Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Identifying Expertise to Extract the Wisdom of Crowds

David V. Budescu, Eva Chen

To cite this article:
David V. Budescu, Eva Chen (2014) Identifying Expertise to Extract the Wisdom of Crowds. Management Science

Published online in Articles in Advance 23 May 2014

. http://dx.doi.org/10.1287/mnsc.2014.1909

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–14
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2014.1909

Identifying Expertise to Extract the Wisdom of Crowds

David V. Budescu
Department of Psychology, Fordham University, Bronx, New York 10458, budescu@fordham.edu

Eva Chen
University of Pennsylvania, Philadelphia, Pennsylvania 19104, evachen@sas.upenn.edu

Statistical aggregation is often used to combine multiple opinions within a group. Such aggregates outperform individuals, including experts, in various prediction and estimation tasks. This result is attributed to the "wisdom of crowds." We seek to improve the quality of such aggregates by eliminating poorly performing individuals from the crowd. We propose a new measure of contribution to assess the judges' performance *relative* to the group and use positive contributors to build a weighting model for aggregating forecasts. In Study 1, we analyze 1,233 judges forecasting almost 200 current events to illustrate the superiority of our model over unweighted models and models weighted by measures of *absolute* performance. In Study 2, we replicate our findings by using economic forecasts from the European Central Bank and show how the method can be used to identify smaller crowds of the top positive contributors. We show that the model derives its power from identifying experts who consistently outperform the crowd.

Data, as supplemental material, are available at http://dx.doi.org/10.1287/mnsc.2014.1909.

Keywords: wisdom of the crowd; forecasting; dynamic modeling; contribution
History: Received January 19, 2013; accepted November 25, 2013, by James Smith, decision analysis. Published online in *Articles in Advance*.

1. Introduction

Nostradamus looked to the stars to foretell disasters, Gallup surveys the populace to model future election outcomes, and sports commentators examine athletes' past performances to predict scores of future games (statistically and otherwise). Whether the discussion centers on the art or the science of forecasting, decades of research have focused on the quality of predictive judgments in various domains such as economics, finance, sports, and popular culture (e.g., Armstrong 2001, Clemen and Winkler 1999, Surowiecki 2004). The literature suggests that individual forecasts are riddled with biases, such as being systematically too extreme or overconfident about reported probabilities, overly anchored on an initial estimate, biased toward the most emotionally available information, neglectful of the event's base rate, etc. (Bettman et al. 1998, Gilovich et al. 2002). A natural remedy is to seek experts in the relevant domains, hoping that they would be less likely to succumb to such biases. Unfortunately, expertise is ill-defined and not always easy to identify. Although experts in some domains (e.g., short-term precipitation forecasts) are highly accurate (e.g., Wallsten and Budescu 1983), generally, this is not the case (see, e.g., Tetlock's 2005 work in the political domain).

An alternative approach is to improve predictive judgment by mathematically combining multiple opinions or forecasts from groups of individuals—knowledgeable experts or plain volunteers—(Clemen 1989, Soll and Larrick 2009) or a consensus derived from interactions among the experts in the group (Sunstein 2006). Surowiecki (2004) has labeled this approach the "wisdom of crowds" (WOC). The claim is that mathematical or statistical aggregates (e.g., measures of central tendency) of the judgments of a group of individuals will be more accurate than those of the average individual by exploiting the benefit of error cancellation. Indeed, Larrick et al. (2011) define the WOC effect as the fact that the average of the judges beats the average judge. Davis-Stober et al. (2014) propose a more general definition, namely that some linear combination of the crowd's estimates should beat that of a randomly selected member of the crowd. Their analysis indicates that WOC is likely to be observed in a wide variety of situations with relatively few exceptions.

The principles of WOC have been applied to many cases ranging from prediction markets to informed policy making (Hastie and Kameda 2005). Budescu (2006) suggests that aggregation of multiple sources of information is appealing and effective because it (a) maximizes the amount of information available for the decision, estimation, or prediction task; (b) reduces the potential impact of extreme or aberrant sources that rely on faulty, unreliable, and inaccurate information; and (c) increases the credibility and validity of the aggregation process by making it more inclusive and ecologically representative. Interestingly, the judges need not be "experts" and can be biased, as long as they have relevant information that can be combined for a prediction (Davis-Stober et al. 2014, Wallsten and Diederich 2001).

Critics of WOC have pointed to instances when the crowd's wisdom has failed to deliver accurate predictions because the aggregate estimate was largely distorted by systematic group bias or by a large number of uninformed judges (Simmons et al. 2011). As an alternative to simply averaging individual judgments, researchers have proposed weighted models that favor better, wiser, more experienced judges in the crowd (e.g., Aspinall 2010, Wang et al. 2011). Such models are a compromise between the two extreme views that favor quality (expertise) and that rely on quantity (crowds). To benefit fully from both the quality of the experts and the quantity of the crowd, the challenge is to identify the wiser members of the crowd and appropriately weight their judgments.

We address this problem and propose a novel measure of "contribution to the crowd" that assesses individual predictive abilities based on the difference of accuracy of the crowd's aggregate estimate with, and without, the judge's estimates in previous forecasts, in the domain of interest, and we illustrate the effectiveness of the new approach. The first case study consists of binary predictions made over nine months regarding the likelihood of current events in five different domains: business, economy, military, policy, and politics. We validate this individual contribution measure of performance and test whether a weighted model based on individual contributions to the crowd is reliably better than simply averaging estimates. The second case study analyzes probability distributions of professional forecasters for two economic indicators over 53 quarters (13 years). We replicate our results and demonstrate that our method can identify the best subset of judges and that one can focus on their average while ignoring all the others. We explore the diversity of the crowd by identifying the best (and worst) individuals within the crowd such that they can be over- (and under-) weighted.

2. Wisdom of the Crowd (WOC)

The premise behind WOC is that individual knowledge (signals) can be extracted, while minimizing (eventually, eliminating) biases or misinformation (noise) by aggregating judgments (Makridakis and Winkler 1983). WOC generates its best results when the judges are knowledgeable, properly incentivized to express their beliefs, and obtain their responses independently of each other, and when there is diversity of knowledge and information in the crowd (see also Larrick et al. 2011).

Successful implementations of WOC have gone out of their way to foster diversity of opinions by (a) selecting judges with different backgrounds, (b) eliciting their inputs independently, and (c) forcefully injecting diverse thoughts to affect their original estimates (Herzog and Hertwig 2009). Diversity is derived not only from the group's composition but also from the method by which information is shared in the group (Lichtendahl et al. 2013). If individuals are not given a chance to think independently before responding, their judgments could be biased by responses from the group (Larrick et al. 2011). Lorenz et al. (2011) demonstrated that even mild social influence can undermine the effect of WOC in simple estimation tasks. In fact, the higher the correlation between individual estimates, the more judges are necessary to achieve the same level of accuracy (e.g., Broomell and Budescu 2009, Clemen and Winkler 1986, Hogarth 1978).

Of course, at least some of the judges must possess relevant information, but in some cases the level of information can be minimal. For example, Herzog and Hertwig (2011) report a study predicting outcomes of three soccer and two tennis tournaments that rely on the recognition heuristic. Predictions based solely on the judges' ability to recognize some of the players' names (through their exposure to different media) gave the group a diverse collective knowledge that was sufficient to consistently perform above chance and as accurately as predictions based on official rankings of the teams and players.

In most WOC forecasting applications, the favorite aggregation method is the simple average of the judgments (Larrick et al. 2011),[1] but this approach may be suboptimal because it neglects external information (e.g., expertise) and, as such, reduces the potential to benefit from the wisdom found in the crowd. For example, low-performing stock market analysts tend to make bolder predictions that drive the average prediction to a more extreme position (Evgeniou et al. 2013). Lee et al. (2011) examined the bids of players on the popular game show *The Price Is Right*. The aggregation models, especially those that took into account strategy and bidding history, outperformed all the individual estimates, and those that used external information outperformed the simple mean. Thus, including the judges' level of expertise has the potential to improve the quality of the crowd's forecasts.

