Advanced Forecasting Methods for Global Crisis Warning and Models of Intelligence

Article · January 1977

1 author:

Paul Werbos
National Science Foundation
164 PUBLICATIONS **11,831** CITATIONS

SEE PROFILE

TO: MIKE CASEY

① Page 30 begins to address more general issues.

② Page 32 on: see markups -- lots of ideas without designs ... only later did I translate the ideas to designs; no one else did, over many years --

GENERAL SYSTEMS
Volume XXII, 1977

ADVANCED FORECASTING METHODS FOR GLOBAL CRISIS WARNING AND MODELS OF INTELLIGENCE*

Paul J. Werbos

I. APPLICATIONS AS GOALS FOR COMPUTER FORECASTING

The Maryland Crisis Warning and Management project [1] has tried to develop and organize a broad set of tools for coping with crises, among which are computer forecasting methods. A complete crisis-management system must exploit a wide variety of forecasting methods, involving man-machine interaction, human intuition, etc. However, like the economists, we still have need of computer forecasting in the classical sense, in which the computer itself fits a quantitative model to a well-structured numerical databank.

(i) *to help us unify and make precise our knowledge of the dynamics of large-scale social trends.* In politics, as in economics, there are many key phenomena which result from continuous changes within a large population. To understand these phenomena, it is not enough to fall back on our intuition about the behavior of individual people; we need to use methods which can exploit the available knowledge about hundreds of political societies in the past. Current trends in Mexico present a clear example of the relevance of this approach to national security: if we wait until Mexico's population/economic problems grow into a political threat to the US, before paying sufficient attention to them, it may be too late for us; also, we need to have a feeling for how the trends work in order to act constructively, to get at the roots of the conflicts instead of increasing contradictions.

(ii) *to stimulate a higher level of relevance in human political analysis.* High-level decision makers need to have assessments of the probabilities of what will happen in the future, if they exercise a given set of policy options. In other words, they need the best available answers to very difficult questions. But, in government and in academia, there are incentives for people to focus on easier questions, on questions which "can be answered." Thus there is a tendency for political analysts to compete with the newspapers in providing passive, factual background information, which can become quickly obsolete; a decision-maker then may prefer to read the *New York Times* instead of an official intelligence report.

Computer forecasting methods can help overcome these negative incentives. If computer forecasts must be passed up to decision-makers on a regular basis, human political analysts can be encouraged to comment on these forecasts. In effect, the analyst can "blame the computer" if the computer offers a frightening, "alarmist" prediction of conflict. When the analyst *evaluates* the computer's prediction, and points to factors which the computer cannot account for, he is *applying* his human knowledge to the problem of *prediction*; he is allowing himself to become more relevant. (This reminds us of traditional Mandarin China, where there was a high level of cultural creativity despite a stifling belief that no modern scholar could improve on the Great Classics; creativity came from "commentaries" and "explanations" of the Classics which went far beyond what was really in the Classics.)

(iii) *to alert us to the unexpected.* Computer forecasting probably does a better job of predicting trends than of predicting anomalous events such as crises. How can it be used then to help us with the short-term warning problem? If it predicts the normal routine flow of events from day to day, how can it help predict a crisis?

Computer models can alert us to any "improbable" *discrepancy* between their predictions and the current flow of events. After all, we cannot say that the flow of events is "out of the ordinary" until we have a good idea of what "ordinary" means. The computer can give us this baseline. Once the alert is given, a different set of analyses can be called into play. Among these may be computer models fitted to daily event data from previous crisis or anomalous situations. Even if human analysts are skeptical about the alert, it would be wise to pass on the information as part of routine daily reports to policy makers. To make all this work, however, one will need to use regular high-frequency data, such as daily satellite data or FBIS condensations. Also, there is a serious problem of security: we consider it unwise even to discuss certain possibilities for indicators, when we know that they could be "jammed" by an aggressor who knows about them. We doubt that this problem would be reduced very much even if a portion of this work were conducted under the usual terms of industrial top secret. For now, the goal is to develop the *methods* themselves, not the models which would be used for alerts.

Further applications may also exist in the areas of artificial intelligence and learning theory; some of these possibilities will become apparent in Section III.

II. A REVIEW OF PRIOR CONCEPTS AND EMPIRICAL RESULTS

II-a. The Classical "Econometric" Approach to Political Forecasting

Despite twenty years or so of quantitative work in international relations, the three "needs" above have not been fully satisfied. Why not? Until recently, it seemed as if we could still point to a simple *lack of substantive knowledge.* It took many years to build up adequate data sources and to pinpoint key variables and interactions.

*The research reported in this paper was supported by the Defense Advanced Research Projects Agency of the Department of Defense and was monitored by ONR under Contract No. N00014-75-C-0846. The views and conclusions herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the DARP Agency of the U.S. Government.

TO: MIKE CASEY

703-413-2220

Very little work had been done in putting the various studies together, to build up unified, dynamic models. Nevertheless, with a bit more work, we could hope to imitate the success of econometrics. Above all, we could hope to fit models to data by trying to minimize least square error. In other words, we could make use of existing computer software which performs multiple regression or nonlinear regression. These methods, at least, seemed highly reliable; the two dominant schools of thought in statistics–"Bayesians" and "maximum likelihood" people–have agreed on this.

In particular, we expected to formulate models which predict the *future* value of each political variable, Y_i as some function, f_i of the *present* (or past) values of other political variables and of external (policy) variables, M_j. If we use "$\hat{Y}_i$" to denote "the predicted value of Y_i," as in engineering, we could write such a model as

$$\hat{Y}_i(t+1) = f_i(Y_1(t), Y_2(t) \ldots Y_n(t), M_1(t) \ldots M_m(t) \quad \text{(for } i = 1 \text{ to } m)$$

Here t means the current time, and t + 1 means the next period of time for which data is available. We use human judgement to guess what the formulas f_i should be. More precisely, we pick a whole set of plausible models where each "model" is a guess for what all these functions f_i should be for good predictions. In each model, the functions f involve unknown parameters which we don't feel we can guess a priori. The computer will treat each function f_i in isolation from the other functions; it will estimate the parameters of f_i by minimizing square error:

$$\sum_t e_i^2(t+1) = \sum_t (Y_i(t+1) - \hat{Y}_i(t+1))^2$$

$$= \sum_t (Y_i(t+1) - f_i(Y_1(t) \ldots M_m(t)))^2,$$

where, in the last expression, we use the *measured, available* databanks to get the values for $Y_i(t+1)$, $Y_1(t)$, etc. Finally, in order to choose the best guess for the function f_i, we simply pick the function which leads to the least square error in predicting $Y_i(t+1)$. (However, if there is another guess for f_i which is almost as successful but much simpler, we prefer it, by Occam's Razor; this is usually interpreted as "deleting terms which are statistically insignificant.")

After this estimation work is done, we could hope to go on to use the model for long-range forecasting. If there are no M_j terms, we could start out with the available data for $Y_1 \ldots Y_n$ at the present time t; we can use this data in calculating the values of $f_i(Y_1(t) \ldots Y_n(t))$ to generate predictions for $Y_i(t+1)$, i.e., for $Y_1(t+1)$ through $Y_n(t+1)$; we can calculate $f(Y(t+1) \ldots Y_n(t+1))$, using our previous predictions as if they were actual data, to predict $Y_i(t+2)$; we can predict $Y_i(t+3)$ similarly; etc. With M_j terms present, we can make long-range forecasts *conditional* upon given future policies.

