April 2006

Findings from evidence-based forecasting: methods for reducing forecast error

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

Forthcoming in the *International Journal of Forecasting*. The author has obtained permission from the publisher to include this material in *Scholarly*Commons@Penn.
Publisher URL: http://www.elsevier.com/wps/find/journaldescription.cws_home/505555/description#description

Findings from evidence-based forecasting: methods for reducing forecast error

Abstract
Empirical comparisons of reasonable approaches provide evidence on the best forecasting procedures to use under given conditions. Based on this evidence, I summarize the progress made over the past quarter century with respect to methods for reducing forecasting error. Seven well-established methods have been shown to improve accuracy: combining forecasts and Delphi help for all types of data; causal modeling, judgmental bootstrapping and structured judgment help with cross-sectional data; and causal models and trend-damping help with time-series data. Promising methods for cross-sectional data include damped causality, simulated interaction, structured analogies, and judgmental decomposition; for time-series data, they include segmentation, rule-based forecasting, damped seasonality, decomposition by causal forces, damped trend with analogous data, and damped seasonality. The testing of multiple hypotheses has also revealed methods where gains are limited: these include data mining, neural nets, and Box-Jenkins methods. Multiple hypotheses tests should be conducted on widely used but relatively untested methods such as prediction markets, conjoint analysis, diffusion models, and game theory.

Keywords
Box-Jenkins, causal forces, causal models, combining forecasts, complex series, conjoint analysis, contrary series, damped seasonality, damped trend, data mining, Delphi, diffusion, game theory, judgmental decomposition, multiple hypotheses, neural nets, prediction markets, rule-based forecasting, segmentation, simulated interaction, structured analogies

Comments
Forthcoming in the *International Journal of Forecasting*. The author has obtained permission from the publisher to include this material in *Scholarly*Commons@Penn.
Publisher URL: http://www.elsevier.com/wps/find/journaldescription.cws_home/505555/description#description

Findings from Evidence-based Forecasting: Methods for Reducing Forecast Error

Forthcoming (after revisions) in the
International Journal of Forecasting

J. Scott Armstrong

Wharton School, University of Pennsylvania

armstrong@wharton.upenn.edu

April 15, 2006 (version 43)

Abstract

Empirical comparisons of reasonable approaches provide evidence on the best forecasting procedures to use under given conditions. Base on this evidence, I summarize the progress made over the past quarter century with respect to methods for reducing forecasting error. Seven well-established methods have been shown to improve accuracy: combining forecasts and Delphi help for all types of data; causal modeling, judgmental bootstrapping and structured judgment help with cross-sectional data; and causal models and trend-damping help with time-series data. Promising methods for cross-sectional data include damped causality, simulated interaction, structured analogies, and judgmental decomposition; for time-series data, they include segmentation, rule-based forecasting, damped seasonality, decomposition by causal forces, damped trend with analogous data, and damped seasonality. The testing of multiple hypotheses has also revealed methods where gains are limited: these include data mining, neural nets, and Box-Jenkins methods. Multiple hypotheses tests should be conducted on widely used but relatively untested methods such as prediction markets, conjoint analysis, diffusion models, and game theory.

Keywords: Box-Jenkins, causal forces, causal models, combining forecasts, complex series, conjoint analysis, contrary series, damped seasonality, damped trend, data mining, Delphi, diffusion, game theory, judgmental decomposition, multiple hypotheses, neural nets, prediction markets, rule-based forecasting, segmentation, simulated interaction, structured analogies.

This paper summarizes what has been learned over the past quarter century about the accuracy of forecasting methods. It relies on empirical studies that compare 'multiple hypotheses' (two or more reasonable hypotheses). This method of reasonable alternatives implies that the current method is included along with other leading methods. Ideally, the hypotheses should specify the conditions in which the findings apply. I refer to this approach as *multiple hypotheses* and to the findings as *evidence-based*.

Evidence-based findings

In judging progress in a field, one might look at new methods and develop a rationale on why they should be useful. Consider an analogy to medical research: one could develop new treatments in the lab based on reasoning about what treatments should be most effective and have them judged by experts. In a like manner, Fildes (2006) examined the most influential new treatments in forecasting. Peer review has supported these approaches. Is this sufficient?

Continuing with the analogy to medicine, Avorn (2004) reports the following, which I have paraphrased: "In a former British colony, most healers believed the conventional wisdom that a distillation of fluids extracted from the urine of horse, if dried to a powder and fed to aging women, could . . . preserve youth and ward off a variety of diseases." The preparation became very popular. Many years later, experimental studies concluded that the treatment had little value and that it caused tumors and blood clots. The former colony is the United States and the drugs were hormone replacement products. The treatment seemed to work because those who used the drug tended to be healthier than those who did not. This was because it was used by people who were more interested in taking care of their health.

I have little faith in the value of forecasting methods until they have been empirically tested. Popular techniques have often failed when subjected to testing. So in this paper, I examine only those methods that have been empirically tested against other methods. As is the case for most research in the social and managements sciences, only a small percentage of papers are concerned with evaluation.

I looked primarily for studies that used real data to compare the *ex ante* forecasting accuracy of alternative methods. When possible, I relied upon published reviews and meta-analyses.

My search for evidence-based findings was intended to include all types of forecasting methods. Using the forecasting methodology tree at forecastingprinciples.com, I examined 17 basic methods: role playing, intentions/expectations surveys, conjoint analysis, prediction markets, Delphi, structured analogies, game theory, decomposition, judgmental bootstrapping, expert systems, extrapolation models, data mining, quantitative analogies, neural nets, rule-based forecasting, causal models, and segmentation. Brief summaries of these methods are available at forecastingprinciples.com with additional details in Armstrong (2001).

While this review focuses on the first 25 years of the International Institute of Forecasters (from its founding in 1981), many of the advances are built upon earlier work. Earlier contributions, such as the classical decomposition of time series (mean, trend, and seasonality) are not discussed if I was unable to obtain evidence from the past 25 years that related to the use of the methods.

The initial base of findings is drawn from Armstrong (2001). In that book, 39 academic researchers in forecasting summarized evidence-based principles in their areas. They were supported by 123 reviewers in an effort to ensure that all relevant evidence on the principles had been included. The principles were initially posted on an open website, forecastingprinciples.com, and appeals were made for peer review as to any information that had been overlooked.

I began to update the review in early 2005 by searching literature, contacting key researchers, and requesting help through various email lists (e.g., the Associate Editors of the *International Journal of Forecasting*, and the authors and reviewers of the *Principles of Forecasting* book). An early version of this paper was presented as a keynote address at the International Symposium on Forecasting in 2005 along with an appeal for peer review. Drafts were posted for months on forecastingprinciples.com along with a request for reviews. I also asked a number of experts to act as reviewers on this paper. I am indebted to the 22 reviewers who provided substantive contributions to the paper as well as to others who made smaller contributions. Some of these reviewers read more than one version of the paper.

Advances in methods for improving forecast accuracy in the past 25 years are summarized below. The review begins with methods that are well established, moves to “promising methods,” proceeds to those that have been tested but found to offer only limited gains, and concludes with methods that have been widely used but not well-tested.

