June 1989

Toward computer-aided forecasting systems: gathering, coding, and validating the knowledge

Fred Collopy
University of Pennsylvania

J. Scott Armstrong
University of Pennsylvania, armstrong@wharton.upenn.edu

Postprint version. Published in George R. Widmeyer (ed.) *DSS-89 Transactions: Ninth International Conference On Decision Support Systems*, Institute of Management Sciences, 1989, pages 103-119.

Toward computer-aided forecasting systems: gathering, coding, and validating the knowledge

Abstract
Direct assessment and protocol analysis were used to examine the processes that experts employ to make forecasts. The sessions with the experts yielded rules about when various extrapolation methods are likely to be most useful in obtaining accurate forecasts. The use of a computer-aided protocol analysis resulted in a reduction in the total lime required to code an expert's knowledge. The implications for overcoming the "knowledge acquisition bottleneck" are considered.

Comments
Postprint version. Published in George R. Widmeyer (ed.) *DSS-89 Transactions: Ninth International Conference On Decision Support Systems*, Institute of Management Sciences, 1989, pages 103-119.

Published in *DSS-89 Transactions: Ninth International Conference On Decision Support Systems*, ed. George R. Widmeyer, Institute of Management Sciences, 1989, pp. 103-119.

Toward Computer-Aided Forecasting Systems: Gathering, Coding, And Validating The Knowledge

Fred Collopy and J. Scott Armstrong
The University of Pennsylvania

Direct assessment and protocol analysis were used to examine the processes that experts employ to make forecasts. The sessions with the experts yielded rules about when various extrapolation methods are likely to be most useful in obtaining accurate forecasts. The use of a computer-aided protocol analysis resulted in a reduction in the total lime required to code an expert's knowledge. The implications for overcoming the "knowledge acquisition bottleneck" are considered.

There are two sides to the problem of delivering knowledge to those who require it. One is getting the expertise into the delivery vehicle, whether it be a book, a presentation, or a computer program. The other side is getting it out. In getting it in, we must deal with several tasks, including acquiring it, coding it, and validating it. This paper deals mainly with acquiring knowledge but, as all of the other aspects of the problem interact with knowledge acquisition, they are addressed as well. We start by positioning this effort in reference to the problem of making accurate business forecasts. Then we describe a pilot study in which two approaches to knowledge elicitation are compared. Coding and validation procedures are then discussed. Finally, we consider implications for further research.

The Problem of Selecting a Forecasting Model

Several empirical studies of forecasting accuracy have been published along with animated discussions and commentaries. One of the themes that punctuates those discussions is that the relative accuracy of forecasts depends on the joint characteristics of the series predicted and the method used (Makridakis and Hibon 1979; Hogarth 1979; Lopes 1983). Gilchrist (1979) proposed that automatic forecasting methods might be improved if a method could be devised for characterizing sets of data.

Makridakis et al. (1982) examined the relationship between model performance and certain characteristics of the time series. The researchers classified the data by subcategories that separated yearly, quarterly, and monthly data, micro and macro data, industry and demographic data, and seasonal and nonseasonal data: While differences in the average rankings of model performance were not statistically significant as far as all data are concerned, the differences were significant when subcategories of data were considered (1982, p. 140). Model performance is dependent in part, they suggest, on the type of data being forecast.

Our research is aimed at producing a better understanding of which method is best for a given set of data. We do this by seeking the knowledge of expert forecasters and forecast researchers, building models to make the knowledge explicit and operational, and validating those models empirically. This will, in turn, aid efforts to develop computer-aided forecasting systems that can assist decision-makers in selecting and employing appropriate methods for various forecasting situations.

In choosing among the available methods, we expect experts in forecasting to employ heuristics, or rules of thumb. These heuristics, if properly coded, might improve selection among candidate extrapolation models. Consistently applied models, based upon roles, can outperform the experts who formulate them (see Armstrong 1985, pp. 274-284 for a summary of much of the evidence).

Knowledge Elicitation

There are two basic sources of knowledge about forecasting: empirical studies and forecasting experts. The growing body of empirical literature on forecasting provides numerous guidelines for selecting among forecasting methods. With meta-analysis, one would use this prior research to test various possible selection rules. Still, much remains to be studied, so one must also rely on the judgment of experts. One way to learn from experts is to have them complete a structured questionnaire. Another way is to conduct in-depth interviews in which they are asked what they do. A third way is to observe them as they make forecasts. In the sessions we describe below, the last two of these are used: in-depth interviews and observation of experts performing forecasting tasks.

One approach to knowledge acquisition is to ask experts to specify the rules that they use. To do this, we used an in-depth interview which we refer to as the direct assessment method. Evidence that this approach can result in models that successfully replicate experts' decisions is provided in the areas of financial forecasting by Larcker and Lessig (1983). This approach has the advantages that it is less expensive than other approaches, it is less likely to be subject to researcher bias, and it has the potential to capture most of the relevant knowledge that an expert possesses. A limitation is that this method is relevant only to situations where the expert has a clear awareness of the problem-solving process.

For observation, we used protocol analysis. That is, we asked the experts to describe what they were doing as they solved actual problems. Protocol analysis is particularly useful where awareness of the problem-solving process is low.