This kind of forecasting would be justified, according to Bayesian statistics, because we have chosen the model which has the highest probability of being true *in light of* the existing data. In other words, we have maximized:

$$\Pr(\text{model and parameters} \mid \text{databank}) = \frac{\Pr(\text{databank} \mid \text{model and parameters}) \Pr(\text{model and parameters})}{\Pr(\text{databank})}$$

Minimizing square error is known to be the same as maximizing this conditional probability, *so long as* we have a "flat prior" (we assume that the a priori probability Pr(model) is essentially the same for all models), and so long as we *interpret* our original model to mean that

$$Y_i(t+1) = \hat{Y}_i(t+1) + e_i(t+1) = e_i(t+1) + f_i(Y_1(t) \ldots M_m(t)),$$

where e_i is a random "error" disturbance which follows a normal distribution. These two assumptions are generally interpreted as reasonable simplifying assumptions; even though the assumptions are not perfectly true, they, like the equations of the model themselves, may be close enough to be reasonable.

The best long-range predictions would always be given by the "true" model. Given that we have picked the model with the highest probability of truth, we expect that this model will also give the best long-range predictions. *Any political model which calculates predictions from numerical data* would be subject to fine-tuning and evaluation by this approach.

II-b. The Failure of Regression and of Advanced Classical Methods

Initially, we hoped that the "econometric" approach would work. However, in a number of regressions run on a variety of models or sub-models, we have confirmed the suspicion that there are some serious difficulties [2, 3]. First, we used the "econometric" method to estimate the rates of population growth and national assimilation in Karl Deutsch's model of nationalism and social communications [4]. The results seemed very encouraging; the R^2 scores indicated 99-99.9% accuracy in prediction. However, when these parameter estimates were used in long-range prediction, the results were quite bad. "Long-range" prediction error was defined as the mean square error in predicting from an initial time period through to about 30-40 years in the future; we used more than twenty sample national time-series, each treated as a separate case. (For the technical details and graphs of the results, see [2]; the dataset was collected by Karl Deutsch, Sheldon Kravitz, Raymond Hopkins, et al.) Median error across different cases was only a bit larger than 10%, but in many cases rose to as high as 20%, and in a number of cases was absurd. Even if we were interested in short-term dynamics, the parameter estimates were absurd, in terms of Deutsch's model. Often there were negative rates of national assimilation (including whites apparently turning into blacks in the US); also the rates were far too large in absolute size, *especially* in cases where the supposed "statistical significance" was good. Certainly there was no "99.9%" accuracy in prediction!

At first we hoped we could interpret this failure in terms of the usual Bayesian or maximum likelihood philosophies. In particular, we noted that the correlation of each variable with prior values of itself fell off very slowly as we considered longer time intervals between the present and past values. From our past work, we recognized this as a sign of *measurement* error. The conventional reasoning, (cf. II-a) assumes that the *measured* values of the data are identical with the *true* values; it

assumes that model "errors" are the result of random disturbances which affect the *true* values.

However, measurement error produces a different situation, and in order to account for it we have to acknowledge that the true values of the variables, Y_i, are different from the *measured* values, Z_i. Two kinds of random disturbance must be recognized: (i) "*process noise,*" which affects the true values; (ii) "*measurement noise,*" which makes the measured date, Z_i, differ from the true value. With the regression approach, we could write our model as

$$Z_i(t+1) = Y_i(t+1) = \hat{Y}_i(t+1) + e_i(t+1),$$

where e_i refers to process noise. To account for the effect of measurement noise, we have a more complex model:

$$Y_i(t+1) = \hat{Y}_i(t+1) + e_i(t+1)$$
$$Z_i(t+1) = Y_i(t+1) + a_i(t+1),$$

where a_i refers to measurement noise. As part of our reported work, we found new methods–perfectly "efficient" methods, in statistical and numerical senses–for estimating the parameters of such a model; these methods are related to methods discussed by the statisticians Box and Jenkins [5], and by the engineer Kashyap ("Vector ARMAX processes"). We also set this up to allow for the possibility of cross-correlations between e_i and e_j or a_i and a_j, where $i = j$.

It was hoped that this strategy would validate the basic philosophy of maximum likelihood statistics. With a better (but still simplified) model, we hoped to get better predictions. Measurement error is not purely "random," as we assumed, but we hoped that allowing for *some* measurement error would make a big difference in a situation where measurement error seemed to be the major source of difficulty in forecasting. Perhaps we could urge political scientists to use this kind of model, instead of the regression model, in political analysis.

However, this approach also failed. It led to a reduction in long-range prediction errors by 10% or less of the original error, from regression; the errors were large in the same countries; the slight improvement appeared to be a random result due to the addition of more parameters in the model. Furthermore, when the comparison against regression was also tried in a study of more sophisticated models of nationalism, tested against a high-grade dataset with more than 1000 observations across approximately 30 years (from the various provinces of Norway), again, the new methods did little to improve long-range predictions.

Early in 1977 we extended our analysis to consider the long-range "econometric" models developed by CACI, Inc., for the Joint Long-Range Strategic Survey. Here, as with the Deutsch model, we found the parameter estimates to be highly unreliable [3], despite the quality of the data and the substantive complexity of the updated world model. In our previous work, we also noted similar difficulties faced by economists; standard econometric forecasting is not an unqualified success, even when judicious fudging is artfully used.

II-c. The Success of a Robust Method

A successful method was found, almost by accident, in this model:

$$Y_i(t+1) = \hat{Y}_i(t+1)$$
$$Z_i(t+1) = Y_i(t+1) + a_i(t+1)$$

It is identical to our more complex model, except that the possibility of "process noise" has been removed; i.e., it assumes that the real world is governed by deterministic laws; all appearances of prediction error are due to incorrect measurements of the data. We called this the "measurement-noise-only" model. This method led to median long-range prediction errors of 4% in predicting the Deutsch variables–less than half that of the other methods. In predicting the *percentage* of population assimilated to the dominant nationality, it was off by 2% or more in only four of the twenty-odd cases, in long range prediction.

To check our conclusions about the new method, we set up twelve different sample "processes" to be studied by three statistical methods: regression, complex classical, and measurement-noise-only; for each process, we simulated ten different sample time series of length 100 and evaluated each of our statistical methods in two ways: (i) if we look at the *average estimate* across all ten examples, how close is it to the true value of the parameter (bias)?; (ii) how close are the individual estimates, in each example, to the average estimate to which they would converge if more data were available (statistical efficiency)? These processes generally involved random process noise, measurement noise, and occasional outliers. The measurement-noise-only method was distinctly superior (less bias and more efficiency) for all processes but two; in these, the three methods were approximately equal. (Again, see [2] for details. "Distinctly superior" meant that errors in parameter estimates were roughly half as much, or less.)

To a well-indoctrinated Bayesian, these results would seem extremely strange. Our simple measurement-noise-only model is just a *special case* of the complex model discussed above in II-b. It cannot be "true" unless the complex model is also "true." When we let the computer pick *any* form of the complex model, it is certain to come up with something which has a higher probability of truth than it would if it has to limit itself to a special case. How, then, can a model with *lower* probability of truth consistently lead to *better* long-range predictions?

Our explanation is that this is a case of *robust estimation*–a relatively new philosophy which Mosteller of Harvard and Tukey of Princeton have promoted to suggest that "least square error" is an inadequate concept. Their specific suggestions are totally different from what we are considering in this paper; our "measurement-noise-only" method can be carried out *either* in terms of least-squares error *or* in terms of the Tukey jack-knife, with equal ease. Their basic philosophy, however, is essential to understanding our results. The fundamental assumption in robust estimation is this: We really *don't* expect any of our

mathematical models to be "true" in an absolute sense. At best, we hope they may be useful *representatives* of a whole set of very complex models, one of which is true but far too complex to handle directly. The true expected value of a future quantity is *not* the same as the expected value given by the *most probable model.* The true expected value is computed by averaging the expected values given by *all possible models,* weighted by the models' probabilities of truth.