Within each of these areas, the methods are organized by those that apply for all types of data, followed by those relevant primarily for cross-sectional data, and then those applicable to time-series data. In assessing improvements, I sought evidence on the percentage reduction in the absolute *ex ante* forecast error. When there was little evidence of error reduction, I report on the percent of the time the specified method improved accuracy. In addition to examining evidence on the accuracy of the methods, I also sought evidence on how research over the past quarter of a century has contributed to a more effective use of the methods.

A research review such as this calls for many judgments. This paper relied heavily on judgments by me, aided by 23 reviewers. Others might look at the same evidence and draw different conclusions. To address this issue, I tried to provide full disclosure about the evidence and invite others to also review this evidence or to supply evidence that might have been overlooked. Alternative viewpoints can be published as open (moderated) peer review at forecastingprinciples.com or as comments or letters to the editors at *Foresight*, the *International Journal of Forecasting*, or the *Oracle*. New evidence can be submitted as papers for *Foresight* or the *International Journal of Forecasting.*

Well-established methods

All types of data

Combining forecasts

Combining forecasts call for developing forecasts from different methods or data, then averaging the forecasts from these methods (typically using a simple average, but sometimes a median or trimmed average). When using simple averages, the absolute error of the combined forecast can be no worse than the average of the absolute errors of the components. In cases where all of the errors for all of the methods all biased in the same direction for all methods, such combining would not improve upon the average of the components. However, this does not apply to all ways of combining (e.g., the mode or median).

Combining is expected to be most useful when the methods or data differ substantially. Batchelor and Dua (1995) found that accuracy improved when the number of methods was increased (up to five) and when different types of data were used.

Equal-weights combining has been shown to be effective at reducing forecast error under most conditions. However, differential weights are occasionally useful when one has good information about which methods are most appropriate for a situation. They were used successfully to tailor the weights to the situation in rule-based forecasting (Armstrong and Collopy, 1992). For example, when uncertainty was high, less weight was placed on trend extrapolation and more on the naïve extrapolation.

A meta-analysis based on 30 studies (24 of which were conducted in the past quarter century) estimated a 12% reduction in error in comparison to the average error of the components (Armstrong 2001b). The reductions of forecast error ranged from 3 to 24%. Since this analysis, Makridakis and Hibon (2000) reported a 4.3% error reduction in the large scale M3-Competition with its 3,003 series. In some studies, combined forecasts were more accurate than even the most accurate of the component methods. Combining forecasts produced similar gains in accuracy for cross-sectional data as for time-series data.

Further research on combining should examine different ways of combining (e.g., mean vs. median) and also the conditions under which differential weights are justified.

Delphi

In the Delphi procedure, at least two rounds of forecasts are obtained independently from a small group of experts. After each round, the experts' forecasts are summarized and reported back to the experts. The experts are also informed about the reasons behind these predictions. All of the inputs are anonymous to reduce group pressure.

The Delphi procedure was developed at RAND in the 1950s and it has been widely used in businesses. However, there have been only a few multiple hypotheses tests on Delphi, with most of these of recent origin; Rowe and Wright (2001) found ten tests in the past 25 years.

Given the problems that traditional groups have with making judgmental predictions, one would expect that Delphi, with its structured approach, would improve accuracy. Rowe and Wright (2001) found that Delphi improved accuracy over traditional groups in five studies, worsened accuracy in one, and was inconclusive in two. Using an alternative benchmark, they found Delphi to be more accurate than one-round expert surveys for 12 of 16 studies, with two ties and two cases in which Delphi was less accurate. Over all of these 24 comparisons, Delphi improved accuracy in 71% and harmed it in 12%.

As might be expected, when the forecasts were in an area in which the panelists had *no* expertise, Delphi was of little value.

Much remains to be done on Delphi. What type of feedback will help to improve accuracy? Does the use of a trimmed mean lead to better results than using an average or median? How much expertise is needed? Under what conditions is Delphi most useful?

Well-established methods for cross-sectional forecasting

Causal Models

To use causal models, one must identify the dependent and causal variables, and then estimate the direction and size of the relationships. This requires much data in which there are substantial variations in each of the variables and the variations in the causal variable are independent of one another. For example, causal models might be used to predict the success of prospective job candidates based on data on the success of previous jobholders by using data on causal variables such as intelligence and prior job success. This works best when the variations in the causal variables are large and somewhat independent of one another.

Studies on the effectiveness of causal models date at least from the 1940s and research has continued since that time. This research shows that causal models reduce errors in comparison to unaided judgments (the most common approach to making forecasts with cross-sectional data). Grove et. al's (2000) meta-analysis, based on 136 studies (primarily from psychology, personality assessments, educational achievement, mental health, and medicine), found that causal models based on regression analysis reduced errors by 10% on average. The causal models were more accurate than unaided judgment in 88% of the comparisons.

Perhaps the most important gain in knowledge over the past quarter century has been in identifying the conditions under which unaided expert judgments are more accurate than the models (Grove et. al 2000). For example, for the few studies in which judgment was more accurate than the models, the judges generally had more information. Jørgensen, (2004b), however, found that structured judgment produces similar gains in some situations

Research is needed on how to gain acceptance for causal models. A few organizations, such as football and baseball teams have adopted causal models with much success. Mostly, however, these findings are met with incredulity by practitioners, who counter with situations where they think that causal models would not improve accuracy.

Judgmental bootstrapping

What if there is insufficient information to develop causal variables either due to a lack of useful data on the dependent variable or to a lack of independent variation in the causal variables? This issue was solved in the early 1900s with a method that is now known as judgmental bootstrapping. It involves developing a model of an expert by regressing his forecasts against the information that he used. The general proposition seems preposterous: It is that the model of the man will be more accurate than the man. But there is some sense to it: The model applies the man’s rules more consistently than he does.

Judgmental bootstrapping has been found to be more accurate than unaided judgment (the normal forecasting method for these situations) in 8 of 11 comparisons, with two tests showing no difference, and one showing a small loss. The typical error reduction was about 6% (Armstrong 2001a). Four of these bootstrapping studies were done in the last quarter century. They have helped to demonstrate the improved accuracy and extended the work to an applied management problem (e.g., advertising), and showed a condition under which it does not help. The failure occurred when experts used incorrect rules; in this case, bootstrapping applied incorrect rules more consistently and thus harmed accuracy (Ganzach et al 2000).

Additional research is needed on the conditions under which judgmental bootstrapping is most useful. However, the primary need is to determine how to most effectively gain acceptance of bootstrapping. Although it can improve accuracy and reduce costs for repetitive forecasts (because the model can be used automatically), judgmental bootstrapping is rarely used. It has had high profile uses by the Dallas Cowboys football team. Many years ago, the owner of the Philadelphia Flyers hockey team told me that he uses it and freely discusses it with other owners, secure in the belief that none of them would use it.