Knowledge acquisition is a time-consuming and costly process. It is frequently identified as the bottleneck of expert systems development (Buchanan et al. 1983; Feigenbaum and McCorduck 1983; Hayes-Roth et al. 1983; Lenat 1983; Fellers 1987). Consequently, studies of expert knowledge tend to rely upon small numbers of subjects. Clarkson's (1962) study of stock selection had a single investment trust officer as its subject. More recently, Grabowski (1988) used three subjects in her study of boat pilots. One focus of our work is to embed the knowledge acquisition process in a decision support system. This will allow us to elicit the knowledge of a substantial number of forecasting experts. In the next section, we describe an experiment that used a first version of the decision support system.

Experimental Procedure

Given the contexts in which protocol analysis is typically used, it is necessary to define exactly what we are looking for. It is useful to make a distinction between an "expert modeling system" and an "expert performance system" (Chambers, Gale and Pregibon 1988). Much of the work for which protocol analysis is used is in the former category-researchers are attempting to develop or validate a model of how experts do something. The current work has no such objective. We are not so much interested in building a model of what the experts do as we are in having them aid us in developing a useful model. It is their rules that we hope to elicit. It is expected that there will be ambiguity and conflict among these rules (both within and among experts). By building programs to test and to reason about them, we hope to move toward computer-aided forecasting systems that exhibit expert performance.

Our procedure followed the general lines given by Ericsson and Simon (1984). In particular, our procedure followed their description for concurrent protocols. The subjects talked while they were performing a task. When they became quiet, the interviewer prompted them to talk as the thought. During the session, the interviewer took notes. He then expanded these notes while listening to tapes of the sessions. Next, he coded the sessions. In doing the coding, he focused on the heuristics and procedures that the subjects employed. The descriptions were then given to the subjects to determine whether they agreed that the summaries reflected the approaches they took and the rules they employed in making their forecasts.

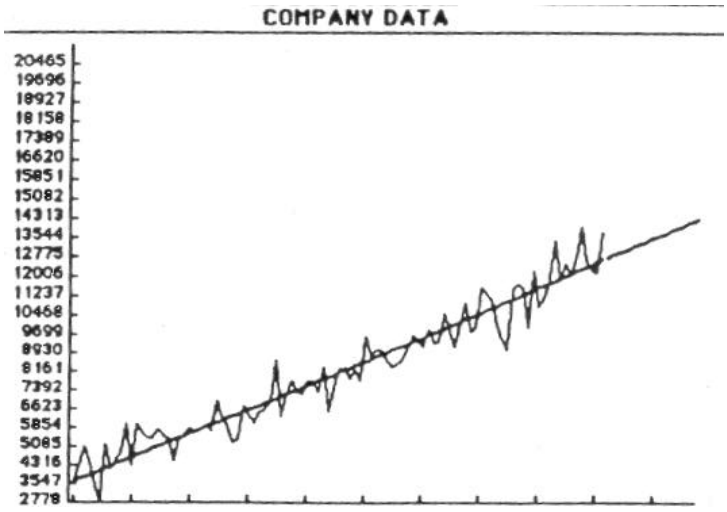
During the sessions, we presented graphic displays of time series to experts in forecasting. The experts were asked to describe what methods would produce the most accurate forecast by the Mean Absolute Percentage Error (MAPE) criteria for various forecast horizons. They were asked to think aloud while considering this and to identify any characteristics of the series they considered or rules they employed. Once they had provided an initial selection and the reasoning for it, the experts were permitted to examine the forecasts by various methods. They

were free to modify their choice of methods and rules in light of this information. Feedback was provided on the performance of their selected method.

The issue of feedback is one that must be addressed in future research. There were instances in the sessions of experts anchoring their assessments of new series on the basis of their performance on a series they had examined earlier in the session. That made little sense to us. The program has been revised to allow us to provide either immediate feedback, as we did in this study, or delayed and summary feedback. This is an area that warrants further research. Armstrong (1985, pp. 380 ff.) summarizes the conflicting evidence of the relative value of summary and case-by-case feedback to learning.

In addition to providing a graphic display of time series, the decision support system allowed subjects to deseasonalize the data and examine each method's fit of the data and its forecasts. Since one of the methods is regression against time, it was easy to examine a least squares fit of the data. Exhibit 1 provides an example. The data being forecast (the holdout data) were not available to the experts).

Exhibit 1. Display Showing a Least Squares Fit of a Deseasonalized Series,



Two experts in forecasting completed both methods. One of them (subject A) first did the direct assessment method, then the protocol analysis. The other (subject B) did the protocol analysis session first, then the direct assessment. The order was reversed in order to get a sense about whether there are order effects that should be taken into account in the design of the larger study. No time limits were imposed. The session with subject A was video-taped; due to equipment difficulties, the session with subject B was only audio-taped.

To assess the performance of the subjects, we used series for which benchmark forecasts were readily available. We selected six monthly series from those used in an empirical comparison of extrapolation methods (Makridakis et al. 1982). These series represented a convenience sample of those in the 111 series. Our desire was to have series that looked substantially different from one another. The series used are shown in Appendix B (the identifications used in the M-competition are MRM17, MNF3, MNM15, MNM52, MNM61, and MNM70 for series A through F respectively).

Details of the Sessions

In the following section we describe some of the details of the protocol and direct assessment sessions. Readers without interest in these details may wish to skip to the Discussion section.

Subject A – Direct assessment method

Subject A spent about 35 minutes on the direct assessment method using four major steps during that time. First, he outlined a general approach to doing an extrapolation. It consisted of the following steps:

Preparing the data

- eliminate errors or outliers-this could be done in an automatic system by fitting a trend line, finding confidence intervals, and bringing values into the 95% confidence interval; it should be done before deseasonalizing, but can also be done again after deseasonalizing.

