An Empirical Test of New Forecasting Methods Derived from a Theory of Intelligence: The Prediction of Conflict in Latin America

PAUL J. WERBOS AND JIM TITUS

Abstract—**The "compromise" method is a new computer-based forecasting tool, available within the conversational CS package on the MIT Multics. Like regression (least squares) or new forms of Box–Jenkins methods, it estimates the parameters of a multivariate dynamic model and may be used for causal analysis or policy impact analysis. Unlike those maximum-likelihood methods, it does not assume that errors are "white noise," random and normal. It follows the newer robust philosophy of trying to minimize estimation errors *on the assumption* that noise will be inextricably dirty. In the case of "strong" dynamic models—models which predict that changes in present variable values lead to comparable changes in future variable values—it may reduce parameter errors by an order of magnitude. Forecasting errors will also be reduced, although the degree of reduction depends on how much randomness exists in the process. When we used the compromise method according to the new "bias" procedure, in order to reestimate the *J*-5 model (a nonlinear multiequation model used by the Department of Defense in long-range forecasting), forecasting errors were reduced by between 0 and 45 percent (with a median of about 20 percent) across different variables, as compared with regression. With simultaneous-equation econometric models, it has reduced them by 50 percent. The procedure has been documented for use by nonprogrammers [1]; it incorporates a new quasi-Newtonian method which can handle many parameters.**

I. Basic Results

IN 1974, we suggested that a new class of robust methods could outperform regression and Box–Jenkins methods in long-range forecasting [2]. These robust methods did quite well in preliminary tests, but until January 1978 they had never been tested in their ability to estimate multiequation models. These methods share a common philosophy with the better known robust methods of Mosteller, Tukey [3], and others: instead of assuming that a model is perfectly "true" in some form, as in conventional maximum-likelihood estimation, where one maximizes an abstract "probability of truth," we try to estimate the parameters of a model in such a way that the model will do a workable job of forecasting *despite* the "dirty noise," the imperfect specification, and the limited coverage of any realistic model in the social sciences. According to maximum-likelihood theory, one can deduce the "proper method" by elegant abstract reasoning purely from first principles; however, the robust philosophy emphasizes the need for empirical tests—like this one—and the need for an operations-research style of thinking, even when one is carrying out careful mathematical analysis.

In our initial test, we have compared regression against the most up-to-date version of the robust method, the "bias" method formulated in October 1977. We used both methods to estimate the equations of the revised *J*-5 model, a model developed by CACI [4], [5] to predict socioeconomic and conflict variables worldwide; forms of this model based on regression are now being used by the Defense Intelligence Agency, *JCS*/*J*-5 and others, to provide long-range forecasts which guide global strategic planning. We found that the bias version of the model outperformed the regression version across virtually all variables for essentially all time intervals of prediction. The median improvement, across variables, was a reduction of roughly 20 percent in the size of errors. The improvement varied over a fairly uniform range, from one case with no improvement (population) to another with a 45 percent cut in error (gross domestic product (GDP)). With population, the bias forecasts seem to be worse, but this is probably a numerical artifact (see the last paragraph of Section III-E). All of this is shown in Table I, where the column for regression "Reg." may be compared with the column for the new bias method "Bias (actual)". The rest of this report deals with the interpretation and analysis of Tables I–IV. More recent results have shown an even stronger improvement over regression (median 50 percent).

In this test, we studied the ten Latin American countries for which we could find highly reliable data from 1950 to 1967: Argentina, Brazil, Chile, Columbia, Ecuador, Guatemala, Honduras, Mexico, Peru, and Venezuela. We focused our attention on the "core" of the *J*-5 model, the part which predicts the following variables: population, gross investment, domestic government spending, defense expenditure, GDP, consumption, imports, exports, tension ratio,

Manuscript received March 23, 1977; revised February 12, 1978 and May 1, 1978. This work was supported by the Cybernetics Technology Office, Defense Advanced Research Projects Agency, Department of Defense, and monitored by the Office of Naval Research under Contract 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 DARPA.

P. J. Werbos is with the Department of Government, University of Maryland, College Park, MD 20742.

J. Titus is with the Department of Economics, University of Maryland, College Park, MD 20742.

TABLE I
MEAN PERCENTAGE FORECAST ERRORS

	Reg.	Bias (actual)	w=1 (p.f.)	w=1	Pure (actual)	Bias (estim.)	Pure (estim.)
Population	1.391	2.072	3.850	3.850	6.314	4.667	142.370
Investment	20.072	15.874	19.753	19.703	16.863	29.645	24.964
Domestic Gov.	21.128	18.507	20.739	20.459	17.192	30.371	172.417
Defense	31.807	26.696	27.458	32.304	28.875	28.480	39.720
GDP	11.934	6.413	8.878	9.624	9.547	15.433	29.290
Consumption	9.914	6.064	6.571	6.895	8.943	11.670	22.228
Imports	31.932	21.135	31.482	32.380	20.203	25.196	42.437
Exports	22.954	21.934	20.137	21.258	23.147	27.294	52.495
Tension Ratio	14.256	12.366	11.081	14.453	13.047	12.009	13.530

Average across country-years of error as a percentage of actual.

