A Sequential Procedure for Aggregation of Expert Judgment Forecasts

John M. Irvine[1], Srinivasamurthy R. Prakash[1], John Regan[1], Drazen Prelec[2]

[1] Draper Laboratory, 555 Technology Square, Cambridge, MA, 02139
[2] Massachusetts Institute of Technology, Cambridge, MA, 02139

Abstract

Policy makers rely on forecasts from experts to inform the decision making process. Merging or aggregating the judgments from multiple forecasters poses methodological challenges. New research is exploring ways of combining judgments from multiple experts to arrive at a better overall decision. For forecasts involving a small set of possible categorical outcomes where expert judgments accumulate over time, we propose a sequential procedure that chooses the best single forecast as soon as the expert judgments indicate sufficient evidence for the specific outcome. We present the formulation of this approach for binary forecasting problems. Using forecasting data for a set of real world events, we demonstrate and evaluate this new method. Comparing the performance of the proposed method to the standard unweighted linear average from the pool of subjects demonstrates the benefits of this approach. Once the outcome of the forecasting problem is known, the Brier score provides an objective measure of performance. Based on the Brier score, this method outperforms the unweighted linear average across a number of forecasting problems.

Key Words: Forecasting, predictions, expert judgments, aggregation methods

1. Introduction

Predictions made through expert judgment are critical to decision making in many fields. Policy makers rely on expert judgment forecasts when formulating strategies for addressing political, economic, and social issues. When multiple experts provide forecast, the merging or aggregation of these judgments presents an interesting challenge. Recent research has shown that combining judgments through averaging leads to poor prediction performance. Individuals are often reluctant to predict s specific outcome with high confidence. If allowed to provide a probability forecast, these individual probabilities are rarely close to zero or one. However, a preponderance of individual forecast that are directionally similar argues for a bolder assertion of the likelihood of a given outcome.

In this paper, we discuss a new aggregation process that considers the directionality of the individual forecasts and assigns a high probability to situations where the majority of forecasters agree on the likely direction of the outcome. We present this new method for combining forecasts. An objective measure of prediction performance compiled from a set of forecasting problems quantifies the benefits of this approach. The forecasting problems span a range of topics including politics, economics, and international affairs. The standard for comparison is the simple average of the individual forecasts, also known as the unweighted linear opinion pool (ULinOP). Overall, the proposed method exhibits significantly better performance than the ULinOP.

Researchers have long understood that aggregate estimations built from the individual opinions of a large group of people often outperform the estimations of individual experts (Surowiecki 2004). The use of the Un-weighted Linear Opinion Pool (ULinOP, or group mean) has proven to be a robust method of aggregating forecasts that often outperforms more complex techniques. Draper Laboratory is participating in the Aggregative Contingent Estimation (ACE) Program sponsored by the Intelligence Advanced Research Projects Activity (IARPA). The goal of the ACE Program is to improve the accuracy of forecasts for a broad range of significant event types through the development of advanced techniques that elicit, weight, and combine the judgments of many subject matter experts. Essentially, our aim is to become more accurate in forecasting events of national interest by aggregating predictions from a large number of analysts and experts.

Our research team is tackling two major research challenges under the IARPA ACE Program: How do we best capture the knowledge and understanding that each forecaster has? And, how do we combine this information to produce the best overall forecasts? To answer the first question requires understanding of human perception and sources of bias. Techniques based on cognitive science give the participants multiple ways to view the forecasting problem and convey their estimates. We are conducting a series of experiments to determine which methods are most effective (Miller, Kirlik, and Hendren 2011;Tsai, Miller, and Kirlik 2011; Poore et al. 2012). To solve the second problem of combining the individual forecasts, we are exploring several avenues of research. For example, it would be useful to know who among the forecasters has the real expertise. When collecting forecasts from participants, additional information is elicited that informs the aggregation process and provides indications of individual expertise (Forlines, et al, 2012).

In this paper, we detail the design of a new aggregation algorithm that meets the goals of

- Being easy to explain to decision makers who have to act upon the aggregate forecast of the group,
- Easy to implement and run on a large collection of forecasts
- Does not require significant effort on the part of the individual forecasters in terms of what information has to be entered.

This new approach builds on sequential methods that have proven to be a useful approach for hypothesis testing. Although our problem is not a formal test of hypotheses, the need to choose between two competing future states of the world has some obvious similarities. As additional forecasters render judgments over time, the accumulation of evidence from these individual predictions forms the basis for deciding between the two outcomes. When the judgments for one outcome greatly outweigh the other, the "aggregate" forecast is to assign a probability close to one for the favoured outcome.