- adjust for known causal effects (e.g. trading days, stockouts, inflation, strikes, etc.) whenever possible.

Seasonality

- determine the number of seasonal cycles

- for series with data on only one cycle, do not use seasonal factors

- if data are available for a number of cycles, increase the emphasis on seasonal factors

- identify regularities or changes in seasonal patterns

- apply Census X-11 decomposition method to deseasonalize

- for series with natural growth, apply multiplicative seasonal factors

- for series that are constrained to values between 0 and 1, such as market share, use additive seasonal factors.

- dampen the seasonal factors for. forecasts, especially for those further out in the future

Level and trend

- estimate level and trend using exponential smoothing (either Holt-Winters' or Brown's model with correction for lag)

Model formulation

- in cases of frequent trend reversals, use a not trend model

- in a region near the previous limit of the data, use a combination of no trend and the previous mean (historical average) from the regression model

- in the presence of uncertainty brought on by short series, irregularities, cycles, or unusual things, use damped trend and no-change models

- when combining, the models should be as different as possible.

Finally, Subject A characterized an ideal extrapolation method. It would have facilities for:

- adjusting for outliers

- using seasonal factors, but modifying their effect

- using a trend factor, but damping its effect

- introducing trend reversals in specified situations.

He identified modified seasonal factors and the trend reversal models as facilities that are lacking in current extrapolation methods. A summary of the session results appears in Exhibit 2.

Exhibit 2. Rules and Procedures. Subject A – Direct Assessment Method.

Preparing the Data
Eliminate errors and outliers
Seasonality
Determine number of seasonal cycles to the data Determine seasonal factors: Apply Census X-11 Identify changes in seasonal patterns Based on horizon: Dampen seasonal factors If natural growth series: Apply multiplicative seasonal factors If bounded series: Apply additive seasonal factors
Level A Trend
Determine level & trend: Apply Holt-Winters or Browns'
Model Formulation
If frequent trend reversals: Apply no trend model If near previous limit: Combine regression & no trend If high uncertainty: Apply damped trend or no change When combining, models should be as different as possible

Subject A – Protocol analysis

In working with the six monthly series, Subject A employed a fairly consistent approach. After he had deseasonalized the data, he assessed the strength and likelihood of continuation of the series' trend. In deciding how to project the trend, he considered several factors. They were:

- the overall trend of the historical series

- the most recent trend (that created by the two latest data points)

- previous interruptions or reversals in the trend

- whether the current value was close to an extreme value in the series.

He then separated the factors as arguments either for or against continuing the trend in future forecasts (he referred to this as the `Ben Franklin approach'). Because the decision support system was limited to the 24 models and combinations of the models used in Makridakis et al. (1982), he combined a trend extrapolation model (Holt-Winters') with a deseasonalized no change model (naive 2) and linear regression to achieve the damping he desired. He determined the amount of damping subjectively, using his assessment of the strength and importance of each of the criteria.

The first series examined was series A (see Appendix B). Subject A indicated that he doubted the analytic result (obtained from ratio-to-moving average) that this was nonseasonal data. He noted that he did not want to pay much attention to the old data, something that exponential smoothing would take care of. The regression, on the other hand, gave equal weight to all data. On looking at a plot of a regression-based fit and forecast, he indicated that he did not believe result. Noting that Holt's looked like a no trend model, he observed that he was leaning toward a no trend model.

The subject then presented a synopsis of what he saw in the data. "It looks like a big decrease, it's been fighting to come back. There's lots of uncertainty. The long term trend is down. The general mean is below the current point. The latest trend is heading down." He believed, therefore, that there was a slight downward trend. He decided that a combination of linear regression and Holt-Winters would capture this. "I'm coming back to the old Ben Franklin approach," he noted. "I have three arguments for a downward trend, none against."

On examining the hold out data, it was evident that the judgment of a downward trend was appropriate. The subject's accuracy was superior to all models compared in Makridakis et al. (1982) except for simple regression against time.

In examining another series, series E, subject A noted it was a series with a multiplicative seasonal pattern, but with atypical things happening with the seasonals. In considering the arguments for extrapolating the trend, he noted that there is a clear, exponentially growing trend, and that the local trend is positive. On the other hand, there are some unsettling points jilt before the end of the series and the current data is at the limits of the historical data. These things made him uneasy. He concluded that the argument for a continuing trend is good but not entirely convincing. Further, there is a great deal of uncertainty at the end of the series.

The forecasts from regression were conservative, so he decided not to include them at all. Instead, he used forecasts consisting of 50% Holt-Winters and 50% no change. The latter he split between seasonally adjusted no change (naive) and a simple random walk model without seasonal adjustment (naive 1)-67% and 33% respectively.

His adjustments helped somewhat. "I got nervous," he noted on seeing the hold out data, "but not nervous enough. Things were changing at the end, and we were at the limit of the data. The fact that the seasonals were going bad should have sent up more red flags."

A summary of subject A's approach to each of the series, as well as the models he used, is given in Exhibit 3. The session took about one hour. The later series took less time (about five minutes) than the early ones (about 13 minutes). In summarizing, subject A observed that his rules were basically rules of conservatism. He pointed out that none of these series would call for a trend reversal.

Exhibit 3. Rules and Procedures. Subject A – Protocol Analysis.
Order in which observations and decisions were made for each of the series.

