                                         N,7.1
```



```
    $N      - NUMBER OF OBSERVATIONS (DIMENSIONLESS)
              <7,18,30>
    DT      - TIME STEP FOR SIMULATION (TIME UNITS)
    $IN     - INCREMENT IN NUMBER OF OBSERVATIONS
              (DIMENSIONLESS) <8,19,31>

$IN.K=PICK(ST,ET,PER)                                              A,8
    $IN     - INCREMENT IN NUMBER OF OBSERVATIONS
              (DIMENSIONLESS) <8,19,31>
    PICK    - PICK FUNCTION (DIMENSIONLESS) <2>
    ST      - START TIME FOR PICK FUNCTION (TIME UNIT)
    ET      - END TIME FOR PICK FUNCTION (TIME UNIT)
    PER     - PERIOD OF DATA FOR PICK FUNCTION
              (TIME UNITS)
```

The number of observations is incremented by one every PER time units between ST and ET. (NB: The term 1E-20 prevents division by zero in Eq. 6 before TIME=ST+PER. The term (DT/DT) in Eq. 7 is necessary only to prevent an "unusual format in level equation" error from DYNAMO.)

```
$SSPE.K=$SSPE.J+(DT/DT)*$SPE.J                                     L,9
$SSPE=0                                                          N,9.1
    $SSPE   - SUM OF SQUARED PERCENT ERRORS (% SQUARED)
              <9>
    DT      - TIME STEP FOR SIMULATION (TIME UNITS)
    $SPE    - SQUARED PERCENT ERROR (% SQUARED) <10>

$SPE.K=PE.K*PE.K*PICK(ST,ET,PER)                                  A,10
    $SPE    - SQUARED PERCENT ERROR (% SQUARED) <10>
    PE      - PERCENT ERROR BETWEEN SIMULATED AND ACTUAL
              VARIABLES (%) <5>
    PICK    - PICK FUNCTION (DIMENSIONLESS) <2>
    ST      - START TIME FOR PICK FUNCTION (TIME UNIT)
    ET      - END TIME FOR PICK FUNCTION (TIME UNIT)
    PER     - PERIOD OF DATA FOR PICK FUNCTION
              (TIME UNITS)

MEND                                                               11
```

The squared percent errors, sampled by the PICK function, accumulate to yield the sum of squared percent errors.

Inequality Statistics Macro

```
MACRO MSE(HV,SV,ST,ET,PER,UM,US,UC)                              12
    MSE     - MEAN-SQUARE-ERROR (UNITS SQUARED) <13>
    HV      - HISTORICAL VARIABLE (UNITS)
    SV      - SIMULATED VARIABLE (UNITS)
    ST      - START TIME FOR PICK FUNCTION (TIME UNIT)
    ET      - END TIME FOR PICK FUNCTION (TIME UNIT)
    PER     - PERIOD OF DATA FOR PICK FUNCTION
              (TIME UNITS)
    UM      - FRACTION OF MSE DUE TO BIAS (FRACTION)
              <14>
    US      - FRACTION OF MSE DUE TO UNEQUAL VARIATION
              (FRACTION) <15>
    UC      - FRACTION OF MSE DUE TO UNEQUAL COVARIATION
              (FRACTION) <16>
```

The MSE macro computes the root-mean-square error between simulated and historical variables and the Theil inequality proportions U^M, U^S, and U^C.

```
MSE.K=$SSE.K/$N.K                                                A,13
    MSE     - MEAN-SQUARE-ERROR (UNITS SQUARED) <13>
    $SSE    - SUM OF SQUARED ERRORS (UNITS SQUARED) <19>
    $N      - NUMBER OF OBSERVATIONS (DIMENSIONLESS)
              <7,17,29>

UM.K=($MSV.K-$MHV.K)($MSV.K-$MHV.K)/(1E-20+MSE.K)                A,14
    UM      - FRACTION OF MSE DUE TO BIAS (FRACTION)
              <14>
    $MSV    - MEAN OF SIMULATED VARIABLE (UNITS) <21>
    $MHV    - MEAN OF HISTORICAL VARIABLE (UNITS) <22>
    MSE     - MEAN-SQUARE-ERROR (UNITS SQUARED) <13>

US.K=($SDSV.K-$SDHV.K)($SDSV.K-$SDHV.K)/(1E-20+MSE.K)            A,15
    US      - FRACTION OF MSE DUE TO UNEQUAL VARIATION
              (FRACTION) <15>
    $SDSV   - STANDARD DEVIATION OF SIMULATED VARIABLE
              (UNITS)
    $SDHV   - STANDARD DEVIATION OF HISTORICAL VARIABLE
              (UNITS)
    MSE     - MEAN-SQUARE-ERROR (UNITS SQUARED) <13>

UC.K=(2)(1-$CORR.K)($SDSV.K)($SDHV.K)/(1E-20+MSE.K)              A,16
    UC      - FRACTION OF MSE DUE TO UNEQUAL
              COVARIATION (FRACTION) <16>
```



```
    $CORR   - CORRELATION COEFFICIENT BETWEEN SIMULATED
              AND HISTORICAL VARIABLES (DIMENSIONLESS)
              <23>
    $SDSV   - STANDARD DEVIATION OF SIMULATED VARIABLE
              (UNITS)
    $SDHV   - STANDARD DEVIATION OF HISTORICAL VARIABLE
              (UNITS)
    MSE     - MEAN-SQUARE-ERROR (UNITS SQUARED) <13>
```

A running measure of the mean-square-error (MSE) is computed as the model moves through time. The fraction of the MSE due to bias is given by the squared difference in the means of simulated and historical series relative to the MSE. The fraction of the MSE due to unequal variation is given by the squared difference in the standard deviation of simulated and historical series relative to the MSE. The fraction of the MSE due to unequal covariation is given by the product $(2)(1-r)(s_S)(s_A)$ relative to the MSE. (A small number is added to the denominator of the inequality proportions to prevent division by zero.)