One possibility for implementation would be to develop software to guide people through this process. This could describe conditions where bootstrapping is relevant, provide instructions for the selection of experts, describe how to collect relevant information, provide a regression program, and offer a report writing template.

Structured judgment

Judgment can be structured in many ways. Some of the key elements involve providing checklists, using systematic and well-summarized feedback on the accuracy of an expert's forecasts (used effectively in weather forecasting), helping experts to focus on relevant information (used by structured employment interviews), asking experts to justify their predictions, using independent experts with a diversity of information (hopefully with different biases), decomposing problems, and using simple heuristics.

Although pop management literature extols the value of intuition or gut feelings, a substantial amount of evidence, much from the past 25 years, suggests that a number of approaches to structured judgments are substantially more accurate than unstructured judgments. For summaries of this evidence, see the section on "Expert Opinions" in Armstrong (2001).

Structure can sometimes improve the accuracy of judgment to the extent that it is comparable to that from causal models. Jørgensen (2004b) examined 12 guidelines for structuring experts' judgmental predictions of time needed to complete software projects. Using findings from 15 studies on software development costs, he found that structured judgments were more accurate than those from causal models in five studies, the same as causal in five, and less accurate in five.

Well-established methods for time-series forecasting

Causal Models

Causal models have been widely used for time-series forecasting. Much research has been done on causal models, including many multiple hypotheses studies. Allen and Fildes (2001) found that where there were good data, causal models were more accurate than non-causal extrapolations. For 80 *ex ante* comparisons from a variety of studies involving medium to long-range forecasts (most related to economics), causal models were more accurate than extrapolative models by almost a 3 to 1 ratio (see their Table A4 on forecastingprinciples.com). Focusing on long-range forecasting, I had found a 7 to 1 advantage for causal models over extrapolation in Armstrong (1985). While I was unable to find a meta-analysis of the expected error reduction, it was quite large a some studies. For example, in a study of the air travel market, an econometric model reduced error consistently as the forecast horizon increased from 1 to 6 years ahead (Armstrong 1985, p. 405); an unweighted average of horizons 2 through 6 (representing 15 ex ante forecasts) showed that the MAPE for the econometric model was 72% less than that for the extrapolation model.

Because they can include policy variables (such as the price of a product), causal methods are useful for forecasting the effects of decisions in government and business. This is particularly true when one has good domain knowledge, accurate data, the causal variable has a strong effect on the dependent variable, and the causal variable will change substantially (as is common when forecasts cover a lead time of many years). For example, causal models should be useful for a situation in which a manager would like to know the effects of a large change in prices.

Econometricians have devoted enormous efforts over the past 25 years in searching for ways to improve econometric methods. Unfortunately, while the complexity of the methods has increased, these efforts have not been shown to improve accuracy. There is a danger that the added sophistication of the methods leads forecasters to rely more on the data analysis and less on prior knowledge. In addition, with some exceptions (e.g., Allen and Fildes 2001), little attention has been paid to studying the conditions under which econometric methods are most useful.

Damped trend

Damped trend involves putting less emphasis on the trend extrapolation as uncertainty increases. Gardner (1985) showed that trend damping improves accuracy for extrapolation models. Gardner's exponential smoothing with damped trends was tested in a study involving the 1,001 series from the M-Competition (Gardner and McKenzie, 1985). Trend damping reduced forecast errors by 10.5% on average over forecast horizons 1-18 when compared with traditional exponential smoothing with a linear trend (Makridakis et al., 1982). In the M3-Competition (Makridakis and Hibon 2000), trend damping reduced forecast errors by 6.6% compared to traditional exponential smoothing. Gardner (1990) also tested trend damping in a Navy distribution system with more than 50,000 inventory items; implementation of trend damping reduced Naval inventory investment by 7% compared to the method that had been used previously, simple exponential smoothing. In addition, compared to reasonable benchmarks, trend damping has been successful on data related to annual consumer product sales (Schnaars 1986), cookware sales (Gardner and Anderson 1997), computer parts (Gardner 1993), and production planning (Miller and Liberatore 1993). However, Fildes et al. (1998), in a study involving 261 telecommunications series for horizons of up to 12 months, found that damping *reduced* the accuracy of Holt's exponential smoothing by 16.8% (based on the two criteria in their table 4); given these unusual findings, an independent replication would be useful. In all, there have been ten multiple-hypotheses studies on damped trend. They have led to an average error reduction of about 4.6%.

Further research would be useful to determine the conditions under which damping is most useful. In addition, there may be other ways to use trend damping.

Promising findings with limited evidence on accuracy

This section contains forecasting methods that show promise, but where evidence is limited. In my search, I favored methods that showed large improvements in accuracy. I avoided one-shot studies with small improvements if there was no prior reason to expect that they would be more accurate. I also avoided murky papers: It is the responsibility of researchers to ensure that their papers are clearly written.

Promising methods for cross-sectional data

Damped causality

In a large-scale study of cross-sectional data, Dana and Dawes (2004) showed that the gain from equal weights is larger when sample sizes are smaller and predictability is poor. These studies compared two extremes: equal weights versus regression weights. (Equal weight applies to variables that have been transformed to standard deviations.) The optimal approach most likely lies in between these two methods.

Damped estimates of causality should also be relevant for time-series data because one must contend with the uncertainty involved in forecasting the causal variables. This suggests damping coefficients toward zero or damping the forecasts of the changes in the causal variables. However, this is speculative as I was unable to find multiple hypotheses tests on real-world data.

Simulated interaction

Simulated interaction is a form of role-playing in which an administrator describes the target situation and the protagonists' roles, then provides a list of possible decisions. Role players adopt a role and read about the situation. They then improvise realistic interactions with the other role players until they reach a decision. The typical session lasts less than an hour. The role players' decisions are used to develop a forecast.

A similar procedure has been used by the military since the 1920s and in jury trials since the 1970s. It has rarely been used in businesses although there have been published reports on its use in marketing and personnel selection. Despite this long-term use, no multiple hypotheses tests were published prior to 1987.

Simulated interaction is expected to be particularly useful in conflicts (such as in buyer/seller negotiations, union/management relations, legal cases, wars, and terrorism) because it is so difficult to think through the many actions and reactions among the parties involved. Simulated interactions allow for a realistic representation. Relative to the current forecasting method (expert judgment) simulated interactions reduced forecast errors by 57% in the eight situations tested to date (Green 2002, 2005). The gains were achieved even though the roles were played by university students who had little knowledge of the types of conflict situations being used. Further information is available at conflictforecasting.com.

Structured analogies

In everyday life, people often refer to analogies when making forecasts. For example, some advised against a military action in Iraq because they saw it as similar to Viet Nam. However, this use of analogies is usually done in an unstructured manner; analogies are likely to be generated in support of a desired outcome. The assumption behind structured analogies is that experts can provide useful information about analogies, but they are not effective at translating this information into forecasts. The latter task should be done in a mechanical manner to avoid biases.

In the structured analogies method, an administrator prepares a description of the target situation and selects experts who have knowledge of analogous situations. Then, the experts identify and describe analogous situations, rate their situation's similarity to the target situation, and match the outcomes of their analogies with potential outcomes in the target situation. The administrator then derives forecasts from the information the experts provided.

