Reproduced with permission from G. Wright and P. Goodwin (eds.), *Forecasting with Judgment*, John Wiley & Sons Ltd., 1998, 269-293.

Integration of Statistical Methods and Judgment for Time Series Forecasting: Principles from Empirical Research

J. Scott Armstrong
University of Pennsylvania, Philadelphia, PA

Fred Collopy
Case Western Reserve University, Cleveland, OH

Abstract

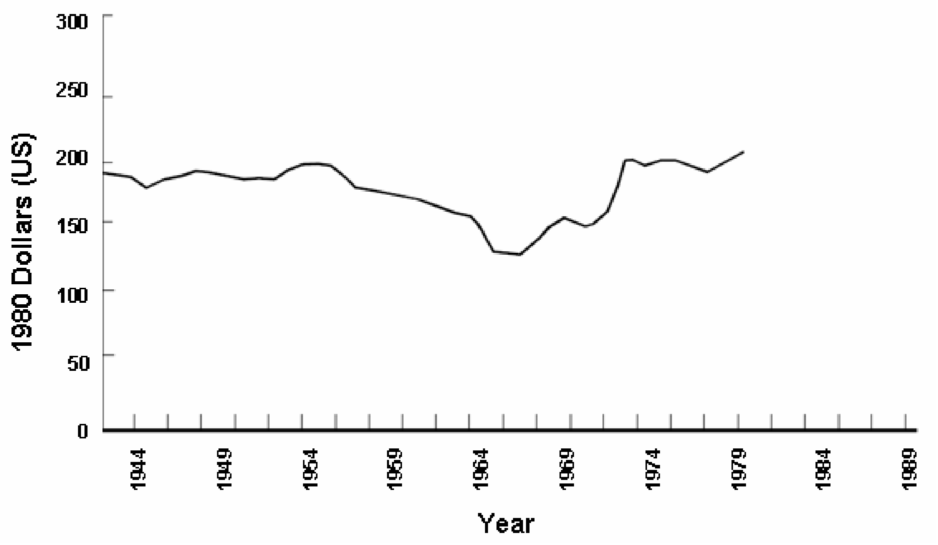
We consider how judgment and statistical methods should be integrated for time-series forecasting. Our review of published empirical research identified 47 studies, all but four published since 1985. Five procedures were identified: revising judgment; combining forecasts; revising extrapolations; rule-based forecasting; and econometric forecasting. This literature suggests that integration generally improves accuracy when the experts have domain knowledge and when significant trends are involved. Integration is valuable to the extent that judgments are used as inputs to the statistical methods, that they contain additional relevant information, and that the integration scheme is well structured. The choice of an integration approach can have a substantial impact on the accuracy of the resulting forecasts. Integration harms accuracy when judgment is biased or its use is unstructured. Equal-weights combining should be regarded as the benchmark and it is especially appropriate where series have high uncertainty or high instability. When the historical data involve high uncertainty or high instability, we recommend revising judgment, revising extrapolations, or combining. When good domain knowledge is available for the future as well as for the past, we recommend rule-based forecasting or econometric methods.

1. Introduction

Assume that you need to make forecasts for time-series, such as the prices of natural resources. Should you use judgment or statistical methods? As an example, the historical series for the price of chromium is shown in Figure 1. The historical data would seem to be valuable and statistical methods can make use of this information. Expert judgment also uses knowledge about the factors influencing the prices of natural resources? Each approach has advantages and disadvantages, so one might also consider integrating statistical methods and judgment.

Judgment is useful because domain experts often have knowledge of recent events whose effects have not yet been observed in a time series, of events that have occurred in the past but are not expected to recur in the future, or of events that have not occurred in the past but are expected for the future. For example, they may know about policy changes that are likely to cause substantial changes over the forecast horizon. While these types of information should be valuable for forecasting, there are also risks in using unaided judgment. Experts may see more in the data than is warranted. And they are subject to a variety of biases such as anchoring, double counting,

Figure 1. Chromiun Prices



and optimism.

Statistical methods are less prone to biases and can make efficient use of prior data. Statistical methods are reliable; given the same data, they will produce the same forecast whether the series relates to costs or revenues, to good news or bad. However, statistical procedures are myopic, knowing only about the data that are presented to them.

Given the relative advantages of judgment and statistical forecasts, it seems sensible to integrate them.[1] As we will show, integration of judgment and statistical methods can lead to substantial gains in accuracy under certain conditions (in comparison to using either of them alone). But when incorrectly applied, as often happens, integration can harm accuracy.

Our conclusions are based on a review of prior literature. The review was aided by previous reviews on judgmental and statistical forecasting by Bunn & Wright (1991), Goodwin & Wright (1993, 1994), and Webby & O'Connor (1996). To ensure that our interpretations of prior research are accurate, we circulated a draft of this chapter to those whose work we cited.[2] We asked them to examine our interpretation of their work, and to indicate if we had misinterpreted their work or if they were aware of additional studies that should be included. Responses were received from eight authors. This led to revisions and to the identification of additional studies.

In all, our review includes 47 relevant empirical studies. Note that research on the integration of judgment and statistical methods is generally recent; all but three of these studies have been published since 1985[3].

In this chapter, we examine when you are likely to gain from integratirlg statistical methods and judgment, and how best to achieve such integration. In the next section, we describe the components that can be integrated. We then propose screening criteria to identify the conditions under which integration is feasible and useful. Most of the chapter is devoted to describing research findings related to various procedures for integrating judgment and statistical methods. These lead to some principles for integration. We conclude with a discussion of thc conditions related to the choice of an integration procedure.

2. Components for Integration: Judgment and Statistical Methods

We consider three ways that judgment can be integrated into time series forecasting.[4] First, there is judgnent about what data are relevant to the forecasting task. Second, the judgment of forecasting experts can be used to determine the approach to be used (e.g. "If you expect that automobile sales will continue to grow, the best way to forecast them over the next 12 months is to use seasonally adjusted exponential smoothing"). Third, experts can incorporate domain knowledge into the forecasts (e.g, "Based on the state af the economy and considering the marketing plan, we expect that Ford's unit automobile sales will increase by 5% over the next 12 months").

The first judgment task facing a forecaster has to do with what data are to be used. By "data" we mean the series of interest (typically called the time series) and any available time series on causal factors. Data can be based on objective or subjective inputs (such as surveys). Exactly what series to use will depend upon the problem being addressed. For example, when studying airline passenger behavior, should one count the number of people who fly, the number of trips they take, the number of segments they fly, the number of airplane embarkations they make, or something else? Should the data be collected by days, weeks, months, quarters or years? These are not trivial issues. They depend on the decisions to be made, the costs of and kinds of data available and the rate at which things are changing. Once a time series is selected, one must decide whether it should be forecast directly or decomposed.

Judgment is needed to decide what statistical procedures to use in a given situation. These statistical procedures can be classified based on the data they use. Extrapolation uses only the historical values of the time series whose future value is of interest. Regression methods estimate the effects of causal variables as well.

