N-1204-DOE

June 1979

MODELS, COMPLEXITY, AND ERROR

David Leinweber

A Rand Note
pre pared for the
U.S. DEPARTMENT OF ENERGY
The work upon which this publication is based was performed pursuant to Contract DE-AC01-79-TE70078 with the U.S. Department of Energy.

The Rand Publications Series: The Report is the principal publication documenting and transmitting Rand's major research findings and final research results. The Rand Note reports other outputs of sponsored research for general distribution. Publications of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.
N-1204-DOE

June 1979

MODELS, COMPLEXITY, AND ERROR

David Leinweber

---

A Rand Note
prepared for the
U.S. DEPARTMENT OF ENERGY

---

Rand
SANTA MONICA, CA. 90406
This Note is derived from energy research and environmental modeling performed at The Rand Corporation in recent years as part of a program of studies for the Department of Energy. The perspectives of modeling should be useful from both a technical and an institutional viewpoint for analysts working in areas characterized by uncertain information.
SUMMARY

Sociopolitical or economic data used in modeling for policy analysis are subject to a wide range of distorting influences. This note discusses the effects of these distortions on the quality of results derived from such data and the appropriate degree of complexity for a model operating in a weak data environment. A probabilistic method of evaluating the quality of those results is described.
CONTENTS

PREFACE ........................................ iii
SUMMARY ....................................... v
FIGURES ....................................... ix

Section
  I. INTRODUCTION ................................ 1
  II. THE DATA ................................... 2
  III. THE MODEL .................................. 4
  IV. THE APPLICATION ............................. 7
  V. THE RESULTS ................................. 8
REFERENCES .................................... 16
FIGURES

1. Figures 1a, 1b, and 1c--Error Versus Complexity........... 11

2. Figures 2a and 2b--Error Versus Complexity,
   Poor Data Case.................................................. 12

3. Diminishing Information Returns................................ 13

4. Results of Study Comparing Coal Models Using
   Identical Inputs.................................................. 14

5. A Differential Impact Map....................................... 15
I. INTRODUCTION

Building and using a model entails a set of related activities. Input data must be collected, the model itself needs to be precisely specified and constructed, as well as a method of interpreting the results. In engineering and the "hard" sciences this process can be fairly straightforward, but when the domain of the model is sociopolitical or economic or both it is considerably more difficult.
II. THE DATA

The input information available to economic and policy modelers is not of a uniformly high quality. In 1950 Oskar Morgenstern discussed these problems in *On the Accuracy of Economic Observations* [1]. He cited multiple sources of error, including the lack of designed experiments, deliberate hiding of information and lying, improperly trained observers, poorly designed questionnaires, lack of proper information definitions, measurement errors, and the inability to allow for unique or changing circumstances.

Things are not much better today. There is an unreported portion of the GNP that may amount to more than $200 billion. This represents activity in the barter economy, off-the-books employment, drug traffic (which reportedly has surpassed tourism as the largest industry in Florida), and other unsavory endeavors. Since $200 billion is 10 percent of the total reported GNP, it seems fair to assume that the accuracy of much of the published information on inter-sector transfers and the like may be suffering as a consequence [2].

There are still ample reasons for deliberate data distortions. A notable example is DoE's multiple-tier oil pricing system, which allegedly turns lies into cash for a large group of newly affluent petroleum dealers [3].

Bureaucratic and institutional inertia often encourage short cuts in meeting federal reporting requirements. It is much easier to use or adjust last year's numbers than to bother with actually going out and measuring what is supposed to be measured. These phenomena were noted by Morgenstern. With ever increasing burdens on respondents
they become an increasingly significant problem [4].

The list of possible problems with economic information is endless. Let it suffice to say that economics is not physics and never will be.
III. THE MODEL

So we have all this distorted data. What does this imply about model design? The answer is found in Fig. 1, which shows a series of graphs of model error versus model complexity [5]. In these figures complexity can be taken as a simple count of the number of arithmetic operations the model performs on the input data. More refined definitions are possible (and necessary in some contexts), but are not required here.