We focus on aggregation of judges who provide probabilities of future uncertain events. Our goal is to combine the, possibly conflicting, probabilistic judgments made by different individuals into one "best" judgment. French (2011) refers to this as "the expert problem." Typically, the judgments are

[1] Jose et al. (2014) make a case for robust measures based on trimmed or winsorized means, and Hora et al. (2013) consider medians.

probabilities or odds, but one could also combine qualitative forecasts (see Wallsten et al. 1993). The most popular weighting schemes are opinion pools of the individuals' judgments: from predictions of volcanic eruptions to risk assessments in the nuclear power industry (Aspinall 2010, Cooke and Goossens 2008). Although there are more general formulas (French 1985, Genest and Zidek 1986), the most common aggregation rules are the weighted linear (arithmetic) and the weighted geometric means. In general, the linear opinion pool, L, is higher than the geometric one, G, which tends to reduce the influence of extremely high values. Variations on these themes include averaging the log-odds, $\log(p_i/(1-p_i))$ and transforming the mean log odds back to the probability scale (e.g., Turner et al. 2014).

The weights in the opinion pools often represent the individuals' relative expertise, but the concept of "relative expertise" is ill-defined and subject to many interpretations (French 2011). One possibility is to assign weights based on "objective" (e.g., historical track record, education level, seniority, and professional status), "subjective" (e.g., ratings of expertise provided by the judges themselves or others, such as peers or supervisors), or a combination of the two. Another approach is to define the weights empirically based on the experts' performance on a set of uncertain "test" events, the resolution of which are unknown to the experts, but known to the "aggregator" (person or software) that assigns the weights in the opinion pool (Bedford and Cooke 2001, Cooke 1991). Clemen (2008) and Lin and Cheng (2009) have compared the performance of Cooke's empirical weights method with equally weighted linear pools ("plain WOC") and the best expert's judgment. The weighted method generally outperformed both the equal weights method and the best expert. However, different scoring rules (and different tests) can lead to different weights. Soll and Larrick (2009) point out that empirically weighted linear pools tend to overweight a few individuals, which can lead to extreme predictions. This may be suboptimal when the test events and the actual events of interest diverge and the correlation between performances of the two is reduced.

In this paper, we develop and illustrate the use of a new empirically weighted linear opinion pool. Unlike Cooke's approach, we do not use an independent standalone set of pretest events to identify expertise. Instead, the weights emerge in the process of forecasting based on the judges' performance relative to others (i.e., contribution to the crowd). We also develop a dynamic version of the model that adapts to changes in the decision environment and the judges' performance as new events are being resolved and included in the model.

3. The Contribution Weighted Model (CWM)

We define an individual's contribution to the crowd to be the change in the crowd's aggregated performance, as measured by some merit function, with and without the target individual. This quantity measures the individual's expertise relative to the crowd, since it captures the effect of inclusion or exclusion of each person in the crowd. Once such individual contributions are calculated, a contribution weighted model (CWM) is devised to be applied in future predictions by the same crowd of judges. To quantify the effects of WOC, we need an appropriate measure of merit or quality of the aggregate (and the individuals). In the context of probability judgment, this measure is typically a proper scoring rule (e.g., Bickel 2007). In this paper we use a quadratic scoring rule (de Finetti 1962), but the proposed approach and procedure can be applied to all other (proper or improper) scoring schemes. The following algorithm determines the contributions of each judge in the crowd:

(1) Let N be the number of events forecasted, and let R_i be the number of categories used in forecasting event i (where $i=1,\ldots,N$). Let m_{ir} be the aggregated (typically, the mean) probability of the crowd for each outcome, r (where $r=1,\ldots,R_i$) of each event ($i=1,\ldots,N$), and let o_{ir} be the binary indicator of the actual outcome for each instance ($0=$ occur or $1=$ not occur). The crowd's score for a given event, S_i, is given by the following:

$$S_i = a + b\sum_{r=1}^{R_i}(o_{ir}-m_{ir})^2.$$

(2) The performance of the crowd is aggregated across all events, based on the quadratic score:

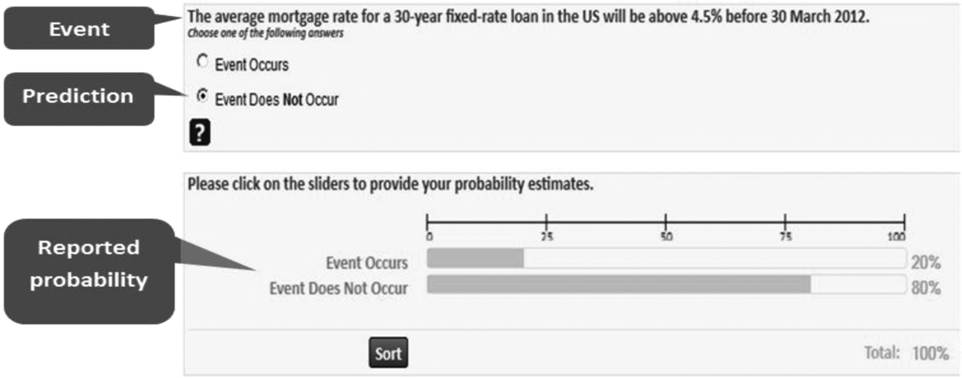
$$S = a + b\sum_{i=1}^{N}\left(\sum_{r=1}^{R_i}(o_{ir}-m_{ir})^2\right).$$

The quadratic score is unique up to a linear transformation. We use constants $a=100$ and $b=-50$ to yield scores ranging from 0 to 100, where 0 indicates the worst possible performance and 100 indicates perfect performance. For binary events ($R_i=2$), the expected "uninformed" score is 75, and it is achieved when a probability of 0.5 is assigned to all events.

(3) The contribution of each judge, C_j (where $j=1,\ldots,J$), is calculated as the average difference between the crowd's scores based on the mean forecasts (m_r), with and without the jth forecaster, across all N_j events answered by the forecaster. We allow for the possibility that not all judges forecast all the events by setting $N_j \le N$. Formally, this is expressed as

$$C_j = \sum_{i=1}^{N_j}(S_i - S_i^{-j})/N_j.$$

Figure 1 **Screenshot of Probability Elicitation**



The average mortgage rate for a 30-year fixed-rate loan in the US will be above 4.5% before 30 March 2012.
Choose one of the following answers

○ Event Occurs
● Event Does **Not** Occur

Please click on the sliders to provide your probability estimates.

Event Occurs 20%
Event Does Not Occur 80%

Total: 100%

Source. http://forecastingace.com (site discontinued).

This contribution, C_j, can be positive (indicating that the judge's forecasts improve on average the crowd's S) or negative (suggesting that the judge's forecasts reduce the average S of the crowd). It is inspired by the statistical literature on measures of influence (e.g., Kutner et al. 2005) that seeks to establish if, and by how much, various parameters and predictions of complex statistical models are affected by specific observations, by eliminating them (one at a time) from the sample. We hypothesize that weights based on C_j would outperform those based on the judge's absolute past performance (track record). The key intuition behind this prediction is that the forecasts of the various judges will typically be highly correlated (see Broomell and Budescu 2009). Thus, there will be many cases where almost everyone in the crowd will have very good scores and, conversely, cases where practically all the members of the crowd will perform poorly. Measures of absolute performance are not likely to be very discriminating in such cases. However, C_j recognizes good performance in a relative sense and places more value on judges who have greater knowledge than the crowd. Judges get a higher C_j if they do well in cases where the majority of the crowd performs poorly, i.e., when they do not follow the wrongful judgment of the crowd. Our weighted aggregated model, CWM, employs only judges with positive C_j in forecasting new events. These C_j are normalized to generate weights such that the aggregated prediction of the crowd is the weighted mean of the positive contributors' probabilities.

4. Study 1: Forecasting Current Events

To validate the weighting procedure and verify that it can identify quality judges in the crowd, we analyzed data from the Forecasting ACE project website.[2] Launched in July 2010, the website elicits probability forecasts from volunteer judges who choose to forecast, at any time, any subset of events from various domains: business, economy, entertainment, health, law, military, policy, politics, science and technology, social events, and sports. We focused on binary events: each event describes a precise outcome that may or may not occur by a specific deadline. On average, 15–20 events are posted at various times every month with various timelines (some as short as three days and some as long as six months) depending on the nature of the event. There are no restrictions on the number of events for which a judge can provide probabilities.