Green and Armstrong (2006) obtained structured analogies forecasts for eight conflict situations (e.g., one of the situations involved a country in the Middle East that built a dam that reduced water to a country downstream; how would the conflict be resolved?) When experts were able to report on two or more analogies, and where a mechanical rule (use the analogy that the expert identified as most similar o the target situation) was then used to make a forecast, there was a 41% reduction in error as compared to using unaided experts to make the forecasts.

Judgmental decomposition

Judgmental decomposition refers to the multiplicative breakdown of a problem. Experts make estimates of each component, and these are multiplied. For example, one could estimate a brand's market share and the total market, and multiply estimates to get a sales forecast. This method is relevant for situations where one knows more about the components than about the target variable. Thus, the analyst should identify segments that are easy to predict.

Decomposition is especially important when there was high uncertainty in predicting the target variable. MacGregor (2001, Exhibit 2) summarized results from three studies (two done since 1988) involving 15 tests and found that judgmental decomposition led to a 42% reduction in error under high uncertainty.

Summary of promising methods for cross-sectional data

Table 1 summarizes the gains in accuracy for the promising methods that relate to cross-sectional data. The table lists the gains that were achieved for the conditions stated. The conditions were narrow. Given the evidence to date, simulated interaction and structured analogies apply only to conflict situations, and the gains for judgmental decomposition apply only when there is high uncertainty and when one has better knowledge of the components than of the global values.

Insert Table 1 about here

Promising methods for time-series data

Segmentation

Segmentation can allow for more effective use of information. Jørgensen (2004) for example, reports that experts prefer the bottom-up approach as it allows them to more effectively use their knowledge about the problem. Segmentation is also advantageous because the forecasting errors in the different segments may offset one another. Assume that you had ten divisions in a company. You might improve accuracy by forecasting each division separately, then adding the forecasts. But there is a problem. If the segments are based on small samples and erratic data, the segment forecasts might contain very large errors.

Armstrong (1985, p. 287) reported on three comparative studies on segmentation that were conducted since 1975. The problems were broken into segments, and then each segment was forecasted either by extrapolation or regression. Segmentation improved accuracy for all three studies. In addition, Dangerfield and Morris (1992), in their study on bottom-up forecasting, found that segmentation was more accurate for 74% of 192 monthly time series from the M-Competition. In a study involving seven teams making estimates of the time required to complete two software projects, Jørgensen (2004) found that the error from the bottom-up forecast was 51% less than that for the top-down approach.

Structured judgmental adjustments

Forecasters often make unstructured judgmental adjustments to times-series forecasts. These can be a source of bias. For example, managers could inflate a sales forecast in the belief that this will motivate employees. Salesmen might deflate a forecast so it is easier to exceed. Therefore, it was not surprising when some early studies concluded that unstructured adjustments often harmed forecasts.

There are several ways to structure judgmental adjustments. These include providing written instructions for the task, soliciting written adjustments, requesting independent and anonymous adjustments from a group of experts, asking for adjustments to be made prior to seeing the forecasts, and recording reasons for revisions.

Goodwin's (2005) review described nine papers published since 1989 with evidence on judgmental adjustments. Judgmental adjustment are useful when:

a) recent events are not fully reflected in the data (e.g., last-minute price reductions for a product). Thus, adjustments might be made to revise the current level of the variable being forecast.
b) historical data are limited.
c) experts possess good domain knowledge about future changes that have not been included in the model as, for example, a sales forecast given a recently planned product improvement. .

Findings to date suggest that minor revisions should be avoided, perhaps because they lead to over-adjustments.

Further research is needed on the conditions under which judgmental adjustments are useful. For example, there may be cases, such as adjusting for recent events, where mechanical adjustments might sufficient.

Rule-based forecasting

Rule-based forecasting (RBF) is a method for weighting and combining extrapolation methods. It integrates judgment and statistical procedures to tailor forecasts to the situation. RBF does this primarily by identifying key features in time series and by capturing managers' knowledge of the domain and expectations about direction of the trend (causal forces).

Empirical results on multiple sets of time series have indicated that RBF produces forecasts that are more accurate than traditional methods and than equal-weights combining of forecasts. RBF is most useful when one has good domain knowledge, the domain knowledge has a strong impact, the series is well behaved, and there is a strong trend in the time-series data. When these conditions do not exist, RBF neither harms nor improves accuracy (Collopy and Armstrong, 1992). Given only a modest amount of domain knowledge, for one-year ahead ex *ante* forecasts of 90 annual series, the MdAPE for RBF was 13% less than that from equally weighted combined forecasts. For six-year ahead *ex ante* forecasts, it had an MdAPE that was 42% less. In comparison with equal-weights combining, RBF was more accurate only for those series for which there was domain knowledge (Collopy and Armstrong, 1992). Adya (2000) replicated these findings after correcting minor mistakes in the rule-base. In the M3-Competition, RBF was run using automatic procedures (Adya et al. 2001) and without any domain knowledge. RBF was the most accurate of the 22 methods for annual forecasts involving 645 series and six-year horizons (Makridakis and Hibon, 2000; Adya, et al. 2000). Its symmetric MAPE (Mean Absolute Percentage Error) was 3.8% less than that for combining forecasts. Vokurka, et al (1996) tested an alternative version of RBF using the same 126 series as used by Collopy and Armstrong (1992); although they did not use domain knowledge, the MdAPE for six-year ahead annual forecasts was 15% less for RBF than that for equally weighted combined forecasts.

Some of the rules can be applied in a simple manner. For example, when managers' knowledge about causal forces (expected direction of trend) conflicts with historical trends, a situation referred to as "contrary series," traditional extrapolation methods produce enormous errors. A simple rule for contrary series is to forecast that there will be no trend. When tested on M-Competition data (Makridakis et al 1982) and with data from four other data sets, the Median Absolute Percentage Error (MdAPE) was reduced by 17% for one-year-ahead forecasts and by over 40% for six-year-ahead forecasts (Armstrong & Collopy 1993). Further information is provided on the Rule-based Forecasting Special Interest Group at forecastingprinciples.com.

Decomposition by causal forces

Complex series are defined as those in which causal forces drive the series in opposite directions. Perhaps the most common situation is when forecasting revenues of a product (such as computer software) where the price is decaying, the number of units and inflation are growing, and market share trends depend on the comparative advantages of the software. If the components of a complex series can be forecast more accurately than the global series, it helps to decompose the problem by causal forces (Armstrong, Collopy and Yokum 2005). For example, to forecast the number of people that die on the highways each year, forecast the number of passenger miles driven (a series expected to grow), and the death rate per million passenger miles (a series expected to decrease), then multiply these forecasts. When tested on five time series that clearly met the conditions, decomposition by causal forces reduced forecast errors by two-thirds. For the four series that partially met the criteria, the errors were reduced by one-half. Although the gains in accuracy were large, there is only a single study on decomposition by causal forces.