Domain experts make judgments based on their knowledge about the product/market, and on their expectations about patterns in the data. This domain knowledge can be used to define the variable of interest, to make revisions in the time series observations, or to adjust for unusual events such as strikes, stockouts or drought. Experts can use their domain knowledge directly to make forecasts, or to estimate starting values and smoothing parameters for models that specify expectations about future effects of causal variables. They can use it to select analogous or related series, which can aid in estimating trends (e.g. Buncan, Gorr & Szczypula, 1993). Domain knowledge can be used to identify the causal factors acting on the series and to make estimates about how they will change over the forecast horizon (e.g. by anticipating a major change in price or a change in product characteristics). It can also be used to specify expectations about future effects, such as price becoming less important to consumers.

It seems to be commonly believed that the more information experts have about a series, the better judgmental forecasts they will be able to make. However, laboratory studies have concluded that additional information can harm accuracy. Further, in some circumstances at least, confidence grows even as accuracy declines (e.g. Davis, Lohse & Kottemann 1994). To avoid this, decision support systems can be used to structure domain knowledge.

3. Feasibility Conditions

To integrate judgment with statistical forecasts, one must be able to produce statistical forecasts. This means that there must be quantitative data available and these must have some relevance for the future. Such data are not always available. For example, when launching a highly innovative product, one often lacks data, even on similar products. In other cases, large discontinuities (new laws, wars, new products, etc.) can render prior data irrelevant for predictions about the future.

To be useful, judgments should incorporate information that is not captured by the statistical forecast and *vice versa.* To a large extent, time series capture the effects of all of the changes in the past, so it is primarily when domain knowledge provides information about recent or pending changes that it may be useful. For example, if management were to decide to phase out a product by removing all marketing support, this knowledge would be useful. Another example would be if substantial price changes are planned for a product that has sold at a constant price in the past. Still another example is when those who are doing the forecasting have good knowledge of the historical data and also much control over the series, such as for management's 1-year forecasts of corporate earnings (Armstrong 1983). On the other hand, if time series are, themselves, the result of many unbiased judgments (as in markets), there may be little benefit from integration.

One problem with human judgment is that it is easily biased. An upturn in sales for one of the company's products can be seen by the product manager as an early indicator of a period of sustained growth, while a downturn of similar magnitude might be dismissed as a transient slump.

When the forecasters have motivations for a particular outcome, inclusion of their judgmental forecasts is likely to add bias to the forecast. This tends to make forecasts less accurate than those produced by a statistical method.

While these conditions might seem obvious, they are sometimes ignored. This is especially so for the condition related to bias. A failure to meet any one of these conditions might cause integration to be of little value or even, to harm accuracy.

4. Procedure for Integration

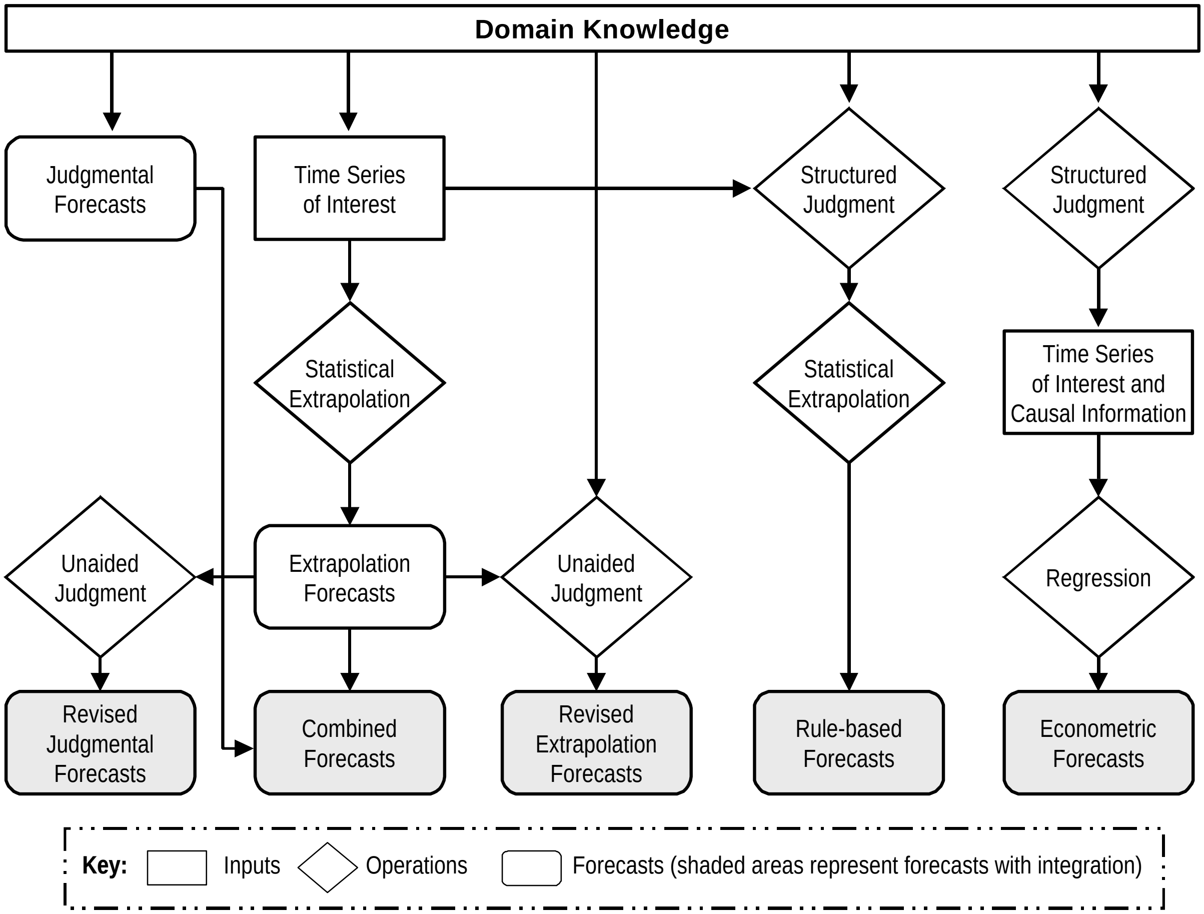
Assuming that the above conditions are met, integration should be considered. All of the methods we will discuss involve integrating*structured* judgment. A variety of decision support procedures can be used to structure judgment. At the simplest level, they may involve ensuring that the arithmetic is done correctly. MacGregor, Lichtenstein & Slovic (1988) concluded that experts often make arithmetic errors when using judgmental decomposition. Another way to structure judgment is to organize historical data for presentation to experts. While it is often assumed that a picture is worth a thousand words, research findings suggest that graphics help only under certain conditions and can sometimes even harm forecast accuracy. Harvey & Bolger (1996) reviewed the research and conducted experiments. They found that graphs led to more accurate judgmental forecasts when the series contained trends, but that otherwise, tables yielded more accurate forecasts. They concluded that the use of graphs of trended series reduced the tendency to underestimate the steepness of the trends when judges used tables. On the other hand, graphs reduced their ability to estimate the level of untrended data due to inconsistency and an over-forecasting bias, which were both greater with graphs.