Figure 1 shows a screenshot of a typical event. The judge first makes a prediction on whether or not the event will occur and then enters the subjective probability of the event's occurrence by moving the slider. The webpage enforces binary additivity by forcing the probabilities of the R_i (in this case 2) possible outcomes to sum to 1. The predictions and probabilities can be revised any time before the closing date, but most judges (90%) do not revise their initial judgments. The current data analysis was conducted only on the last reported probability for every judge for any given event.[3]

The judges are scored based on their participation (number of forecasts performed) and accuracy of prediction. A leaderboard serves as the only explicit incentive provided by the website. In addition to providing forecasts, judges are encouraged to complete a background questionnaire[4] covering their self-assessed knowledge of the domains, the hours they spend reading the news, education level, and experience in forecasting.

4.1. Data Collection

Between the launch date of the site and January 2012, 1,233 judges provided forecasts for 104 events. The judges answered an average number of 10.4 events (SD = 12.64). We analyze only those judges who

[2] http://forecastingace.com, last accessed May 2013 (site discontinued); see https://www.facebook.com/ForecastingACES.

[3] Additional analysis showed virtually no correlation between timing (from the time a forecast is submitted to the resolution date) and C_i.

[4] The correlations between C_i and these measures are low.

Table 1 Alternative Aggregation Models Compared to CWM

Model	Description	Justification
ULinOp	Equally weighted S for all 1,233 judges (the crowd).	Test CWM against unweighted S of entire data set.
UWM	Equally weighted S for the subset of 420 forecasters who answered 10 or more events.	Test CWM against unweighted S of the same subset.
Contribution	Equally weighted S of all *positive contributors* from the subset of 420 forecasters who answered 10 or more events.	Compare the advantage of weighting contributors.
BWM	Weights are calculated with S for all of the 420 judges. The weights depend only on the judge's past performance (S).	Compare CWM with weighted model based on absolute past performance.
xBWM	Same as BWM, but using a percentage of positive contributors similar to CWM.	Compare CWM with weighted model based on absolute past performance with the same number of positive contributors.

answered 10 or more events ($n = 420$). This threshold is used to reduce the possibility that C_j capitalizes on chance, which can easily happen in probabilistic forecasting.[5] In fact, the proper measurement of the accuracy of an individual forecaster or a crowd aggregate should be performed over a substantial number of events and possibly, over an extended period of time.

These 420 judges responded to a mean number of 23 events, reported a mean general knowledge of current issues at approximately five on a seven-point scale ($1 =$ no knowledge, and $7 =$ extremely knowledgeable), and reported spending on average 23 minutes a day reading the news. Their level of education ranged from high school (4%) to Ph.D. (10%), and most of them (64%) have at least a bachelor's degree. Most judges were novices, with only 37% of the judges having experience in forecasting with an average of five years.

4.2. Comparison of Aggregation Models

The performance of the CWM was compared with the five competing models listed in Table 1. The first two models (ULinOp and UWM) are unweighted means and serve as a baseline to all other weighted models. UWM is a "trimmed" version of the ULinOp that includes only those judges who have answered 10 or more events. The Contribution model is an unweighted version of the CWM model, which uses only the positive contributors and assesses the effect of choosing judges using the new metric. BWM and xBWM are weighted models built with the judges' past S and, unlike CWM, the weights are independent of the performance of the other members of the crowd. BWM uses all 420 judges, whereas xBWM, similar to CWM, uses the top 220 judges to compute the weighted model.

Table 2 Performance of the Models Compared (in Terms of Their Scores)

			$\bar{S}$ of models across all events				
Model	Judges included	Mean positive contributors	Min	Median	Mean	Max	SD
CWM	420	220	39.93	**91.90**	**88.26**	99.56	12.06
Contribution	420	220	39.52	89.55	86.46	99.50	11.82
UWM	420	—	41.58	87.45	83.73	98.25	11.51
ULinOp	1,233	—	42.81	87.64	83.62	98.67	11.76
xBWM	420	220	9.46	89.16	80.07	99.49	20.92
BWM	420	—	25.31	82.84	77.35	97.93	17.65

To maximize the information used to compute C_j and yet avoid overfitting, we cross-validated the models by eliminating one event at a time (jackknifing). The CWM used all events except for the one being eliminated to compute the weights, and the aggregated forecast of the jackknifed event was determined as a weighted average of forecasts from positive contributors.[6] Thus, all predictions being considered are "out of sample."

A summary of the performance of the competing models is provided in Table 2, in which the models are listed according to their mean scores, $\bar{S}$. CWM produces the highest $\bar{S}$, which is highlighted in bold, with Contribution running a close second. Only CWM and Contribution beat the unweighted models, UWM and ULinOp, (which are almost equally good). Our metric of improvement is defined as 100[(difference in the two $\bar{S}$ being compared)/(100 − $\bar{S}$ of the baseline model)].

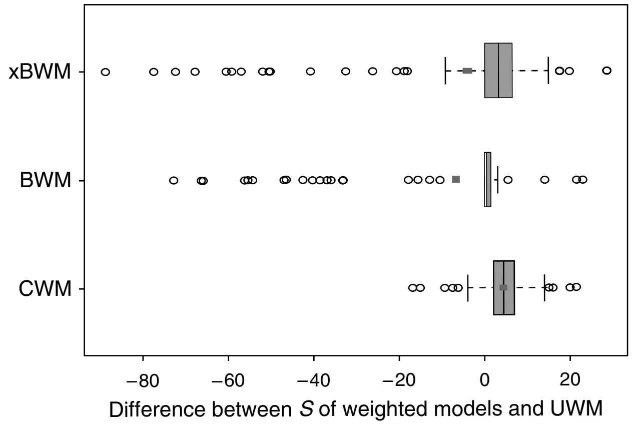
CWM beats the unweighted mean (UWM) values by approximately 28% on this metric. The models that weighted judges by past performance performed worse than the UWM with a decline of 39% for BWM and 22% for xBWM.

The CWM's superior performance stems from its ability to identify the judges with specific knowledge. The Contribution model (giving equal weights to positive contributors) produced 17% improvement over the UWM. Thus, 60% of the CWM impact can be attributed to identifying expertise and 40% to overweighting those who perform better than the crowd on average. To illustrate this point, we simulated forecasts for each event from a beta distribution with the same mean

[5] Imagine, for example, a sports league wherein the home teams win 60% of the games and a forecaster who predicts the results of 10 independent games. There is a probability of 0.37 (calculated by the binomial distribution with $p = 0.6$ and $n = 10$) that the home teams will lose a majority (six or more) of the games.

[6] BWM and xBWM were cross-validated following the same jackknifing procedure.

Figure 2 Comparison of the Performance of CWM, BWM, and xBWM Relative to the Simple Mean



and variance as the crowd's estimates for that event. This model performed well above the "uninformed" baseline (75) but below the CWM. The key difference is that, although each event is predicted equally well, the contributions to the crowd are randomly distributed across judges.

Figure 2 presents boxplots of the difference between S of three weighted models (CWM, BWM, and xBWM) and the baseline model (UWM) for all the events. The figure shows that there are fewer outliers with CWM because C_i is calculated relative to the crowd and is less sensitive to individual performances, which have higher variances (due to cases where most members of the crowd are wrong).

As a sensitivity analysis, we repeated all the calculations using a logarithmic scoring rule (instead of the quadratic one). The key results were replicated: the CWM model had a $\bar{S}$ of 87.17, a 7.21% improvement over the UWM model ($\bar{S} = 86.18$).

4.3. Domain Analyses of CWM

We applied the CWM model to the five major domains separately, using only judges answering 10 or more events in each domain, and cross-validated by jackknifing each event. Table 3 shows that the crowd (not necessarily the same judges in each domain) performed better using the CWM weighting than the UWM. The CWM excelled in the domain of policy, with an improvement of approximately 46% (in bold), and fared worst in economic events (with an improvement of approximately 18% improvement). The overall improvement, weighted by the number of events in the five domains, is 35%. We caution that this is not directly comparable to the general model (with approximately 28% improvement) because some events were included in multiple domains. Figure 3 plots the relationships between the contribution scores in the various domains. All the correlations are positive, especially those involving military, policy, and politics events, suggesting that the presence of an underlying general expertise factors in geopolitics among our best judges.