In the maximum likelihood approach, we concentrate on the goal of *absolute statistical efficiency*; we make total use of all the data, at the price of assuming that the model is perfectly "true" in some form. An alternative goal is that of *absolute consistency*: we can require that our estimates will converge to *exactly* what we want them to be, as the quantity of data goes to infinity. In the case of long-range forecasting, what we want are the parameter estimates which minimize long-range prediction error. In order to be certain that our estimates will converge to these values, with infinite data, we may simply *minimize directly what we want to minimize*—long-range prediction error itself.

Looking again at the "measurement-noise-only" model, we can see that fitting this model is really the same as minimizing long-range errors directly:

$$Y_i(t+1) = \hat{Y}(t+1)$$
$$Z_i(t+1) = Y_i(t+1) + a_i(t+1)$$

In order to "fit" this model, we must somehow estimate the true values Y, since we only have data for the measured values Z. We *estimate* the true values of $Y_i(0)$ (i.e., for the different variables Y_i at the initial time). Then we calculate the later values simply by calculating $\hat{Y}_i(t+1 = f_i(Y_i(t) \ldots)$ over and over again, for values of t from 0 to the end of the data. This must be done to make sure that the upper equation is satisfied *exactly*, as this method demands. But this is *exactly* what we do in making long-range predictions. The $Y_i(t)$ are essentially *long-range* predictions, projected forwards from the data at the initial time 0.

In minimizing the sum of a_i^2, we are minimizing the difference between the *actual measured* values Z_i and the *long-range* predictions; we are minimizing *directly* the long-range prediction errors. With regression, however, we were minimizing errors in predicting time t + 1 from data at time t; in other words, minimizing prediction errors over the *shortest possible* period of time. It should be no surprise, then, that the "measurement-noise-only" method leads to better long-range forecasts. Also, if key "feedback" terms are estimated badly, or other parameters are grossly misestimated, we would expect *very large* cumulative errors in long-range prediction; when we minimize the long-range prediction errors themselves, we may expect fewer random estimation errors of this type.

The robust method was proposed in 1973 and reported in 1974. Recent work on "smoothing," in engineering, has echoed similar mathematics. Hartley, in economics, is said to have proposed a similar method, but with features that make it impractical to estimate. Our own report discusses new numerical procedures which make it feasible to estimate these models even in cases of enormous complexity and nonlinearity.

The forecasting problems cited above have existed for decades. Therefore, people doing practical studies have invented dozens of ad hoc fixes for trying to reduce the problems. Space prevents our discussing here all the complexities of these many methods. By adhering to the "robust" strategy of minimizing *directly* the long-range predictions, from the beginning to the end of our dataset, we may be sure of two key things: (i) the procedure is general and can be applied to *any* predictive model, not just to special cases, such as linear models: (ii) we know that we are directly minimizing the errors we want to minimize, instead of something else which has a vague or muddled relation to these errors.

One ad hoc alternative may come to mind for those who have relied heavily on regression: "If you want to predict 30 years in the future, why not simply set up a model with 30-year time lags in it? What you are doing is really minimizing the *average* prediction errors for prediction intervals from the minimum time interval in the data up to the maximum; you are trying to predict *all* the future history of a system from data at the first time interval, or at least from an estimate at the first time interval. But to minimize least squares error for a 30-year prediction, directly, you would use regression with a 30-year time lag."

Problems in crisis management present a dramatic example of what is wrong with this approach. Suppose that we want to have 30-day advance warnings of likely crises. As noted in Section I, the real world crisis dynamic is likely to require a knowledge of day-to-day changes in events in order to achieve such warnings. Thus we are saying that the crisis *evolves* dynamically through the 30-day period. McClelland's work with WEIS indicators strongly supports this conclusion [6]. If we use one month lags, there is no way to tap these day-to-day processes. We would simply bypass any changes (or predicted changes) which occur between time t_0 and t_{30}. If we do not get at those dynamics, we can only expect to provide reasonable one month advance warnings under special circumstances. For example, there might be systematic, real world, one month time lag between the precipitating circumstances and the resultant event because of built-in bureaucratic delays.

The mathematical statistician would consider this an example of obvious, general limitations of the ad hoc approach: a low level of statistical efficiency and a model-specification problem. The robust method, however, does try to account for the day-to-day fluctuation of events; it tries to predict *tomorrow* as a function of *today*, but to do it in such a way that our model is good for 30-day forecasting. (For the present, it is more realistic to talk about yearly data and 30-year forecasts, but the same principles apply.) Also, because the usual random errors in regression estimates lead to large cumulative errors in long-range prediction, the robust method may even be more reliable in estimating the parameters appropriate for *short-term* prediction. Our simulation studies support this expectation.

The Compromise Method: A Generalization of the Robust Approach

The pure robust method, in its original form, is still not the right tool to use in crisis management.

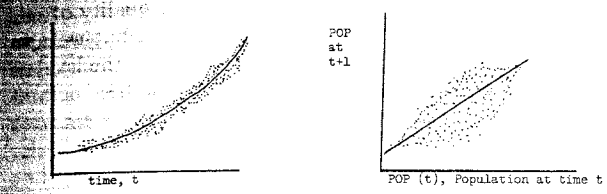

Fig. 1

In order to get a better feeling for its strengths and weaknesses, consider the example shown in Figure 1 on the left. What we are really doing with the pure robust method is fitting the *curve* as close as possible to the dots. (In II-c, the dots were called Z, the "measured values"; the curve represents a set of values for the Y.) With regression, one also tries to fit a curve to dots, as in the graph on the right. However, with regression, the dots represent only the relations between pairs of measured data, at time t versus time t + 1; knowledge about longer-range regularities in the data has simply been thrown away when we plot these dots. But with the robust method, on the left, we are plotting against *time,* and we retain the whole time history. This is not the same as simple "trend analysis," where we might try to regress population against time; we are trying to pick the best possible curve from the set of curves which represent *possible* histories of the true values of the variables, assuming that our model is *exactly correct* for the true values. We are trying to pick the "solution trajectory" as close as possible to the actual, measured history (dots) of the process. We pick parameters for our model which make the solution trajectory as close as possible to the measured history. (For certain simple models, it is feasible to find these trajectories by doing a complex nonlinear regression against time; however, that approach is unnecessarily difficult, confusing and limited to special cases.)

In the example of Figure 1, the curve (left) and the dots stay reasonably close together. This was also possible in the more complex examples we have studied empirically. *So long as this remains true,* the goal of minimizing square error (the distance between dots and the curve, in the vertical direction) will involve a real consideration of the dots as individuals; the ordinary fluctuations between t and t + 1 are significant in size, compared with the average distance between the dots and the curve, so that they have not been "drowned out." In effect, *all* the data is being accounted for; we should not be surprised that our empirical tests have shown a high level of "statistical efficiency" in this kind of situation.

There is another way of looking at this: the curve in Figure 1 has really "captured" the big shifts, over time, in the process being studied; the remaining errors in "long-range prediction," between the dots and the curve, are small enough to be compared with the small fluctuations and error between one time period and the next. If our curve and model describe the past history of the process so well, it is reasonable for us to project this curve ahead into the future.