A comprehensive approach to aiding judgment was developed by Edmundson (1990). Using a computer-aided approach, judges decomposed judgments about the level, trend and seasonality of a series. This approach proved superior to the use of statistical methods alone.

The various ways in which statistical methods and domain knowledge can be integrated for time series forecasting are illustrated in Figure 10.2. The five procedures to integrate judgment and statistical methods are shown in the shaded boxes. They include revising judgmental forecasts after seeing statistical extrapolations, combining judgmental and extrapolation forecasts, using judgment to revise extrapolations, rule-based forecasting and econometric methods.

Integration procedures are expected to be successful to the extent that the judgments and the statistical forecasts are independent. Such independence can be achieved by carefully structuring the procedures used for integration, such as by specifying any weighting in advance of seeing the statistical forecasts. A review of evidence supporting this is provided in Armstrong (1985, pp. 52-7). The methods on the right side of the Figure, econometric methods in particular, are the most highly structured.

Figure 2. Integration of Judgment and Statistical Methods



4.1 Revised Judgmental Forecasts[5]

One way to integrate judgment and statistical methods is for experts to make judgmental forecasts, and then revise them based on statistical extrapolations. Carbone & Gorr (1985) asked 14 subjects to make judgmental forecasts for 10 time series from the M-competition data (Makridakis et al., 1982). Subjects were then allowed to make extrapolation forecasts and use these to revise their judgmental forecasts. These revised forecasts were more accurate.

Lim & O’Connor (1996b) suggest that decision makers get overwhelmed by information, but if the information is structured it improves performance. Using simulated data, Lim &

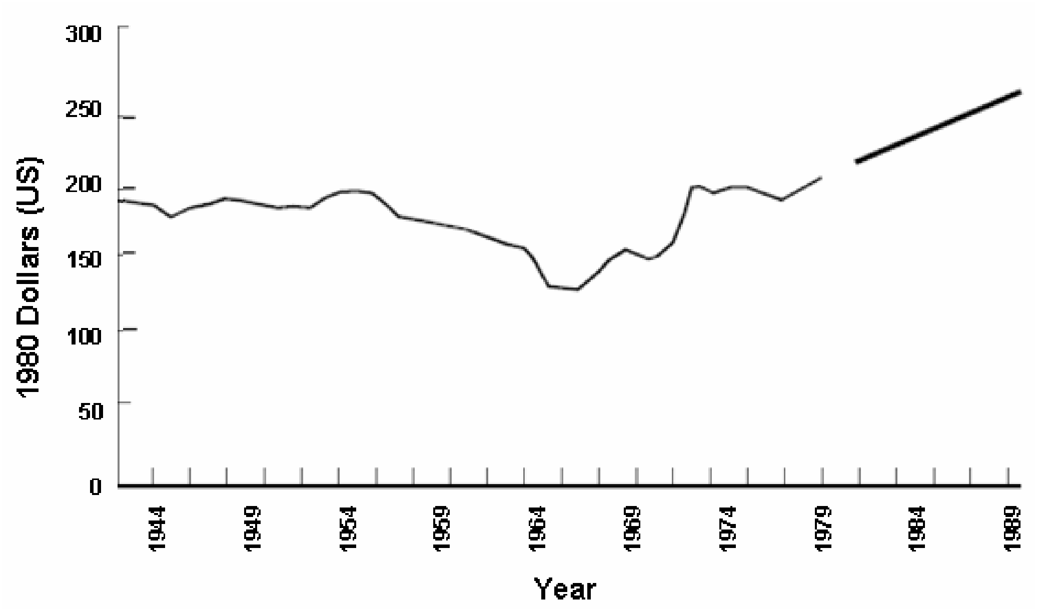
O'Connor (1996a) asked subjects to make judgmental extrapolations and then provided them with forecasts from a damped trend exponential smoothing model. The subjects' adjustments improved accuracy. This gain was enhanced if they were also provided with relevant causal information.

Thus, accuracy was improved when the forecaster followed a structured procedure in which a preliminary forecast was made, the data were reviewed and the forecast was then revised.

Judges may be resistant to new information provided by statistical forecasts. Lim & O'Connor (1995) found that subjects presented with statistical forecasts (after having made a judgmental forecast) put too much weight on their initial judgmental forecast when combining the two. This persisted even when they were presented with information about the reliability of the statistical forecasts and the inaccuracy of their initial forecasts.

Consider again the time series showing the price of chromium. Experts would be asked to make a judgmental forecast. Following this, they would be provided with a statistical extrapolation like that in Figure 3, which is based on Holt's exponential smoothing, and asked to

Figure 3. Chromiun Prices: Holt's Forecast



adjust their initial forecast. The evidence suggests that the adjusted forecasts would be superior if the time series adds information beyond what is in the judgments.

4.2 Combined Forecasts

If one could know which method, statistical or judgmental, would produce the more accurate forecast, then that method should probably be used, or at least weighed more heavily. In practice, alternative forecasts nearly always contain some added information. Combining them aids accuracy to the extent that the added information is valid and reliable but different.

Mechanical combinations have at least three advantages. First, they are more objective, avoiding the introduction of biases or political manipulation. Second, it is easier to disclose fully the process that produced them. Third, they tend to be more accurate because they use knowledge more effectively. These conclusions about combining are drawn partly from Clemen’s (1989) review of over 200 empirical studies on combining statistical forecasts. Similar conclusions have been obtained from research on combining judgmental forecasts (Armstrong 1985, pp. 91-6). Below we review the evidence that these conclusions apply as well to combinations of statistical and judgmental forecasts. Such combinations seem especially useful because they may have different biases that cancel one another.

Lawrence, Edmundson & O’Connor (1986), using 68 monthly series from the M-competition (Makridakis et al., 1982), combined eyeball judgmental extrapolations with statistical extrapolations from exponential smoothing. The judges had no domain knowledge. Alternative presentations of the data (tables vs. graphs) and various combinations of the forecasts were compared. Overall, combining reduced the mean absolute percentage error (MAPE) by about 7% (compared with the accuracy of the average component forecast), with a tendency for more improvement for shorter horizons (an 8.9% gain for 1-6 months ahead forecasts) than for longer-term (6.2% for 13-18 months ahead). This gain was slightly better than the gain for equal-weights combining of six statistical extrapolation forecasts for these series (6.1% gain on average).