```
$N.K=$N.J+(DT/DT)*$IN.J                                           L,17
$N=1E-20                                                        N,17.1
    $N      - NUMBER OF OBSERVATIONS (DIMENSIONLESS)
              <7,17,29>
    DT      - TIME STEP FOR SIMULATION (TIME UNITS)
    $IN     - INCREMENT IN NUMBER OF OBSERVATIONS
              (DIMENSIONLESS) <8,18,30>

$IN.K=PICK(ST,ET,PER)                                             A,18
    $IN     - INCREMENT IN NUMBER OF OBSERVATIONS
              (DIMENSIONLESS) <8,18,30>
    PICK    - PICK FUNCTION (DIMENSIONLESS) <2>
    ST      - START TIME FOR PICK FUNCTION (TIME UNIT)
    ET      - END TIME FOR PICK FUNCTION (TIME UNIT)
    PER     - PERIOD OF DATA FOR PICK FUNCTION
              (TIME UNITS)
```

The running total of the number of observations is calculated exactly as in the RMSPE macro.

```
$SSE.K=$SSE.J+(DT/DT)*$SE.J                                    L,19
$SSE=0                                                        N,19.1
    $SSE   - SUM OF SQUARED ERRORS (UNITS SQUARED)
               <19>
    DT     - TIME STEP FOR SIMULATION (TIME UNITS)
    $SE    - SQUARED ERROR (UNITS SQUARED) <20>

$SE.K=(SV.K-HV.K)(SV.K-HV.K)*PICK(ST,ET,PER)                   A,20
    $SE    - SQUARED ERROR (UNITS SQUARED) <20>
    SV     - SIMULATED VARIABLE (UNITS)
    HV     - HISTORICAL VARIABLE (UNITS)
    PICK   - PICK FUNCTION (DIMENSIONLESS) <2>
    ST     - START TIME FOR PICK FUNCTION (TIME UNIT)
    ET     - END TIME FOR PICK FUNCTION (TIME UNIT)
    PER    - PERIOD OF DATA FOR PICK FUNCTION
               (TIME UNITS)
```

The squared errors, sampled by the PICK function, accumulate in the sum of squared errors.

```
$MSV.K=MEAN(SV.K,ST,ET,PER,$SDSV.K)                            A,21
    $MSV   - MEAN OF SIMULATED VARIABLE (UNITS) <21>
    MEAN   - MEAN OF INPUT SERIES (UNITS) <28>
    SV     - SIMULATED VARIABLE (UNITS)
    ST     - START TIME FOR PICK FUNCTION (TIME UNIT)
    ET     - END TIME FOR PICK FUNCTION (TIME UNIT)
    PER    - PERIOD OF DATA FOR PICK FUNCTION
               (TIME UNITS)
    $SDSV  - STANDARD DEVIATION OF SIMULATED VARIABLE
               (UNITS)

$MHV.K=MEAN(HV.K,ST,ET,PER,$SDHV.K)                            A,22
    $MHV   - MEAN OF HISTORICAL VARIABLE (UNITS) <22>
    MEAN   - MEAN OF INPUT SERIES (UNITS) <28>
    HV     - HISTORICAL VARIABLE (UNITS)
    ST     - START TIME FOR PICK FUNCTION (TIME UNIT)
    ET     - END TIME FOR PICK FUNCTION (TIME UNIT)
    PER    - PERIOD OF DATA FOR PICK FUNCTION
               (TIME UNITS)
    $SDHV  - STANDARD DEVIATION OF HISTORICAL VARIABLE
               (UNITS)
```


The means and standard deviations of the simulated and historical variables are calculated over the relevant range of data by the MEAN macro (below).

```
$CORR.K=(($SPSH.K/$N.K)-$MSV.K*$MHV.K)/(1E-20+
$SDSV.K*$SDHV.K)                                               A,23
    $CORR  - CORRELATION COEFFICIENT BETWEEN SIMULATED
               AND HISTORICAL VARIABLES (DIMENSIONLESS)
               <23>
    $SPSH  - SUM OF PRODUCTS OF SIMULATED AND HISTORICAL
               VARIABLES (UNITS SQUARED) <24>
    $N     - NUMBER OF OBSERVATIONS (DIMENSIONLESS)
               <7,17,29>
    $MSV   - MEAN OF SIMULATED VARIABLE (UNITS) <21>
    $MHV   - MEAN OF HISTORICAL VARIABLE (UNITS) <22>
    $SDSV  - STANDARD DEVIATION OF SIMULATED VARIABLE
               (UNITS)
    $SDHV  - STANDARD DEVIATION OF HISTORICAL VARIABLE
               (UNITS)

$SPSH.K=$SPSH.J+(DT/DT)*$PSH.J                                 L,24
$SPSH=0                                                       N,24.1
    $SPSH  - SUM OF PRODUCTS OF SIMULATED AND
               HISTORICAL VARIABLES (UNITS SQUARED)
               <24>
    DT     - TIME STEP FOR SIMULATION (TIME UNITS)
    $PSH   - PRODUCT OF SIMULATED AND HISTORICAL
               VARIABLES (UNITS SQUARED) <25>

$PSH.K=SV.K*HV.K*PICK(ST,ET,PER)                               A,25
    $PSH   - PRODUCT OF SIMULATED AND HISTORICAL
               VARIABLES (UNITS SQUARED) <25>
    SV     - SIMULATED VARIABLE (UNITS)
    HV     - HISTORICAL VARIABLE (UNITS)
    PICK   - PICK FUNCTION (DIMENSIONLESS) <2>
    ST     - START TIME FOR PICK FUNCTION (TIME UNIT)
    ET     - END TIME FOR PICK FUNCTION (TIME UNIT)
    PER    - PERIOD OF DATA FOR PICK FUNCTION
               (TIME UNITS)