TABLE II
t RATIOS FOR MEAN PERCENTAGE FORECAST ERRORS

	Bias (actual)	w=1 (p.f.)	w=1	Pure (actual)	Bias (estim.)	Pure (estim.)
Population	-4.540	-6.601	-6.601	-9.532	-4.908	-3.015
Investment	1.658	0.971	1.239	1.164	-3.037	-1.232
Domestic Gov.	2.412	3.343	5.530	4.191	-4.604	-3.101
Defense	2.122	2.002	-4.774	1.831	1.496	-2.604
GDP	5.367	3.973	4.265	1.886	-2.741	-3.788
Consumption	4.030	3.197	4.025	0.861	-1.421	-4.400
Imports	4.688	0.976	-1.822	4.686	3.173	-3.619
Exports	0.889	1.544	3.408	-0.054	-2.640	-5.035
Tension Ratio	1.999	2.506	-4.559	1.443	1.895	0.370
Turmoil	1.398	0.572	2.260	0.121	0.920	-0.469
Conflict	4.789	2.524	-0.594	-1.984	5.000	-1.190
Coups	0.408	-0.687	1.994	-2.734	-0.397	-1.061

Tests hypothesis that regression is worse than the given alternative method for errors cited in Table I.

TABLE III
AVERAGE OF WEIGHTED PERCENTAGE FORECAST ERRORS

	Reg.	w=1 (p.f.)	w=1	Bias (actual)	Pure (actual)
Population	1.161	3.733	3.733	1.936	6.183
Investment	14.380	14.074	14.021	16.352	16.775
Domestic Gov.	16.281	15.706	15.961	18.256	15.824
Defense	18.281	15.375	18.597	16.606	16.602
GDP	10.075	7.043	7.723	7.604	9.862
Consumption	9.495	6.477	6.951	6.826	9.384
Imports	30.414	30.178	30.954	19.067	19.298
Exports	28.656	20.594	26.032	26.475	20.525
Tension Ratio	14.175	11.056	14.381	12.231	12.871
Turmoil	86.936	88.054	86.633	87.791	88.116
Conflict	280.808	255.447	281.716	215.779	279.642
Coups	80.805	81.585	80.470	79.024	88.367

Average across time of "weighted percentage error," defined as mean absolute value of error as a percentage of mean value of actual values.

TABLE IV
t RATIOS FOR MEAN ABSOLUTE VALUE OF FORECAST ERRORS

	Bias (actual)	w=1 (p.f.)	w=1	Pure (actual)	Bias (estim.)	Pure (estim.)
Population	-3.406	-3.681	-3.681	-4.326	-5.897	-6.154
Investment	-0.811	1.187	1.277	-0.861	-1.530	-1.148
Domestic Gov.	-1.376	2.104	2.312	0.568	-3.513	-3.656
Defense	1.421	1.822	-3.953	1.877	1.593	-2.608
GDP	3.240	4.115	4.478	0.182	0.896	-2.597
Consumption	3.536	3.551	4.045	0.122	3.251	-3.973
Imports	3.137	0.239	-1.328	3.339	2.352	-2.882
Exports	1.335	3.251	2.977	1.962	-2.132	-2.417
Tension Ratio	2.192	2.650	-4.717	1.611	2.096	0.603
Turmoil	-0.798	-0.575	1.241	-0.430	0.097	-0.598
Conflict	4.914	2.976	-0.144	0.048	4.997	0.076
Coups	-0.100	-0.886	1.262	-1.963	-0.931	-1.350

Tests hypothesis that regression is worse than the given alternative method in terms of absolute values of error.

coups, international conflict, and internal turmoil. We fitted the *J*-5 model to the data from 1950 to 1961 without accounting for later data in any way. Then we used this early data, and the model, to generate forecasts only as far as 1967, the last year for which we had complete data. For the target years, 1962 to 1967, we compared the forecast values of the variables against the actual values. (1961 was chosen as the base year, well in advance, because an empirical data span twice the length of the forecast span seemed desirable as a way of being sure that there would be no complications due to gross instability in parameter values.) These forecasts—averaging only three and a half time periods forward in time—are a far cry from the many-period long-range forecasts with which we have been associated in the past (see the forecasts of nationalism in [2, ch. VI]; see [6]). Nevertheless, the superiority of the bias model appears fairly uniform across all target years, even for one-year forecasts. For this reason, we have chosen to summarize our results in Tables I–IV, which demonstrate the average results across all countries and target years for each target year studied. Separate such tables have also been constructed for each individual target year; these tables have been submitted with the version of this report filed with the Defense Advanced Research Projects Agency (DARPA) [1].

Note that the test of performance here is quite different from the conventional r^2 test of performance used in most regression studies. In essence, r^2 measures the ability of a model to predict time $t + 1$ on the basis of actual data from time t over the *same* data which are used to fit the model in the first place. Here, we are testing the ability of the model to predict *several* years into the future on the basis of initial real data; we are testing the fit between actual and predicted values for a *new* set of data, to which the model was not fitted. Given that the *actual application* of forecasting models

hinges on their ability to predict new sets of data—the future—this test is much more realistic than the usual test. Also, this test is a more difficult test to pass. A multiple R of 99.5 percent sounds very good in regression. However, such an R implies an r^2 of 99 percent, an error variance 1/100 the size of the variance of the variable predicted, and a standard deviation of error equal to 1/10 of the standard deviation of the variable. In other words, a multiple R of 99.5 percent implies an actual error size of 10 percent, even when one predicts only one year into the future; if errors have a persistent tendency to accumulate, despite the usual correction procedures, this could imply as much as a 50 percent error in predicting only five years into the future.

Table I shows essentially what we have said above, that the bias method "Bias (actual)" outperforms regression to a substantial degree. (Again, there are other things in this table, too, which we will explore, item by item, in Section III.) However, one may ask whether this result could be a coincidence. Is the difference in errors statistically significant? For each target variable, and for each method other than regression, we performed a classical t test for paired samples (see [7, Section 8-3-c]); the results are shown in Table II. With a sample of 60, as in this case, a t ratio of 1.671 or better indicates that regression is worse than the method we are comparing it against at a 95 percent level of significance. (In other words, if the two error distributions were identical, the probability would be 5 percent that regression would appear to be that much worse as a result of coincidence.) For most of the 12 target variables, the bias method passes this test. A t ratio of 3.46 indicates significance at a level of 99.95 percent; with 4 of the 12 target variables—including international conflict—the superiority of the bias method is substantially greater than this. In brief, the statistics indicate that we may be confident of the validity of the general impression we obtain from Table I.