2. Measuring Forecasting Performance

The measure the accuracy of a probability forecast can be quantified by the Brier score, computed as the average squared deviation between predicted probabilities for a set of events and the (eventual) outcomes (Brier 1950):

$$B = \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{r}(f_{ti} - o_{ti})^2$$

where:

f_{ti} is the forecast probability
o_{ti} is the binary indicator of the event outcome
r is the number of possible outcomes
t is the number of forecast instances

The range of the Brier score is [0,2] where 0 indicates a 100% accurate prediction and 2 indicates a completely inaccurate prediction. Applying the Brier scoring rule requires knowledge of the actual resolution for the forecasting problem. Consequently, Brier scoring can only be performed after the forecasting problem has closed and truth is known. To assess performance on a set of forecasting problems, we compute the Brier scores for each individual forecasting problem (IFP) for two competing aggregation methods: the ULinOP and our new sequential procedure method.

3. SP♠DE System Overview

To address the ACE Program goals, we have developed the System for Prediction, Aggregation, Display, and Elicitation (SP♠DE), which elicits individual forecasts and related information from a pool of over 1000 participants and generates daily forecasts about a wide variety of world events. The forecasting data collected under the first year of the program forms the basis for the analysis presented here. We performed retrospective analysis on this collection of forecasts with the aim of developing aggregation approaches for use in the next year of the program.

The elicitation methods used in SP♠DE acquire a rich set of information to characterize and model the forecasters and the individual forecast problems (IFPs). A series of experiments have explored the distribution of knowledge among the forecasters, the relationship between knowledge and forecasting accuracy, and the irreducible uncertainty associated with each IFP (Tsai, Miller, and Kirlik 2011; Poore et al. 2012; Miller, Forlines, and Regan 2012). The sequential procedure relies only on the individual forecasts provided by each participant. An active area of investigation is how to improve performance by incorporating ancillary information into the aggregation process.

Using a web-based interface, the SPADE System elicits forecast and related information from approximately 900 – 1,000 active participants. For each individual forecasting problem (IFP), participants provide judgmental forecasts:

- Will the event occur?
- Probability of the event occurring
- Meta-forecast: What will others predict?
- How would the forecasts improve with access to the knowledge of all participants?

Participants are able to update forecasts, as desired. If news reports indicate a change in conditions related to the forecasting problem, it may be wise to adjust one's predictions based on the emerging story. However, very few participants actually provide updates.

Identifying and recruiting participants with relevant subject matter expertise was a challenge. The participants in this study were recruited through targeted advertisements

on numerous, topically relevant announcement boards and academia websites. The team identified and reached out to subject matter experts associated with topical blogs, think tanks, news outlets, and academic institutions. To maximize the effectiveness of these interactions, we employed a three-tiered approach seeking to

1. Stimulate the prospective participant's interest and address their questions about joining the study,
2. Encourage the individual to pass recruitment literature to their colleagues with relevant backgrounds,
3. Invite the individual to share his or her insight about novel venues or mediums which could be used to connect with potential recruits.

Utilizing this three-tiered approach proved successful in achieving the recruiting needed to support the study. All participants are U.S. citizens. The gender balance was approximately two-thirds male. The mean age is 36.5 years and the standard deviation is 13.2. About 88% of participants are college graduates and more than half have advanced degrees.

Table.1. Gender distribution of participants

	Count	Percent
Female	632	37%
Male	1059	63%
Total	1691	100%

To gain a deeper understanding of each forecaster's expertise, we ask participants a variety of addition questions. One question, which we call the meta-forecast, elicits the participant's best estimate of what others in the study are likely to predict. Another question considers their perceptions about the distribution of knowledge among forecasters. In particular, we ask participants how their prediction would change if they had access to all of the knowledge available among the pool of participants. The participants that indicate their forecasts would be unchanged by this additional information are implying that they already have the knowledge and expertise needed to make a good forecast.