On the other hand, consider Figure 2. What if *none* of the possible solution trajectories can get close to the historical data? A flippant answer would be, "If you can't even explain the past, how can you expect to predict the future? In a case like this, you know that your model is grossly inadequate. There is no way you can make good forecasts with a bad model, no matter *how good* your estimation technique." Still, in Figure 2, we can see that a simple exponential model for population growth makes sense *most of the time*; the model breaks down only in the middle, where an external factor (World War II) produces unexpected changes. In such cases, a flippant answer is not good enough. A *perfect* model of population growth should predict such things as World War II; however, in the real world, as we try to move from ignorance to *better* models on our *way* to far-distant perfection, we need to have techniques which work well on imperfect models. If we use the pure robust method on the example in Figure 2 and try to fit an exponential growth model, we would wind up with the curve C_1, which really does not represent the normal rate of population increase.

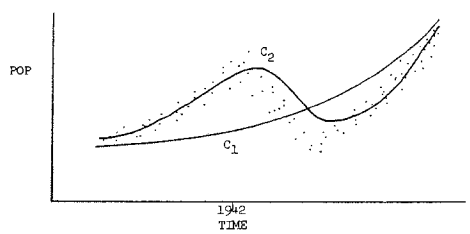

Fig. 2

Figure 2 illustrates a basic paradox which our research has sought to resolve. It is an example of a situation where "process noise"—real random factors in the real world—are too important to be ignored. The impact of this noise does not die out with time; it does not "average out" enough to let us use the pure robust method. This paradox is vital to long-range strategic planning [7], where we need models which meet two tests at once: (i) they *must* allow for real-world uncertainties and assess the probabilities involved; (ii) they must "hold up" over time so that they will be valid for both long-range forecasting and planning. We have noted that classical methods do *not* hold up well enough over time. On the other hand, the pure "robust" method does not account for real uncertainties; in situations such as in Figure 2, it loses its statistical efficiency, because the gap between the curve and the dots is very large and depends only on gross characteristics of past history. When both methods are inadequate, what do we do?

In 1974 we suggested a "compromise method" to generate forecasts in this kind of situation. The "compromise method" is based on the concept of filtering. In ordinary regression, we tried to generate "good" predictions of the measured values $Z_i(t \pm 1)$ by minimizing:

$$e_i(t+1) = Z_i(t+1) - \hat{Z}_i(t+1) = Z_i(t+1) - f_i(Z_1(t) \ldots M_m(t)).$$

In other words, we plugged in the *measured* values at time t in order to generate predictions. On the other hand, with the pure robust method we minimized:

$$e_i(t+1) = Z_i(t+1) - \hat{Z}_i(t+1) = Z_i(t+1) - f_i(Y_1(t)\ldots M_m(t)).$$

In other words, we plugged in the Y_i, *long-range predictions* of the values at time t, in order to generate new predictions. These two methods can be considered as two ways of estimating what the *true* values ($X_i(t)$) were at time t. The *measured* value is one clue to the true value, but we can also get a clue from what we would predict at time t from our estimates of the true values at *earlier times*. Instead of choosing one clue or the other, we can achieve a synthesis by accounting for *both* sources of information:

$$X_i(t) = (1-r)\hat{Z}_i(t) + rZ_i(t).$$

In other words, we can estimate the true value at time t by taking a weighted sum of $\hat{Z}$ (what we predict from previous information) and of Z (what we measure at time t). We can expect that the X_i will be a better estimate of the true values at time t than *either* Y or Z, because they account for more information. We may go on to estimate the parameters in our model by trying to minimize the square of errors defined as

$$e_i(t+1) = Z_i(t+1) - \hat{Z}_i(t+1) = Z_i(t+1) - F_i(X_1(t)\ldots M_m(t)).$$

(Note that we can estimate probabilities by assuming that the e_i represent random normal noise.) If the constant r is very small, this will be close to the pure robust method. However, the curves we fit to the data are the curves of X_i; like the curve C_2 in Figure 2, these curves will *move back* towards the measured data Z whenever the predicted values start to be far away from the actual values for a moderately long period of time. The constant r may be thought of as a "relaxation constant," which represents how far we are willing to "relax" the curve.

Another way of interpreting the "compromise method" is that we are trying to figure out what to do when we are forced to abandon our goal of deterministic long-range forecasting. In the pure robust method, we were summing up errors for prediction errors across all time-intervals t; *each* time interval contributed equally. Here we are effectively *discounting* the importance of prediction error over longer intervals of time; we are applying a kind of interest rate, r, to reduce the emphasis on more distant times, because the errors over longer time intervals have been too large and erratic to cope with. (Note that r may be different for different variables.)

But how do we decide what value of "r" to pick? This is a basic problem considered in our research. There are dozens of ad hoc methods one might think of. For example, our filtering equation above is exactly like the Kalman filtering equation [8] for the case of one variable. We cannot use the Kalman equation to tell us what the filtering constant should be until we *already* have estimated a white-noise model of the process under study; however, we might try to fit such a model, then filter, then refit, etc. Unfortunately, that strategy takes us directly back to the white-noise maximum-likelihood approach, discussed in II-b; we already know that that strategy "relaxes" too much and fails to be robust. Thus, to pick "r" correctly, is a difficult problem. Somehow, we want to make sure that we reduce the impact of longer time intervals to the point where shorter time intervals are not totally drowned out; also, in improving the quality of our model, we want to reduce prediction errors, but we also want to be able to live with lower values for r; the level of foresight is a measure of the success of the model.

III. NEW METHODS

III-a. A New Context for Estimation

At the start, we knew that we had only one method in hand—the "compromise method" or "filtering method"—capable of meeting the basic demands of crisis management. Yet, even this method was not rigorously complete, because there was no explicit procedure for choosing the filtering constant, r. Moreover, from a previous work in artificial intelligence, we strongly suspected that one could do better than the "compromise method" itself in any form. Even though these forecasting problems are practical, empricial problems, we felt that we should avoid approaching them with a naive "fishing expedition"; therefore, we began the project with a thorough review of the theoretical possibilities and vicissitudes of robust estimation and we considered a wide number of questions:

1. Can we come up with a procedure for picking r which makes general intuitive sense?
2. How can we extend the notion of "robustness" here to involve *utility* in decision-making instead of just accuracy in forecasts?
3. How can we be sure these procedures give us good probabilities instead of just forecasts?
4. How do we cope with the interdependences among the "error" terms in a complex nonlinear situation?
5. How should our computers deal with the old problem of "symmetry": whether to treat different nations as different processes, or as different examples of the same general process, or something in-between? (This is related to the old problems of "object identity" and "object permanence" in psychology.)
6. How can we imagine the human brain copes with these problems? (Recall that the "minimum time unit" with the human brain is a tenth of a second, at most, and yet the human brain can easily and naturally think hours into the future without being deterministic. If it weren't for this example, we might have given up this research, at certain times, as an impossible task.)
7. Can our forecasting abilities be improved if we try to mimic the human faculty of "syncretism," of prediction by *analogy* to individual past experiences of a similar nature?
8. How can we maximize the reliability of the numerical convergence methods we use, in estimating the parameters?
9. What is necessary to translate all this into practical computer software, either for general use or to carry out further empirical special studies?
10. Can we use the concept of "entropy" (i.e., "information content") to help us "drain" from the available databank all the information of interest to us?
11. Does it help to conceive of forecasting as the problem of trying to predict an entire future history (or calculate probabilities for possible future histories) conditional upon a given past history?