MEND                                                           26
```

The "hand computation" formula is used to calculate the correlation coefficient between simulated and historical series as the model moves through time. The hand computation formula is based on the definition of the correlation coefficient

$$r = \frac{COV(S,A)}{s_S s_A}$$

and the following formula for covariance

$$COV(S,A) = \frac{1}{n}\sum[(S_t-\bar{S})(A_t-\bar{A})]$$

$$= \frac{1}{n}\sum(S_t A_t) - \bar{S}\,\bar{A}.$$

Mean and Standard Deviation Macro

```
MACRO MEAN(IS,ST,ET,PER,SD)                                        27
    MEAN    - MEAN OF INPUT SERIES (UNITS) <28>
    IS      - INPUT SERIES (UNITS)
    ST      - START TIME FOR PICK FUNCTION (TIME UNIT)
    ET      - END TIME FOR PICK FUNCTION (TIME UNIT)
    PER     - PERIOD OF DATA FOR PICK FUNCTION
              (TIME UNITS)
    SD      - STANDARD DEVIATION OF INPUT SERIES (UNITS)
              <33>
```

The MEAN macro computes running means and standard deviations over a specified range and periodicity of data.

```
MEAN.K=$SIS.K/$N.K                                                 A,28
    MEAN    - MEAN OF INPUT SERIES (UNITS) <28>
    $SIS    - SUM OF INPUT SERIES (UNITS) <31>
    $N      - NUMBER OF OBSERVATIONS (DIMENSIONLESS)
              <7,17,29>
```

The mean is defined as the sum of the sampled input series over the number of observations.

```
$N.K=$N.J+(DT/DT)*$IN.J                                            L,29
$N=1E-20                                                          N,29.1
    $N      - NUMBER OF OBSERVATIONS (DIMENSIONLESS)
              <7,17,29>
    DT      - TIME STEP FOR SIMULATION (TIME UNITS)
    $IN     - INCREMENT IN NUMBER OF OBSERVATIONS
              (DIMENSIONLESS) <8,18,30>
```



```
$IN.K=PICK(ST,ET,PER)                                              A,30
    $IN     - INCREMENT IN NUMBER OF OBSERVATIONS
              (DIMENSIONLESS) <8,18,30>
    PICK    - PICK FUNCTION (DIMENSIONLESS) <2>
    ST      - START TIME FOR PICK FUNCTION (TIME UNIT)
    ET      - END TIME FOR PICK FUNCTION (TIME UNIT)
    PER     - PERIOD OF DATA FOR PICK FUNCTION
              (TIME UNITS)
```

The number of observations is computed exactly as in the RMSPE and RMSE macros.

```
$SIS.K=$SIS.J+(DT/DT)*$IS.J                                        L,31
$SIS=0                                                            N,31.1
    $SIS    - SUM OF INPUT SERIES (UNITS) <31>
    DT      - TIME STEP FOR SIMULATION (TIME UNITS)
    $IS     - SAMPLED INPUT SERIES (UNITS) <32>

$IS.K=IS.K*PICK(ST,ET,PER)                                         A,32
    $IS     - SAMPLED INPUT SERIES (UNITS) <32>
    IS      - INPUT SERIES (UNITS)
    PICK    - PICK FUNCTION (DIMENSIONLESS) <2>
    ST      - START TIME FOR PICK FUNCTION (TIME UNIT)
    ET      - END TIME FOR PICK FUNCTION (TIME UNIT)
    PER     - PERIOD OF DATA FOR PICK FUNCTION
              (TIME UNITS)
```

The sampled input series is summed for computation of the mean.

```
SD.K=SQRT(MAX(0,($SISS.K/$N.K)-MEAN.K*MEAN.K))                     A,33
    SD      - STANDARD DEVIATION OF INPUT SERIES
              (UNITS) <33>
    SQRT    - SQUARE ROOT
    MAX     - MAXIMUM FUNCTION
    $SISS   - SUM OF INPUT SERIES SQUARED
              (UNITS SQUARED) <34>
    $N      - NUMBER OF OBSERVATIONS (DIMENSIONLESS)
              <7,17,29>
    MEAN    - MEAN OF INPUT SERIES (UNITS) <28>

$SISS.K=$SISS.J+(DT/DT)*$ISS.J                                     L,34
$SISS=0                                                           N,34.1
    $SISS   - SUM OF INPUT SERIES SQUARED
              (UNITS SQUARED) <34>
    DT      - TIME STEP FOR SIMULATION (TIME UNITS)
    $ISS    - INPUT SERIES SQUARED (UNITS SQUARED)
              <35>
```

```
$ISS.K=IS.K*IS.K*PICK(ST,ET,PER)                                    A,35
    $ISS   - INPUT SERIES SQUARED (UNITS SQUARED) <35>
    IS     - INPUT SERIES (UNITS)
    PICK   - PICK FUNCTION (DIMENSIONLESS) <2>
    ST     - START TIME FOR PICK FUNCTION (TIME UNIT)
    ET     - END TIME FOR PICK FUNCTION (TIME UNIT)
    PER    - PERIOD OF DATA FOR PICK FUNCTION
             (TIME UNITS)