II. The Model

Before going on, we should mention a few interesting substantive results. We did not intend, initially, to look for any "new" substantive explanations of conflict, apart from what is already implied in the CACI J-5 model. We intended to focus only on the core of this model, on a set of equations which are mathematically independent of the rest. However, the original CACI model was based on a cross section of one year's data; with many years' data, and slightly different definitions of variables, we found that the model did not hold up very well at all. In 1976 [8] our project noted that the CACI parameter estimates were not stable over time, over the United Nations data base then being prepared; however, here we found that the choice of terms in the model itself was a problem, over new and more reliable data. Our initial regressions recorded low predictive power (r^2) for all of the political equations in the model, reaching as low as 0.0016 in the case of international conflict; while we were willing to accept the idea that some prediction is better than none, 0.0016 seemed more like "none" to us, even for a purely methodological study. Fortunately, however, when we were searching for data, we had looked at the McIlroy thesis [9] on the value of "strain" as a predictor of conflict, as well as at similar findings by Feierabend and others suggesting that measures of social development were crucial in predicting conflict. In addition to urbanization data, therefore, we collected data on the fraction of the population enrolled in primary education ("per_primary") as an index of social development. This turned out to be the single most important predictor of international conflict: more education means less conflict. This result holds up in all versions of the J-5 model which we estimated. (Note, however, that we had to treat per primary as an exogenous variable, like "STRAIN" as it appears in the original equations reported by CACI; see [9, p. 7, eq. (24)].) In like manner, we added exports to this equation, largely because a recent report from the Cross-National Crisis Indicators project (CNCI) [10] indicates a strong association between different components of foreign-policy-behavior-sent. Also, many independent variables which seemed important in the CACI equations had t ratios here of 0.1, −0.08, 0.3, etc.; for this reason, we had to drop them from the equations or replace them by similar terms which represented the same concepts in a more accurate way, with a better t ratio. It was unpleasant having to drop average foreign military assistance from the equations, but with t ratios such as −0.08, there was not much choice; the variables simply did not seem to have a measurable effect on Latin American politics. In general, terms of substantive importance were kept in the model, but only if they had t ratios above 0.7. As a result of these changes, the r^2 of the regression equations was lifted up to meaningful levels for our sample of 110 observations ($r^2 = 6$ percent for conflict, 23 percent for turmoil, 22 percent for coups, and "high"—mostly above 90 percent—for all others). Furthermore, the degrees of freedom of the model were reduced in number. These changes were made in order to give regression a fair chance. Also, to be honest, they were made because our theoretical analysis indicates that the bias method will show greater relative improvement for models which are already "strong" to begin with. A strong model is one which tells us that changes in the present state of the world will lead to significant changes in the far future. All of these modifications were made during the regression phase of our analysis, before a single robust estimation of any kind had been done on any of these data. The final regression model is shown in Fig. 1. The robust versions are shown in the appendices of [1].

III. Analysis of Tables I–IV

A. Regression Versus Bias for Economic Variables

As economic variables, we include gross investment (I), domestic government spending (dom), defense expenditure (defx), GDP, consumption (C), imports (imp), and exports (tex), all defined in terms of 1973 real dollars. Bias reduces the average percentage error in predicting every one of these variables. In one case (exports), the percentage of error is reduced by only about 5 percent of its original size with regression (see Table I), but for the others we obtain error reductions from about 15 to 45 percent. With the exception

```
pop(t)=1.02955*pop(t-1);
I(t)=.069376*gdp(t-1)+.0685683*I(t-1);
dom(t)=-.001928*gdp(t-1)+.003267*pop(t)+1.025*dom(t-1);
defx(t)=.792*defx(t-1)-.083*tml(t-1)+.441*rivdex(t)
      -.303*rivdex(t-1)+.000515*pop(t)+.000627*gdp(t-1);
C(t)=.747*C(t-1)+.009882*pop(t)+.19*gdp(t);
imp(t)=.769*imp(t-1)+.0202*gdp(t)-.001843*pop(t);
tex(t)=1.055*tex(t-1)+.095*gdp(t)-.102*gdp(t-1);
gdp(t)=C(t)+I(t)+dom(t)+defx(t)+tex(t)-imp(t);
adex(t)=.8*adex(t-1)+.2*defx(t);
tr(t)=100*(defx(t)/adex(t));
tml(t)=.312*tml(t-1)+5.846*conf(t-1)+.237*tr(t)
     -16.223+.000002*tr(t)*gdp(t);
conf(t)=1.126-4.512*((gdp(t)/pop(t))-(gdp(t-1)/pop(t-1)))
  -75.339*per_primary(t)+.000195*tex(t)-.000312*imp(t-1);
coup(t)=.003673*tml(t)-.002609*tml(t-1)+151*conf(t)
  -.031*conf(t-1)+.0022*gdp(t)*tml(t)/pop(t)
  +.059*gdp(t)/pop(t);
```

Fig. 1. Model as estimated by regression (note that units are important in interpreting this model; economic figures are in millions of dollars, population is in thousands, per-primary is percentage enrollment divided by 1000, and conflict variables are about as big as their means).

of exports, the error reduction is not only large but statistically significant; the t ratios are greater than the 95 percent level of 1.671, except for investment, which is close enough at 1.658 (see Table II). Better prediction of exports would probably require a new model, a model which accounts for economic conditions outside the exporting country itself; dyadic modeling of that sort might require additions to our computer package, however, depending on the type of models being considered. The biggest and most significant reduction in error is for the most important single economic variable, GDP.