Lobo & Nair (1990) examined combinations of judgmental and statistical extrapolations of corporate earnings for eight years for 96 firms. Judgmental forecasts were made by two groups with domain knowledge (security analysts who specialized in those stocks) and two statistical forecasts were made (one using random walk with drift and one using ARIMA). We re-analyzed results from their Table 1 and concluded that, on average, the combination of judgmental and statistical forecasts led to a 5.2% reduction in MAPE. This exceeds the gain from combining their two statistical methods (2.1%) or their two judgmental methods (0.6%).

Blattberg & Hoch (1990) combined judgmental and statistical forecasts and achieved statistically significant gains in accuracy (tested using R). They did not, however, estimate the magnitude of the error reduction.

Sanders & Ritzman (1989, 1990, 1992), studied judgmental forecasts by warehouse planners. The experts had much domain knowledge and the forecast horizon was short (1 day ahead). The conditions, then, were ideal for the use of judgment. They concluded that the combination of judgmental and extrapolation forecasts led to improved accuracy when applied to series of medium to low variability. Otherwise, it was best to rely on judgmental forecasts alone. Overall, however, their combined forecasts were about 15.8% more accurate than the typical component forecasts (based on our analysis of Table 1 in Sanders & Ritzman, 1989).

Lim & O'Connor (1995) conducted an experiment where judges had a decision support system that provided feedback on the relative accuracy of judgment and statistical procedures.

Subjects placed too much weight on their own judgment, so that the forecasts were not as accurate as they might have been. One might expect this to harm accuracy. Indeed, Bretschneider et al. (1989), in a field study of revenue forecasting by state governments, concluded that state governments in the USA that used formal procedures for combining had more accurate forecasts than states that used subjective weighting.

A reasonable starting point for combined forecasts would be to use equal weights (Clemen, 1989). In some conditions, however, it may help to use differential weights. For example, because experts are relatively good at estimating current levels, it makes sense to weight judgment more heavily for levels. Consistent with this, Webby & O'Connor's (1996, pp. 97- 9) review concluded that judgmental forecasts were superior to statistical ones, especially for short-term forecasts (where the level estimate is most important) and when the experts had good domain knowledge.

If differential weights are used, they should be developed in advance and should be based on research findings, not on the opinions of experts involved in the forecasting task. The weights should also be recorded so that analyses may determine how to improve the weighting scheme.

4.3 Revised Extrapolation Forecasts

The most common way to integrate statistical methods and domain knowledge is to revise extrapolations. According to Sanders & Manrodt's (1994) survey of forecasters of 96 US corporations, about 45 % of the respondents claimed that they always made judgmental adjustments to statistical forecasts, while only 9 % said that they never did. The main reasons they gave for revising quantitative forecasts were to "incorporate knowledge of the environment" (39

%), “incorporate product knowledge” (30 %) and “incorporate past experience” (26 %). While these reasons seem sensible, there is a problem in that adjustments are often made by biased experts. In a survey of members of the International Institute of Forecasters, 269 respondents were asked whether they agreed with the following statement:“Too often, company forecasts are modified because of political considerations.”[6] On a scale from 1 = “disagree strongly” to 7 = “agree strongly,” the mean response was 5.4. Fildes & Hastings (1994), using a survey of 45 managers in a large conglomerate, found that 64% of them believed that“forecasts are frequently politically motivated.”

Judgmental revisions might improve accuracy if the forecaster is able to identify patterns that are missed by the statistical procedure. Most extrapolation methods do not, for example, deal with discontinuities (Collopy & Armstrong, 1992b). This might also be true for other important pattern changes. A number of studies have examined subjective adjustments where subjects did not have relevant domain knowledge, suggesting that the improvements were due to their ability to recognize pattern changes that were not handled by the extrapolation methods. Willemain (1989) showed that subjective adjustments led to improvements in the accuracy of extrapolation for artificial time series; however, in another study involving 24 series from the M-Competition, he found no advantage for subjective adjustments (Willemain, 1991). Sanders (1992), in a study involving simulated data, concluded that judgmental revisions of extrapolation forecasts led to some gains in accuracy for series that had low noise, especially when there was a step function in the historical series. Carbone et al. (1983) had subjects make revisions for 25 time series from the M-competition (where subjects had only the title and dates of the series in addition to the series itself), In general, these adjustments tended to harm accuracy.

Judgmental revisions might also improve accuracy if the forecaster were able to take advantage of causal information that the statistical method had not used. Wolfe & Flores (1990) and Flores, Olson & Wolfe (1992) found improvements when judgmental adjustments were made to corporate earnings series that had high variability. However, experts were no more accurate than non-experts, suggesting that these improvements had more to do with the patterns of data than with the causal information.

Based on their study of a situation where the judges had domain knowledge, Mathews & Diamantopoulos (1989) concluded that judgmental revisions of quantitative forecasts led to improved accuracy. These adjustments were made by individual experts.

It seems appealing to make subjective adjustments. Consider the forecast of chromium prices. In Figure 3, we provided an extrapolation produced by Holt's exponential smoothing. One could start with the forecast, then revise it judgmentally. Given good domain knowledge and a structured process, judgmental revisions are likely be useful.

Subjective adjustments contradict our advice that judgment should be used as an input to statistical methods, rather than to revise the statistical forecasts. The exceptions that one might consider, such as correcting errors, adjusting for interventions or failure to include an important variable, are better dealt with by producing a new forecast than by revising the old one.

Judgmental revisions are risky. Given the danger of double counting and the difficulty of avoiding biases, one way to make subjective adjustments would be to have experts decide what adjustments would be appropriate for the model (e.g. add 2 % to the forecast for each horizon) *before seeing the forecasts.* We are unaware of any tests of this approach.

In summary, revisions of extrapolation forecasts seem most relevant where forecasters have good domain knowledge and the revisions are based on structured judgment. Lacking such conditions, judgmental revisions might harm accuracy. Given that the research findings to date are limited and difficult to interpret, other factors might be involved, adding to the risk of judgmental revisions.

4.4 Rule-based Forecasts

Rule-based forecasting uses structured judgmental inputs to statistical procedures. Rule-based forecasts depend upon an assessment of the conditions faced in the forecasting task. The basic idea is that the forecasting methods must be tailored to the situation and that a key aspect of this is domain knowledge. Domain knowledge can help to identify the direction of causal forces on the series of interest, the functional form that the time series is likely to follow, and the presence of any unusual patterns or observations.

The rules are instructions on how to weight the forecasts from a set of simple forecasting methods, and the weights vary according to features of the series. The rules for weighting the forecasts are derived from experts and from prior research. Rule-based forecasting uses expert judgments about the characteristics of series and about causal forces as inputs to extrapolation methods.