Table 3 Domain-Specific Comparisons of CWM and UWM

Domain	No. of events	Mean positive contributors	Mean S of UWM	Mean S of CWM	Mean difference in S	% improvement
All	104	220	83.73	88.26	4.53	27.86
Policy	32	52	84.33	91.51	7.18	**45.84**
Business	23	23	83.52	90.00	6.48	39.32
Politics	45	77	86.36	91.67	5.31	38.93
Military	19	33	84.78	87.71	2.93	19.27
Economy	16	10	77.85	81.73	3.88	17.51

4.4. A Dynamic Version of the CWM

As a test of the CWM's true predictive powers, we introduced 90 new events (posted between January 2012 and April 2012) and recomputed the weights in a dynamic fashion. We used the original set of 104 events to compute the initial weights for predicting the probabilities of the first new event. Weights were then recomputed with every new event that was resolved for predicting the next one, and so on. The CWM dynamic model based on new events showed an overall improvement of 39%, despite the fact that the response rate of the new events was lower (39 compared to 127 judges per event in the initial set).

The dynamic model did better at the aggregate level, and the $\bar{S}$ of the CWM was better than the $\bar{S}$ of the UWM in 71 of 90 events (79%). This split is significantly better than chance (50%) by a Wilcoxon signed rank test with $p < 0.001$. We also implemented the dynamic model for each domain using the same recursive procedure. Table 4 summarizes the results. The CWM improved the $\bar{S}$ in all domains, except for economy. The superiority of the CWM over the UWM model was significant in three of the five domains.

5. Study 2: Forecasting Inflation and GDP in Europe

In this section we apply the CWM in a purely dynamic setting while addressing some limitations associated with special features of the ACES data set. Responses in the ACES example are provided in a somewhat arbitrary fashion—forecasters choose which events to forecast, when to enter forecasts, and in what order to forecast various events. The project involves multiple domains and volunteer forecasters (so one could argue that most are not experts in most domains). The first goal of this second study is to validate the approach in a single, well-defined domain with forecasters who are recognized experts, and where predictions are made on the same events at the same time (in a framework with a more rigid and systematic temporal structure).

This also provides an opportunity to study the impact of selectively reducing the number of positive contributors to improve the CWM. The creation of a smaller

Figure 3 **Distribution of Contribution Measures in Five Domains and Intercorrelations**

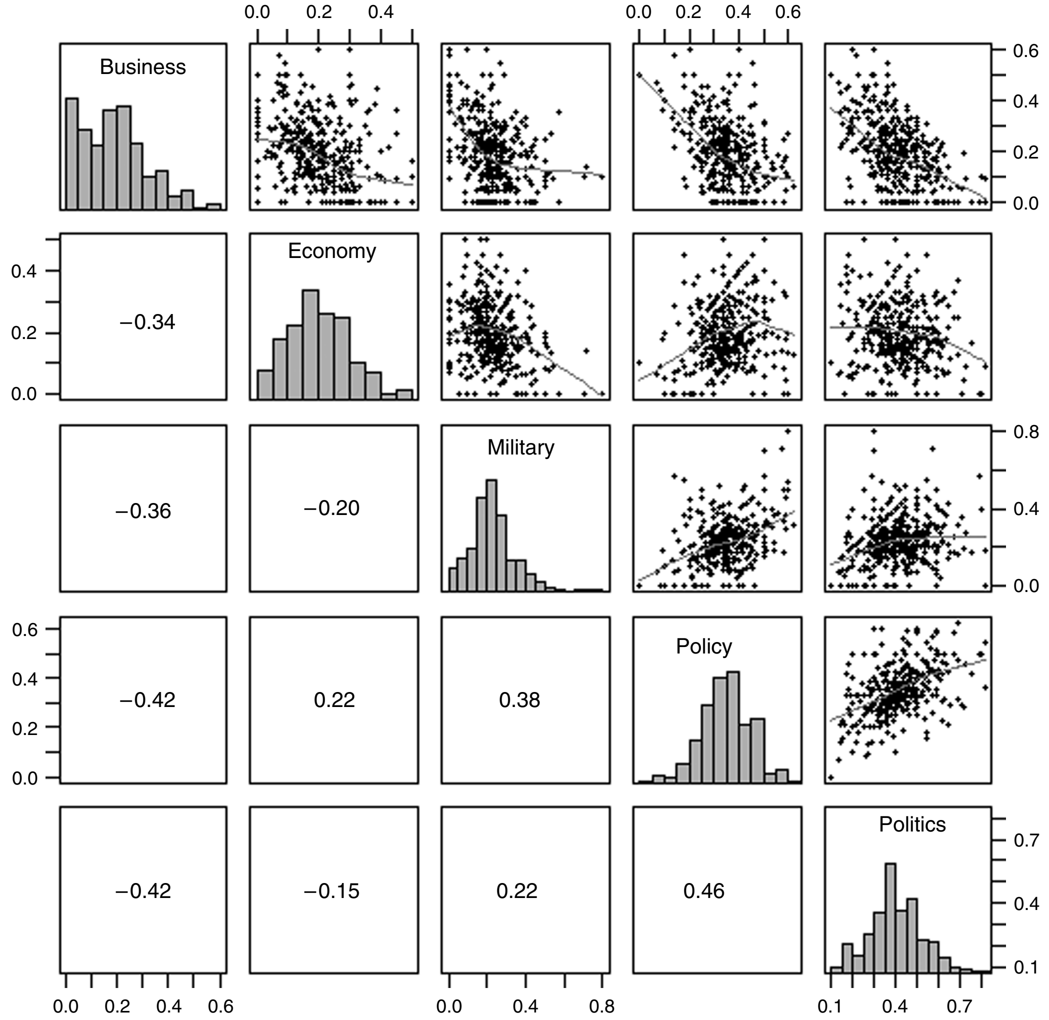


Table 4 **Summary of the Performance of the Dynamic CWM for Each Domain**

Domain	No. of events	Mean No. of judges	Mean positive contributors	$\bar{S}$ of UWM	$\bar{S}$ of CWM	Difference in $\bar{S}$	% of improvement
All	90	39.30	17.46	83.56	87.90	4.35	39.40*
Military	15	47.00	13.53	85.41	92.46	7.05	54.12*
Politics	49	36.80	14.84	84.26	91.13	6.87	53.36*
Policy	25	40.96	13.28	81.74	84.35	2.61	30.48*
Business	19	44.68	10.74	82.75	82.97	0.22	14.76
Economy	16	37.31	9.75	83.73	78.59	−5.15	−10.13

*Significant ($\alpha = 0.05$) by a sign test.

crowd of positive contributors is based on the idea proposed by Mannes et al. (2014), who stipulate that averaging the opinions of a small crowd of *properly selected* (three to six) experts can perform as well as averaging the entire crowd. We examine the effectiveness of using the judges' contribution to identify this "small crowd."

5.1. Data Collection

We analyzed quarterly forecasts of the real GDP growth and inflation rate in Europe using the European Central Bank (ECB)'s Survey of Professional Forecasters. The forecasts are in the public domain (http://www.ecb.int/stats/prices/indic/forecast/html/index.en.html). The entire data set consists of over 72,000 forecasts for three EU economic indicators (inflation, real GDP growth, and unemployment rate) across six time horizons by professionals from the financial industry and academic institutions. We focused on forecasts of (a) inflation (i.e., the year-on-year percentage change of the Harmonised Index of Consumer Prices, as defined by the

ECB) and (b) GDP growth for nine months from the date of the survey. The data include 2,386 forecasts for GDP and 2,491 for inflation from the first quarter of 1999 to the last quarter of 2011. Each forecaster provided a probability distribution over seven to 22 categories (in the -6% to 4.9% range)[7] depending on the survey period. We collapsed all the probability distributions to the same seven categories for all periods. The quadratic scoring rule was applied to the multinomial distributions. The score ranges from 0 (worst) to 100 (perfect forecasts). The S of an "uninformed" judge who assigns equal probabilities to the seven categories is: $100 - 50[(1 - 1/7)^2 + 6(0 - 1/7)^2] = 57.15$, which serves as a baseline for analysis.