Some of these questions clearly cannot be answered if we think of "forecasting" as a problem in isolation, independ-

ent of how we plan to *use* the forecasts and independent of where we get our data. For example, we cannot talk about "utility" without expanding our problem definition. The human brain provides us with an example to show that these forecasting problems basically can be solved. It does not matter whether the forecasts are generated by algebraic "models," by (neuron) network circuits which implement the same mathematics, or even by networks of human beings in a political institution; the problem of forecasting requires a reliable way to set up correct relations between our past input (independent variables) and predictions, which are output somewhere in the system. Nevertheless, the human brain performs forecasting *in the context* of pursuing external goals and recognizing patterns. We have no assurance that forecasting can be done this well in isolation.

In trying to organize the different strands of thought concerned with robust forecasting, it became clear that we would have to establish a more general context before going further. Our goal in "forecasting" is really just "estimation": to estimate the parameters of models and to choose models in such a way that they are useful in decision making. "Estimation" may be thought of as *one element* or subsystem of the more general problem of *intelligent decision making.* A decision-making system needs three interrelated subsystems:

1. a *pattern recognition* system, to select the variables which will be available to the models;
2. an *estimation system,* to make models of the world and forecasts;
3. an *optimization* subsystem, which *uses* these models to help it calculate the best choice of actions to maximize some utility function provided from elsewhere. (Often, we subdivide this into two smaller subsystems, one to choose a measure of strategic utility and one to pick actions to maximize that.)

In the beginning of this research, we developed two new techniques to help define the context for forecasting: "pattern analysis," to perform pattern recognition; and "dual heuristic programming," to help perform optimization in a complex, nonlinear stochastic environment. Also, we documented "heuristic dynamic programming," a related optimization method.

By "pattern recognition," we do not refer to the myriad things this word sometimes means in artificial intelligence. Our goal is simply to *provide variables* for use in the "forecasting" or "estimation" system, which in turn will be useful to the optimization system. In ordinary regression analysis, for example, one often finds that the raw data are unsuitable as inputs to one's model. Usually, there are simply too many variables available and they seem to be somewhat redundant. A separate body of methods—"factor analysis" or "principal components analysis"—is used to *reduce* the number of variables, *prior* to forecasting proper. We interpret this as a way of accounting for the *interdependence* of the different variables for which we have data; in other words, it is a way of answering question 4 above. "Pattern analysis" is a more general way of accounting for such interdependence, *within the context* of ordinary econometric forecasting approaches. It allows us to *merge* the pattern recognition and estimation tasks into a single analytic procedure. In particular, it allows for the possibility of *nonlinear* relations between the raw data and the processed variables. It may sometimes *increase* the number of variables, in the nonlinear case, but decrease the apparent *information content,* by singling out variables which equal zero ("pattern not present" or "pattern not changed") about 90% of the time.

The mathematics of pattern analysis is summarized in Appendix A. Applications of the two optimization methods in strategic planning are discussed in [7]. Appendix B reviews only the mathematics.

III-b. Three Strategies for Robustness in the New Context

This new context provided a dramatic change in our approach to robust estimation. Estimation, like pattern recognition, is now subordinated to pattern analysis, a single overarching technique. But pattern analysis in its original form is strictly an econometric-style, maximum likelihood method. It pays very close attention to probabilities and entropy scores. This makes it easy to get good probabilities out of pattern analysis, but it takes us back to the old problems of robustness: how can it be brought back, in the context of pattern analysis? Most approaches to improving forecasts here have turned out to be either invalid or secondary to more powerful approaches. Although we could hope that a better choice of higher-level variables, as in pattern analysis, would itself increase "robustness," we have singled out only three strategies which are appropriate to achieving robustness in the general context. From a formal point of view, these are not really alternative, but complementary, strategies; we would expect that a complex, well-rounded forecasting system (like that in the human brain) would use all three together.

1. *Bias: weighing the importance of different target variables in prediction according to their relevance and their actual variance instead of the variance in prediction error.* Maximum likelihood tells us to weigh them according to the variance of current prediction errors. However, the reasoning above (II-c) pushes us towards minimizing errors as weighted by *our interest* in the variables or by their *dynamic importance,* if we want to achieve robustness.

Within pattern analysis, the *filtered* version of a measured variable is itself *another* variable in our system. With "bias," we may shift attention to the *filtered* variable instead of the original variable. This produces an effect as if trying to minimize the sum of the squares of the original prediction errors (as in the robust "compromise method") *multiplied* by the square of the filtering constant, r. Strictly speaking, we may multiply by $r^2 + 1/T^2$, where T is the length of the average time-series in our data; this keeps us away from the rather anomalous minimum at $r = 0$, except in cases where the pure robust method is strongly favored. Thus we deduce one possible strategy for picking r in our original compromise method: pick r, and *all other* model parameters, to minimize this product. *This will be the standard version of the "robust method" for our next round of empirical tests.* A conservative strategy along the same lines is to multiply error by r itself, not r^2,

as a kind of 50-50 compromise between an r^2 multiplication and no multiplication. Note that these procedures would lead to the same estimates for the other parameters in our model as we would have had before, *for a given value of r*; the novelty is that we now know how to pick r.

Minimizing $(r^2 + 1/T^2)$-times-long-range-prediction-error also has a more intuitive argument in its favor. If we picked r so as to minimize prediction error itself, we would in effect be penalizing more ambitious models. For example, if model A predicts *tomorrow* with an error rate of 10%, while model B predicts the *next century* with an error rate of 12%, model B would probably be a better model on all counts; it is unfair to compare 10% against 12% unless we can find a way to *weight* these numbers to indicate how much more *difficult* is the task attempted by model B. (Recall that a different choice of r gives us a different effective definition of what long-range-prediction-error to pick. A standard definition, a priori, is equivalent to picking r a priori, without regard to the properties of the system being studied.)

How then can we find a fair way to weight these numbers? How can we measure the "ambitiousness" of a model? The constant r is supposed to measure the rate of decay of the value of past information; if the value of past information decays by a factor of $(1 - r)$ per time period, then, over all future times, the sum of the value of past information should be proportional to $1 + (1 - r) + (1 - r)^2 + \ldots = 1/r$, a measure of how much our model is trying to do. (Again, it is assumed that the time-series is long enough for the sum to converge normally; if not, with $r = 0$, the sum comes out to T, and it can be seen why we add a factor of $1/T$. Fancier procedures are possible for the $r = 0$ limit, but they require a different kind of analysis to fine-tune them and are probably not worth the effort.) We can evaluate the significance of the standard deviation of the long-range prediction error by asking how large it is as a fraction of "potential error," assumed to equal $1/r$ times some constant. This leads us to r-times-error-variance as a *weighted* measure of relative error variance. Again, this "derivation" is purely intuitive, but it helps to assure that the more formal "bias" concepts with pattern analysis make sense.

Put in the terminology of II-d, we will pick r and all the parameters of our model by minimizing:

$$(r^2 + 1/T^2) \sum_t e^2(t) = (r^2 + 1/T^2) \sum_t (Z(t) - \hat{Z}(t))^2,$$

in the univariate case. In the multivariate case, we will initially try to use the same r for all variables and minimize:

$$(r^2 + 1/T^2) \sum_t \frac{1}{\sigma_i^2} \sum_t (Z_i(t) - \hat{Z}_i(t))^2.$$

This differs from the standard "multinormal" approach, but it better reflects the "bias" approach suggested above.

A more general measure of the relevance of a target variable in pattern analysis is the mean square of the derivative of error with respect to that variable, plus the mean square of the derivative (lambda) of "strategic utility" with respect to that variable (see Appendix B and [7]), plus the usual reciprocal of the error variance if the variable represents raw data. This has the right dimensional properties for a measure to multiply square error; dimensional analysis indicates that there are few alternatives.