	Series A	Series B	Series C	Series D	Series E	Series F
Changes & Irregularities						
Old data is not relevant to current pattern	1					
Uncertainty in series	2				7	5
Irregular periods in the data		5				
No reversals or other unusual things			7			
Unsettling periods near end of data series					6	
Type of Trend						
Long series of data with very regular growth		1				
Additive trend		2				
Gradual exponential growth			3			
Direction of General Trend						
Down	3			1		
Up		8			4	3
Direction of Latest Trend						
Down	5	6		3		
Up					5	2
Location of Current Data						
Above previous mean	4	7				
Below previous mean				4		
At limits of previous data						4
Seasonality						
Deseasonalized the data		4	2	4	3	1
Regular seasonal pattern			1			
Changing or irregular seasonals				5	2	
Wide variation in seasonals			4			
Multiplicative seasonal pattern					1	
Model Formulation						
Model slight downward trend	6					
Model damped seasonals			6			
Model damped trend		3	5			
Model Used						
Naive1			30%	10%	16%	33%
Naive2		30%		40%	34%	33%
Linear regression	50%			25%		
Holt-Winters	50%	70%	70%	25%	50%	34%

Subject B – Protocol method

Subject B was asked first to examine some specific series and discuss how he would forecast them. He examined the same six series as subject A. His first rule is one he called the rule of conservatism-a general reluctance to extrapolate trends at their apparent rates, particularly in the presence of noise or uncertainty. He often combined Brown's linear and simple exponential smoothing, sometimes adding naive 2 to further dampen the trend. He usually examined the forecasts produced by each of these three methods "looking for differences." From doing this he got a sense about how much noise there was in the series.

On series A, subject B noted that he was most influenced by the horizontal movement. He discounted the early downward sloping trend (the first two years or so of the data). He indicated that he believed the analytic result that this was nonseasonal data and accepted the movement as noise. He indicated that the would favor simple

exponential smoothing. On applying it, he noted that it looked as if the model were using an alpha level of 1 (puts all of the weight of the forecast on the most recent data point) and said he would accept that. He suggested that he would pull this down, perhaps 5% or so. His forecast, then, was 95% of naive 1 (naive 1 and simple exponential smoothing with alpha = 1 amount to the same thing). He said that he was dropping his forecast slightly because he believed that pessimism is better than optimism. Also, he noted that the middle period of the data appeared to be down.

Subject B told us that he was using the rule that exponential smoothing is preferable to moving averages. He didn't see any indication of trend. "If there is a trend it doesn't have any basis in history. If somebody were to pick a trend model in this case, they'd be guessing."

The literature, he said, suggests that combining methods is better than using any individual method. But in this case he thought that the best models were exponential smoothing and naive 2, and that they are essentially the same in this case.

As with most of the series, subject B first started his examination of series D by deseasonalizing it. He then noted that the series was highly uncertain. From simply examining the plot, he felt that this series could go anyplace. He looked at simple exponential smoothing and naive 2. Since they are essentially the same, either could be used. He noted that he does not consider it necessary to look at the fits. Everything he has seen in the literature indicates that fits are misleading. "I'm using a far more intuitive approach than examining fit."

"I'm not going to do it, but I have this feeling, and I don't know where it comes from-I suspect from the downward slope-that this one is, might be, going back up again. 1 feel if anything, the chance is it's going to rise rather than decline further."

He decided to select a seasonally adjusted no change model (naive 2), though, because of his concern that it looks like a series that's "all over the place." It has, he noted a good deal of variance and changing trends. When asked about analytic methods for identifying such series he cautioned that counting trend changes would not be enough. It is essential to take into account the magnitude of the trends. He thought the human eye might be able to identify them.

The downward trend did continue, so the naive 2 forecast was higher than the actual values. Subject B noted that he is using almost a contrarian strategy. Since this series was going down slightly, he believed that there was some opportunity for it to go up.

Subject B examined only five of the six time series. This was because he noted that the display of the sixth series included white space at the top, suggesting a continued upward trend (this problem is discussed in a later section). A summary of subject B's approach to each of the series, and the models he used, appears in Exhibit 4. The protocol sessions took about 50 minutes.

Exhibit 4. Rules and Procedures. Subject A **– Protocol Analysis.**
Order in which observations and decisions were made for each of the series.

	Series A	Series B	Series C	Series D	Series E	Series F
Changes & Irregularities						
Irregular periods in data	4	2				
Previous leveling of series		6				
High variance				3		
Variance is increasing					3	
Trend changes				4		
Uncertainty in series				2		
Type of Trend						
Overall horizontal movement (no trend)	1					
Exponential growth in series					2	
Direction of General Trend						
Up		3				
Gradual upward trend			2			
Direction of Latest Trend						
Down		5				
Seasonality						
Non-seasonality of series	2					
Deseasonalized the data		1	1	1	1	
Model Formulation						
Trend will turn down		4				
Model slight downward trend			3			
Model no change but discounted a bit	3					
Model damped trend					4	
Model Used						
Naive1	95%					
Naive2		33%		100%		
D. Single exponential smoothing		33%	50%		50%	
D. Brown's exponential smoothing		34%	50%		50%	

Subject B – Direct assessment method

Subject B said he looks first for the presence or absence of trend and for variability regardless of trend. If there's a great deal of variability (if the series looks random-that is, appears to show no clear trend), he would pick a random walk model.