In our detailed discussion of methods, we will emphasize that bias yields the biggest improvements over regression in the case of strong models, models in which errors in parameter estimation would lead to *cumulative* error in long-range forecasting. The economic variables are the main source of strength in the *J*-5 model as a whole. Thus we would expect that bias leads to a *direct improvement* in the economic forecasts; one would expect an indirect improvement in the forecasts of noneconomic variables, especially into the far future, only because the quality of those forecasts depends on the economic forecasts which provide the key independent variables. It should be no surprise, therefore, that the economic forecasts seem to be the area of greatest improvement with bias. In order to improve the quality of conflict forecasts, we would have to improve the *J*-5 conflict equations very carefully, probably by adding more and better socioeconomic data; however, it will be equally important to improve and strengthen the social and economic equations which are used to predict the independent variables of any new conflict equation. With stronger models of this kind, we expect that the bias method will have even greater advantages over regression than it does here.

The discussion above is based on Tables I and II. In Tables III and IV, however, the improvement is not so clear cut. If we look at the *absolute value* of error, as tested by Table IV, we obtain three variables with significant improvement, two improvements of borderline significance, one *worsening* of borderline significance, and one insignificant worsening. (By borderline, we mean a t ratio between 1 and 1.6.) Why the difference?

Strictly speaking, the t tests of Table II are more valid than the t tests in Table IV. In both cases, we are trying to evaluate the mean and standard error of a variable called "improvement," which equals error for regression minus error for the bias method. The classical t test assumes that this variable is normally distributed, more or less. In the case of *percentage* errors, this is generally reasonable. However, in the case of *absolute* errors, there is "heteroscedasticity;" in other words, the largest nations (Brazil and Mexico) dominate the analysis. This means that the effective sample size is much smaller and that the variance of the variable appears larger than it really is. With a smaller effective sample size, we obtain less significant results, even when the true differences are equally large; furthermore, the larger standard errors lead to larger random errors in our measurement of model improvement. A simple paired comparison like this is extremely susceptible to heteroscedasticity.

Tables III and IV were initially constructed to indicate whether heteroscedasticity might be a problem with the original model estimation. Both with regression *and* bias, we followed the conventional practice of minimizing error across *all* nations, *without* adjusting for the relative sizes of the nations; this is justifiable on grounds that the bigger nations are more important and do represent a bigger sample in some ways. Tables III and IV are subject to more random error than Tables I and II because of heteroscedasticity, but they do provide some indication of how well the model performs on big nations, the nations to which the model was fit most closely. If we compare Tables I and III, we can see that the bias method was not thrown off by heteroscedasticity; the percentage errors across *all nations* (Table I) are just as good, on the whole, as the errors in the big nations (Table III). Indeed, they even seem slightly better, as our analysis of Table IV has indicated. However, the difference is not statistically significant; it is probably due to random noise affecting Table III, which we have just discussed. One would certainly expect a model to do just as well on the nations it pays more attention to as on those it does not pay as much attention to, if there were enough data to evaluate both categories of performance accurately. There is a corollary to this argument: with more data, we would expect that Table III would indicate uniform large improvements with the bias method, just as Tables I and II do now.

B. Regression Versus Bias for Noneconomic Variables

Regression does seem to outperform bias for one of the variables in Table I, population. But there is excellent reason to believe that this is a numerical artifact, due to the low level of error with all methods in predicting population (see the last paragraph of Section III-E. for a detailed analysis of this point).

For two of the variables, tension ratio and international conflict, bias did significantly better in all of the tables.

However, the errors in predicting international conflict were very large in Table III; in Table I, they were so large that they could not even be printed on the table. The reason for this is not that conflict is totally inscrutable. After all, the t tests do indicate substantial improvement—implying some knowledge—in going from regression to bias predictions of conflict both in Table II and in Table IV. The unpleasant fact, in the real world, is that we must make do with the best conflict indicators we can obtain, however much noise there is in observing the process; improvements in our prediction capabilities (here, 25 percent reduction in error) are important, even if they do not lead us in one step to nirvana. More advanced models of conflict, based on better data, will be needed to reduce these errors much further.

Nevertheless, the actual errors are not nearly so bad as these 200 percent figures indicate. For Tables I and II, percentage error was defined as

$$\frac{|\text{predicted minus actual}|}{\text{actual value}}.$$

Because the actual value was sometimes zero, we added 10^{-13} to the "actual" to prevent numerical catastrophe. In the cases of conflict, coup, and turmoil, the scores were often zero; for coups and conflict, they were sometimes zero for the *whole* of Latin America. When we computed simple percentage error, then the score was *entirely* dominated by the false-alarm rate, the rate of predicting coup or turmoil when no coups or turmoil took place. A t test comparing false-alarm rates, of course, is just as valid as a t test comparing ordinary percentage error. In terms of the false-alarm rate, bias was uniformly superior to regression over all three conflict variables subject to this problem: with international conflict the superiority of bias was highly significant, with turmoil it reached only a borderline significance, and with coups it was insignificant. The t ratios in Table II correspond exactly to the numbers in Table I.

Table IV gives us an evaluation of error in *absolute* terms for cases where it is illogical to compute error as a percentage of the actual value (again, for coups, turmoil, and conflict; for the tension ratio, the two tables, II and IV, are essentially the same). More precisely, it compares the mean absolute size of errors for different methods across all years and nations. In effect, it tells us how well the methods did in terms of *average* error in predicting coups, while Tables I and II tell us how bad the false-alarm rate is. The *average* error in predicting international conflict is better with the bias method by a very significant margin. However, with coups and turmoil, bias actually did worse. Still, this difference was not statistically significant, and indeed, in the case of coups, it was virtually nil. Realistically, in the case of coups, an entropy measure of error would be more appropriate since the "expected number of coups" may be better interpreted as "probability of coup;" however, such elaborate measures of error will have to await future studies.