Damped seasonal factors

Miller and Williams (2003, 2004) developed a procedure for damping seasonal factors. Given uncertainty and errors in the historical data, their procedure damps the seasonal factors (e.g., multiplicative factors are drawn towards 1.0 and additive seasonal factors towards zero). This is useful because estimated seasonal factors are affected by errors in the data. The Miller-Williams procedures reduced forecast errors by about four percent in tests involving the 1,428 monthly time series from the M3-Competition. The damped seasonal forecasts were more accurate for 68% of these series.

Bunn and Vassilopoulos (1999) damped seasonal estimates by averaging those for a given series with seasonal factors estimated for a set of related series. This approach reduced forecast error by about 20%. When Gorr, Oligschlager & Thompson (2003) pooled monthly seasonal factors for crime rates for six precincts of a city, the forecasts were 7% more accurate than when the seasonal factors were estimated individually for each precinct. On average then, averaging seasonal factors across series led to a 13.5% error reduction.

The estimation of seasonal factors might be improved through the use of domain knowledge, especially for short series. For example, experts can avoid the use of seasonal factors in areas where there is no reason to expect seasonality (such as in the stock market) and rely on them in areas where seasonal factors are obvious, such as for ice cream sales. Finally, because uncertainty increases over time and seasonal influences might change, increased damping might improve accuracy for longer time horizons.

Summary of promising methods for time-series data

Table 2 summarizes the gains for the promising methods for time series. The gains apply only for the conditions stated. In other words, the gains would be expected to be less as one departs from these conditions.

Insert Table 2 about here

Tested areas with little gain in accuracy

Evidence–based forecasting can also show what does not work. Some areas with much research efforts have shown limited gains in accuracy. In some cases, this may be due to the small number of papers testing multiple hypotheses. In other cases it might simply be that the effects on accuracy are so small that they are difficult to measure.

Time-series forecasts

Data mining

The key assumption of data mining is that, given large amounts of data, statistical analysis can determine patterns that will aid in forecasting. Similar to the approach used in stepwise regression, data mining ignores theory and prior knowledge. It searches for patterns in the data.

Data mining is popular in forecasting. In March 2006, a Google search using the term “data mining” and either “prediction” or “forecasting,” Google produced 2.2 million sites. The interest in data mining is aided by the availability of large data sets such as those obtained from scanners in retail stores.

Keogh and Kasetty (2002) conducted a comprehensive search for empirical comparisons of data mining. They criticize the failure of data mining researchers to test alternative methods. To address this problem, they found procedures from two dozen papers on data mining, which they then tested on 50 real-world data sets. They found little gain from data mining. Keogh (personal correspondence) concluded:

> “[Professor X] claimed to be able to do 68% accuracy. I sent them some "stock" data and asked them to do prediction on it, they got 68% accuracy. However, the "stock" data I sent them was actually random walk! When I pointed this out, they did not seem to think it important. The same authors have another paper in [the same journal], doing prediction of respiration data. When I pointed out that they were training and testing on the same data and therefore their experiments are worthless, they agreed (but did not withdraw the paper). The bottom line is that although I read

> every paper on time-series data mining, I have never seen a paper that convinced me that they were doing anything better than random guessing for prediction. Maybe there is such a paper out there, but I doubt it."

In general, methods that have ignored theory, prior evidence, and domain knowledge have had a poor record in forecasting. For example, stepwise regression has not been shown to improve forecasting accuracy.

Advocates of data mining have argued that the area encompasses many methods and that without a good definition, one cannot evaluate the methods. However, in my search for evidence I have asked for evidence relating to any approach to data mining. I have been unable to find any evidence.

Neural nets

Neural nets, which are designed to pick up nonlinear patterns from long time series, have been an area of great interest to researchers. Wong, Lai & Lam (2000) found over 300 research papers published on neural nets during 1994-1998. Early reviews on the accuracy of neural nets were not favourable; Chatfield (1993) refers to some of these. However, Adya and Collopy (1998) found eleven studies that met the criteria for a comparative evaluation, and in 8 of these (73%), neural nets were more accurate. There were no estimates of the error reductions versus alternative methods although Liao and Fildes (2005), in a test involving 261 series, 18 horizons, and 5 forecast origins, found impressive gains in accuracy for neural nets with an error reduction of 56% compared to damped trends. Chatfield (personal correspondence) suspects that there is a 'file-drawer problem,' saying that he knew of some studies that failed to show gains and were not submitted for publication, and there is a well-known bias by reviewers against papers with null results. Also, the comparisons were sometimes against less effective forecasting methods, not against damped trend or combining.

Because all studies are not equal, I turned to the large-scale M3-Competition with its 3,003 varied time series. Here, neural nets were 3.4% less accurate than damped trends and 4.2% less accurate than combined forecasts.

Given, the mixed results on accuracy and the difficulties in using and understanding neural nets, my conclusion is that too much research effort is being devoted to this method. On the other hand, the impressive findings from Liao and Fildes (2005) deserve further attention in an effort to discover the conditions under which neural nets are useful. For the latest on neural nets, see the special interest group at forecastingprinciples.com.

Box-Jenkins methods

Researchers have published an immense number of studies using Box-Jenkins methods for the extrapolation of time series. Fildes (2006) identified this as one of the most influential innovations in forecasting (based on number of citations and on expert opinion). The interest has spread beyond academics: a March 2006 Google search on “Box-Jenkins” and either “forecasting” or “prediction” produced 120,000 hits.

Some early studies showed promise. However, using the M-Competition study (Makridakis et al. 1982), I compared average MAPE for Box-Jenkins (BJ) with combining over the 18 forecast horizons for the 111 series in which there were comparisons. In this analysis, the BJ forecasts were 1.7% less accurate. In the M2-Competition, with 29 series (23 from companies), I examined the MAPE over 15 horizons for the 3-year period; the BJ forecasts were 27% less accurate than either damped trend or combined exponential smoothing (Makridakis, et al. 1994). For the M3-Competition, none of the BJ models yielded forecasts that were as accurate as the combined forecasts for any of the ten forecast horizons reported in the table. As a crude measure, I averaged the symmetric MAPE errors across the 3,003 series and 18 forecasts horizons and found that the four BJ models were, on average, 7.6% less accurate than damped trends and 8.3% less accurate than combining (Makridakis and Hibon 2000, Table 6).

Widely used methods that have been subject to little testing

The following methods that have been widely used, but there has been little testing. Multiple hypotheses studies are needed in these areas.

All types of data

Prediction markets

Studies by psychologists in the early 1900s showed that accuracy could be improved by aggregating across a large number of people. This had been obvious many years before as people had used markets to predict what would happen in politics and sports. If you want to get an unbiased forecast of future events by judgment, create a market and let people bet on possible outcomes.

Most comparative testing of prediction markets (information markets) has been done in financial and commodities markets and in sports. While I was unable to find a meta-analysis in these areas, a large number of studies have been published in which forecasters have struggled in vain since the 1920s to develop methods that are superior to financial markets. In addition, small-sample studies show that betting markets were more accurate than political polls for forecasting political elections (Wolfers and Zitewitz 2004).