2. *"Multiple filtering": the use of two or more filtered versions of the same variable, to sort out long-term versus short-term fluctuations.* Within the context of pattern analysis, or the human brain, it is not natural to think of our simple filtering procedure (the "compromise method") as a built-in special system. Rather, it is natural to think of the *filtered version* of a variable as a new, abstract variable, whose value is *calculated* as a weighted sum of the present value of the raw data and of some function of the past values of the filtered variable and others. "Filtering" is just one application of our ability to set up *recursive models*, in which internal variables may be affected by their own past values. Filtering constants may be treated like any other parameters in our model itself; we may pick them to minimize some measure of global error, *so long as* we pick a global measure which leads us in the direction of "robustness." (If we did not pick a "biased" measure of global error, then our "compromise method" would lead us back to the maximum likelihood white-noise model, which failed our empirical tests. The calculations to show this explicitly are straightforward but tedious.) If we are allowed to pick *any* recursive model, we can just as easily have two filtered versions of any variable, or three, or more. Even without any bias factor in our global error measure, we can hope that this procedure will lead to greater and greater effective foresight; with luck, our time horizon may grow exponentially with the number of filters. We are not talking about classical filters as in electronics, which are designed to respond to predetermined frequencies; all the filtering constants are to be *estimated* by econometric-style procedures. Also note that our earlier work [2] already shows how to compute the derivatives needed in estimating the parameters of a recursive model. Strictly speaking, multiple filtering is not a new mathematical method, but a secondary strategy, a natural corollary of pattern analysis.

3. *Syncretism: a special system to exploit memory of unique events.* Questions 5 and 6 above are extremely subtle and difficult, but we have concluded that a system of "syncretism" is enough to close the major gap in our system of methods, as a kind of theory of intelligence. In principle, a system of syncretism is necessary (if we wish to achieve maximal "statistical efficiency") to exploit all the relevant information from our historical databank. The system which finally emerges, for computer forecasting, is much simpler than the logic which points towards it.

In artificial intelligence, it is common to try to predict a dependent variable ("pattern classification") by comparing the present values of the independent variables against past sets of values; one's prediction is simply a weighted sum of what the dependent variable turned out to be in the past, *weighted* according to the closeness of the sets of values for the independent variables [9]. In effect, this carries our notion of "absolute consistency" even further than was done with the pure robust method. Here, we do not even assume the truth of the *predictive* part of a model; instead of producing a curve or

SYNCRETISM

predictions from a model, we produce it by smoothing off the curve of actual recorded data. Thus we may hope to achieve an even higher level of robustness than before. However, as with the pure robust method, this method used in pure form would create a serious problem with efficiency.

There is an obvious compromise between the pure syncretic method (as above) and the modelling approach. *After* we fit our model, we can keep our original data available and add a record of what the *prediction error* of the model was in each case. When encountering a new situation, we can make our prediction by calculating the prediction of our model and then *adding* a prediction for the *error* of the model, as based on the syncretic method with model-error as the dependent variable. This can be done for every dependent variable in our system, both raw data and abstract "pattern" variables. This means that we think of our historical records as forming separate datasets, one set for each dependent variable and the independent variables used in predicting that dependent variable.[1]

Before we can use "syncretism" to predict model errors, we need to figure out what *weight* to place on a given past experience. The choice of weights is usually fairly arbitrary in artificial intelligence. Here, if we have a measure of "distance" between the past and present sets of values for the independent variables, we can start out by saying that the weight will equal e^{-kd}, where d is the distance and k is some constant; then we can adjust the weights by dividing each one by the sum of all the weights, so that they add up to one. Initially, we can pick k so that, on the average, we expect a constant, small handful of other experiences with initial weights larger than e^{-1}. "Distance" is measured, formally, by taking the square root of the sum of squares of the differences between the two situations along each independent variable. However, here we may weight each independent variable according to the square of the regression coefficient, if our prediction is made on a linear basis; if it is not linear, we can use, instead, the mean square of the derivative of the prediction of the model with respect to the independent variable. These procedures work fairly well when there is little past experience available to choose k and the weights of the components of d; with more experience, the obvious procedure is to adjust these constants as if they were *model parameters*: i.e., to minimize the overall error in predicting the dependent variable we are concerned with. Note that this general procedure introduces a new functional relation, at every time, between the independent variables and the overall prediction; therefore, when we evaluate the derivative of the prediction with respect to the independent variables, theory tells us to account for *all* these aspects of our prediction procedure.

This kind of procedure may be a bit too expensive, at present, in its original form. Certainly, in a device like the human brain, one would expect severe approximations (such as clustering chemical records of past experience into cells which represent the entire cluster as if it were one experience) to reduce costs; we may be forced to use approximations in order to use syncretism. Another problem is that, in pattern analysis, we do not just predict the *expected* value of the dependent variables; we also create a measure of *uncertainty* in its value. Certainly, if a new event reminds us of an unexpected past trauma, it may increase our feeling of uncertainty, not just our expectations of what is most likely to happen.

For each past record of dependent variable and independent variables, we may define the *primary error* as the actual value of the dependent variable minus the value which would have been predicted by our general model in its current form. We may define the *secondary error* as the primary error, minus the value for the primary error which we would have predicted by syncretism, if we used our other data records in making this prediction. The secondary error corresponds to the actual error in predicting the dependent variable, when the model and the past records are both used, as they normally would be. The *variance* in the actual error may be predicted as the sum of (i) the mean variance of the secondary error, historically; (ii) the weighted sum of the square of the secondary error minus the mean variance of the secondary error, across similar past cases, using the same weights as before. The existence of arbitrary parameters in this kind of procedure may be related to the existence of interpersonal differences in the operation of human brains.

To reduce the cost of such a system, one may simply throw out past data-records which meet two tests: they involve relatively little primary error; they are not very similar to other records which involve a high level of primary error. In such case, one would then normalize the weights above by accounting for both the explicit weights and the weights one might have expected for the "silent majority" of experiences which have been virtually assimilated into the ego.[2]

SYNCRETISM

III-c. Research Strategy in Using the New Methods

If our theoretical analysis is as complete as we like to believe, the application and refinement of the tools described here should be sufficient for as long as we are

1. The generalized predictive model here may be compared with the "ego" of Freudian psychology; the specific records and their influence may be compared with the "id". In econometric analysis, we try to fit our general model better and better to the data, by going over complete, global records of the past. When such records do not exist, or even when they do, one might try to fit the "ego" to *simulated* data, which essentially reconstructs the past, generated by the ego and id together in the absence of external stimuli. This could be done as part of the simulations which we need anyway as part of optimization (see Appendix B). The analogy to dreaming should be obvious.

2. Analogies with the human brain go further than one might imagine. It would be inappropriate to discuss the details here, but a few points may be of interest. The state of "deep sleep" could be interpreted as a time when *clusters* of many experience-records are updated and even transferred between nearby cells; on the other hand, it could be interpreted as a time when individual prediction functions are updated to reflect their own local records, without reference to the global consistency of these adjustments. The specifically human "trance state" could be interpreted as a state in which *social stimuli* are joined with an individual's memory to produce simulated experience which is then remembered as if it were real (e.g., tribal dances after the hunt). This would allow the transfer of experience from individual to individual, more than would be possible if other individuals were perceived solely as noise-producing objects. The human brain may not yet be fully adapted to further possibilities in this direction.

requires human effort in formulating, comparing, and upgrading alternative models. Still, it is much better than treating each contingency as totally unique.[5] It is possible in principle, to allow computers to scan a wide variety of algebraic forms as possible models. This may be worse than human quality control, but it is probably a lot better than treating each contingency as unique.