With the one conflict variable that never got to be zero, the tension ratio, the bias method did significantly better than regression by the test of absolute error (Table IV), just as it did by the test of percentage error (Table II).

Table III was put together to try to do for Table IV what Table I does for Table II: to give us some feeling for the magnitude of differences in error. However, the correspondence between Tables III and IV is not as exact. Instead of printing out the mean absolute value of errors, which may mean very little to most people, we created a percentage by dividing the absolute error in each year by the mean value of the target variable in that year; this kind of percentage was calculated for each target year and then averaged up across all target years to produce Table III. For all of the variables except conflict, coup, and turmoil, the table can be thought of simply as showing "percentage error, weighted by the size of the nation predicted." With turmoil, the table provides a balanced evaluation of performance, but the numbers are all somewhat inflated because in 1964, when the actual level of turmoil was temporarily very low across all of Latin America and the predictions were moderate in size, the errors looked huge *as a percentage of* that low mean; nevertheless, the numbers are still comparable, and we can see that the choice of method has little impact in predicting turmoil. For the other two variables, coup and international conflict, the actual mean value was sometimes zero across Latin America; in these years, we calculated the percentage on the basis of a fictitious mean of 1, which is very large. Thus the figures for coups and conflict do not represent *average error* in Table III, but rather average errors across cases where there is no false alarm. Note that predictions of international conflict are *much* better for the bias method in these cases. From Table II, we also know that the false-alarm rate is much less with the bias method; therefore, the forecasts of international conflict are *substantially* better in all respects than those of regression. In the case of coups, the difference in prediction quality is negligible in these cases. Again, the numbers themselves may be inflated, but the comparison between methods remains meaningful. Once again, the t ratios in Table IV do not depend in any way on percentage calculations; they are definitely more meaningful than the percentages in Table III.

In summary, bias did substantially and significantly better than regression in predicting tension ratios and international conflict, by any measure. In predicting coups and turmoil, the differences between the two methods were both small and statistically insignificant. Bias seems to reduce the false-alarm rate for internal turmoil, but this result is of borderline significance. The other differences between methods are both insignificant and equivocal in this case.

C. Errors in Parameter Estimates

Tables I–IV indicate a significant reduction in forecasting errors when we change over from regression to the bias method. However, forecasting error is not the only criterion for judging the value of a model. Models have two major roles to play in public decisionmaking: 1) forecasting the background conditions which policymakers must anticipate; 2) describing "how the world works," so that a policymaker can assess the *changes* in future conditions which would result from *changes* in present policy. The first of these applications depends on the quality of forecasts as

such. However, the second application depends on the *quality of the parameters*. If one knows exactly the *probability* of alternative outcomes, one can use decision analysis to make a correct rational decision; even if one cannot forecast exactly what *will* happen, a correct statistical model will provide correct probabilities. Correct model parameters imply correct probabilities, not exact predictions of the future. (Needless to say, however, no probabilities are completely correct until the correct statistical estimates have also been combined with human judgment and any other source of information which goes beyond the historical time-series data; the discussion here is focused on the question of how to correctly assess what we find in the statistics.)

Unfortunately, it is very difficult to measure errors in parameter estimation. One has to know the "true" values of parameters before one can measure the errors in parameter estimation; this is possible, in general, only for known simulated processes, not for the real world. Still, in Section III-D we will argue that bias leads to a moderate reduction in forecasting errors, because of a reduction in parameter errors which may be something like an order of magnitude, for crucial parameters. Given that we have observed the predicted moderate reduction in forecasting errors, it seems likely that parameter errors—if we could measure them—have been reduced much more. In our past work, on studies of simulated data (see [2, ch. IV]), we did find that the improvements in parameter estimates were much more dramatic than the improvements in forecasting errors as such.

D. Alternative Methods Tested: Background

In the column headings of Tables I–IV, "Reg." refers to regression, and "Bias(actual)" is the specific form of robust estimation which we now advocate. However, we have yet to define precisely what the bias method entails.

The bias method is a specific strategy for making use of the more generalized method which we have called the compromise method [6], [2]. All of the other methods cited in Tables I–IV, except regression, are different strategies for making use of the compromise method; however, these other strategies are closely related to more conventional methods, such as the full-information maximum-likelihood method, and will provide some evidence of the superiority of bias over those methods.

The compromise method has been described at great length in the other contexts cited above; for now, we will merely summarize the method briefly to help the reader make sense of Tables I–IV. We can picture the compromise method as a three-step process.

1) For each equation of your model, regress the dependent variable on the independent variables in the conventional manner.

2) Use the resulting model to "filter" those variables which the model tries to predict. Filtering means estimating the true underlying values of the variables by use of the equation

$$\tilde{X}_i = (1 - w_i)\hat{X}_i + w_i X_i.$$

(Note that X_i refers to the *measured data* for variable number i, $\tilde{X}_i$ refers to the *estimated* "true" value of the variable, and $\hat{X}_i$ refers to the *predicted* value of the variable for the current time period, based on applying the *current form of the model* to the estimated true values in the previous time period. w_i is not another variable, but a "filtering constant" or "weight," to be discussed below. w_i was referred to as r in our earlier papers.)

3) Replace each variable by its filtered version, where possible, *in its appearances as an independent variable* in the model equations. Then go back to step 1) and redo the regressions for the new versions of the variables. For example, if I (investment) appears both as a dependent variable ($I(t)$) and as an independent variable ($I(t-1)$), we use the filtered version of I to give us a new version of $I(t-1)$.