Over the past quarter century, little research has been done to improve our knowledge about the use of prediction markets. It would be useful to test prediction markets against other structured group methods, such as Delphi. In addition, we know little about the conditions under which prediction markets are most useful. It seems likely, for example, that if people know little about a situation, there would be little to gain from a prediction market. In addition, if their knowledge were generally wrong, there might be little benefit. For example, contrary to a large body of empirical research, most people believe that a minimum wage law is good for the economy and for poor people. What could a prediction market add if a forecast was needed on the impact of a change in the minimum wage?

According to Surowiecki (2004), prediction markets are being used for forecasting within companies. It seems reasonable to expect them to be more accurate than traditional meetings, but this be expected for nearly any structured method.

Cross-sectional data

Conjoint analysis

In conjoint analysis, people are asked to state their preferences from pairs of offerings. For example, various features of a laptop computer, such as price, weight, and battery life might be varied to develop a set of offerings. The values for the features are varied so that they do not correlate with one another. These offerings would then be presented to a sample of potential customers to assess the likelihood that they would purchase each offering. Their responses can be analysed by regressing their choices against the product features. The method is called "conjoint analysis" because respondents *consider* the product features *jointly*. This forces them to consider tradeoffs among the various features of the computer.

Conjoint analysis is analogous to judgemental bootstrapping except that one is examining customers' preferences for each offering rather than experts' judgments about the sales volumes. The approaches can be used on similar problems, such as for new product forecasting. However, I have been unable to find any comparisons of these methods. Certainly the judgmental bootstrapping approach should be much less expensive as needs only about 5 to 20 experts, where as conjoint analysis might require hundreds of potential customers.

Although conjoint analysis seems to be based on solid principles, there have been no tests against alternative methods, despite repeated calls for such research (Wittink & Bergestuen 2001).

Game theory

Game theorists have studied the behavior of subjects in various games, such as the Prisoners' Dilemma. A number of researchers and consultants have suggested that the behavior in such games can be used to predict behavior in the real world. In April 2006, A Google search of "game theory" and either "forecasting" or "prediction," produced 950,000 sites. While game theory has much intuitive appeal, attempts to find comparative studies on the value of game theory for forecasting have been unsuccessful.

In a related study, however, Green (2005) asked game theorists to use game theory to make predictions for eight conflict situations. The game theorists were also expected to benefit from their long experience with conflict situations as well as by their ability to use game theory. As it turned out, their predictions were no more accurate than those made by university students.

Structured judgmental adjustments

Judgmental adjustments of cross-sectional predictions are common. For example, one might have a model to decide whether someone should undergo a medical operation. In contrast to time-series forecasting, however, judgmental adjustments do not seem to improve cross-sectional predictions. Meehl (1956), in reviewing the evidence on predictions about people, concluded " . . . it almost looks as if the first rule to follow in trying to predict the subsequent course of a student's or patient's behavior is to carefully avoid talking to him, and the second rule is to avoid thinking about him." This conclusion also applies to personnel predictions because employers over-ride the forecasts with irrelevant information. Grove et al. (2000), in their meta-analysis, found further support; when judges had access to interviews with the subject, their predictions were less accurate.

Meehl's advice was followed with great success by the general manager of the Oakland Athletics baseball team (Lewis 2004). He intentionally avoided watching games so his evaluation of a player would not be affected by his judgment; instead, he used statistics to make his decisions.

Time series data

Diffusion models

Diffusion models assume that a series will start slowly, begin a rapid rise, and then slow and gradually approach a saturation level. This has great intuitive appeal as one can see when plotting the historical sales of products such as refrigerators, TVs, and personal computers. However, despite the substantial research on diffusion models, there have been few tests of comparative forecast accuracy to date, and these tests have produced mixed results. Meade and Islam (2001), in their review of the accuracy of alternative methods, found that no one diffusion model dominated and that relatively simple diffusion models were about as accurate as more complex diffusion models.

Discussion

In November 2005, I searched for "forecasting OR predicting" among the titles and topics on the Social *Science Citation Index (SSCI)*: This yielded over 15,000 papers. In addition, Armstrong and Pagell (2003) estimated that only 42% of the papers that contribute to forecasting are found by key word searches in the title or topic. Thus, there may have been about 35,000 papers relevant to forecasting (that excludes the *Science Citation Index,* which has substantially more forecasting studies than the SSCI).

I made a rough estimate that this paper draws upon 300 multiple hypotheses studies related to assessing forecast accuracy; this represents less than one percent of the total papers published on forecasting. Perhaps there is an additional one percent of multiple hypotheses studies related to other aspects of forecasting such as assessing uncertainly or gaining acceptance. Using a much different approach, Armstrong and Pagell (2003) estimated that only 3% of the papers in forecasting assessed which methods contribute to the development of forecasting principles.

By comparing this paper with Fildes (2006), you will find little overlap between papers that are influential among academics and those that use multiple hypotheses testing of forecasting procedures. This is not to imply that the advances noted by Fildes will not be shown to be useful, but merely that the jury awaits findings from evaluation research. It does imply a serious gap between research and practice.

Despite extensive help on this paper from colleagues, relevant papers have no doubt been overlooked here. In addition, I expect that there are useful methods that have not yet been properly tested. I plan to put this paper on forecastingprinciples.com to allow for open peer review and thus to allow for additions and challenges to my conclusions.

The methods that I referred to as promising are in need of replication. As has been shown in other fields, promising findings often fail when attempts are made at replication. My expectation is that additional research will identify conditions under which these methods fail.

A number of seemingly useful methods (such as prediction markets and conjoint analysis) will continue to be applied even though research lags. Given that these methods are consistent with basic forecasting principles, I applaud their use. Nevertheless, we could learn much about the conditions under which they are most useful.

The findings over the past quarter century are not so surprising as might seem at first glance. The successful methods follow some basic generalizations in forecasting:

- *Be conservative when uncertain*: Damping methods are based on the need to be conservative in the face of uncertainty.
- *Spread risk*: Decomposition, segmentation and combining are based on spreading risk, while efforts to find the single best method often lead one astray.
- *Use realistic representations of the situation*: For example, simulated interaction and structured analogies represent situations in a realistic manner, while game theory does not.
- *Use lots of information*: Methods that use more information (e.g., combining, prediction markets) are superior to those that rely on a single source (e.g., exponential smoothing).

• *Use prior knowledge:* Methods based on prior knowledge about the situation and relationships (e.g., econometric methods) are superior to those that rely only on the data (e. g., exponential smoothing, Box-Jenkins, and data mining).

• *Use structured methods:* Structured methods are more accurate than unstructured methods. This shows up, for example, in the judgmental adjustments to forecasts.

Speculation

What new areas might lead to improved accuracy over the next 25 years? Two areas stand out in my judgment: index methods and combined methods.

Index methods

Consider a situation where you have many variables that affect an outcome. This occurs, for example, when forecasting growth in national incomes. By my count, there are at least 50 causal variables that can have an important effect on growth. For this problem, there is not enough data to obtain estimates for traditional causal models (such as those estimated by regression). However, such problems can be approached through the use of index methods.