Of the available models, there are some that the would never use. Brown's quadratic smoothing is one of them, because previous experience has inaccurate forecasts, overreacting to small changes in the series proven it produces as though they were changes m trend. He would also eliminate the Bayesian method, AEP filtering, and, to a slightly less extent, adaptive response rate exponential smoothing. He does not believe that automatic methods work. He sees no indication that they do anything but provide a better fit. He eliminated simple linear regression. It simply fits a straight line to the data, not allowing for any changes. He would not use Box-Jenkins, either automatic or manual. He eliminated moving averages in favor of exponential smoothing; which places more emphasis on the recent data, something that is intuitively appealing.

Subject B would usually make choices between simple smoothing and the linear smoothing methods. Brown's and Holt's methods are similar. "For no good reason," he prefers Brown's, although Holt's is more widely used. He indicated that he has more experience with Brown's and that it has worked well for him. He does not deal

with seasonal series, so he has no real experience with Holt-Winters. He would consider damped smoothing, especially for annual series that show a trend. He employs the heuristic that "what goes up, has to come back down, and the faster it goes up, the faster it comes back down." For series that are rising quickly, one is better betting against the trend, rather than with the trend.

The random walk is a good performer. In the lack of any other information or when uncertainty is high, he favors using a random walk. In fact, he noted, exponential smoothing is a random walk when alpha is 1. As the forecast horizon gets longer, subject B tends to gravitate more towards a random walk model. The session lasted about 15 minutes. A summary is presented in Exhibit 5.

Exhibit 5. Rules and Procedures. Subject B– Direct Assessment Method

Level and Trend
Identify presence/absence of trend
Variability
Identify variability
Model Formulation
Choose between simple smoothing R linear smoothing If great deal of variability: Apply random walk model If annual and trend: Apply damped smoothing For rapidly rising series: Apply no trend or damped trend When uncertainty high: Apply no change In absence of other Information: Apply no change As horizon gets longer: Move toward no change

Discussion

The study sample is too small and the methods not yet sufficiently refined to allow drawing strong conclusions. Nevertheless, much was learned from this pilot study.

Comparison of the two methods

The direct assessment method took less time than the protocol method did (an average of 23 and 55 minutes, respectively). To summarize the direct assessments took about an hour each. To summarize and code the protocols took about four hours each. Although the computer-aided protocols were more time consuming than the direct assessments (5 hours vs. 1.5 hours), the total time required for them seems to be small relative to what is suggested in the general protocol analysis literature where the process is frequently referred to as the "knowledge acquisition bottleneck."

The time spent in the protocol sessions was greater for the first few series than for the later series. It appears that a shorthand developed, as did comfort with the decision support system. The sessions have helped us to improve the decision support system. For example, experts will now be able to view combined forecasts during their sessions. We also automated many of the tasks that the interviewer and the subjects had to do manually.

Direct assessment provided some rules that did not arise during protocol sessions. Four of the rules that subject A identified during the direct assessment session did not arise during the protocol session. This occurred because the protocol sessions failed to use a broad enough variety of forecasting situations. An example is subject A's rule that additive seasonal factors should be used on bounded data (no bounded data were included). We are

currently in the process of identifying additional sources of data. For example, we have annual data on naval reenlistment rates; these are bounded and have shown a tendency to regress to the mean.

Only the protocol sessions provided information about weights that the subjects attached to the various features they identified. The protocol method, by virtue of the repetitive nature of the task, also furnished information about how frequently the forecaster considers various factors. For example, subject A often noted the direction of the latest trend, and both subjects began the analysis by deseasonalizing the data for most of the series.

Changes and irregularities in the series received much attention in the protocol sessions of both subjects. On the other hand, except for a few general comments about dealing with outliers and cleaning up the data, they were not discussed during the direct assessments.

The decision support system promises to greatly reduce the cost of doing protocol analysis. By reducing the time it takes experts to produce a forecast. we can increase the number of forecasts they can do. If we reduce it so that they can do dozens of series in a one- or two-hour session, we can develop a significant body of expert experience. This will greatly increase the value of protocol analysis for this kind of task.

Media Considerations

We thought that video might prove useful in capturing reference to particular series and parts of graphs. This information, however, was always adequately identified on the auto tape, which is less expensive to use and somewhat easier to code.

Our graphical presentation of the data led to certain problems. Some series were plotted with white space at the top because the y-axis was scaled before the holdout data were removed from the series. Subject B mentioned this while completing his fifth series. While it does not appear to have affected either subject prior to this point, the possibility must be allowed for. The discovery of such embarrassments as this is one of the purposes of pilot studies. The purpose of relating it is to spare others a similar "discovery." The decision support system has been changed in order to provide a similar amount of white space at the top and bottom of each historical series.

Performance of the Experts

We are concerned in this study with the rules that experts are using in producing forecasts, not in the forecasts themselves. Future studies will examine how well the experts' rules do in terms of predictive validity. To illustrate the validation process, though, we examine the forecasts that the experts produced in the process of the protocol sessions in this pilot study.

To maintain consistency with the analysis in the M-competition, the mean absolute percentage errors (MAPEs) averaged over the 18 periods being forecast are reported (see Exhibit 6). Unfortunately, the series in which the MAPEs are large will have a greater impact on the average than ones with small MAPEs. For that reason, we also present the data using relative MAPEs (see Exhibit 7). There, the MAPE of each method is related to that of Naive 2 for the same series. The Spearman rank correlation coefficient for the two accuracy measures was 0.724.