Collopy & Armstrong (1992a) found that rule-based forecasting was more accurate than the use of equal weights in 1 to 6-year forecasts for 126 annual economic time series from the M-competition. The gains were greater for series that involved good domain knowledge, significant long-term trends, low instability and low uncertainty. In an extension using the same 126 series, Vokurka, FIores & Pearce (1996) assessed some of the characteristics automatically. They achieved substantial improvement over using statistical procedures alone, although not quite as large as those in Collopy & Armstrong (1992a). This suggests that at least some of the gain from using judgment can be incorporated into the statistical forecasting procedures. Vokurka, Flores & Pearce found further improvements when they used a decision support system. It had a graphical interface that allowed the user to view a graph of the time series, the forecast method selected by the system, and forecasts from alternative methods. Users had the capability to change the forecasting method or the forecasted values. Adya, Collopy & Kennedy (1997) used rules to identify the presence of many of the features for the series examined in Collopy & Armstrong (1992a) and obtained similar results.

The chromium series that we have been using as illustration was part of a 1980 challenge that Julian Simon posed to ecologists: "Pick any natural resource and any future date. I'll bet the [real] price will not rise" (Tierney 1990). He based this on long-term trends, and the reasoning that there had been no major changes in the long-term causal factors. Paul Ehrlich, an ecologist from Stanford University, accepted the challenge; he selected five metals (copper, chromium, nickel, tin, and tungsten) and 10 years. The prices for these metals had been rising over the recent past. In general, the causal forces for the prices of resources are "decay." This is due to improved procedures for prospecting, more efficient extraction procedures, lower energy costs, reduced transportation costs, development of substitutes, more efficient recycling methods, and more open trade among nations. A proposed force that might lead to increased prices is that the resource is exhaustible; however, this seldom has a strong effect because new sources are found. With respect to petroleum reserves, for example, Ascher (1978, pp. 139- 41) showed that forecasts of the ultimate available petroleum reserves increased from the late 1940s to the mid-1970s. Such changes seem common for resources because of improvements in exploration technology. Thus, the overall long-term causal force seems to be decay, and prices of metals would be expected to decrease.

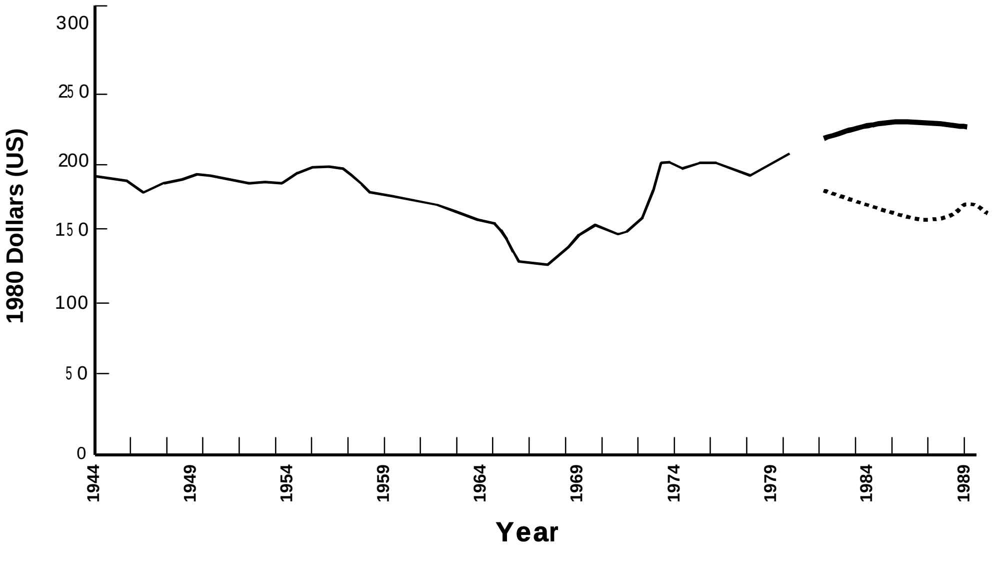
Rule-based forecasting is especially useful when domain knowledge indicates that recent trends may not persist. In the case of metals prices, Ehrlich assumed that recent price trends would continue. We had implemented Ehrlich's assumption in Figure 10.3 by using Holt's exponential smoothing to extrapolate recent trends for one of his five metals, chromium. This led to a forecast of sharply rising prices. In contrast, although the rule-based forecast initially forecasts an increase in prices (because it allows that short-term trends might continue), over the 10-year horizon the forecasts become dominated by the long-term trend, which is downward and consistent with the causal forces. This same pattern was found for each of the five metals forecasts made in 1980.[7] A rule-based forecast for chromium prices is shown in Figure 4. This forecast resulted from specifying the causal forces acting on the series as decay and the functional forrn as additive. The actual data for 1981 – 1990 are also shown in this Figure. Simon won the bet on the price of chromium as well as for the other four metals.

4.5 Econometric Forecasts

When judgmental inputs are used to identify a model and regression is used to obtain estimates for the coefficients of this model, we refer to the integration as an econometric model. Econometric models provide the most highly structured approach to integrating judgment. In addition to functional form, judgment is used to select causal variables and to specify the directions of effects (Dawes & Corrigan, 1974). Prior research (summarized by Armstrong, 1985, and Fildes, 1985) indicates that, when judgment is based on good domain knowledge, econometric models are typically more accurate than alternative procedures when large changes are involved.

Stepwise regression is an approach in which statistical procedures are used to identify the model. It does not call for integration with judgment, so we exclude it from our discussion of econometric methods. In fact, given that it ignores domain knowledge, we recommend that stepwise regression should not be used in forecasting. Indeed, there is little support for its use (Armstrong 1985, pp. 52-7).

Figure 4. Chromium Prices: Rule-based Forecast (bold) and Actuals (dotted)



Econometric models are more accurate than judgment for cross-sectional forecasts. Gove & Meehl (1996) reached this conclusion after a review of 136 comparative empirical studies; 64 studies found the model’s forecasts to be superior to the experts, there were 64 ties, and the experts were superior in eight studies. No similar analysis has been done for time series forecasts, but our impression is that judgment is superior to econometric forecasts under certain conditions, such as for short-range forecasts.

Studies involving cross-sectional forecasts have found that subjective revisions are of little benefit and they tend to harm accuracy. For example, Griffith & Wellman (1979) found that subjective adjustments of quantitative forecasts reduced accuracy for six hospitals during the period 1967 – 1971. In Harris (1963), experts' revisions of football forecasts from an econometric model reduced their accuracy. Weinberg (1986) found that judgmental revisions of forecasts for arts performances did not improve accuracy.

In contrast, subjective adjustments to econometric models often improve the accuracy of short-term time series forecasts (McNees, 1990). Armstrong (1985, pp. 237-8) reviewed four empirical studies and concluded that subjective revisions will be useful primarily to the extent that they adjust for recent changes in the current level. Comparable gains can be made by mechanica1 adjustments based on the most recent error of the model. For example, one useful procedure is to take half of the error in forecasting the latest period, and add that to the forecast for the next period. (In effect, this adjusts the current status or level.) Vere & Griffith (1995) and Donihue (1993) provided additional evidence on this issue. However, McNees (1996) concluded that subjective adjustments are superior to mechanical adjustments.