A total of 113 forecasters made at least one forecast during the 52 periods, which we treated as 52 events in the analysis. We analyzed only the 90 forecasters who made at least two predictions. The mean number of forecasts per quarter is 47 ($\text{SD} = 5$). We implemented the dynamic CWM model at a yearly level: C_i was first computed based on the probability distribution for the four quarters. The CWM then used the first year's C_i (the mean C_i of the four quarters) to compute the aggregate predictions for the second year. In the third year, the CWM applied the mean C_i from the first and second years to determine the aggregates of that year. The model proceeded thusly for 12 years of predictions.

5.2. Analysis of the CWM and the Contributions

Table 5 summarizes the performance of the CWM model and its close variant, Contribution, relative to the UWM for the 12 years. The results for inflation replicate the results of the first study: (1) the CWM beats the UWM on average and in most quarters, (2) the Contribution model (that ignores the differential weights) performs almost as well as the CWM. However, (3) the experts were very poor at predicting GDP growth: neither the UWM nor the CWM beat the uniform distribution, implying that there was little knowledge for the models to extract. We compared the judges' performance in the two domains and found only low levels of agreement: 27% of judges are positive contributors in both domains, 32% are negative in both, and most (41%) are positive in one domain but not the other.

An in-depth analysis of contributors in the prediction of inflation rates provides some insight into the mechanics of the model. We calculated for each of the forecasters the proportion of quarters in which they were assigned a positive C_j (and were part of the CWM). Figure 4 summarizes this information in the form of a cumulative distribution. There is a distinct cluster of forecasters who have positive C_j values in most cases (31% of forecasters in over 80% of the periods)—the "global" experts—and, at the other end of the distribution, a slightly larger cluster of forecasters who are almost never included (36% of forecasters in less than 20% of the periods).

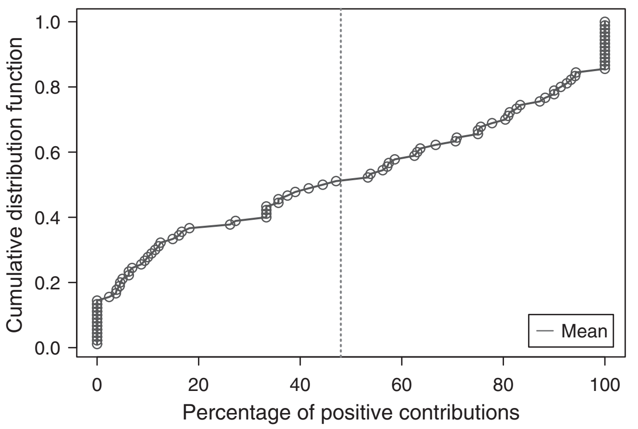
Table 5 **Performance of Dynamic Models Predicting Inflation and GDP in Europe**

	UWM	Contribution			CWM		
Indexes	$\bar{S}$	$\bar{S}$	% of improve.	No. of quarters > UWM	$\bar{S}$	% of improve.	No. of quarters > UWM
Inflation	57.86	60.31	5.81*	3.08	60.97	7.36*	2.83
GDP	53.39	54.43	2.23	2.33	54.48	2.34	2.42
Both	55.54	57.19	3.71*	2.84	58.74	7.20*	2.96

*Significant ($\alpha = 0.05$) by a sign test.

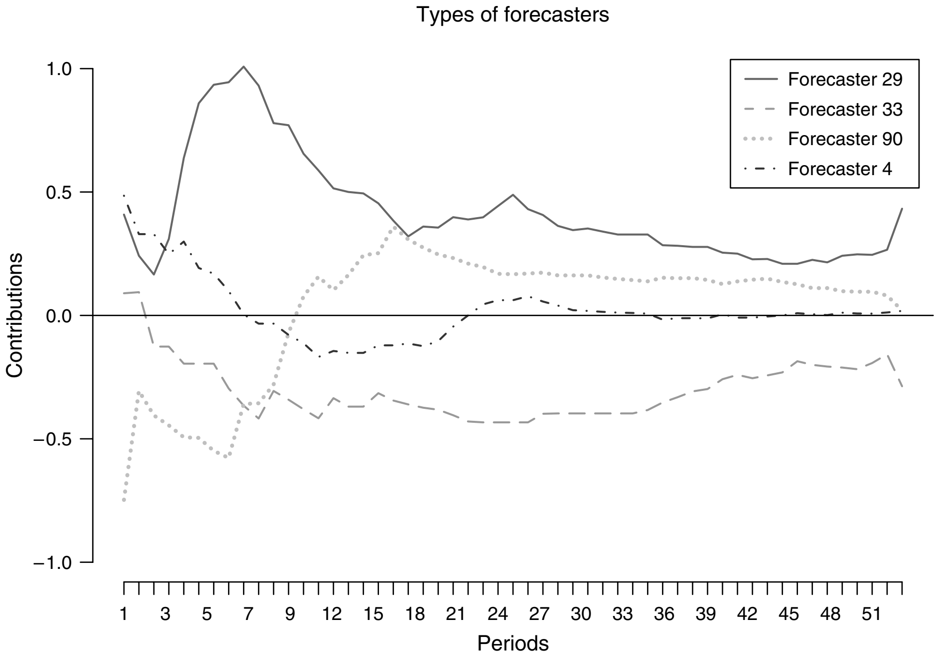
We classified the 76 judges who made at least 10 predictions into four groups based on their S being above/below the "uninformed" level (57.15) and their mean C_js being positive/negative. Figure 5 presents the pattern of C_j values of distinct types of judges (who made predictions in a majority of the periods). There are 20 (26%) judges who are "global" experts: their $\bar{S}$ is 60.8, their mean C_j is 0.17, and they get positive C_j in 92% of the cases (e.g., forecaster 29 in Figure 5). Moreover, their mean C_j value and S values are positively and significantly correlated ($r = 0.47$). Conversely, there are 39 judges (51%) that perform poorly by both measures—their $\bar{S}$ is 48.0, their mean C_j is -0.13, they make positive contributions in only 16% of the cases (e.g., forecaster 33 in Figure 5), and their mean C_i values and S values are positively and significantly correlated ($r = 0.56$). There are only two judges (3%) with $\bar{S}$ values above the baseline and mean negative C_j values. However, there is an interesting subgroup of 15 judges (20%) with $\bar{S}$ (51.0) below the baseline and mean positive C_j values (0.04) who are

Figure 4 **Distribution of Percentage of Positive Contributions in the Model Predicting Inflation**



[7] The number of categories and their definition are determined by the ECB for every quarter.

Figure 5 **Illustration of Several Types of Forecasters**

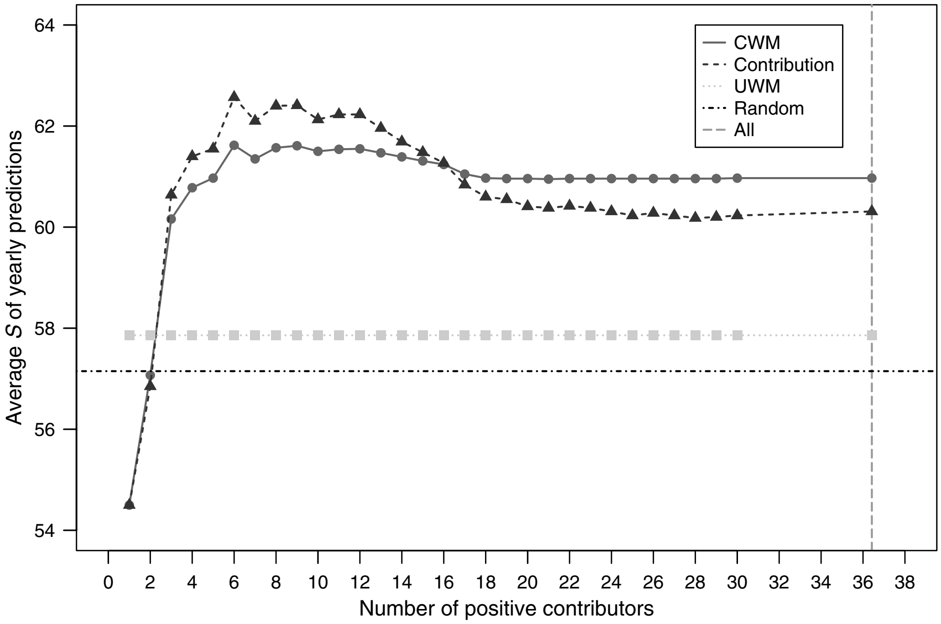


assigned positive weights in a majority (63%) of the cases. In this group, the $\bar{S}$ values and mean C_j values are uncorrelated ($r = 0.04$). Clearly, their C_j values vary systematically as a function of the period, presumably reflecting reliance on distinct cues that were particularly diagnostic and valid at various times. Thus they display “local” expertise. Some do very poorly first and improve later (e.g., forecaster 90 in Figure 5); others start very well but their performance deteriorates over time (e.g., forecaster 4 in Figure 5). It is notable from Figure 5 that the weights have a clear temporal pattern rather than a haphazard and random structure.