This process is circular, of course; after we redo the model, we have to recalculate the predictions $\hat{X}_i$, recalculate the filtered values $\tilde{X}_i$, and then go back and redo the model again. Actually, the predicted and estimated values must be calculated forward in time, starting from estimates either at time $t = 1$, the first actual observation, or at $t = 0$, the previous time-period. For long time series, we would recommend the following order of calculation: $\tilde{X}(t=0)$, $\hat{X}(t=1)$, $\tilde{X}(t=1)$, $\hat{X}(t=2)$, $\tilde{X}(t=2)$, $\cdots$. Fig. 2 may help to explain this process. After we redo the regressions, we should then recalculate the estimates, redo the regressions a third time, recalculate the estimates a third time, and so on. In principle, we should continue to go around and around the circle until our estimates change very little from iteration to iteration. Unfortunately, step 1) by itself could be very expensive, because our model may require nonlinear regression; convergence by this method could be very expensive and possibly unreliable. However, there is an easier way to proceed, which is mathematically equivalent and more flexible in other respects. For each set of trial values of the parameters, we can calculate the predictions and the estimates together and then add up the square errors (or other loss measure) between the predicted and actual values; we can use a generalized minimization subroutine to keep guessing sets of trial values until it is satisfied that it has minimized error. When the available time series are very long, we can estimate the values of $\tilde{X}_i(0)$ in each nation as extra parameters of the model; however, when the time series are shorter, as in the runs reported here, it seems more appropriate to begin with measured data, with $\tilde{X}_i(1)$ set to the measured value $X_i(1)$.

Note, by the way, that this procedure will be no different from regression if there is no variable which appears both as a dependent variable and as a (lagged) independent variable; however, such a situation is impossible, for all practical purposes, for models capable of yielding long-range forecasts.

There is one big problem with the compromise method: we need a specific procedure to tell us how to pick w, the filtering constant. This is exactly like the situation with ridge regression, where one has to specify a special parameter k.

The compromise method, in general, has several important advantages. First of all, the filtering equation cited above is just the Kalman filtering equation, specialized to the simplest kind of process—a process where every variable

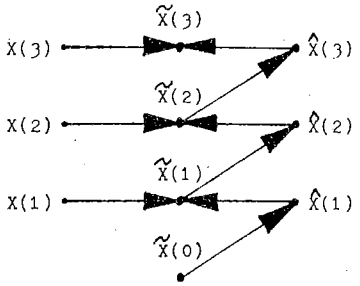

Fig. 2. Order of calculation of $\tilde{X}_i(t)$ and $\hat{X}_i(t)$.

is independent of the others and where the parameters of the system do not change.[1] This equation is well known in applied mathematics to yield the best possible estimates of the true values of the underlying variables; also, to minimize error in this fashion is consistent and efficient, even from the viewpoint of classical maximum-likelihood theory, when it is appropriate to filter. In this approach, we would pick w to minimize a maximum-likelihood measure of error; in other words, we would simply add w to our list of parameters to be estimated. Filtering is appropriate whenever there is likely to be transient error—either measurement error, noise, or imperfect relation between concept and indicator—in the available data. Certainly, this is always true for political data! Economists have made many suggestions, some generalized and some ad hoc, for coping with "errors in data" such as this; however, it has long been known in applied mathematics that filtering is the rigorous solution to the problem of random measurement noise,[2] according to maximum-likelihood theory. The Box–Jenkins and Hibbs methods are essentially smaller steps in this direction. Note, by the way, that our simple w filtering equation is similar to the Mosteller–Busch equation to describe learning in animals, an equation which has often been tried out in the field of artificial intelligence.

Second of all, if we set $w = 0$, the compromise method is equivalent to what we have called the pure robust method. In regression, one picks parameters of the model so as to minimize "error" defined as the gap between actual values and predicted values which were predicted from actual values in the immediately preceding time period; in effect, one minimizes errors in *short-range* prediction, prediction from actual values in one period up to the very next period. In the pure robust method, one does the opposite. One minimizes the errors in *long-range* prediction. To evaluate the errors of one's model, one first constructs a stream of predictions (see Fig. 2, but remember that $w = 0$), starting from estimated data in the first period *without* accounting in any way for the measured data in later periods; one then compares these predictions for the whole sweep of history against the measured data. If we plan to *apply* a model by generating long-range trajectories of predictions (as with the U.S. government use of the CACI *J*-5 model), then it makes sense to *estimate* the model so as to be sure that the long-term trajectories it *would* have predicted in the past *did indeed fit* these data in the past. In regression, as in other forms of maximum-likelihood estimation, one argues that minimizing the short-run errors will maximize the probability that the model is completely "true" (assuming a "flat prior"). In the pure robust approach, we deny the possibility that a statistical model will be completely "true;" instead, we pick the model which does best *over past data* what we want it to do over the future: minimize the separation between *long-range* trajectories of predicted and actual data. The pure robust method does have some relation to simple extrapolation in the univariate case; however, when it is applied to multiequation models, it is considerably *less* trivial than regression to estimate. The pure robust method is the same as the method used by Gillespie *et al.* [12] in their successful studies of arms race models. It is also the method we used before in predicting national assimilation and population variables; in that study, the long-range errors over time intervals averaging about 30 years were reduced by about half using either regression or Box–Jenkins methods. Furthermore, it is equivalent to the ad hoc estimation procedure used by Penner and Icerman in forecasting energy demand and by K. Hubbert in forecasting energy supply, both discussed in [13]. Those authors seemed disappointed that regression did so poorly by comparison for their models, but did not appear to realize that their ad hoc procedures could be generalized and made available as a standard method for nonmathematicians.