To use an index method, an analyst prepares a list of all variables that might have an effect of the variable to be forecast (e.g., national income), then determines for each observation (e.g., France) whether each variable is favorable (+1), unfavorable (-1) or indeterminate (0). In its simplest form, he could then add the scores and use the total in making forecasts. Thus, each variable has the same weight. This use of judgmental indexes has also been referred to as "experience tables."

This simple, easily understood method is expected to aid forecasting in situations where there are many causal variables, good domain knowledge about which variables are important and about the direction of effects, and limited data. These conditions apply where discrete choices must be made, such as for the selection of personnel, retail sites, investment opportunities, product names, or advertising campaigns.

Lichtman (2005) reported that his "Keys model," based on an equally weighted index of 13 variables, picked the winner of every U. S. presidential election since 1860 (retrospectively through 1980 and prospectively from 1984-2004). This record cannot be matched by any of the traditional quantitative models, which are based on regression analyses involving three to five variables.

Combined methods

There may be benefits in combining forecasting methods. For example, judgmental bootstrapping might be used along with regression analysis to develop a hybrid causal model. Judgmental bootstrapping could be used to obtain estimates for variables for which there is insufficient information (for example, a causal variable such as price might have remained constant over a given time series).

The integration of judgment and quantitative methods is promising. Indeed, this has been one of the key areas for research over the past quarter century. A review by Armstrong and Collopy (1998) found 47 papers related to the integration of judgment and quantitative methods; all but 2 of these studies were published in the last quarter century.

Conclusions

Studies that have used multiple reasonable hypotheses for important problems have led to much progress. These studies are especially helpful when they describe the conditions under which hypotheses apply.

Over the past quarter century, evidence from comparative studies has led to seven well-supported forecasting methods. Two of these methods apply to all types of data: combined forecasts with an estimated 12% error reduction, and Delphi, which improved accuracy in of 19 of 24 comparisons (79%). Three methods apply to cross-sectional data: causal models with a 10% error reduction, judgmental bootstrapping at 6%, and structured judgment for which I had no estimate of the error reduction. Two methods apply to time series data: damped trend with a 5% error reduction, and causal models with improved accuracy over 3/4 of the ex ante comparisons for medium to long-range forecasts. Practitioners should implement these well-established methods.

Researchers should give particular attention to the methods designated as promising. The evidence is sparse for these promising areas (e.g., there is only one study on decomposition by causal forces). However, the size of the gains suggests that they are worthy of use in firms. The gains in these promising areas ranged from 4% to 67%. In addition, each of these methods is consistent with basic principles of forecasting. In judging the value of these promising methods, one should consider the conditions. For example, while damping with analogous series produced only a 5% error reduction, the conditions are quite common. The same applies to the 4% error reduction when using damped seasonality.

The testing of multiple reasonable hypotheses has also identified areas that offer little promise even after much research. These include neural nets, data mining, and Box-Jenkins methods. Perhaps the major flaw in this methods is that they do not effectively use domain knowledge.

Some widely used methods, such as prediction markets, conjoint analysis, and game theory, would benefit from multiple hypotheses testing. Do they improve accuracy, and if so, under what conditions? What are the most effective ways to use these methods?

Multiple hypotheses studies have proven to be useful for advancing knowledge on the accuracy of forecasting methods. We need more of these studies, especially when they produce surprising results for important problems.

Method	Conditions	General research effort	Multiple hypotheses studies	Multiple hypotheses tests	Percent error reduction
Simulated interaction	conflicts with 2 or more parties	low	6	9	57
Structured analogies	conflicts with 2 or more analogies per expert	very low	1	8	41*
Judgmental decomposition	very large or small numbers; knowledge of components	low	3	15	42

*Based on single study

Table 2: Promising methods for time-series data

Method	Primary Conditions	General research effort	Multiple hypotheses studies (tests)	Alternative methods	% error reduction (% better)
Segmentation	none	low	6	top down; global	51* (74)*
Structured adjustments	recent events, limited data & domain knowledge	moderate	14	unadjusted forecasts	NA
Rule-based forecasting/ causal forces	domain knowledge, annual data & long-term	low	5/1	combining	42
Decomposition by causal forces	conflicting forces	very low	1 (5)	global forecasts	67*
Damped trend with analogous data	small samples	low	1	undamped trend	5*
Damped seasonality	uncertainty; analogous data	low	3	undamped seasonality	4*; 13* (68)*

* Based on single studies

References

[All papers by Armstrong are available in full-text at jscottarmstrong.com]

Adya, M. (2000). Corrections to rule-based forecasting: Results of a replication. *International Journal of Forecasting*, 16, 125-127.

Adya, M., Armstrong, J. S., Collopy, F. & Kennedy, M. (2000). An application of rule-based forecasting for a situation lacking domain knowledge. *International Journal of Forecasting*, 16, 477-484

Adya, M. & Collopy, F. (1998). How effective are neural networks at forecasting and prediction? A review and evaluation. *Journal of Forecasting*, 17, 491-485.

Adya, M., Collopy, F., Armstrong, J. S. & Kennedy, M. (2001). Automatic identification of time series features for rule-based forecasting. *International Journal of Forecasting*, 17, 143-157.

Allen, G. & Fildes, R. (2001). Econometric forecasting, in J. S. Armstrong (ed.), *Principles of Forecasting*. Boston: Kluwer Academic Publishers.

Armstrong, J. S. (1985). *Long-range Forecasting.* New York: John Wiley.

Armstrong, J. S. (2001). *Principles of Forecasting*. Boston: Kluwer Academic Publishers.

Armstrong, J. S. (2001a). Judgmental bootstrapping, in J. S. Armstrong (ed.), *Principles of Forecasting*. Boston: Kluwer Academic Publishers.

Armstrong, J. S. (2001b). Combining forecasts, in J. S. Armstrong (ed.), *Principles of Forecasting*. Boston: Kluwer Academic Publishers.

Armstrong, J. S. & Collopy, F. (1993). Causal forces: Structuring knowledge for time-series extrapolation. *Journal of Forecasting*, 12, 103-115.

Armstrong, J. S. & Collopy, F. (1998). Integration of statistical methods and judgment for time-series forecasting: Principles from empirical research, in G. Wright and P. Goodwin (eds.), *Forecasting with Judgment.* John Wiley, 69-293

Armstrong, J. S., F. Collopy, F. & Yokum, T. (2005). Decomposition by causal forces: A procedure for forecasting complex time series. *International Journal of Forecasting*, 21, 25-36.

Armstrong, J. S. & Pagell, R. (2003). Reaping benefits from management research: Lessons from the forecasting principles project. *Interfaces*, 33 (5), 1-21.

Avorn, Jerry (2004), *Powerful Medicines*. New York: Alfred A. Knoff

Batchelor, R. & Dua, P. (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science*, 41, 68-75.

Bunn, D.W. & Vassilopoulos, A. I. (1999). Comparison of seasonal estimation methods in multi-item short-term forecasting. *International Journal of Forecasting*, 15, 431-443.