We can follow a similar procedure in dealing with dynamic programming. We construct a "model" of J. More precisely, we can set up computer programs designed to input a model of J as an algebraic expression with certain parameters in it identified as requiring estimation. As in nonlinear regression, we can allow the user to put in his or her own initial values as a matter of choice. Also, we can print out the "degree of fit" for the final model, as in regression. (Here, the "degree of fit" is measured as the expected value of U across future time which would result from trying to maximize the user's version of J.)

In Howard's version of dynamic programming [11], we generate J by successive approximations. In each step, we reset $J(x(t))$ to equal $U(x(t)) + \text{MaxE}(J(x(t+1))) - \bar{U}$, where the latter value of J is determined by the old estimate of J. In each step, we also pick a new set of actions, to maximize the expected value of $J(x(t+1))$. This procedure can be adapted easily to our purpose. We can *fit* the parameters of $J(x(t))$, as in statistics, to be as close as possible to our previous estimate of $U(x(t)) + \text{MaxE}(J(x(t+1))) - \bar{U}$. In theory, we could attempt to find the optimal set of actions in each step for each x(t) tried out or simulated. In practice, we would probably derive the actions from an "action model," whose parameters may also be estimated as part of this process. With the latter strategy, we can afford to use simple simulation to give us the equivalent of a carefully-computed expectation value. (Note that the constant $\bar{U}$ is not too critical here. In each iteration, scale factors for J can be stored, both additive and multiplicative, to make sure that nothing diverges. This will handle the kind of crossroads problems discussed in [7].)

In theory, this system can only look ahead one extra unit of time per iteration. However, if we estimate these parameters by computing the gradient in each iteration and plug it into a conservative version of Broyden's sparse quasilinear numerical method, convergence will be much faster in practice, yet still practical in cases with many parameters in one problem. To compute this gradient inexpensively, with a complex network model, we recommend the use of the "dynamic feedback" algorithm, discussed in [2, Ch. II]. The above method we would call "heuristic dynamic programming."

Let us suggest another method, which is more efficient for highly complex situations. It is similar in some respects to differential dynamic programming, developed by Jacobson and Mayne [12]. Instead of estimating J(x) for a raw input vector x, we first derive a vector y which is a function of x. We also make sure that U itself is included as one of the components of y. Then, for each component of y, y_i, we estimate $\lambda_i(x)$ as a function. $\lambda_i(x)$ represents the *derivative* of J with respect to y_i. In effect, it represents the "shadow price" of y_i. Since there are many y_i, this would mean many functions to estimate, and it *could* mean many parameters. However, since these λ_i are really interdependent, we could formulate "network" models in which different λ_i share many parameters and terms.

In each time cycle, our method proceeds as follows. From a given situation x(t), we carry out a simulation of x(t + 1) by first simulating the set of random numbers w(t) required by our stochastic model of reality. Then, for those values of w(t), we get a sample value of the gradient of likelihood with respect to the parameters of $\lambda_i(t)$ by trying to fit each function $\lambda_i(t)$ to match our estimate of

$$\sum_j \frac{\partial^+ y_j(t+1)}{\partial y_i(t)} \lambda_j(t+1) \qquad \text{(plus 1 if "}y_i\text{" refers to "U")}$$

This computation can be done inexpensively, with complex network models, by the dynamic feedback method mentioned above. Note that we added a plus sign in the derivative, to indicate that we wish to measure influence *forward* in time, as formalized by our concept of "ordered derivative," the mathematical basis of the dynamic feedback method. Once again, we can update or optimize action strategies, and use a variant of Broyden's method to estimate all the parameters.

This method, which we call "dual heuristic programming," is particularly suited to complex dispersed systems like the human brain; also, it is capable of supporting action models which are slightly faster to react than those with heuristic dynamic programming, at least for a real-time system, because one does not have to wait for feedback to trickle down from the highest levels.

It is extremely important that y_i may be a very complex function, itself to be estimated in this process. In principle, y_i itself could equal J, if this estimate were highly successful. In the human brain, we would speculate that the "dynamic feedback" calculations are performed by the well-known "retrograde" chemical transmissions, flowing back from cell to cell along small tubes inside the brain cells.

Note that two competing strategic models can be weighted in either of these schemes by plugging in a weighted sum of the two J candidates (with the weight itself a parameter to be estimated) into the computer. In this respect, "heuristic dynamic programming" and "dual heuristic programming" are again comparable to regression methods in modelling.

The practical use of these methods [7] requires the prior availability of stochastic predictive models of the global environment. Such models could come either from statistics or from "judgmental models" on Bayesian lines. However, the development of good judgmental *models* will require the development of Bayesian techniques to a higher level than has been considered in the past. This will require careful studies of the effectiveness of variants of these techniques in cases where the measures of performance have a high variance.

REFERENCES

1. This work was supported by the Cybernetics Technology Office of the Defense Advanced Projects Agency, through the Maryland Crisis Warning and Management Project at the Department of Government and Politics at the University of Maryland, College Park. It constitutes a first draft of technical Report #3 from the advanced forecasting group of this project, for the 1976-1977 contract period (starting mid-1976).
2. WERBOS, PAUL J. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.* Ph.D. thesis, Harvard University, Cambridge, Mass. Microfiche copies available from Archives Dept., Widener Library, Harvard. A few text copies may also be available from the author, c/o the Department of Government and Politics, University of Maryland.
3. MCCORMICK, DAVID and WERBOS, PAUL. "Instability in the Results of Regression When Applied to Global Political Forecasting." Report #2, Maryland Crisis Warning and Management Project (see [1]).
4. DEUTSCH, KARL W. *Nationalism and Social Com-*

5. If our model for J includes the pattern analysis circuitry with "syncretism" built in (as in III-b), we can *bring back* an ability to account for those special unique contingencies which appear to merit such special treatment, without reducing our ability to handle normal contingencies in a more generalized way.

interested in any kind of forecasting, by machine or by mind. For the time being, the key fact is that we have a method (r^2 bias) for picking our filtering constant with the compromise robust method. This gives us something immediately useable to re-evaluate conflict models, like the CACI model [3], with a better methodology. As in working with the Deutsch model, we may fit the parameters of various models to the first half of our time-series, using classical and robust methods, and then see which does better in predicting the second half. At present, we have set up a 20-year databank covering most nations of the world, suitable for interactive computer analysis.

Our own numerical methods could perform these analyses efficiently, in theory [2]. However, because of the difficulties and development costs involved in trying to write general software in a university department, we borrowed from existing nonlinear programming [10]. According to the authors, this program works better on dynamic control problems than do the complex Riccati equation and matrix methods which dominate most of the literature. It can be used to fit parameters to minimize our "robust" measure of model error if it is generalized in certain ways: (i) a "model compiling" routine is needed to translate a simple, user-specified model into an object subroutine which calculates model error as a function of parameter values; (ii) there must be provision for "multiple sector" estimation, to allow estimating the material values for endogenous variables in *different* countries, without waiting for computer time; (iii) certain contingencies must be planned for which the NASA routine did not consider. We have now reprogrammed most of the NASA routine in ANSI PL/1, with their additional features, so that it can run interactively on the MIT Multics, which we are using over the ARPANET. These routines, like the whole of our project, are in the "government-related public domain."

Unfortunately, this routine uses only the Fletcher-Powell method for convergence; a similar method, the "Broyden method," would have allowed the use of a "sparse information matrix," which in turn would allow the use of more parameters.[3] In the present situation, we will have to use human labor, to pick groups of about twenty parameters in the model, fine-tune them, then pick another group, and so on, until all the parameters are optimized. Regression analysis should provide adequate initial values for these parameters. (The "r^2 bias" method should eliminate the convergency problems one might otherwise expect with these initial values, if we set r initially to 1, which corresponds with regression.)