Exhibit 6. Performance of Various Forecasting Methods.
Average MAPEs (Mean Absolute Percentage Errors) for Forecasts of Horizons 1 to 18

	Series A	Series B	Series C	Series D	Series E	Series F	Average
D. Simple regression against time	12	1	14	10	23	12	12.0
Combining B	29	2	17	6	35	8	16.2
D. simple exponential smoothing	29	2	17	6	30	12	16.3
Combining A	30	1	12	7	41	7	16.3
Subject A	20	1	15	12	42	9	16.5
Subject B	22	2	14	12	41	*	16.7
D. ARR exponential smoothing	30	2	20	8	23	17	16.7
D. Brown's exponential smoothing	27	2	10	5	53	5	17.0
Naïve 2 (D. Random walk)	29	2	13	12	36	13	17.3
Holt-Winters' method	29	5	14	7	46	5	17.5
Automatic AEP filtering	34	2	8	12	44	6	17.7
D. Holt's exponential smoothing	39	3	11	8	54	4	18.2
D. Moving Average	39	2	19	16	30	14	18.3
D. Quadratic exponential smoothing	22	12	8	12	60	3	19.5
Bayesian method	32	8	8	11	73	3	22.5
Box-Jenkins	37	4	6	7	78	14	22.6
Average	27	3	13	9	44	9	17.6

D = Deseasonalized

* Assumes an average MAPE for series F

Exhibit 7. Performance of Various Forecasting Methods.
Relative MAPEs (Mean Absolute Percentage Errors) for Forecasts of Horizons 1 to 18

	Series A	Series B	Series C	Series D	Series E	Series F	Average
D. Simple regression against time	0.41	0.50	1.17	0.83	0.64	0.92	0.70
Combining B	0.93	1.00	0.83	0.42	1.47	0.38	0.75
D. simple exponential smoothing	1.03	0.50	1.00	0.58	1.14	0.54	0.75
Combining A	0.69	0.50	1.25	1.00	1.17	0.69	0.84
Subject A	1.00	1.00	1.42	0.20	0.97	0.62	0.87
Subject B	1.17	1.00	0.67	1.00	1.22	0.46	0.87
D. ARR exponential smoothing	1.00	1.00	1.42	0.50	0.83	1.08	0.93
D. Brown's exponential smoothing	0.76	1.00	1.17	1.00	1.14	*	0.94
Naïve 2 (D. Random walk)	1.00	2.50	1.08	0.58	1.28	0.38	0.96
Holt-Winters' method	1.03	1.00	1.67	0.67	0.64	1.31	0.99
Automatic AEP filtering	1.00	1.00	1.00	1.00	1.00	1.00	1.00
D. Holt's exponential smoothing	0.76	6.00	0.67	1.00	1.67	0.23	1.03
D. Moving Average	1.10	4.00	0.67	0.92	2.03	0.23	1.04
D. Quadratic exponential smoothing	0.93	2.00	0.50	0.58	2.17	1.08	1.04
Bayesian method	1.00	1.50	0.92	0.67	1.50	0.31	1.07
Box-Jenkins	1.00	1.00	1.58	1.33	0.83	1.08	0.00
Average	0.92	1.04	1.13	1.08	1.06	0.70	0.89

D = Deseasonalized

* Assumes an average MAPE for series F

While the results of a small sample of series are not statistically significant, they do provide a sense of how we can validate the rules. The subjects performed well, relative to the other methods. The combinations made by the two experts did *not,* though, lead to results superior to those derived from the mechanical combinations.

We had assumed that the application of heuristics to model formulation would be a productive avenue of pursuit. Interestingly, the only other methods in which human intervention was employed (Bayesian and Box-Jenkins) were among the least accurate on these six series (Makridakis et al. 1982). Those methods used human judgment directly to manipulate the model, while ours used it primarily to select among models. We do not know from this small sample whether this is an important distinction in how judgment should be used.

Simple regression against time performed unusually well on these series. It did not perform as well in the study of 1001 series in Makridakis et al. (1982). This suggests that our selection of series may have favored ones that show a strong overall trend. Clearly a larger number and a wider variety of series are desirable for testing.

Some Heuristics for Further Research

One of our major objectives is to capture knowledge that can be coded as rules. Some rules that emerged warrant further examination. These include:

1. reduce the trend component if there have been frequent interruptions or reversals of trend;

2. reduce the trend component if the most recent point is near one of the extreme values in the historical data;

3. reduce the trend component if the direction of the most recent trend is opposite to that of the long term trend;

4. reduce the trend if it has been rising rapidly;

5. reduce the seasonal factors if the seasonals appear to be changing;

6. reduce the trend component as uncertainty increases;

7. reduce the seasonal component as uncertainty increases;

8. reduce the trend component as the forecast horizon lengthens;

9. reduce the seasonal component as the forecast horizon lengthens;

10. in the region near a previous limit of the data, use a combination of no trend and the previous mean (historical average), or use the mean of similar data.

We will code these and similar heuristics. The coded rules will then be presented to the experts along with their performance on additional time series. The experts will be invited to refine their rules. The process will be repeated until there is no longer any significant change.

We plan to examine extensions to different types of time series, in particular to bounded series and series for which some extrinsic information is available. Both of the experts in our study indicated that additional information about each series would be useful. "Of course, I'd want to know what the company makes," observed subject B. Subject A commented that knowledge of planned promotions, incentives and similar things would condition his forecasts if the data related to sales.

To be useful, the rules must be broadly applicable. The validation process must. therefore, include the application of the rules to series beyond those used in their formation.