Error, as applied to input data, refers here to the initial uncertainties of measurement arising due to sources ranging from the uncertainty principle to misreporting for tax purposes. These errors will combine and compound in a mathematically predictable manner as uncertain inputs are transformed into uncertain results. Using the formal relation between input errors and output errors [5] is the most straightforward interpretation for "error in the results," and is sufficient for this discussion. The error of model specification is not derived in this precise manner, but refers to the deviation of the model's results from what is known about the subject under study, hopefully from sources independent of the modeling process. The reference for comparison will almost certainly be subject to the same distorting effects that degrade input information, but the ultimate resolution of circularities in perception is beyond the scope of almost everything, this piece humbly included.

Fig. 1a shows one kind of model error--the error of specification. This is the error in the model's prediction if it were working with perfectly accurate ideal data for inputs. The only
legitimate reason for making a model more complex is to increase its accuracy. Complicating a model without an increase in accuracy is, generally, an unacceptable strategy. This doesn't mean it doesn't happen, only that in an ideal world it wouldn't. The downward curve of $e_s$, the error of model specification, reflects this idealized notion. Real models may be worse, but not better. The shape of the curve says that at some point further complications of the model yield only meager improvements in predictive ability. For some models (usually physical) the error of specification may effectively reach zero and the curve of $e_s$ terminates at that point.

The second curve, in Fig. 1b, shows $e_m$, the compounded error of measurement in a model. This is the error that arises from the unavoidable errors in real data that compound as more arithmetic operations are performed on them. As the models grow larger and more complex, the compounded error in the prediction increases.

Adding the two curves shown in Fig. 1a and 1b produces the curve of total error, $e_t$, shown in Fig. 1c. The important point is that this curve has a minimum. There is an optimal model complexity for attacking a problem based on real, i.e., imprecise, data. Further complicating the model buys nothing. It does impose an additional cost in accuracy of prediction, as well as the cost of resources required for a more complex analysis.

An important point is made by considering what happens when the quality of the data gets worse. Fig. 2a shows $e_{m2}$, the compounded error of measurement curve for a set of data worse in quality than was used in Fig. 1. The original data from Fig. 1 are shown on the same axes as $e_{m1}$. 
Adding both \( e \) and \( e \) to \( e \), the error of specification, produces two total error curves, \( e \), identical to that shown in Fig. 1c, and \( e \), the total error for the poor data case. Notice that the minimum point on the curve moves to the left. As the data get worse in quality the optimal level of model complexity decreases.

Put more succinctly: Weak data require simpler models.

When this notion is taken to the limit, with the quality of the data becoming increasingly poor, the optimal complexity for the model using that data (denoted by \( C \) in Fig. 2b) approaches zero. A model of zero complexity is an intuitive guess. When the data are nonexistent or worthless this is the best model available.

There is good evidence that many real models are more complex than they should be, given the quality of the data they run on. Much of the detail in models is based on presumed detail in the data which, on closer inspection, is not really there. An overly detailed and complex model may be well past the point of diminishing information returns with regard to the domain under investigation [6]. It may still say something about large models (see Fig. 3). The unchecked growth of models—which may become "systems" as their budgets head into seven figures—brings with it a range of institutional forces affecting and arising from the modeling activity. One such force is the pressure to apply the model in the wrong way to the wrong questions.
IV. THE APPLICATION

It should go without saying that a model should be appropriate to its application. For new models this should be fairly well assured; for existing models it is not so certain. There is a great temptation to apply models in areas beyond their intended range, to ask questions about issues to which the model is essentially blind. The situation is worst in the case of large, expensive, complex models that represent a substantial investment by the sponsors and may even become the sole source of income for their tenders [7]. This kind of institutional momentum leads to the development of a myriad of add-on modules, adjustment factors, complex feedback systems and other contrivances which add little or no value to the information derived from the model. They do help extend the bureaucratic visibility and organizational reach of their sponsors. This is the unacceptable situation referred to previously, in which the complexity of the model increases without a reduction in the error of specification. Incorporation of weak or marginally relevant relationships based on unsubstantiated data may have a detrimental effect on the results.
V. THE RESULTS

What happens after the data have been collected, and the model has been designed, implemented, and applied to the problem? The issue becomes one of deciding how to interpret the results.