5.3. Identification of a Small Crowd

Can we use the C_j index to identify an optimally small crowd of forecasters by using only the top K ($K = 1, \ldots, J$) contributors, and how is the group’s performance affected by this reduction in size? Figure 6 plots the $\bar{S}$ of the CWM and Contribution models for inflation with a diminishing number of contributors.

Figure 6 **Performance of the Models as a Function of the Number of Top Contributors Selected**



To put things in proper perspective and facilitate interpretation, we also include the baseline of respondents (who assign equal probability to the seven categories) and the UWM. At the right end of the plot, "All" includes all forecasters with positive C_j. The average number of positive contributors over 12 years is 36.42 (SD = 3.23). The UWM ($\bar{S}$ of 57.86) is only slightly better (1.66%) than the baseline. The CWM ($\bar{S}$ of 60.97) and Contribution ($\bar{S}$ of 60.31) beat the UWM (by 7.38% and 5.81%, respectively).

As the inclusion criterion becomes more stringent and group size is reduced, the performance of the CWM and Contribution models reveals a remarkable pattern. As the number of selected top contributors drops below 16, the Contribution model (simple unweighted mean of positive contributors) performs better than the weighted model. Performance peaks with the six best contributors, where the Contribution ($\bar{S} = 62.57$, SD = 18.32) is 11.18% better than the UWM and 2.48% better than the CWM ($\bar{S} = 61.62$, SD = 18.19). Finally, when the number of positive contributors drops below 3, both models fare worse than the UWM or even the baseline.

These results support the claim of Mannes et al. (2014) that averaging a small crowd of properly selected experts can do just as well as, or even better than, averaging all judges (UWM) or using just the top-ranking judge. The measure we used to identify these top forecasters was C_j, as opposed to absolute past performance used by Mannes et al. (2014), which is based on the Mean Absolute Error. Study 1 indicates that C_j is more stable over time than absolute measures. The improvement of S between using a small crowd (three to six experts) and all positive contributors (approximately 36 experts) is less than 2.43% for the CWM.

6. General Discussion

There are two distinct approaches in the quest for the most accurate probabilistic forecasts. One approach seeks to identify individual expertise, and the other seeks to aggregate multiple opinions from a crowd without differentiating among its individual members. The key insight of WOC is that the aggregation process reduces the effects of individual biases, and that the central tendency of the crowd's opinions can be used to forecast the target events (Armstrong 2001, Clemen 1989, Wallsten et al. 1997). Our approach combines the two philosophies by: (a) identifying the experts in the crowd and (b) aggregating their opinions, while ignoring the estimates of the nonexperts. This can also be seen as a compromise between the two approaches. The major contribution of the current paper is the development and validation of our new measure for identifying experts *in a crowd* by measuring their contribution to the crowd's performance.

We often assume that simply relying on past performance on similar tasks is sufficient to identify expertise. Indeed, if at some point in the process one were asked to choose a single expert, we cannot think of a way of selecting one that would beat this intuitive metric of *absolute quality of* performance. However, if one continues to rely on a crowd (or at least a subset of its members), our results show that one can do considerably better by relying on the proposed measure of *relative quality*, C_j. We illustrated this approach in two longitudinal studies. By simply isolating the experts—those who make positive contributions to the crowd—the mean accuracy score improved in both studies compared to the average of the crowd. When using the weighted model (CWM) the performance improved even further. This is not to say that every event, period, or game predicted in the studies was improved by using only the forecasts of these experts, but over time the variance decreased and the model proved significantly better than the simple (unweighted) mean(s) and weighted means using all the forecasters.

Various sensitivity analyses confirmed the robustness of the CWM model. We found that (1) the model was successful in setups with sparse responses, (2) its performance improved when applied separately to various domains of expertise, (3) the measures of individual contributions outperformed simulated judges with identical mean performance at the item level, (4) the results were replicated with different probability distributions (binomial in Study 1 and multinomial in Study 2), and (5) the results were replicated with a different scoring rule (logarithmic). All these findings indicate that C_j reflects real expertise. Indeed, the comparison of the CWM with simulated experts that predict equally well at the item level but whose expertise varies randomly from one event to another confirms this conclusion.

Our selection of experts is not based on the best performers (highest S) because their performances can be skewed by one or a few extreme predictions (Denrell and Fang 2010) and be nondiagnostic in many cases. We pick those who consistently outperform the group, and our model is updated dynamically to reduce variance due to chance results and to reflect "true" expertise that emerges in the process.

The success of our approach is quite intuitive, once one realizes that judges are usually highly correlated (see Broomell and Budescu 2009) because they share many assumptions and/or have access to the same information. Consequently, crowds often behave like herds, as almost everyone expects certain events to happen (or not). In some cases, when judges choose to forecast events that most people in the crowd predict quite confidently and correctly, no one will get high C_j values because the crowd is quite accurate. Conversely, in other cases, when judges forecast events that most people in the crowd predict incorrectly, no one will get low C_j values because the crowd is inaccurate. Such events do not much affect the CWM model, which assigns high (positive or negative) contributions to cases where judges deviate from the majority of the crowd. In this respect the CWM differs from the weighting schemes that are based on absolute performance.

Consider, for example, the recent case when prediction markets "failed" to predict the U.S. Supreme Court's decision regarding the Affordable Care Act. (Prediction markets estimated a 75% chance that it would not be upheld by the court.[8]) CWM identifies and overweights the predictions of those judges who do not necessarily follow the crowd in such cases and perform well in moving away from the crowd. We identify these consistently positive contributors and use weighting to reap the benefits of (large or small) crowds without predetermining the size of such a crowd. As we show in Study 2, positive contributors can be of two types. One group consists of undisputed experts who beat the crowd consistently because of their superior knowledge and/or ability to identify the relevant cues from the environment and combine them correctly. They are positively weighted in every period. The second group involves judges who perform very well in some, but not all, environments. The internal mental models that lead to their predictions are imperfect, so they may not always identify and/or properly weight relevant cues. Such judges do very well in some circumstances but perform worse in other cases, so they may be in or out of the subset of positively weighted forecasters over various periods of time.

An interesting theoretical issue is what makes the CWM work—its ability to identify the experts or their differential weighting. Our results clearly suggest that it is primarily the model's ability to identify the experts to be positively weighted (or in other words, its ability to identify those members of the crowd who should be excluded) that is responsible for most of the model's improvement. This is not surprising, as the relative insensitivity of the model to departures from optimal weighting is well recognized in the literature (e.g., Broomell and Budescu 2009, Davis-Stober et al. 2010, Dawes 1979). In fact, once the smallest subset of positive contributors is identified, there is a penalty associated with differential weighting (see Figure 6), and a simple unweighted mean of the carefully selected subset of judges provides the most accurate predictions.

The dynamic implementation of the CWM is probably the most attractive feature from a practical point of view. Our results demonstrate that the CWM can easily adapt to new events (producing 39% improvement over the UWM in Study 1) by including new experts or discarding old ones as their C_i values drop. For Study 1, the dynamic model was especially useful in correlated domains like military, policy, and politics (where predictions were enhanced by 54%, 53%, and 31%, respectively) for judges possessing knowledge that was adaptable to all three. The strength of the dynamic model is that no forecaster is ever totally and irreversibly eliminated from the crowd. The model overweights positive contributors, but as the environment or expertise changes, it learns and adapts to a decline or incline in relative performance. Judges who have negative C_i values (and are ignored for a while) can find their way back if their performance improves as a result of a change in the environment or a deterioration in the performance of others (see Figure 5). The beauty of the dynamic model is that it can accommodate such changes in the environment and adjust the inclusion status and the weights attached to each judge. We illustrate this point in a small simulation described in the appendix.