Once the individual heuristics are coded, it will be necessary to resolve conflicts and inconsistencies among them. An attempt will be made to find a set of "consensus rules" from the experts. To do this, it will be necessary to develop methods for resolving differences in their sets of rules.

Another objective of this study is related to the development of a low-cost approach to knowledge elicitation. Consistent with prior research, the results suggest that protocol analysis is a useful technique for eliciting the kinds of rules that we require. Reducing the cost of protocol analysis is, therefore, an important direction for future research efforts. User-interface and decision support issues are at the center of this task. To explore them, we are developing a demonstration program. In addition to serving as the knowledge elic itation system, it serves as the test and validation system for the coding of the heuristics.

Conclusions

This pilot study suggests that it is feasible to do useful protocol analysis of forecasting experts. The use of interactive computing allows the process to be structured so that expert experience can be gathered rapidly. It required only about five hours to collect and summarize the results from an expert.

Protocol analysis was more useful than direct assessment in providing information that is codable. However, the protocol rules were incomplete because we failed to provide enough variety in the series.

Audio, due to its lower cost, portability, and easier use, was superior to video in recording the process.

Once we have a set of coded and validated rules, we can write computer programs to reason about them. Can the rules of thumb of forecasting experts outperform statistically sophisticated forecasting methods (Box-Jenkins, Bayesian forecasting), automatic methods used statistically optimized parameters, and combinations of methods using automatic weighting? If so, how do they do it? If not, why not? In reasoning about the heuristics of forecasting experts we will be able to answer such questions.

Acknowledgments. We thank the Navy Personnel Research and Development Center, San Diego, CA, and the U.S. Coast Guard Research and Development Center, Groton, CT (contract number DTCG39-86-C-80348) for partial funding of this project. Data were provided by Spyros Makridakis, Michelle Hibon, and Everette Gardner. We also wish to thank Steven Kimbrough and Steven Schnaars for their useful comments on earlier drafts, and Chris Chatfield for his suggestions regarding comparative metrics.

Appendix A. Instructions for the direct assessment method.

Make a list of all of the rules that you employ (or might employ) when you must select an extrapolation method to forecast a monthly time series. In particular, assume that you are presented with a time series like one of those found in the M-Competition and that you are provided with:

- the list of historical data points
- a plot of the historical data
- the period of the data (monthly, quarterly, or yearly)
- the beginning and ending dates of the data series
- the seasonality of the data (none, or of period four or twelve)

- the general source of the data (macro, micro, firm, industry, nation)
- a brief description of the data (e.g. `Aluminum production Netherlands')
- plots of the least square fit of any of the available extrapolation methods

Indicate what rules you would employ to select for accuracy (using mean absolute percentage error of the forecasts) among the following extrapolation methods:

- D. Random walk (Naive 2) forecast
- D. Moving average
- D. Single exponential smoothing*
- D. Adaptive response rate exponential smoothing*
- D. Holt's (two parameter) exponential smoothing*
- D. Brown's linear (two parameter) exponential smoothing*
- D. Brown's quadratic (three parameter) exponential smoothing
- D. Linear regression trend fitting (against time)
- Holt-Winter's linear and seasonal exponential smoothing*
- Automatic AEP filtering (Carbone- Longini)*
- Bayesian method (Harrison-Stevens)
- Box-Jenkins method
- Combining of six methods indicated about (*)-equal weights
- Combining of six methods indicated above (*)-weighting based on errors of model fitting

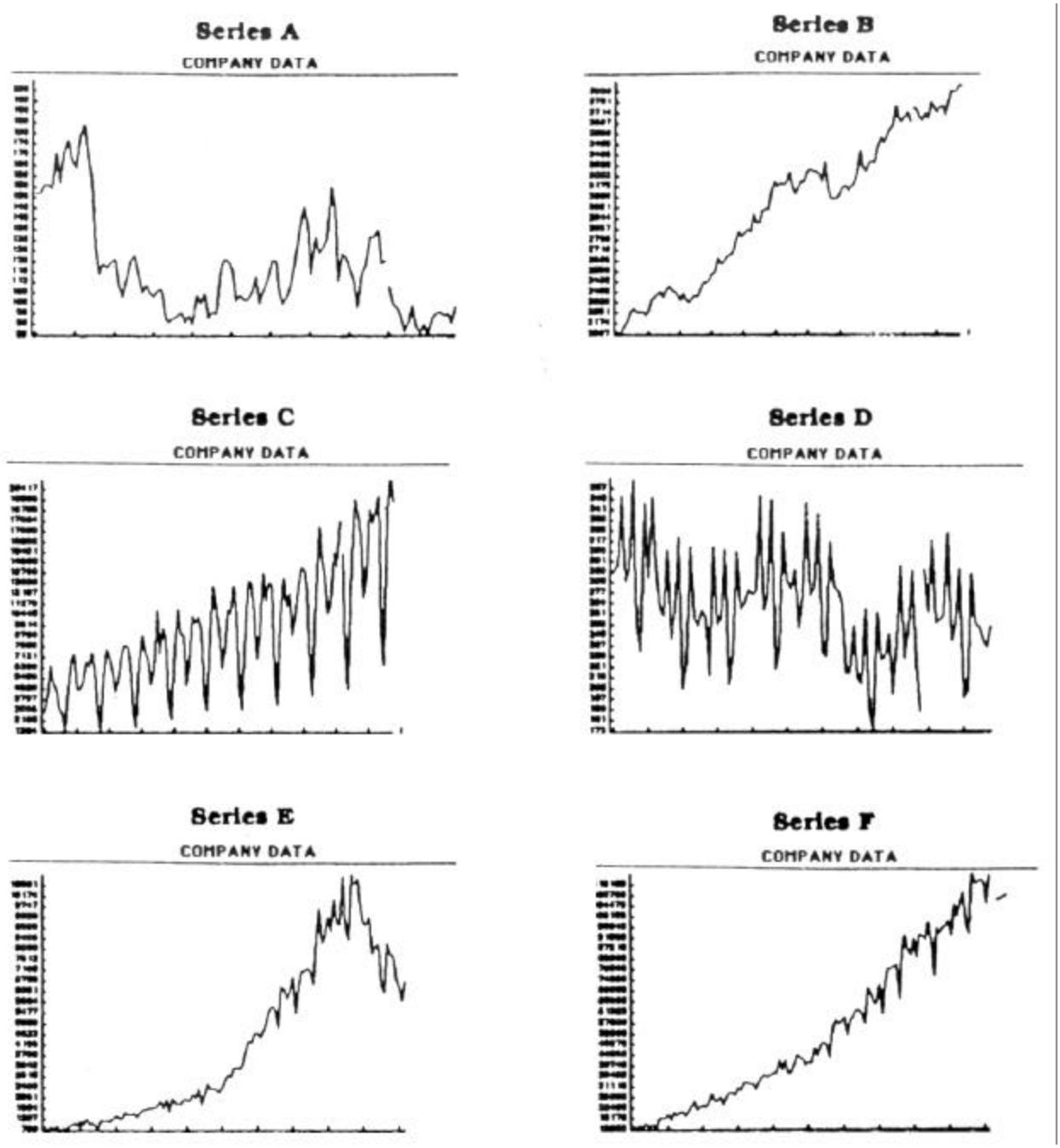
Formulate rules for making forecasts for horizons of 1 to 18 months.