The model's predictions are in the form of numbers which, hopefully, in some way describe what the future will be. Placing a lot of credence in the absolute level of the prediction, the value itself, can be a risky proposition, especially for times more than five or ten years into the future. The Energy Modeling Forum at Stanford University applied a group of coal models to a standardized problem and came up with a remarkably broad set of predictions [8]. The curves from the different models tended to move together, but the absolute values of the predictions varied widely (see Fig. 4). The implication is that it may be better to interpret the results of modeling analyses in a differential mode, i.e., to look at the changes in the predictions for one case relative to another. In policy analysis this is often the kind of question one is interested in anyway. If there are two or more alternative actions under consideration the basis for a differential comparison clearly already exists. For most problems a base case, or business-as-usual scenario, can be constructed to provide a reference point for the differential comparisons. The use of this technique is shown in Fig. 5, a map showing estimated 1985 sulfur oxide emissions under the original Carter energy plan relative to estimated emissions if the plan were not in effect [9].

Differential mode interpretation is probably the safest way of
looking at model results in use today, but there is still room for improvement. There are uncertainties and errors in the model's results, as well as in its inputs. They are somewhat more difficult to quantify since it is a fairly intractable mathematical procedure to trace changes in the distribution of the input variables as they churn through any (but the simplest) model and become output variables. If this could be done, one could then interpret the results in a differential probabilistic sense, and obtain measures of the statistical validity of one's results, along with the results themselves. Generating and convolving the distributions of predictions or approximations of these distributions allow the modeler to distinguish between a real statistically significant effect and one that is just random noise.

The map in Fig. 5 corresponds (partially) to the request: "Show me those regions which will experience an increase greater than 10 percent in sulfur oxide emissions." A probabilistic differential map would correspond to requests similar to: "Show those regions which will experience an increase greater than 10 percent in sulfur oxide emissions, with a likelihood of at least 80 percent." For a complex model it is essentially impossible to do this precisely, but it can be done in an approximate way. If reasonable assumptions are made concerning the accuracy of components of the model's final results or if actual data on the accuracy of some components are known, it is usually possible to envision a practical way to combine the known and assumed uncertainties to produce some kind of approximation to the distribution of predictions, and to calculate the distribution of the difference in predictions, yielding an approximate probabilistic
differential result. Some of this must be done on an ad hoc basis; it may be best to use a range of assumptions concerning the uncertainties in the models, ranging from very optimistic to very pessimistic, and see how this affects the confidence intervals for the results. If one must make heroic assumptions everywhere before anything with a likelihood beyond the 10-percent level appears, it is likely that the model is not yielding very much real information. If 70-percent likelihoods are seen in the results with moderate assumptions one may well be observing a real effect [10].

Important policy decisions should not be based on noise. While it is far from clear that such decisions actually are closely coupled to modeling, it is worthwhile to make some attempt to determine the validity of one's conclusions, whatever their ultimate application may be.
Figure 1a. Error or Specification

Figure 1b. Compounded Error of Measurement

Figure 1c. Total Error
Figure 2a. Compounded Errors of Measurement

Figure 2b. Total Errors
Figure 3. Diminishing Information Returns
Figure 4. Results of Study Comparing Coal Models Using Identical Inputs
1985 INITIATIVE SOX BY AQCR

Figure 5. A Differential Impact Map

PERCENT OF 1985 NOMINAL

- ABOVE 110
- 90 TO 110
- 0 TO 90
REFERENCES