After initial work with the r^2 bias method, we intend to investigate the r bias method, multiple filtering, and then perhaps other possibilities suggested above. The choice between alternatives suggested here cannot be sorted out on a purely theoretical basis, because "robustness" is essentially an empirical issue, as is the choice of models. When more empirical examples are available, and when we have an idea of what an adequate theory would show, perhaps than we can start to figure out how we might have guessed our results before doing any tests. At this stage, however, it would be dangerous to make too many a priori assumptions about what works and what doesn't. Indeed, the specific combination of filters and models which works best in crisis warning may turn out to be unique to that subject. A long period of strictly experimental work lies ahead.

APPENDIX A

PATTERN ANALYSIS AS A MAXIMUM LIKELIHOOD METHOD

In Chapter II of *Beyond Regression* [2], "pattern analysis" was suggested as a new approach to the problem of pattern recognition. In this approach, pattern recognition is treated as a system to help support prediction and optimization. "Pattern analysis" attempts to extricate the key variables which underly the dynamics of the environment one is trying to analyze; it is necessary, as part of effective forecasting, because the original raw data contains *nonlinear interdependence* which cannot be analyzed efficiently or explicitly by conventional direct methods. Figure 3 indicates the difference between conventional econometric approaches and pattern analysis. In pattern analysis, we construct three systems of formulas, each of which looks like an econometric model.

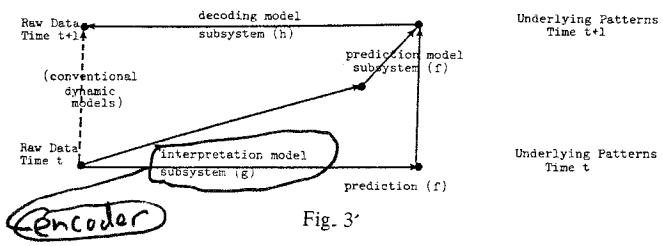

Fig. 3

1. The "prediction model" gives probablistic forecasts of the underlying patterns in the *future*, as a function of *past* and *present* conditions.

2. The "interpretation model" gives a probabilistic description of what the *underlying patterns* are in the *present*, as a function of *direct observations* in the present and of general conditions in the past.

3. The "decoding model" gives a probabilistic description of what *direct observations* to expect at any time, as a function of the underlying patterns at that time, and also as a function of observations or patterns for earlier times.

This technique is oriented towards dynamic systems; however, depending on how the model is specified, and depending on what we call "past" information (e.g., nothing?), it can be applied in situations with a different dimensional structure.

Our hope was to blend the advantages of control theory, nonlinear regression, and maximum likelihood factor analysis into a single technique, which would contain none of the arbitrary aspects of classical artificial intelligence methods. In particular, all of the parameters of the three models (the three subsystems) were to be estimated by maximum likelihood methods, with perfect "statistical efficiency": no information from our past experience would be

3. We have carefully studied the possiblity of more advanced convergence methods, building on those we have already suggested [2], as they would be plugged into a Broyden-style "front-end." However, major changes would come only by allowing explicit models of the convergence process itself, models with internal estimation and convergence requirements. It would seem better just to alter Broyden's methods in minor, ad hoc ways (e.g., size cutoffs in responsiveness changes and fractional powers of the changes recommended by Broyden), while stating that an intelligent system may attempt to treat some of its internally-generated derivatives as explicit variables for explicit study. This is like Deutsch's idea of "self-consciousness."

munications. New York: MIT Technology Press, 1953.
5. BOX, GEORGE E. P. and JENKINS, GWILYM. *Time-Series Analysis: Forecasting and Control.* San Francisco: Holden Day, 1970.
6. MCCLELLAND, CHARLES A. "Warnings in the International Event Flow: EFI and ROZ as Threat Indicators." Report to Cybernetics Technology Office, Defense Advanced Research Projects Agency, July 1976.
7. WERBOS, PAUL J. "Strategic Planning for Global Survival." Report #1, Maryland Crisis Warning and Management Project (see [1].
8. BRYSON, AUTHUR E. and HO, YU-CHI. *Applied Optimal Control.* Blaisdell, 1968.
9. DUDA, RICHARD O. and HART, PETER E. *Pattern Classification and Scene Analysis.* New York: Wiley Interscience, 1973
10. JOHNSON, IVAN L., JR. "The Davidon-Fletcher-Powell Penalty Function Method: A Generalized Iterative Technique for Solving Parameter Optimization Problems." NASA Technical Note TN D-8251; program code obtained from Ivan Johnson, Lyndon B. Johnson Space Center, Houston, Texas; originally written by Analytical Mechanics, Inc.
11. HOWARD, RICHARD. *Dynamic Programming and Markov Processes.* Cambridge, Mass.: MIT Press, 1960.
12. JACOBSON, DAVID H. and MAYNE, DAVID Q. *Differential Dynamic Programming,* New York: Elsevier, 1970.

...n. It would be interesting to see what capabilities would lost thereby. Below, we will use the assumption of variable riance, for the sake of generality.) Then, for each underlying riable R_i, we simulate a value Q_i by computing g_i plus $\sigma_{e_i}^{(i)}$ times a ndom number. (A random number of unit variance, mean zero.) e plug in the *simulated* value of $R_i(t+1)$, plus the real values of evious direct observations, into the decoding model. This gives us edictions of the $x_j(t+1)$. Overall, we try to minimize the sum of o error terms:

(i) the sum over all raw observation variables of the decoding or after our simulation; we compute this essentially as

$$\Sigma_i \log \Sigma_t (x_i - h_i(Q))^2;$$

← LIKE $\sum_i \log (x_i - \hat{x}_i)^2$

(ii) the sum, over all underlying variables, of the prediction or, defined as the *correction* entropy from the predicted tribution to the interpreted distribution. We compute this integral

$$(\log \sigma_{e_i}^{(p)} - \log \sigma_{e_i}^{(i)} + 1/2(\frac{g_i - f_i}{\sigma_{e_i}^{(p)}})^2 - 1/2(1 - \frac{\sigma_{e_i}^{(i)2}}{\sigma_{e_i}^{(p)2}})).$$

$\sigma_{e_i}^{(p)}$ APPEARS HERE, BUT NOT IN THE NEW VERSION!

e general case of variable variance may allow some non-uniqueness the final solution but may upgrade the quality of solutions over hat we would expect with fixed-variance models.

For the decoding parameters H_j, the first of these two error ms is the only one which is operational. The simulation process scribed above is a valid Monte Carlo procedure for estimating the pected value of

$$\frac{\partial E_C}{\partial H_j} = \frac{\partial E_R}{\partial H_j},$$

the following reasons. Let $\hat{p}_i(x|R)$ be the probability distribution $x(t+1)$, given $R(t+1)$, implied by the decoding model. Let (R) be the probability distribution for $R(t+1)$, given past ormation, as per the prediction model. Let $\hat{p}_3(R|x)$ be the erpretation model. From our definition above,

$$-E_R = \int p(x) \log \hat{p}(x)\, dx,$$

THE NEW LOSS FUNCTION IN THE HANDBOOK MAY BE DESCRIBED AS "A RELATIVE ENTROPY MEASURE SIMILAR TO EQUATION 46 OF THE HANDBOOK, WHICH DOES NOT REQUIRE AN ESTIMATE OF $\sigma_{e_i}^{(p)}$ IN PREDICTING R FROM PREVIOUS OR EXOGENOUS INFORMATION"