The approach we proposed is very general and flexible and can be easily refined by adjusting its various features. Two directions that we think are especially worth exploring in future research are (1) replacing the exclusive reliance on the mean when calculating the judges' contributions by alternative measures such as the median or weighted means that over- (or under-) weight recent (or old) events and (2) replacing the exclusive reliance on each judge's complete history of forecasts and focusing primarily on the most recent forecasts (e.g., by using a rolling history of the last X periods).

Another avenue for refinement is the starting weights assigned to each person. In our applications we focused on forecasters who established a track record by having made a minimal number of forecasts (e.g., more than 10% of predictions in Study 1) for computing contributions. The rationale for this choice was to make sure that the contributions are based on knowledge and not mere chance. In some cases this requirement may be too strong and impractical. One simple and sensible alternative rule is to let the contribution at period t be a dynamically weighted average of the contribution based on the previous $(t-1)$ forecasts and some constant K (which could be an equal weight, $1/N$, or some prior measure).

$$C_{it} = W_t C_{i(t-1)} + (1 - W_t)K, \quad \text{with } 0 \le W_t \le 1.$$

The weight assigned to the prior contributions, W_t, can increase from 0 on the first period to 1, as a function of the amount of information that one considers allocating to past performance, where a weight of 1 relies exclusively on past performance.

7. Conclusion

We proposed a new measure of individual contribution that is simple, reliable, easily interpreted, and useful for assessing a forecaster's performance relative to a crowd. We showed that, in addition to identifying

[8] See Leonhardt (2012).

the experts (who always do well and better than the crowd) and eliminating the persistent and consistently poor performers, the method derives strength from being able to identify "local expertise" and properly rewarding those instances where judges rely effectively on cues and information that are diagnostic in particular circumstances (but not others). This illustration also highlights the flexibility of the dynamic model that we advocate. We tested our model in two contexts, and in all cases it outperformed models built solely on past individual performance and on the simple average of the crowd. It works well when there is longitudinal, categorical data even in cases of sparse data, and it identifies the experts relatively quickly.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mnsc.2014.1909.

Acknowledgments

The authors thank all members of the Forecasting ACE and Good Judgment teams for guiding and shaping the authors' thinking on this problem, and the anonymous reviewers for constructive feedback on earlier versions of this paper. This work was supported, in part, by the Intelligence Advanced Research Projects Activity via the Department of Interior National Business Center [Contract D11PC20059].

Appendix. Simulation of Diversity

To illustrate the effect of information diversity, we simulated a sequence of 60 binary events. The "true" probability of

Figure A.1 **Number of Positive Contributors of Type 1, Type 2, and Type 12**

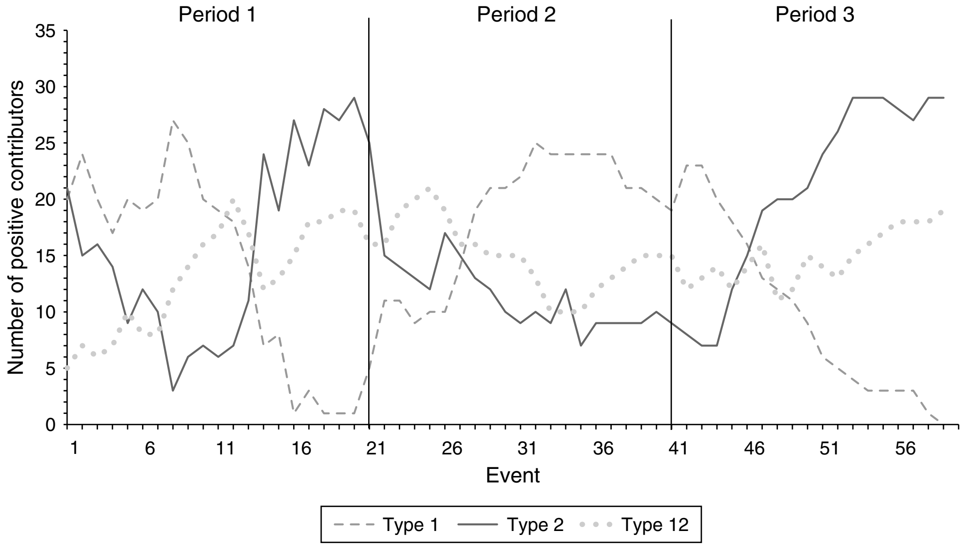
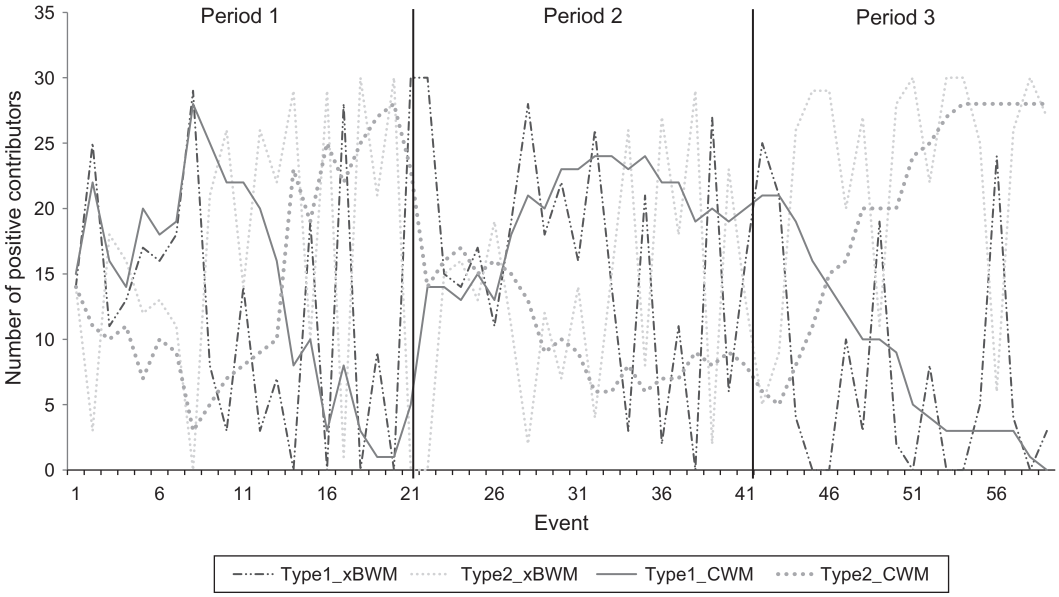


Figure A.2 **Number of Positive Contributors of Type 1 and Type 2 for xBWM and CWM**



each event is a logistic function of two cues, $X1$ and $X2$:

$$\Pr(Event_i) = \frac{\exp(\beta_1 X_{1i} + \beta_2 X_{2i})}{1 + \exp(\beta_1 X_{1i} + \beta_2 X_{2i})}$$

The two cues are sampled from a bivariate normal distribution with equal means ($\mu x1 = \mu x2 = 0.4$), equal variances ($\sigma x1 = \sigma x2 = 1$) and a relatively low intercorrelation, $\rho x1$, $x2 = 0.2$.[9]

The key manipulation is that there are three distinct "periods" (or regimes) defined by different parameters:

- In period 1 (events 1–20) both cues play an equal role ($\beta_1 = \beta_2 = 1$).
- In period 2 (events 21–40) we set $\beta_1 = 1$ and $\beta_2 = 0$, so only $X1$ matters.
- In period 3 this is reversed and only $X2$ drives the probability ($\beta_1 = 0$; $\beta_2 = 0$).

We define a population of 90 forecasters of three "types" defined by the cues to which attend: type 1 forecasters ($n1 = 30$) only consider $X1$, type 2 forecasters ($n2 = 30$) have access to only $X2$, and type 12 forecasters ($n12 = 30$) consider both cues (of course, the individual judgments are perturbed by random errors). If our account holds, we expect that judges of type 12 will do relatively well throughout, but we expect judges of types 1 and 2 to surge and be overweighted by the model in periods 2 and 3, respectively.