The paragraph above seems to imply that regression is better for short-range prediction, while the pure robust method is better for long-range prediction. However, in our earlier work, we noted that regression will often give even the wrong signs for small but critical feedback terms, terms which dominate the long-term dynamics of a system (see the "rates of assimilation" with regression for the Deutsch–Solow model, tabulated in [2, ch. VI]). For example, consider the model

$$\text{population } (t+1) = \text{population } (t) + g^* \text{ population } (t)$$

where g is the growth rate of population. In regression, the standard error of a term such as g might well be 0.02 or so—large enough to flip the estimate from $+0.01$ to -0.01. The impact of such an error on *short-range* prediction would not be very great. Therefore, the random factors and the standard errors would be large compared with the coefficient, and we may expect very small t ratios (circa one) for normal data samples. Not only did we see this in our earlier work on national assimilation and population variables, but we have also seen the same effect at work, even more strongly, with the *J*-5 model, although not for the particular example of population. The t ratios were down in the range of one or two for many critical predictive factors; while this may indicate a coefficient significantly different from zero, it also indicates a very large expected error in our parameter estimates. On the other hand, errors in such a key

[1] See [11, p. 361]. In the univariate case, the strange matrices there are just scalar constants, and the formula reduces to ours quite simply.

[2] From filtering theory, as discussed in [11], we know that the probability of $x(t)$ conditional upon the estimates at time $t-1$ should be independent of all prior measurements, with a "lag = 1" model and correct filtering. This implies that overall log probability as given by the model may be decomposed into the sum of squared error. A given w corresponds to certain values of the noise parameters which would require this value of w; thus all parameters may be estimated in this way.

parameter as g, above, would lead to very large *cumulative* errors in forecasting; if $g = -0.01$, the long-range forecasts would spiral down to zero while the actual values grow. Because the pure robust method minimizes the gap between long-range forecasts and actual values, it would minimize errors of this sort; it will have a much narrower distribution of possible parameter estimates. In other words, it will be much more efficient and reliable in estimating critical parameters. Instead of being better *merely* at long-range prediction, it would *also* be more accurate in estimating the parameters of the model; one may expect more accurate short-range forecasts, as well, if this reasoning is correct. This reasoning assumes, of course, that the parameters do have "true values" which apply to short-range prediction and long-range prediction, both, even though the model itself is not "true" in the strong sense demanded by maximum-likelihood theory (i.e., that all noise is random and normal, etc.). The all-pervasive presence of dirty noise is what makes the robust method more efficient in parameter estimation.

With $w = 0$, however, the model predictions may always deviate so far from actual data that the efficiency of the model is lost. This seems clear from the theory, but we have not documented the effect empirically until now. Late in October 1977, we formulated a new bias method—based on minimizing error variance divided by $(1 - |w|)^2$—which generalizes the notion of minimizing long-range prediction errors directly to the case of highly stochastic systems. (This renders obsolete the r^2 and r bias methods discussed before.) In effect, this method replaces the key assumption of maximum-likelihood methods, that errors are always independent and random, by Murphy's law, that errors will always find a way to accumulate if they possibly can. The details are discussed in [1, Appendix A].

In all robust estimations reported here, we have carried one step further the idea of measuring "error" in a way which reflects the prospective application. Instead of adding up the logarithms of error variance across different variables, we have tried to minimize the sum of "percentage error," defined as error variance divided by the variance of the variable times $(1 - |w|)^2$. In retrospect, our October 1977 analysis also suggests a further step: to add up the square roots of the terms we added up, in order to arrive at a more practical loss function. This probably would have avoided the "fluke" case of population, a variable which was relatively disregarded when we used the square loss function.

Let us emphasize that these new loss functions specified by the bias method will still encourage us to pick the ordinary parameters of a model so as to minimize error variance; however, when w is added as a parameter of the model, these formulas "bias" us towards a much smaller value of w than we would have picked if we had been engaging in maximum-likelihood filtering.

E. Alternative Methods: Conclusions from the Tables

First of all, Tables I–IV make it very clear that our theoretical reasoning was right even for such "relatively deterministic" processes as the one we have studied here. The pure $w = 0$ method did break down, in comparison with the new bias method, and even in comparison with regression, all across the board.

Second, we have demonstrated the efficiency of the bias method as a *different approach to estimation*, not as a form of filtering which can be understood within the conventional maximum-likelihood approach; in other words, it is truly a robust method, in the sense defined by Mosteller and Tukey. If the bias method were better only because it is a back-door way of generating good filtered estimates of the variables, then forecasts based on the *estimated* data for 1961 ought to perform better than forecasts based on the *actual* data. However, if our theoretical argument were correct, then the filtering equation ought to be interpreted *not* as part of the model, but *rather* as a device to give us more accurate parameter estimates. Certainly, in the $w = 0$ extreme, it is clear that we would risk very large significant errors if we used the model on the estimated 1961 data instead of the actual data; the "estimated true values" are just scaffolding that we use to help us in estimating the parameters. The empirical results in our tables support our theoretical argument. When the bias method is used on *actual* data in 1961, it performs substantially better than it does on estimated data; in other words, the errors in the columns for "Bias(actual)" are smaller than those for "Bias(estim.)." The improvements given by bias do *not* follow the pattern one would have expected if bias were merely a watered-down version of maximum-likelihood filtering. Note that we are not attacking the value of classical filtering here; indeed, we have set up our computer package to allow maximum-likelihood filtering estimation, robust estimation, or a combination of the two. We are exploring the value of our new bias method for models which *may or may not* have filtering equations within the model proper; when such equations are justified, they are another addition above and beyond the w filtering equations used by bias. Classical filtering and bias are not two competing alternatives, but rather two complementary methods; the former can add more sophistication to one's model proper, while the latter is strictly an approach to estimating such models.