Chatfield, C. (1993). Neural networks: Forecasting breakthrough or passing fad? *International Journal of Forecasting*, 9, 1-3

Collopy, F. & Armstrong, J. S. (1992). Rule-based forecasting. Development and validation of an expert systems approach to combining time series extrapolations. *Management Science*, 38, 1394-1414.

Dana, J. & Dawes, R. M. (2004). The superiority of simple alternatives to regression for social science predictions. *Journal of Educational and Behavioral Statistics*, 29(3), 317-331.

Dangerfield, B. J. & Morris, J. S. (1992). Top-down or bottom-up: Aggregate versus disaggregate extrapolations. *International Journal of Forecasting* 8, 233-241.

Fildes, R. (2006). xxxxxxxxxx *International Journal of Forecasting*, (this issue).

Fildes, R., Hibon, M., Makridakis, S. & Meade, N. (1998), Generalizing about univariate forecasting methods: Further empirical evidence. *International Journal of Forecasting*, 14, 339-358.

Ganzach, Y. A., Kluger, N. & Klayman, N. (2000). Making decisions from an interview: Expert measurement and mechanical combination. *Personnel Psychology*, 53, 1-20.

Gardner, E.S., Jr. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting*, 4, 1-28.

Gardner, E.S., Jr. (1990). Evaluating forecast performance in an inventory control system. *Management Science*, 36, 490-499.

Gardner, E. S., Jr. (1993). Forecasting the failure of component parts in computer systems: A case study. *International Journal of Forecasting*, 9, 245-253.

Gardner, E. S., Jr. & Anderson, E.A. (1997). Focus forecasting reconsidered. *International Journal of Forecasting*, 13, 501-508.

Gardner, E. S. & McKenzie E. (1985). Forecasting trends in time series. *Management Science*, 31, 1237-1246.

Goodwin, P. (2005). How to integrate management judgment with statistical forecasts. *Foresight: The International Journal of Applied Forecasting*, 1 (1), 8-11

Gorr, W., Olligschlaeger, A. & Thompson, Y. (2003). Short-term forecasting of crime. *International Journal of Forecasting*, 19, 579-594.

Gough, H. G. (1962). Clinical versus statistical prediction in psychology, in L. Postman (ed.), *Psychology in the making*. New York: Knopf, pp 526-584.

Green, K. C. (2005). Game theory, simulated interaction, and unaided judgement for forecasting decisions in conflicts: Further evidence. *International Journal of Forecasting*, 21, 463-472. In full text at conflictforecasting.com.

Green, K. C. (2002). Forecasting decisions in conflict situations: A comparison of game theory, role-playing, and unaided judgement. *International Journal of Forecasting*, *18*, 321-344. In full text at conflictforecasting.com.

Green, K. C. & Armstrong, J. S. (2006). Structured analogies for forecasting. Department of Marketing Working Paper, The Wharton School.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E. & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12, No. 1, 19-30.

Jørgensen, Magne (2004a), Top-down and bottom-up expert estimation of software development effort. *Journal of Information and Software Technology*, 46 (1), 3-16.

Jørgensen, Magne (2004b), A review of studies on expert estimation of software development effort.

Journal of Systems and Software, 70 (1-2), 37-60.

Keogh, E. & Kasetty, S. (2002). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 102-111.

Lewis, M. (2003). *Moneyball*. New York: London.

Liao, K.P. & Fildes, R. (2005). The accuracy of a procedural approach to specifying feed forward neural networks for forecasting. *Computers & Operations Research*, 32, 2151-2169.

Lichtman, A. J. 2005. The keys to the White House: Forecast for 2008. *Foresight: The International Journal of Applied Forecasting*, 3, 5-9.

MacGregor, D. G. (2001). Decomposition for judgmental forecasting and estimation. In J. S. Armstrong (ed.), *Principles of Forecasting*. Boston: Kluwer Academic Publishers.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, R. & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1, 111-153.

Makridakis, S., Chatfield, C., Hibon, M. Lawrence, M. Mills, T. Ord, K. & Simmons L. F. (1993), The M2-Competition: A real-time judgmentally based forecasting study (with commentary). *International Journal of Forecasting*, 5-29.

Makridakis, S. & Hibon, M. (2000). The M3-Competition: Results, conclusions and implications. *International Journal of Forecasting*, 16, 451-476.

McGuickin, R. H. & Ozyildirim (2004). Real-time tests of the leading economic index: Do changes in the index composition matter? *Journal of Business Cycle Measurement and Analysis*, 1 (No. 2), 171-191.

Meade, N. & Islam, T. (2001). Forecasting the diffusion of innovations: Implications for time series extrapolation, in J. S. Armstrong (ed.), *Principles of Forecasting*. Boston: Kluwer Academic Publishers.

Meehl, P. E. (1956). Wanted: A good cookbook. *American Psychologist*, 11, 263-272.

Miller, D. M. & Williams, D. (2004). Shrinkage estimators for damping X12-ARIMA seasonals. *International Journal of Forecasting*, 20, 529-549.

Miller, D. M. & Williams, D. (2003). Shrinkage estimators of time series seasonal factors and their effect on forecasting accuracy. *International Journal of Forecasting*, 19, 669-684.

Miller, T. & Liberatore, M. (1993). Seasonal exponential smoothing with damped trends: An application for production planning. *International Journal of Forecasting*, 9, 509-515.

Rowe, G. & Wright, G. (2001). Expert opinions in forecasting: The role of the Delphi technique, in J. S. Armstrong (ed.), *Principles of Forecasting*. Boston: Kluwer Academic Publishers, pp. 125-144.

Schnaars, S.P. (1986). A comparison of extrapolation models on yearly sales forecasts. *International Journal of Forecasting*, 2, 71-85.

Surowiecki, J. (2004). *The Wisdom of Crowds*. New York: Doubleday.

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.

Wittink, D. & Bergestuen, T. (2001). Forecasting with conjoint analysis. in J. S. Armstrong (ed.), *Principles of Forecasting*. Boston: Kluwer Academic Publishers.

Wolfers, J. & Zitewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, 18 (2), 107-126.

Wong, B. K, Lai, V. S. & Lam, J. (2000). A bibliography of neural network business applications research: 1994-1998. *Computers & Operations Research*, 27, 1045-1076.

Acknowledgements

Monica Adya, P. Geoffrey Allen, Anirvan Banerji, Roy Batchelor, Fred Collopy, Sven Crone, Jason Dana, Robert Fildes, Everette S. Gardner, Jr., Kesten C. Green, Paul Goodwin, Wil Gorr, Rob Hyndman, Magne Jørgensen, Don MacGregor, Nigel Meade, Don Miller, Keith Ord, Gene Rowe, Nada Sanders, Len Tashman, Tom Yokum and Lars-Erik Öller made useful suggestions on early drafts. The reviewers did not always agree with my interpretation of the evidence. April Anderson and Bryan LaFrance provided editorial assistance.

File: IJFProgress25R43 (folder: IJF)