4. Description of the Sequential Procedure

Consider binary forecasting problems with possible outcomes A and B. Let $P(i,t)$ be latest personal prediction from person i at time t. Compute a statistic $S(t) = S[P(1,t), \dots, P(K,t)]$. Then, the decision is:

$$S(t) > C_A \qquad \text{predict outcome A (Forecast} = 1)$$

$$S(t) < C_B \qquad \text{predict outcome B (Forecast} = 0)$$

$$C_B < S(t) < C_A \qquad \text{predict intermediate value (Forecast} = P^*)$$

The test procedure depends on choices for C_A and C_B, and computation of $S(t)$. Define test statistic $S(t)$ by:

$$S(t) = \log\left(\frac{\sup\{P\ f(X_i,\theta): \theta \in w_0\}}{\sup\{P\ f(X_i,\theta): \theta \in w_1\}}\right)$$

Hence:

$$S(t) = N_A \log(P_A) + N_B \log(1-P_A) - N_A \log(P_B) - N_B \log(1-P_B)$$

Where N_A is the number of forecasts favouring outcome A and N_B is similarly defined

$$P_A = \max\{P^*, 0.5\} \qquad P_B = \min\{P^*, 0.5\}$$

where $P^* = \frac{N_A}{[N_A+N_B]}$.

Define C_A and C_B, based on allowable type 1 error denoted α and type 2 error denoted β. Furthermore, we assume $\alpha = \beta$. We compute C_A and C_B by:

$$C_A = \log\left[\frac{1-\alpha}{\alpha}\right] \qquad \text{and} \qquad C_B = \log\left[\frac{\beta}{1-\beta}\right]$$

In general, large values of α and β will lead to quick decisions at the increased risk of an incorrect decision. We opted for a more conservative approach, with $\alpha = \beta = 0.0001$.

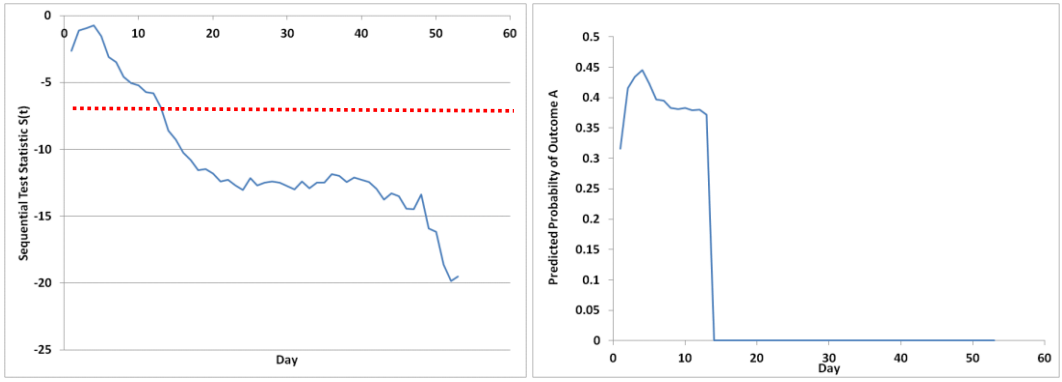

Figure 1: Decision Thresholds. The figure on the left shows the sequential statistic$S(t)$. When $S(t)$crosses the threshold represented by the dotted line, the decision is to set the aggregated forecast to zero, as shown on the right side.

4. Retrospective Performance Analysis

Using the individual forecasts collected for 72 distinct forecasting problems over the first year of the program, we performed retrospective analysis to assess the benefits of the sequential procedure compared to the ULinOP. Typical performance for the sequential procedure and ULinOP appears in Figure 2. Over the life of the forecasting problem, evidence accumulates until the test statistic

exceeds the relevant threshold and the aggregate prediction drives to an extreme value. The specific forecasting question in this case was:

Before 13 April 2012, will the Turkish government officially announce that the Turkish ambassador to France has been recalled?