Finally, there is one other aspect of filtering which is not entirely clear from the theory *a priori*. What should we do when some of the model equations predict variables at time $t + 1$ as a function of other variables *also* at time $t + 1$? In computing our predictions of the former variables, should we insert the *estimated* or the *predicted* values of the latter variables into the model equations? Maximum-likelihood theory clearly favors the former strategy (i.e., using the estimates); it favors "filtering in series." In series filtering, one filters *each* variable in turn immediately after we predict its value and assess the error in this prediction. On the other hand, with parallel filtering, one waits and filters *all* the variables together after one has processed an entire observation's worth of predictions. Consider the following example. If one variable $y(t + 1)$ can be predicted very well from $x(t + 1)$, but neither can be predicted well from data at time

t, then *series* filtering will involve low errors in predicting y (i.e., it predicts from the observed $x(t+1)$), while *parallel* filtering will indicate high errors (i.e., the forecast for $y(t+1)$ is based on observed data at time t, which is used to make an (inaccurate) *prediction* of $x(t+1)$). In this example, the errors in predicting $x(t+1)$ and in predicting $y(t+1)$ from data at time t are clearly highly correlated with each other; this correlation is what makes the errors with sequential filtering appear low. In fact, in full-information maximum likelihood, cross correlations between prediction errors for different variables would reduce the determinant of the error matrix and require at least as much "credit" as series filtering gives to such cases.[3]

Because of maximum-likelihood considerations, series filtering has been built into our model-analysis program. However, the *robust* philosophy permits an argument in favor of parallel filtering. When errors are correlated across variables, this is analogous to our concerns about errors accumulating across time. The robust approach would suggest that we look at the "complete trajectory" of forecasts, from t to $t+1$, across all variables and then assess error and filter *later* (this is also the only practical approach for the implicit models so popular in economics). Therefore, to see if this idea had any merit at all, we tested out parallel filtering with $w = 1$ (see the column labelled "p.f." in Tables I–IV) to see if it would compare well with $w = 1$ and series filtering. Empirically, this method did perform very well in reducing error in comparison with regression; it did about as well as bias did. Nevertheless, more research into this issue is needed; it is conceivable that when w is chosen by bias and when the loss function chosen is based on the square root of the terms we now use, the advantage of parallel filtering will disappear.

Note that bias and parallel filtering are *independent* strategies for improving on different aspects of $w = 1$ maximum-likelihood estimation. The two strategies are in no sense competitive. They exploit different effects in order to get better estimates, and it is a straightforward matter to use both together. In retrospect, we should have done this. Recently, the *combination* of bias and parallel filtering (with the square-root loss function) has indeed turned out to be synergistic in estimating a more refined model of the U.S. economy, based on a model suggested by Kuh, on National Bureau for Economic Research (NBER) quarterly data. Median reduction in forecast error is about 50 percent, and for the better half of the variables, median error reduction was 75 percent (i.e., a *factor of four* reduction). Forecasts were made for twelve quarters ahead from the base year. Note, however, that the use of implicit models or of parallel filtering now requires some programming knowledge on the part of the user, at least for complex models; they require that the user edit the output of our "model compiler" routine. Also note that the issue of parallel filtering is irrelevant for models which predict variables at time $t+1$ as a function of *previous* values of endogeneous variables and as a function of exogenous variables; such models have the advantages of parallel filtering automatically built in.

Tables I–IV contain figures for one other method: $w = 1$ with conventional (series) filtering. Normally, this method is *exactly* equivalent to regression. In this case, however, we did "solve" the equations of the CACI model, algebraically, so that the model would be "explicit." This is necessary in using our computer package as it stands. We actually had to solve only one equation here, the equation for GDP as a function of past values of other variables. This led to a slight perturbation of the economic part of our model, since we did ask for a comparison between the actual and predicted values of GDP; however, as Table I makes clear, the errors with $w = 1$ are almost identical to those with regression, as we might expect. The differences, although very small, do appear to be very significant statistically; this implies a very tiny improvement in the model, an improvement which is nevertheless fairly uniform across all prediction targets. Note that this inclusion of the GDP in the $w = 1$ model yields estimates based on assumptions very close to those of full-information maximum likelihood, the ideal case which two-stage least squares is supposed to approximate, according to Johnston [14]. When representatives of CACI presented this model at the 1976 meeting of the International Studies Association, they stated that CACI had published results based on ordinary least squares instead of two-stage least squares because they found little or no improvement when they tried the latter.

Before discussing other aspects of the $w = 1$ model, we should mention one other conventional econometric method which we did not have time to evaluate here: the Aiken or Cochrane–Orcutt procedure. This procedure, despite its popularity, is basically just an autoregressive method. Its results depend only on the matrices of correlations with lags one and two and cannot exploit the higher order cumulative effects which the bias method exploits. The Aiken approach often has value, as a way of formulating what amounts to a more sophisticated model, but to estimate this model in the conventional manner leaves one open to all the same hazards as with regression. Even if the "errors in data" took the form of classical white noise, they would convert the observed data into a Box–Jenkins process, not an Aiken model; if the Aiken procedure is too close to regression even for the case of white noise, then there is little basis for expecting noteworthy robustness in handling the dirtier real-world forms of noise.

There is one big exception to our statement that $w = 1$ yielded almost the same errors as regression: the case of population forecasts in our tables, where $w = 1$ is far worse. This we have traced to a different estimate of population growth. Yet mathematically, the estimate with $w = 1$ should be exactly the same. In going over the computer runs, we

[3] See discussion of FIML in [14]. With the Wishart distribution, estimating the true error covariance matrix as the sample covariance matrix leaves the determinant which they share as the key factor determining likelihood. In a simple two-by-two positive symmetric matrix, it is obvious that big off-diagonal terms—either positive or negative—will reduce the determinant.