MEND                                                                 36
```

The "hand computation" formula for the standard deviation is used to calculate the standard deviation without prior knowledge of the mean. The hand computation formula follows from the definition of variance:

$$\text{VAR}(A) = \frac{1}{n}\sum [(X-\bar{X})^2]$$

$$= \frac{1}{n}\sum (X^2) - [\frac{1}{n}\sum (X)]^2$$

$$= \frac{1}{n}\sum (X^2) - [\text{MEAN}(X)]^2.$$

However, the hand computation formula (and the hand computation formula for the correlation coefficient above) are subject to more round-off error than the computations based on the definitions of variance and covariance. (The error is larger because the hand computation involves small differences of large numbers.) To guard against the possibility that the difference $SISS.K/$N.K-MEAN.K*MEAN.K is negative, a MAX function is inserted in Eq. 33. Round-off error has not been a problem in actual applications to date. Sterman 1981 describes an alternative approach to computing the statistics described here that involves less round-off error but is more cumbersome to use.

NOTES

1. I am indebted to Ernst R. Berndt, Jack B. Homer, and George P. Richardson for many useful comments and criticisms; of course, all errors are mine.

2. E.g., Forrester 1961, Ch. 13; Ansoff and Slevin 1968; Forrester 1968; Nordhaus 1973; Forrester 1973; Forrester et al. 1974; Forrester and Senge 1980; Richardson and Pugh 1981.

3. Forrester 1961, p. 115.

4. Naylor and Finger 1967, p. B-97, McKenney 1967, p. B-102, Hermann 1967, pp. 217ff, Lilien 1975, Pindyck and Rubinfeld 1976, p. 315. Greenberger et al. 1976, pp. 62-63 and 70-74.

5. Forrester 1961, p. 123.

6. The t-statistic and other common tests of significance are discussed in any introductory econometrics text, e.g., Pindyck and Rubinfeld 1976.

7. Pindyck and Rubinfeld 1976, p. 37.

8. Mass and Senge 1978, p. 451.

9. Mass and Senge 1978 demonstrate that a moderate amount of measurement error causes insignificant t-statistics in OLS estimation of a model with the same specification as the model used to generate the data in the first place. One way to recover from such errors (if they are detected) is to employ a more sophisticated estimation procedure, such as full-information maximum likelihood estimation using Kalman filtering (Peterson 1980). A simpler and often more illuminating approach is to subject the model to the behavior anomaly test (Table 1): does anomalous behavior arise if the assumption is deleted? Senge 1978 uses such behavior tests to establish the necessity of various hypotheses in an investment function whose statistical performance was only slightly better than that of the neoclassical function.

10. Hausman 1978 presents a test of model specification.

11. Forrester 1973, pp. 24-31. The operator/observer distinction is an important one that accounts for much of the disagreement on validation.

12. Churchman 1973, p. 12.

13. Greenberger et al., pp. 70-71.

14. Naylor and Finger 1967 propose "multistage verification" as a confidence-building procedure (see also Naylor 1971). Hermann 1967 proposes a five-stage confidence-building approach; Emshoff and Sisson 1970 also emphasize an iterative and variegated approach to confidence building; Schrank and Holt 1967 stress "the criterion of usefulness," (p. B-105).

15. Naylor and Finger 1967 (p. B-97) endorse Cyert's version of the Behavior Reproduction test, which lists eight attributes of similarity, the least important of which is "exact matching of values of variables."