Forecasters may be biased and they may put too much emphasis on their opinions. For example, Weinberg (1986) found that experts were correct about the direction of the revisions on 12 of 15 arts performances, but they were too aggressive, such that on average there was no improvement. McNees (1990) found that economic experts were too aggressive and that this harmed accuracy. To address these problems, adjustment factors should be independent of the forecasts. This could be done by asking experts what adjustments they would make for a given model before they see the forecasts. These would be used to adjust the level and the trend estimates; these revisions would be recorded for further evaluation.

Another problem with judgmental revisions is that record keeping is often poor (Turner, 1990). The lack of a record means that forecasters are unlikely to get well-summarized feedback of the effect of their adjustments, and, as a result, they will have difficulties learning how to improve them, To address this, judgmental revisions should be done in a structured and fully disclosed manner.

6. Principles for Integration

Although much research has been done recently, the number of studies does not yet allow for a quantitative integration of their results. Furthermore, the failure of many of the studies to specify conditions has compounded the problem. Nevertheless, we draw some conclusions about principles for integrating judgment and statistical methods.

One generalization is that judgment is most effective if used as an input to statistical forecasting, less effective if used independently, and least effective if used to revise quantitative forecasts. Another generalization is that structure helps; the more structured the inputs and the more structured the integration procedures, the more accurate the forecasts.

Equal weighting of statistical and judgmental forecasts is a good starting point. Departures from an equal weighting of judgment and statistical models should be based on good domain knowledge. When uncertain about which forecasting procedure will be most accurate, use equal weights on judgment and statistical forecasts. However, in cases where one method can be expected to do better than the other, it should be weighted more heavily. For example, in cases where there have been substantial recent disruptions for reasons that are known to the experts, and where the experts have received good feedback on the accuracy of their forecasts, one might put more weight on the judgmental forecasts. On the other hand, as the amount of high quality and relevant data is increased, more emphasis should be given to statistical methods. As domain knowledge increases, put more weight on judgement relative to statistical forecasts. When the amount and quality of the data improves, more weight should be placed on statistical forecasts.

Experts are not effective in integrating large amounts of information that might be relevant to forecasting changes. This implies that judgment is more effective for estimating where one is now than for predicting change, Another way of saying this is that experts are better at' "nowcasting" than at forecasting. Evidence supporting this position is summarized in Armstrong (1985, p. 237). As a result, it may help to decompose the forecasting problem into level and change. Judgment would be weighted more heavily for estimating levels. With increases in the amount of change anticipated, it is advisable to place more weight on the statistical methods relative to judgment (Armstrong, 1985, pp. 393-402).

Judgmental revisions of forecasts are likely to harm accuracy if done by biased experts. Furthermore, revision does not appear to be the most effective way to use judgment. It is distressing that software developers have been making it easier for forecasters to make unstructured judgmental adjustments. Instead, software should encourage prior inputs of judgment and describe how to structure this information (e.g. information about levels or about causal forces). If judgmental revisions are used, they should be carefully monitored (what revisions were made, when, by whom, and why). The evaluation should compare the accuracy of the forecasts with and without the revisions.

Our discussion has focused on forecasting expected values. We have not addressed the issue of estimating the uncertainty associated with prediction intervals. To date, little research has been done on the integration of judgment and statistical methods to estimate prediction intervals. However, Armstrong & Collopy (1997) found that judgment could be valuable in the estimation of prediction iotervals; in particular, the use of causal forces helped to identify series where the log errors were expected to be asymmetric.

7. Conditions Under Which Integration Procedures Are Effective

Assuming that the feasibility conditions have been satisfied (quantitative data available, time series and judgment each contribute different information, and judgment is not obviously biased), one might conclude that some form of integration is relevant. But which? While our recommendations focus on accuracy, other criteria, such as ease of understanding the forecasting method, might be as important as accuracy (Yokum & Armstrong, 1995).

The choice of an integration procedure is dependent upon the conditions. Time series conditions can be separated into those related to domain knowledge and those related to the historical time series, and the interaction of these:

(1) Domain knowledge

- Unusual past event (e.g. knowledge about when a strike occurred).
- Planned intervention.

(2) Historical time-series

- Significant long-term trend.

- Uncertainty (e.g. coefficient of variation about the trend; differences in direction of longand short-term trends).
- Instability (e.g. discontinuities or last observation unusual).

(3) Conflicts between domain knowledge and time series

- Contrary series (when the expected direction of the trend conflicts with the short-term extrapolation).

The determination of these conditions can be done judgmentally, although some of the historical conditions can also be done statistically, as shown in Collopy & Armstrong (1992a), and extended by Adya, Collopy & Kennedy (1997) and by Vokurka, Flores & Pearce (1996).

Research has provided little direct evidence on how these conditions affect the accuracy af various methods, except fc r rule-based forecasting, where the conditions are an integral part of the procedure. Applying general principles from above, however, permits us to speculate about the relative effectiveness of the various integration methods.

For situations where little historical domain knowledge exists, the selection of an integration approach is not critical. But if there is good domain knowledge, econometric forecasting can permit it to be modeled, and thereby integrated effectively with the statistical data from the time series itself. Rule-based forecasting can also integrate domain knowledge, although not as well.

If experts have up-to-date information that has not been integrated into the historical data, then their information should be used in the forecasts. This can occur with respect to recent events whose effects have not been fully incorporated into the data or with respect to planned interventions, such as a major price change.

For situations involving high degrees of uncertainty or instability, revisions and simple combinations can be helpful. In these conditions the methods on the left hand side of Figure 2 are likely to do at least as well as the more sophisticated methods. Since they will generally be easier to implement, they should be favored.

When there is knowledge of future changes, structured approaches to integration are likely to have a particularly high payoff. When contrary trends are encountered, their extrapolation is dangerous (Armstrong & Collopy, 1993). In such cases, econometric methods offer a more promising approach. Although not directly tested, we also expect that econometric forecasts will be more accurate than rule-based forecasting if one has good information about the causal relationships, and if the causal factors can be accurately forecast, or if they can be set by decision makers. Econometric forecasts are likely to be superior to rule-based forecasts because they give explicit attention to the effects of each causal variable. Of course, one must also consider cost as a factor, especially for situations requiring thousands of forecasts.

If experts have knowledge about large recent changes, judgmental adjustments of the current status are likely to improve accuracy. These judgements should be made by unbiased experts and they shou1d be fully disclosed..

8. Conclusions

We have presented three conditions under which the integration of judgment and statistical methods should be considered. The conditions involve having relevant quantitative data, judgmental inputs that provide different information, and unbiased judgments.

Given those three conditions and uncertainty about which method is likely to produce the best forecast, integration is expected to improve accuracy, although the improvements in accuracy will depend upon the extent to which the judgmental inputs are well structured. Of particular importance is that judgment be used as an input to the statistical methods, rather than to revise their output.