The prediction was confirmed when we ran the dynamic model. The top panel of Figure A.1 presents the proportion of judges of each type that had positive weights at each stage. Period 1 is characterized by high fluctuations because of the small number of events but, on average, the three types fare equally well (i.e., on average 43 judges (48%) are assigned positive contributions and the three types are almost equally represented by 33%, 37%, and 30%, respectively). However in period 2, judges of type 1 dominate (18/45 positive contributors = 40%), and in period 3, forecasters of type 2 are a clear majority (20/44 = 45%).

This pattern was even more pronounced when we reran the model with only the 60 judges of types 1 and 2 for both the CWM and xBWM (using the top 50% of judges as measured by their S), as shown in Figure A.2. The CWM model does a much better job than the xBWM model of detecting differential performance of subgroups of forecasters in changing regimes and of putting this information to good use. The CWM is less influenced by the variance in performance over the three periods.

References

Armstrong JS (2001) *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Kluwer Academic, Boston).

Aspinall W (2010) A route to more tractable expert advice. *Nature* 463(7279):294–295.

Bedford T, Cooke R (2001) *Probabilistic Risk Analysis: Foundations and Methods* (Cambridge University Press, Cambridge, UK).

Bettman JR, Luce MF, Payne JW (1998) Constructive consumer choice processes. *J. Consumer Res.* 25(2):187–217.

Bickel E (2007) Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Anal.* 4(2):49–65.

Broomell S, Budescu DV (2009) Why are experts correlated? Decomposing correlations between judges. *Psychometrika* 74(3):531–553.

Budescu DV (2006) Confidence in aggregation of opinions from multiple sources. Fiedler K, Juslin P, eds. *Information Sampling and Adaptive Cognition* (Cambridge University Press, Cambridge, UK), 327–352.

Clemen RT (1989) Combining forecasts: A review and annotated bibliography. *Internat. J. Forecasting* 5(4):559–609.

Clemen RT (2008) Improving and measuring the effectiveness of decision analysis: Linking decision analysis and behavioral decision research. Kugler T, Smith JC, Connolly T, Son Y-J, eds. *Decision Modeling and Behavior in Complex and Uncertain Environments* (Springer, New York), 3–31.

Clemen RT, Winkler RL (1986) Combining economic forecasts. *J. Bus. Econom. Statist.* 4(1):39–46.

Clemen RT, Winkler RL (1999) Combining probability distributions from experts in risk analysis. *Risk Anal.* 19(2):187–203.

Cooke RM (1991) *Experts in Uncertainty* (Oxford University Press, Oxford, UK).

Cooke RM, Goossens LHJ (2008) TU Delft expert judgment data base. *Reliability Engrg. System Safety* 93(5):657–674.

Davis-Stober CP, Dana J, Budescu DV (2010) A constrained linear estimator for multiple regression. *Psychometrika* 75(3):521–541.

Davis-Stober CP, Budescu DV, Dana J, Broomell SB (2014) When is a crowd wise? *Decision* 1(2):79–101.

Dawes RM (1979) The robust beauty of improper linear models in decision making. *Amer. Psychologist* 34(7):571–582.

de Finetti B (1962) Does it make sense to speak of "Good probability appraisers"? Good IJ, et al., eds. *The Scientist Speculates, An Anthology of Partly-Baked Ideas* (Basic Books, New York), 357–364.

Denrell J, Fang C (2010) Predicting the next big thing: Success as a signal of poor judgment. *Management Sci.* 56(10):1653–1667.

Evgeniou T, Fang L, Hogarth RH, Karelaia N (2013) Competitive dynamics in forecasting: The interaction of skill and uncertainty. *J. Behavioral Decision Making* 26(4):375–384.

French S (1985) Group consensus probability distributions: A critical survey. Bernardo JM, DeGroot MH, Lindley DV, Smith AFM, eds. *Bayesian Statistics*, Vol. 2 (North-Holland, Amsterdam), 183–201.

French S (2011) Expert judgment, meta-analysis and participatory risk analysis. *Decision Anal.* 9(2):119–127.

Genest C, Zidek JV (1986) Combining probability distributions: A critique and annotated bibliography. *Statist. Sci.* 1(1):114–148.

Gilovich T, Griffin D, Kahneman D (2002) *Heuristics and Biases: The Psychology of Intuitive Judgment* (Cambridge University Press, Cambridge, UK).

Hastie R, Kameda T (2005) The robust beauty of majority rules in group decisions. *Psych. Rev.* 112(2):494–508.

Herzog SM, Hertwig R (2009) The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psych. Sci.* 20(2):231–237.

Herzog SM, Hertwig R (2011) The wisdom of ignorant crowds: Predicting sport outcomes by mere recognition. *Judgment and Decision Making* 6(1):58–72.

Hogarth RM (1978) A note on aggregating opinions. *Organ. Behav. Human Performance* 21(1):40–46.

Hora SC, Fransen BR, Hawkins N, Susel I (2013) Median aggregation of distribution functions. *Decision Anal.* 10(4):279–291.

Jose VRR, Grushka-Cockayne Y, Lichtendahl KC Jr (2014) Trimmed opinion pools and the crowd's calibration problem. *Management Sci.* 60(2):463–475.

Kutner MH, Nachtsheim CJ, Neter J, Li W (2005) *Applied Linear Statistical Models*, 5th ed. (McGraw Hill, Irwin, CA).

Larrick RP, Mannes AE, Soll JB (2011) The social psychology of the wisdom of crowds. Krueger JI, ed. *Frontiers of Social Psychology: Social Judgment and Decision Making* (Psychology Press, New York), 227–242.

[9] We thank one of the reviewers for suggesting this approach.

Lee MD, Zhang S, Shi J (2011) The wisdom of the crowd playing the price is right. *Memory and Cognition* 39(5):914–923.

Leonhardt D (2012) When the crowd isn't wise. *New York Times* (July 7), http://www.nytimes.com/2012/07/08/sunday-review/when-the-crowd-isnt-wise.html.

Lichtendahl KC Jr, Grushka-Cockayne Y, Pfeifer PE (2013) The wisdom of the competitive crowds. *Oper. Res.* 61(6):1383–1398.

Lin S-W, Cheng C-H (2009) The reliability of aggregated probability judgments obtained through Cooke's classical model. *J. Model. Management* 4(2):149–161.

Lorenz J, Rauhutb H, Schweitzera F, Helbing D (2011) How social influence can undermine the wisdom of crowd effect. *Proc. Natl. Acad. Sci.* 108(2):9020–9025.

Makridakis S, Winkler RL (1983) Averages of forecasts: Some empirical results. *Management Sci.* 29(9):987–996.

Mannes AE, Soll JB, Larrick RP (2014) The wisdom of select crowds. *J. Personality Soc. Psych.* Forthcoming.

Simmons J, Nelson LD, Galak J, Frederick S (2011) Intuitive biases in choice vs. estimation: Implications for the wisdom of crowds. *J. Consumer Res.* 38(1):1–15.

Soll JB, Larrick RP (2009) Strategies for revising judgment: How (and how well) people use others' opinions. *J. Experiment. Psych.: Learn., Memory and Cognition* 35(3):780–805.

Sunstein CR (2006) *Infotopia: How Many Minds Produce Knowledge* (Oxford University Press, New York).

Surowiecki J (2004) *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations* (Little, Brown, London).

Tetlock PE (2005) *Expert Political Opinion, How Good Is It? How Can We Know?* (Princeton University Press, Princeton, NJ).

Turner BM, Steyvers M, Merkle EC, Budescu DV, Wallsten TS (2014) Forecast aggregation via recalibration. *Machine Learn.* 95(3):261–289.

Wallsten TS, Budescu DV (1983) Encoding subjective probabilities: A psychological and psychometric review. *Management Sci.* 29(2):151–173.

Wallsten TS, Diederich A (2001) Understanding pooled subjective probability estimates. *Math. Soc. Sci.* 41(1):1–18.

Wallsten TS, Budescu DV, Zwick R (1993) Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Sci.* 39(2):176–190.

Wallsten TS, Budescu DV, Erev I, Diederich A (1997) Evaluating and combining subjective probability estimates. *J. Behavioral Decision Making* 10(3):243–268.

Wang G, Kulkarni SR, Poor HV, Osherson DN (2011) Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Anal.* 8(2):28–144.