16. Forrester 1961, app. K.

17. Phelps Brown 1972, pp. 5-6.

18. The Mistaken Identity Test is described in Forrester 1973, pp. 53-54. For examples, see Naill 1977, app. A and Runge 1976, Ch. 5. The Mistaken Identity Test is similar to McKenney's 1967 proposal to employ a Turing Test as "an adequate method of validation." The Turing Test or imitation game (Turing 1950) was originally proposed as a sufficient test for artificial intelligence: in Turing's view, if a panel of human interrogators cannot distinguish the performance of a machine (model) from that of a human (real system), then the machine (model) <u>is</u> an artificial intelligence (valid model).

19. Zellner 1980, p. 567.

20. The philosophy of parameter estimation in system dyanmics is described in Forrester 1961, e.g., pp. 171-172, Forrester 1980, pp. 559-560, and Richardson and Pugh 1981, pp. 230-240. Estimation of parameters "below the level of aggregation" of a model is discussed in Graham 1980. The fallacy in using aggregate data to evaluate structural relationships is illustrated by Nordhaus 1973 and exposed by Forrester et al. 1974.

21. The maintained hypotheses of most single-equation techniques are violated by the multiloop, nonlinear nature of complex feedback systems with measurement error. However, optimal filtering (Peterson 1980) offers a promising approach to formal estimation of system dynamics models.

22. See, e.g., Pindyck and Rubinfeld 1976, pp. 157ff.

23. The summary statistics described here may be useful even if the data are used to estimate the parameters. Relative to an <u>ex post</u> forecast, such a case, like an in-sample simulation of an econometric model, is a weaker test of a model if it passes and a stronger test if it fials.

24. The RMSPE and other error measures are discussed by Pindyck and Rubinfeld 1976, pp. 314-320. Other normalizations include the root-mean-square error as a percent of the mean,

$$\frac{\sqrt{\frac{1}{n}\sum (S_t - A_t)^2}}{\frac{1}{n}\sum A_t}.$$

Theil 1966, pp. 27-28, divides the MSE by the mean of the squared actual values to define what he calls U, the "inequality coefficient";

$$U^2 = \frac{\frac{1}{n}\sum (S_t - A_t)^2}{\frac{1}{n}\sum A_t^2}.$$

These and other normalizations have individual strengths and weaknesses in particular situations. As usual one cannot apply statistics blindly without considering the purpose of the analysis or the nature of the data.

25. The description of the Theil statistics is reproduced from Theil 1966, Ch. 2.3-2.5.

26. It would also make one suspect the model had been "fine-tuned" with exogenous variables to match the phasing. Most models of cyclical phenomena must include some randomness to excite the latent oscillatory modes, but inclusion of noise will certainly cause the turning points to differ from historical behavior in the same way that two samples drawn from the same distribution will differ point by point. In practice, the error shown in Figure 3c is unlikely to arise.

27. Note that when the purpose of a model abstracts from a cycle, the cycle becomes noise: that part of the decisionmaking process that is not modeled. In such a case there will be high serial correlation in the residuals, but the presence of such autocorrelation does not compromise the model.

28. Sterman 1981, Sterman 1982.

29. Note that $RMSPE^2 \neq MSE$, so one cannot multiply the $RMSPE^2$ by U^M, U^S, or U^C to yield the "RMSPE due to bias," etc.