In general, equal-weights combining provides a good way to integrate and it should be viewed as the benchmark. To the extent that the historical series involve good domain knowle6ge, significant trends, low uncertainty and low instability, we recommend the use of rule-based forecasting. If, in addition, the future conditions involve interventions and contrary trends, we recommend econometric methods.

While the recent surge of interest in the integration of judgment and statistical methods is promising, we expect that our ability to draw generalizations will continue to be limited unless researchers report on the conditions involved in their studies. We have proposed some historical and future conditions that we think will be helpful in organizing and testing knowledge in this area, but suspect that they can be expanded upon. Given the importance to decision makers of incorporating judgment into their forecasts, and the importance to businesses and society of unbiased and accurate forecasts, this seems to be a most promising area for further research. We believe that a research program should be oriented to identifying the conditions under which a given type of integration should be used.

Acknowledgments

Paul Goodwin, Nigel Harvey, Stephen Hoch, Marcus O'Connor, Nada Sanders and Thomas Willemain provided useful comments on early versions of this paper. Monica Adya contributed to the analysis. Editorial assistance was provided by Colleen Gepperth.

References

Adya, M., Collopy, F. & Kennedy, M. (1997), "Heuristic identification of time series features: an extension of rule-based forecasting." Working paper available from Monica Adya (adya@umbc.edu).

Armstrong, J,S. (1983), Relative accuracy of judgmental and extrapolative methods in forecasting annual earnings, *Journal of Forecasting*, 2, 437-447.

Armstrong, J.S, (1985), *Long-range Forecasting,* 2nd edn. Wiley, New York.

Armstrong, J.S. & Collopy, F. (1993), "Causal forces: structuring knowledge for time series extrapolation," *Journal of Forecasting,* 10, 147-149.

Armstrong, J.S. & Collopy, F. (1997), Prediction intervals for extrapolation of annual economic data: evidence on asymmetry corrections. Working paper.

Ascher, W. (1978), *Forecasting: An Appraisal for Policy Makers and Planners.* Johns Hopkins University Press, Baltimore.

Blattberg, R.C. & Hoch, S.J. (1990), Database models and managerial intuition: 50 % model + 50 % manager. *Management Science,* 36, 887-899.

Bretschneider, S.I., Gorr, W.L., Grizzle, G. & Klay, E. (1989), Political and organizational influences on the accuracy of forecasting state government revenues. *International Journal of Forecasting*, 5, 307-519.

Bunn, D. & Wright, G. (1991), Interaction of judgmental and statistical forecasting methods; issues and analysis. *Management Science*, 37, 501-518.

Carbone, R., Andersen, A., Corriveau, Y. & Corson, P.P. (1983), Comparing for different time series methods the value of technical expertise, individualized analysis and judgmental adjustment. *Management Science,* 29, 559-566.

Carbone, R. & Gorr, W.L. (1985), Accuracy of judgmental forecasting of time series. *Decision Sciences,* 16, 153-160.

Clemen, R. (1989), Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, 5, 559-583.

Collopy, F. & Armstrong, J.S. (1992a), Rule-based forecasting: development and validation of an expert systems approach to combining time series extrapolations. *Management Science,* 38, 1394-414.

Collopy, F. & Armstrong, J.S. (1992b), Expert opinions about extrapolation and the mystery of the overlooked discontiuities. *International Journal of Forecasting,* 8, 575-582.

Davis, F.D., Lohse, G.L. & Kottemann, S.E. (1994,) Harmful effects of seemingly helpful information on forecasts of stock earnings. *Journal of Economic Psychology*, 15, 253-267.

Dawes, R. & Corrigan, B. (1974), Linear models in decision making.*Psychological Bulletin,* 81, 95-106.

Donihue, M.R. (1993), Evaluating the role judgment plays in forecast accuracy.*Journal of Forecasting,* 12, 81-92.

Duncan, G., Gorr, W.L. & Szczypula, J. (1993), Bayesian forecasting for seemingly unrelated time series: application to local government revenue forecasting.*Management Science,* 39, 275-293.

Edmundson, R.H. (1990), Decomposition: a strategy for judgmental forecasting. *Journal of Forecasting,* 9, 301-314.

Eichorn, P. A Yankhauer, A. (1987), Do authors check their references? A survey of the accuracy of references in three public health journals.*American Journal of Public Health,* 77, 1011-1012.

Fildes, R. (1985), The state of the art: econometric models. *Journal of the Operational Research Society,* 36, 549-586.

Fildes, R. & Hastings, R, (1994), The organization and improvement of market forecasting. *Journal of the Operational Research Society*, 45, 1-16.

Flores, B.E., Olson, B.L. & Wolfe, C. (1992), Judgmental adjustment of forecasts: a comparison of methods. *International Journal of Forecasting*, 7, 421-433.

Goodwin, P. (1996), Statistical correction of judgmental point forecasts and decisions.*Ornega,* 24, 551-559.

Goodwin, P. & Wright, G. (1993), Improving judgmental time series forecasting: a review of guidance provided by research. *International Journal of Forecasting,* 9, 147-161.

Goodwin, P. & Wright, G, (1994), Heuristics, biases and improvement strategies in judgmental time series forecasting. *Omega*, 22, 553-568.

Grove, W.M. & Meehl, P. (1996), Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical – statistical controversy. *Psychology, Public Policy, and Law*, 2, 293-323.

Griffith, J.R. & Wellrnan, B.T. (1979), Forecasting bed needs and recommeoding facilities plans for community hospitals: a review of past performance.*Medical Care,* 17, 293-303.

Harris, J.G. Jr. (1963), Judgmental vs. mathematical prediction: an investigation by analogy of the clinical vs. statistical controversy.*Behavioral Science,* 8, 324-335.

Harvey, N. & Bolger, F. (1996), Graphs versus tables: effects of data presentation format on judgmental forecasting.*International Journal of Forecasting,* 12, 119-317.

Lawrence, M.J., Edmundson, R.H. & O'Connor, M.J. (1986), The accuracy of combining judgmental and statistical forecasts.*Management Science*, 32, 1521-1532.

Lim, J.S. & O'Connor, M. (1995), Judgmental adjustment of initial forecasts: its effectiveness and biases, *Journal of Behavioral Decision Making,* 8, 149-168.

Lim, J.S. & O'Connor, M. (1996a), Judgmental forecasting with time series and causal information. *International Journal of Forecasting,* 12, 139-153.

Lim, J.S. & O'Connor, M. (1996b), Judgmental forecasting with interactive forecasting support systems. *Decision Support Systems,* 16, 339-357.

Lobo, G.J. & Nair, R.D. (1990), Combining judgmental and statistica1 forecasts: an application to earnings forecasts. *Decision Sciences,* 21, 446-460.

MacGregor, B., Lichtenstein, S. & Slovic, P. (1988,) Structuring knowledge retrieval: an analysis of decomposed quantitative judgments. *Organizational Behavior and Human Decision Processes,* 42, 303-323.