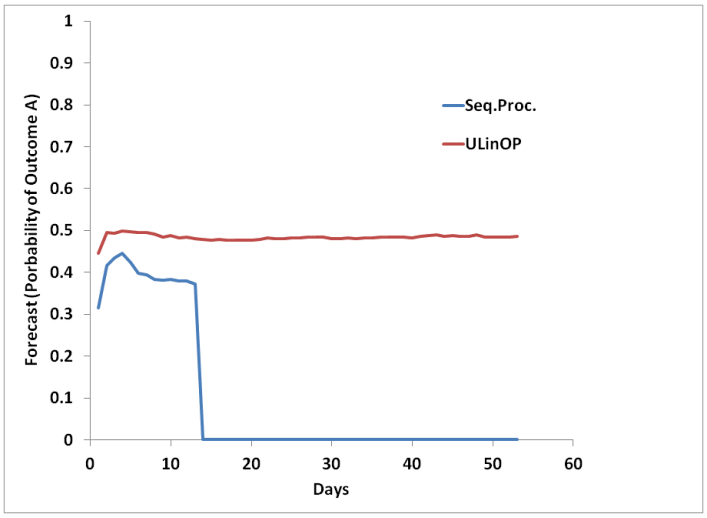

Figure 2: Prediction comparison for selected forecasting problem

In the example given above, the pattern of predictions is evident. By comparison, the forecasting problem depicted in Figure 3 exhibits different behaviour. In short, the evidence starts to drive to an extreme forecast. Circumstances can change, however, and the sequential statistic drops back below the threshold indicating a new level of uncertainty in the outcome. The forecast question for this example was:

Will Italy restructure or default on its debt by 31 December 2011?

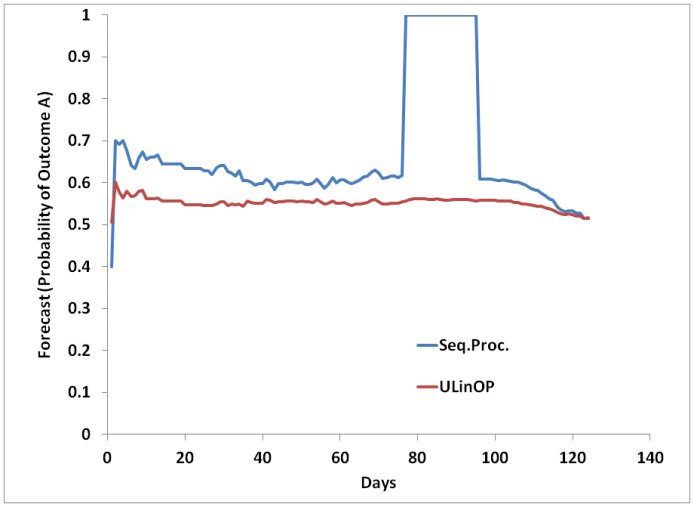

Figure 3: Prediction comparison for selected forecasting problem

Applying the sequential procedure to all 72 forecasting problems, we computed the mean Brier score across all days. The analogous computation was performed form the ULinOP and the percent difference quantifies the relative performance of the two aggregation methods. For the vast majority of forecasting problems considered in this study, the sequential procedure greatly outperforms the ULinOP (Figure 4).