REFERENCES

Ansoff, Igor and Slevin, Dennis, "An Appreciation of Industrial Dynamics," Management Science 14(7), March 1968, pp. 383-397.

Churchman, C. W., "Reliability of Models in the Social Sciences," Interfaces, 4(1) November 1973, pp. 1-12.

Emshoff, James and Sisson, Roger, Design and Use of Computer Simulation Models. New York, MacMillan, 1970.

Forrester, Jay W., Industrial Dynamics. Cambridge: The MIT Press, 1961.

Forrester, Jay W., Industrial Dynamics--A Response to Ansoff and Slevin," Management Science, 14(9), May 1968, pp. 601-618.

Forrester, Jay W., "Confidence in Models of Social Behavior--With Emphasis on System Dynamics Models," Working paper D-1967, System Dynamics Group, MIT, December 1973.

Forrester, Jay W., "Information Sources for Modeling the National Economy," Journal of the American Statistical Association, 75(371), September 1980, pp. 555-574.

Forrester, Jay W. et al., "The Debate on World Dynamics--A Response to Nordhaus," Policy Sciences 5 (1974), pp. 169-190.

Forrester, Jay W. and Senge, Peter M., "Tests for Building Confidence in System Dynamics Models," TIMS Studies in the Management Sciences 14 (1980), pp. 201-228.

Graham, Alan K., "Parameter Estimation in System Dynamics Modeling," Elements of the System Dynamics Method (Jørgen Randers, ed.). Cambridge: The MIT Press, 1980, pp. 143-161.

Greenberger, Martin et al., Models in the Policy Process. New York: Russell Sage Foundation, 1976.

Hausman, J. A., "Specification Tests in Econometrics," Econometrica, 46 (1978), pp. 1251-1272.

Hermann, C. F., "Validation Problems in Games and Simulations," Behavioral Science, 12 (1967), pp. 216-231.

Lilien, Gary L., "Model Relativism: A Situational Approach to Model Building," Interfaces, 5(3) (May 1975), pp. 11-18.

Mass, Nathaniel J. and Senge, Peter M., "Alternative Tests for the Selection of Model Variables," IEEE Transactions on Systems, Man, and Cybernetics, SMC-8, no. 6, June 1978, pp. 450-460.

McKenney, James L., "Critique of 'Verification of Computer Simulation Models,'" Management Science, 14(2) (October 1967), pp. B-102 - B-103.

Naill, Roger F., Managing the Energy Transition. Cambridge: Ballinger, 1977.

Naylor, T. H. and Finger, J. M., "Verification of Computer Simulation Models," Management Science, 14(2) (October 1967), pp. B-92 - B-101.

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

Nordhaus, William D., "World Dynamics: Measurement Without Data," The Economic Journal 83, pp. 1156-1183.

Peterson, D. W., "Statistical Tools for System Dynamics," Elements of the System Dynamics Method (Jørgen Randers, ed.). Cambridge: The MIT Press, 1980, pp. 224-245.

Phelps Brown, E. H., "The Underdevelopment of Economics," The Economic Journal, 82(325) (March 1972), pp. 1-10.

Pindyck, Robert S. and Rubinfeld, Daniel L., Econometric Models and Economic Forecasts. New York: McGraw Hill, 1976.

Richardson, George P. and Pugh, Alexander L., Introduction to System Dynamics Modeling with DYNAMO. Cambridge: The MIT Press, 1981.

Schrank, W. E., and Holt, C. C., "Critique of 'Verification of Computer Simulation Models,'" Management Science, 14(2) (October 1967), pp. B-104 - B-106.

Senge, Peter M. The System Dynamics National Model Investment Function: A Comparison to the Neoclassical Investment Function. Ph.D. Dissertation, M.I.T.: 1978.

Sterman, John D., The Energy Transition and the Economy: A System Dynamics Approach. Ph.D. Dissertation, M.I.T.: 1981.