Makridakis, S., Anderson, A., Carbone, R., Fildes, R., Hibon, M., Newton, J., Parzen, E. & Winkler, R. (1980), The accuracy of extrapolation (time series) methods:Results of a forecasting competition. *Journal of Forecasting,* 1, 111-153.

Mathews, B.P. & Diamantopoulos, A. (1989), Judgmental revision of sales forecasts:A longitudinal extension.*Journal of Forecasting,* 8, 129-140.

McNees, S.K. (1990), The role of judgment in macroeconomic forecasting accuracy. *International Journal of Forecasting,* 6, 287-299.

Sanders, N.R. (1992), Accuracy of judgmental forecasts: a comparison.*Omega*, 20, 353-364.

Sanders, N.R. & Manrodt, K.B. (1994), Forecasting practices in US corporations: survey results. *1nterfaces,* 24(2); 92-100.

Sanders, N.R. & Ritzman, L.P. (1989), Some empirical findings on short-term forecasting: technique complexity and combinations.*Decision Sciences,* 20, 635-640.

Sanders, N.R. & Ritzman, L.P. (1990), Improving short-term forecasts. *Omega,* 18, 365-373.

Sanders, N.R. & Ritzman, L.P. (1992), The need for contextual and technical knowledge in judgrnental forecasting. *Journal of Behavioral Decision Making*, 5, 39-52.

Tierney, J. (1990), Betting the planet. *New York Times Magazine,* December 2, p. 52.

Turner, D.S. (1990), The role of judmnent in macroeconomic forecasting.*Journal of Forecasting,* 9, 315-345.

Vere, D.T. & Griffith, G.R. (1995), Modifying quantitative forecasts of livestock production using expert judgments. *Journal of Forecasting,* 14, 453-464.

Vokurka, R.J., Flores, B.E. & Pearce, S.L. (1996), Automatic feature identification and graphical support in rule-based forecasting: a comparison. *International Journal of Forecasting,* 12, 495-512.

Webby, R. & O'Connor, M. (1996), Judgmental and statistical time series forecasting: a review of the literature. *Internaiional Journal of' Forecasting* 12, 91-118.

Weinberg, C.B. (1986), Arts plan: implementation, evolution, and usage.*Marketing Science*, 5, 143-158.

Willemain, T.R. (1989), Graphical adjustment of statistical forecasts.*International Journal of Forecasting*, 5, 179-185.

Willemain, T.R. (1991), The effect of graphical adjustment on forecas accuracy.*International Journa1 of Forecasting*, 7, 151-154.

Wolfe, C. & Flores, B. (1990), Judgrnental adjustment of earnings forecasts*Journal of Forecasting,* 9, 389-405.

Yokum, J.T. & Armstrong, J.S. (1995), Beyond accuracy: Comparison of criteria used to select forecasting methods. *International Journal of Forecasting,* 11, 591-597.

NOTES

1. We do not include procedures that manipulate judgmental forecasts statistically, such as judgmental bootstrapping procedures.

2. According to a study by Eichorn & Yankauer (1987), authors frequently make mistakes in their summaries of prior research.

3. This overstates currency somewhat, in that some of the review papers cited works that had been published before 1985.

4. We do not address the use of judgment as data. For example, intentions surveys involve people's judgments of people about how they will behave.

5. We recognize that when forecasts are used for decision making, managers may make adjustments. For example, given a sales forecast of 100 units per month, the manager might initially produce 110 units per month in order to build up inventory. Our concern in this chapter is only with the forecast, in this case the 100 units per month, so we do not examine decision-making adjustments. Such adjustments might be useful, of course, as is discussed by Goodwin (1996).

6. This survey was conducted in 1989 by Thomas Yokum and Scott Armstrong. The responses to this question were analyzed depending on whether the respondent was a decision maker, practitioner, educator or researcher. While the practitioners stated the strongest agreement, there were no statistically significant differences among these groups.

7. The forecasts were prepared by Monica Adya, using a version of rule-based forecasting that is described in Adya, Collopy & Kennedy (1997). The data were obtained from *Metals Week*.

Comments on Armstrong and Collopy (1998)

Nada R. Sanders, Wright State University
Published in *International Journal of Forecasting*, 15 (1999), 345-346.

Much of the forecasting literature has historically viewed judgmental and statistical forecasting methods as distinctly separate from one another, often point to the shortcomings of judgmental forecasting. Today it is generally accepted that judgmental and statistical methods each have unique strengths they can bring to the forecasting process. Judgment of domain experts has value as it is often based on up-to-date knowledge of changes and events occurring in the environment that can affect the variable being forecast. On the other hand, statistical methods are consistent, can efficiently process large amounts of information, and are not subject to human biases. To improve forecast accuracy, it makes sense to bring together the advantages of each method by integrating them to reach a final forecast.

Armstrong and Collopy address the issue of integrating statistical methods and judgment for time series forecasting based on empirical research from 47 studies, almost all published since 1985. The paper draws conclusions from those studies in an effort to unite findings from different studies conducted under varying conditions. Five procedures for integration are identified: revising judgment; combining forecasts; revising extrapolations; rule-based forecasting; and econometric forecasting. Research supporting each of these procedures is presented. The authors also identify components that can be integrated and feasibility conditions when benefits from integration are most likely. Finally, principles for integration are developed and conditions under which integration procedures are effective are identified.

Integration is shown to improve accuracy when the experts have good domain knowledge and when significant trends are present. On the other hand, integration can harm accuracy when judgment is biased or when judgment is unstructured. Armstrong and Collopy identify three conditions under which integration should be considered. First, one needs quantitative data that have some relevance for the future. Second, judgmental inputs should provide additional relevant information beyond that contained in the statistical model and vice versa. Third, the judgment should be unbiased. Integration is most effective when judgment is used as an input rather than to revise the statistical output. Also, structuring the judgmental inputs and the integration process contributes to improved accuracy. A good starting point is to use equal weighting of statistical and judgmental forecasts as a benchmark, particularly when there is high uncertainty or instability in the series.

The choice of integration approach was found to have a substantial impact on forecast accuracy. When there is high uncertainty or instability in the historical data, the authors recommend revising judgment, revising extrapolations, or combining. Rule-based forecasting is recommended when good domain knowledge is available, when significant trends are present, and there is low uncertainty. When future conditions are expected to contain trends that are contrary to expectations, econometric models are recommended.

This is a well-written and timely paper, that will certainly serve as a reference for years to come. Drawing generalizations from studies conducted under varying conditions was a challenging task.

To ensure accuracy of interpretations, Armstrong and Collopy had sent a draft of this paper for review to the authors of the 47 papers cited. Though important conclusions are drawn, additional questions emerge and future research needs are identified. The authors stress the importance of identifying effective ways for decision makers into incorporate judgment into their forecasts and improve forecast accuracy. One important area of research is identifying the conditions under which a given type of integration should be used. Researchers are encouraged to contribute to this knowledge by reporting on conditions involved in their studies.