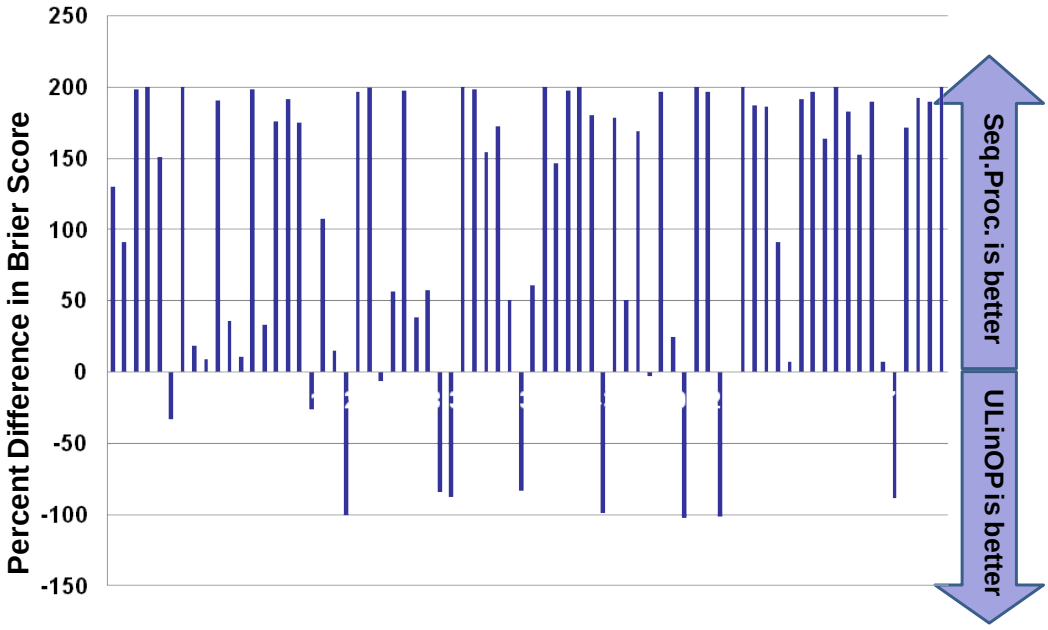

Figure 4: Brier Score comparison for all IFPs

Brier score is an asymmetric loss function which assigns high penalty to the incorrect forecasts, but rewards correct predictions fairly modestly. The penalty is especially severe for extreme forecasts, i.e., values close to 0 or 1. To minimize the adverse effects of this scoring procedure, we consider moderating the aggregated forecasts by capping the values at some level. Thus, if the evidence points to outcome A, rather than predicting A with probability 1, we assign the value $1 - \Delta$ to the aggregate forecast. Similarly, if the evidence points to outcome B, we predict a with some probability $\Delta > 0$.

Proceeding empirically, we consider varying values of Δ and compute the mean Brier score across all forecasting problems (Figure 5). The analysis shows that $\Delta = 0.15$ yields the best overall performance. Table 2 shows the mean Brier scores.

Table 2. Mean Brier Score across 72 IFPs for 2 values of Δ

	Mean Brier Score
Δ = 0	**0.171**
Δ = 0.15	**0.154**

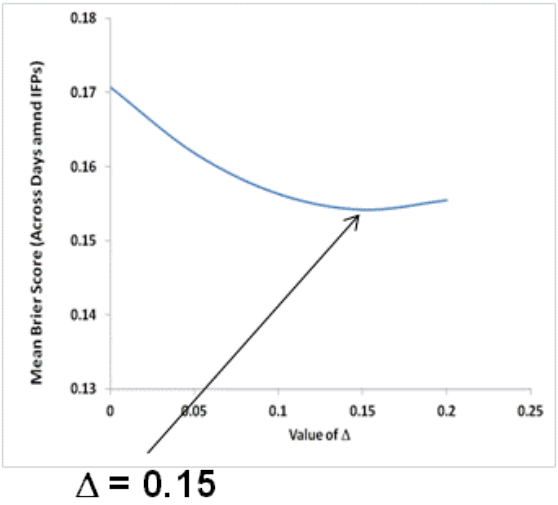

Figure 5. Mean Brier scores for varying values of Δ

6. Conclusions and Future Research

We have presented a new method for aggregation of expert forecasts based on a sequential procedure that assesses the current weight of evidence over time. This procedure will drive to a specific outcome when subjective judgments indicate sufficient evidence. Comparison of this method to the ULinOP, based on retrospective analysis, show substantial performance benefits as measured by the Brier score.

The current approach has some clear limitations and future research will investigate ways to address these limitations. In particular, the current method treats all forecaster equally. Information acquired when we elicit the forecasts includes self assessment of various measures of knowledge and expertise, which can be used to weight individual forecasts. Similarly, information derived from the meta-predictions could be used to modify the relative importance of individual forecasts. And, finally, the timing of the forecasts could be incorporated into the procedure by giving more weight to the most recent forecasts.

Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20058. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1-3.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavailable injustice*. New York: Oxford University Press.

Surowiecki, J. (2004). The wisdom of crowds. New York: Doubleday

Sarah Miller, Alexl Kirlik, and Nathan Hendren (2011) "Applying knowledge and confidence to predict achievement in forecasting" *Human Factors and Ergonomic Society Meetings*, September 19-23, 2011, in Las Vegas, NV

Jennifer Tsai, Sarah Miller, and Alex Kirlik (2011) "Interactive Visualizations to Improve Bayesian Reasoning" *Human Factors and Ergonomic Society Meetings*, September 19-23, 2011, in Las Vegas, NV

Joshua Poore, John Regan, Sarah Miller, Clifton Forlines, John Irvine "Fine Distinctions within Cognitive Style Predict Forecasting Accuracy" *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting*, Boston, MA, October 22-26, 2012. (in press)

Sarah Miller, Clifton Forlines, John Regan, "Exploring the Relationship Between Topic Area Knowledge and Forecasting Performance" *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting*, Boston, MA, October 22-26, 2012. (in press)

Clifton Forlines, Sarah Miller, Srinivasamurthy Prakash, John Irvine, (2012) "Heuristics for Improving Forecast Aggregation" *Machine Aggregation of Human Judgment: AAAI-12 Fall Symposium* (in press).