Sterman, John D., "Economic Vulnerability and the Energy Transition," Energy Systems and Policy. Forthcoming 1983.

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

Turing, Alan M., "Computing Machinery and Intelligence," Mind 5(1950), pp. 433-460.

Zellner, Arnold, "Comment" on Forrester's 'Information Sources for Modeling the National Economy,' Journal of the American Statistical Association 75(371), September 1980, pp. 567-569.

Table 1. Tests for Building Confidence In System Dynamics Models*

Tests of Model Structure	Question Addressed by the Test
Structure Verification	Is the model structure consistent with relevant descriptive knowledge of the system?
Parameter Verification	Are the parameters consistent with relevant descriptive (and numerical, when available) knowledge of the system?
Extreme Conditions	Does each equation make sense even when its inputs take on extreme values?
Boundary Adequacy (Structure)	Are the important concepts for addressing the problem endogenous to the model?
Dimensional Consistency	Is each equation dimensionally consistent without the use of parameters having no real-world counterpart?
Tests of Model Behavior	
Behavior Reproduction	Does the model endogenously generate the symptoms of the problem, behavior modes, phasing, frequencies, and other characteristics of the behavior of the real system?
Behavior Anomaly	Does anomalous behavior arise if an assumption of the model is deleted?
Family Member	Can the model reproduce the behavior of other examples of systems in the same class as the model (e.g., can an urban model generate the behavior of New York, Dallas, Carson City, and Calcutta when parameterized for each)?

Surprise Behavior	Does the model point to the existence of a previously unrecognized mode of behavior in the real system?
Extreme Policy	Does the model behave properly when subjected to extreme policies or test inputs?
Boundary Adequacy (Behavior)	Is the behavior of the model sensitive to the addition or alteration of structure to represent plausible alternative theories?
Behavior Sensitivity	Is the behavior of the model sensitive to plausible variations in parameters?
Statistical Character	Does the output of the model have the same statistical character as the "output" of the real system?
Tests of Policy Implications	
System Improvement	Is the performance of the real system improved through use of the model?
Behavior Prediction	Does the model correctly describe the results of a new policy?
Boundary Adequacy (Policy)	Are the policy recommendations sensitive to the addition or alteration of structure to represent plausible alternative theories?
Policy Sensitivity	Are the policy recommendations sensitive to plausible variations in parameters?

* Adapted from Forrester and Senge 1980, esp. p. 227, and from Richardson and Pugh 1981, esp. pp. 313-319.

Table 2: Summary of Energy-Economy Model Boundary

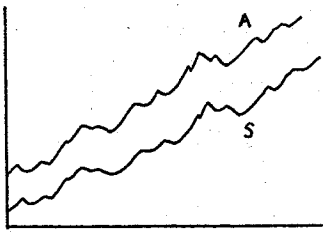
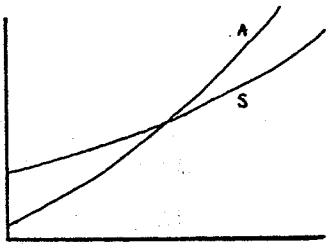
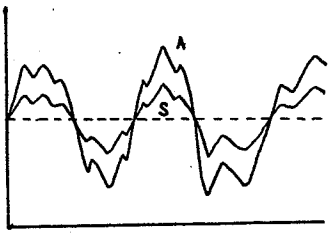
ENDOGENOUS	EXOGENOUS
GNP	Population
Consumption	Technological change
Investment	Historical OPEC Price (endogenous after 1982)
Savings	
Prices (Real and Nominal)	
Wages (Real and Nominal)	
Inflation Rate	
Labor Force Participation	
Employment	
Unemployment	
Interest Rates	
Money Supply	
Debt	
Energy Production	
Energy Demand	
Energy Imports	

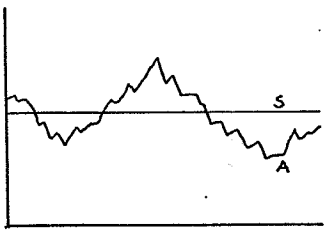
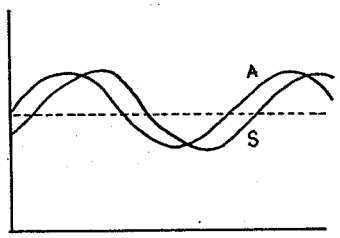
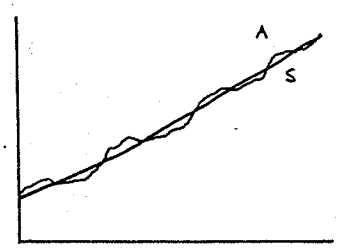
Table 3: Error Analysis of Energy-Economy Model