Your rules should have a form like: *Given certain characteristics in the data, use (favor) a certain extrapolation method. You* may identify characteristics of extrapola tion methods, rather than identifying particular ones. Identify all features that you consider important in associating time series and extrapolation methods.

While you are writing your list, talk out loud about considerations involved in compiling it. Describe any adjustments, analyses, and transformations you would perform on the data.

Appendix B. The six monthly series used in the knowledge elicitation sessions.

These displays show both the historical data used to select and tit the forecasting models and the 18 months that were to be forecast. The subjects did not see the last 18 months (after the break in the graphs) until after making their decisions about how to forecast them.



References

Armstrong, J. Scott, *Long-Range Forecasting: From Crystal Bull to Computer,* 2nd ed., John Wiley, New York, 1985.

Buchanan, B. G. et al., "Constructing an expert system" in Hayes-Roth, F., D. Waterman and D. B. Lenat *(eds.), Building Expert Systems,* Addison-Wesley, Reading, MA, 1983, 127-168.

Chambers John M., William A. Gale and Daryl Pregibon, "On the existence of expert systems," *Statistical Software Newsletter,* 14 (1988), 63-66.

Clarkson, Geoffrey, P. E., *Portfolio Selection.* Prentice-Hall: Englewood Cliffs, NJ, 1962.

Ericsson, K. Anders and Herbert A. Simon, *Protocol Analysis: Verbal Reports as Data.* MIT Press: Cambridge, MA, 1984.

Feigenbaum, E. A. and P. McCorduck, *The Filth Generation.* Addison-Wesley, Reading, MA, 1983.

Fellers, Lack W., "Skills and techniques for knowledge acquisition: A survey, assessment, and future directions;" in DeGross, Janice I. and Charles H. Kriebel (eds.), *Proc. Eighth International Conf. Information Systems,* 1987, 118-132.

Gilchrist, W. G., "Discussion of the paper by Professor Makridakis and Dr. Hibon," J. *Roy. Statist. Soc.,* 142, 2 (1979), 126-12%.

Grabowski, Martha, "Knowledge acquisition methodologies" in DeGross, Janice I. and Margrethe H. Olson (eds.), *Proc. Ninth International Conf. Information Systems,* 1988, 47 - 54.

Hayes-Roth, F., et al. *Building Expert Systems.* Addison-Wesley: Reading. MA, 1983.

Hogarth, Robin, "Discussion of the paper by Professor Makridakis and Dr. Hibon," *J. Royal Statistical Society,* 142, 2 (1979), 136.

Larcker, David F. and V. Parker Lessig, "An examination of the linear and retrospective process tracing approaches to judgment modeling," *Accounting Rev.,* 58 (1983). 58-77.

Lenat, Douglas B., "The role of heuristics in learning by discovery: Three case studies;" in Michalski, Ryszard S., Jaime G. Carbonell, and Tom M. Mitchell (eds.), *Machine Learning: An Artificial Intelligence Approach.* Morgan Kaufmann: Los Altos, CA, 1983, 243-306.

Lopes, Lola L., "Pattern, pattern-who's got the pattern?," *J. Forecasting,* 2 (September 1983), 269-272.

Makridakis, S. et al., "The accuracy of extrapolation (time series) methods: Results of a forecasting competition;" *J. Forecasting,*1(1982), 111-153.

Makridakis, S. and M. Hibon, "Accuracy of forecasting: An empirical investigation;" J. *Royal Statistical Society, Ser. A.* 142, 2 (1979), 97-145.