Variable	Root Mean Square Percent Error RMSPE (%)	Mean Square Error MSE (units^2)	Inequality Statistics[a] U^M (fraction of MSE)	U^S (fraction of MSE)	U^C (fraction of MSE)
Real GNP (Billion 1972 $/year)	3.2	9.7×10^{2}	.10	.00	.90
Real Consumption (Billion 1972 $/year)	4.7	19×10^{2}	.54	.29	.17
Consumption Fraction[b] (fraction)	3.6	8.7×10^{-4}	.46	.01	.53
Real Private Investment (Billion 1972 $/year)	11.7	3.5×10^{2}	.02	.10	.88
Real Wage (Thousands of 1972 $/person/year)	5.4	9.0×10^{-2}	.10	.23	.67
Workforce Participation Fraction (fraction)	2.5	2.2×10^{-4}	.75	.16	.09
Primary Energy Consumption (Quads/year)	4.0	4.2×10^{0}	.04	.05	.91
Primary Energy Production (Quads/year)	7.6	9.1×10^{0}	.14	.26	.59
Energy Import Fraction (fraction)[c]	13.9	1.8×10^{-4}	.12	.01	.87
Real Energy Price (1972 $/MMBTU)	14.0	4.9×10^{-3}	.58	.00	.42
Net Energy Consumption (Quads/year)	9.7	2.2×10^{1}	.16	.62	.22

[a] Totals may not add due to rounding.
[b] Real Consumption/Real GNP
[c] Computed from 1960 to 1977

Figure 1. Interpretation of the Theil Inequality Statistics

		U^M	U^S	U^C	Characterization	Interpretation
a)		1	0	0	$S_t=A_t+BIAS$ (S is equal to A translated by a constant BIAS.)	Systematic error.
b)		0	1	0	$S_t-\overline{S}=k(A_t-\overline{A})$ where $\overline{S}=\overline{A}$. ($S_t$ is a stretched version of A_t about their common mean.)	Systematic error: S and A have different trends.
c)		0	1	0	Same as (b).	Systematic error: S and A have the same phasing but different magnitude fluctuations.

		U^M	U^S	U^C	Characterization	Interpretation
d)		0	1	0	$S_t=k$; $A_t=k+f(t)$ where $\overline{f(t)}=0$ (A has cycles or noise not present in S.)	The error is unsystematic unless purpose of model is to study the cycles in A.
e)		0	0	1	$A_t=\overline{A}+Ksin(wt)$; $S_t=\overline{S}+Ksin(wt+p)$ where $\overline{A}=\overline{S}$. (S is a translation in time of A by a phase margin.)	Same means and variances but phasing differs: probably unsystematic error.
f)		0	a	1-a	$S_t=f(t)$; $A_t=f(t)+e_t$ where $\overline{e}=0$ (S equals A except for different values of the 'error' term.)	S and A have the same mean and trends but vary point by point: Unsystematic error unless purpose is to study the cycles in A.

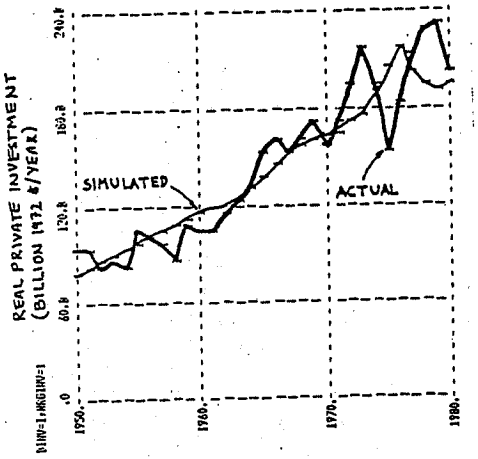


Figure 2a: Real Private Investment

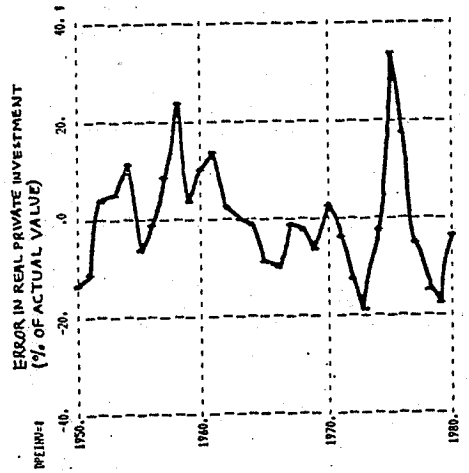


Figure 2b: Real Private Investment: Residuals

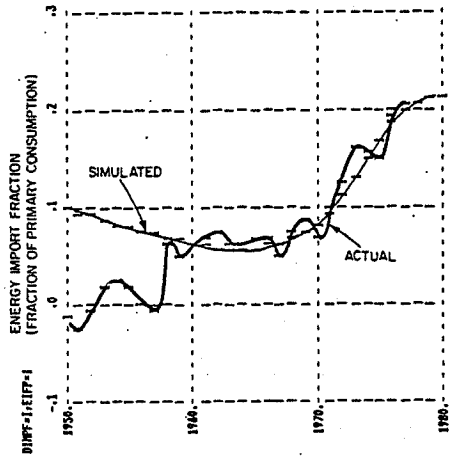


Figure 3a: Energy Import Fraction

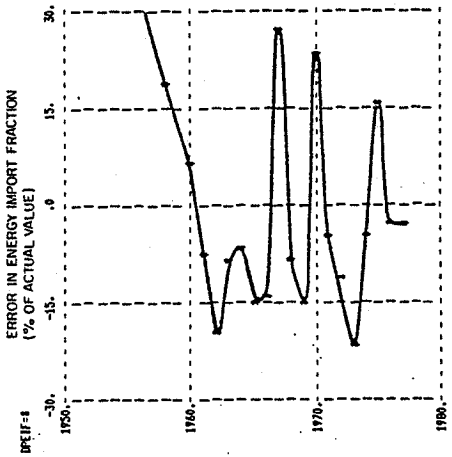


Figure 3b: Energy Import Fraction: Residuals

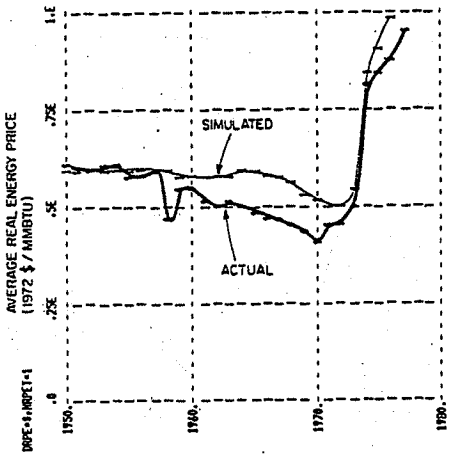


Figure 4a: Real Energy Price

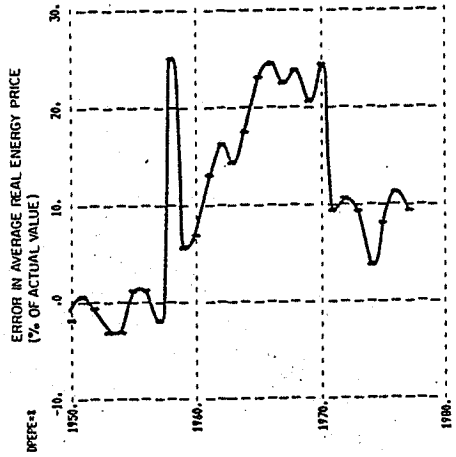


Figure 4b: Real Energy Price: Residuals

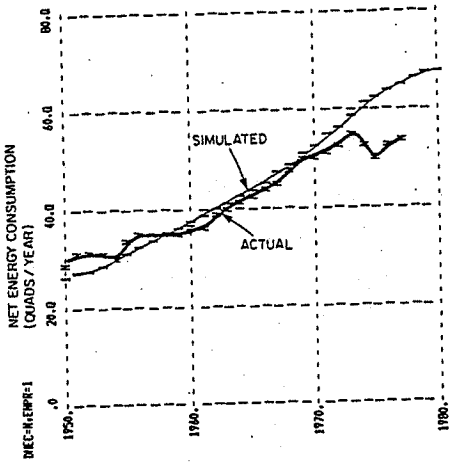


Figure 5a: Net Energy Consumption

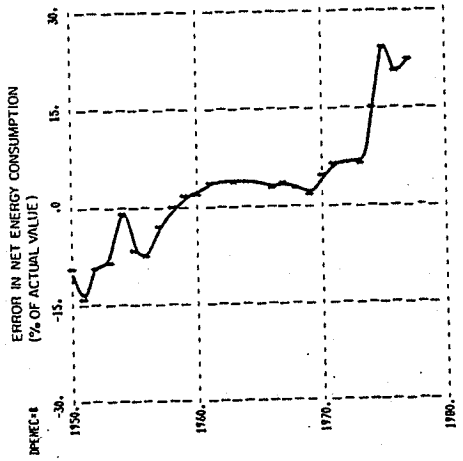


Figure 5b: Net Energy Consumption: Residuals

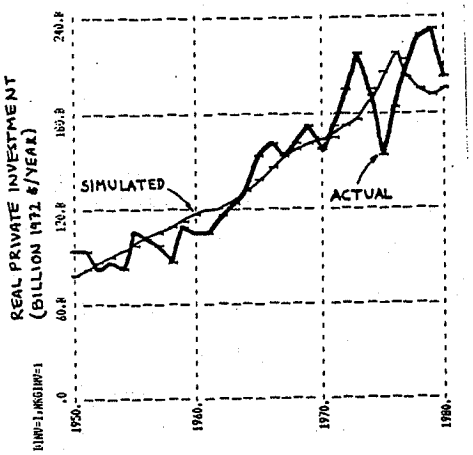


Figure 2a: Real Private Investment

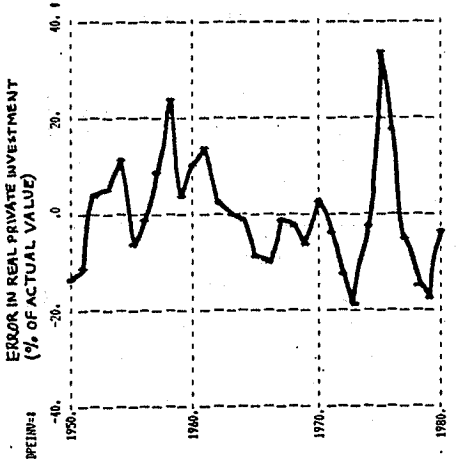


Figure 2b: Real Private Investment: Residuals

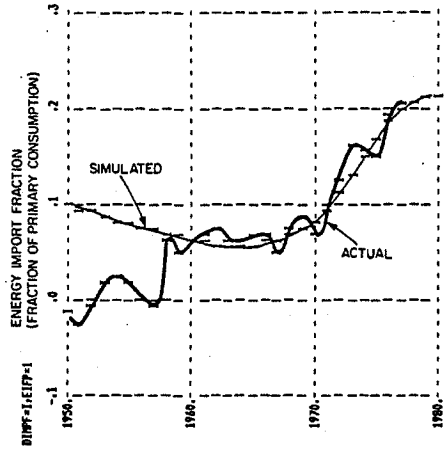


Figure 3a: Energy Import Fraction

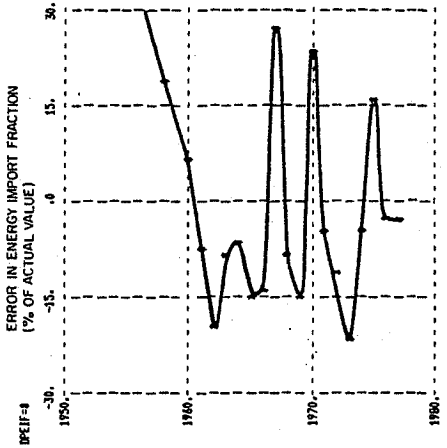


Figure 3b: Energy Import Fraction: Residuals

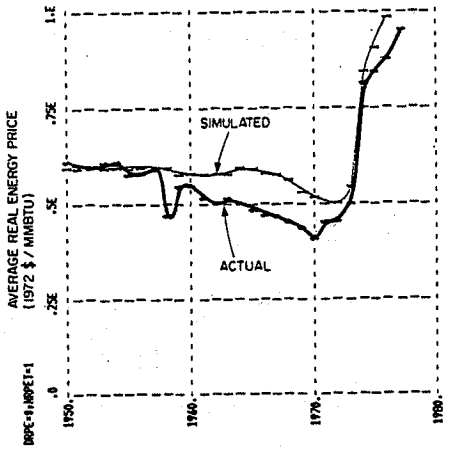


Figure 4a: Real Energy Price

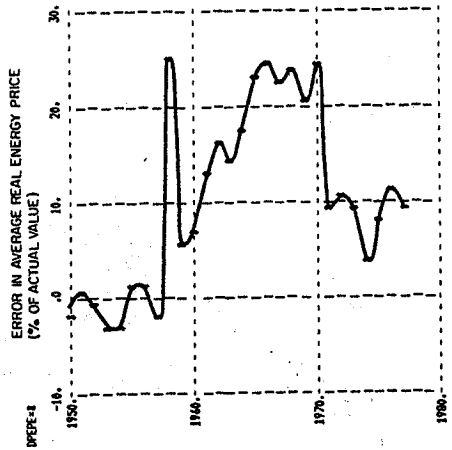


Figure 4b: Real Energy Price: Residuals

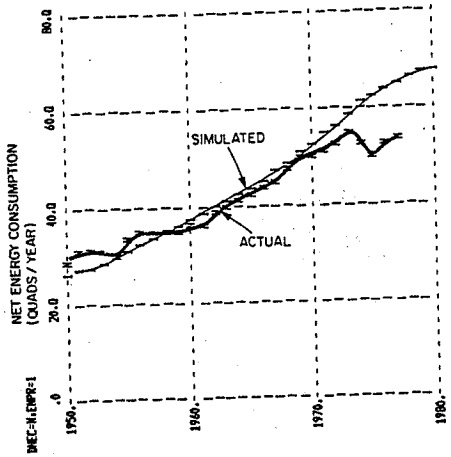


Figure 5a: Net Energy Consumption

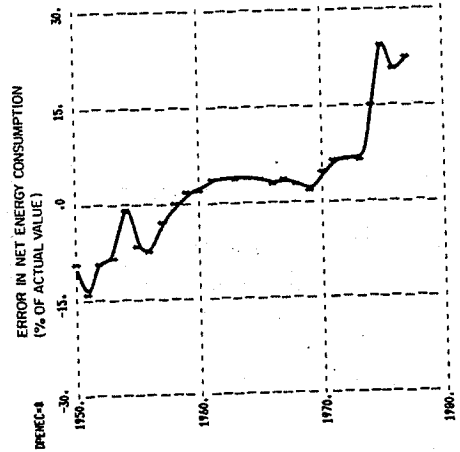


Figure 5b: Net Energy Consumption: Residuals