Assessing the Accuracy of Geopolitical Forecasts from the US Intelligence Community's Prediction Market

Seth Goldstein[1]; Rob Hartman[2]; Ethan Comstock[3]; Thalia Shamash Baumgarten[4]

Abstract: Since 2010, the United States Intelligence Community (IC) has run a crowdsourced forecasting platform called the IC Prediction Market (ICPM) on its classified network. This effort has been funded in part through the Intelligence Advanced Research Projects Activity's (IARPA's) Aggregative Contingent Estimation (ACE) program, which has also spawned the Good Judgment Project (GJP). More than 4,300 ICPM users or "traders" have produced more than 190,000 trades on a large array of geopolitical questions, producing the largest dataset on the accuracy of analytic judgments in the history of the IC. Drawing on a large set (N=139) of geopolitical forecast questions posted to both the ICPM and GJP systems from one year of the ACE forecasting tournament, this paper characterizes the ICPM's absolute and comparative accuracy. Across this corpus of forecasting questions, the ICPM was directionally accurate on an average of about 82% of question-days, and this performance was comparable to that of a GJP prediction market hosted on the open internet. The ICPM also is discernibly more accurate than a GJP opinion pool but less accurate than a GJP method that has had the benefit of substantial research dollars spent to enhance its accuracy.

1. Introduction

Whereas the avoidance of strategic surprise is one of the IC's core functions[5], failures to anticipate such surprise are the subject of considerable attention and collective societal hand-wringing. Postulated causes of anticipatory "intelligence failure" are numerous. Unfortunately, attempts to avoid strategic surprise frequently devolve into simply trying to avoid overcorrections made in response to the most recent, high-profile predictive mistake—whether overprediction or underprediction (Tetlock and Mellers 2011). One plausible cause of strategic surprise is "stovepiping" (Hersh 2003), in which intelligence is produced and passed to top level decision makers without proper vetting. For instance, the U.S. government's official report on pre-Iraq War intelligence activities cited stovepiping as a significant contributor to U.S. decision to invade (United States 2004). Another problem, to which the 9/11 attacks have been partially attributed, is the failure to share intelligence between agencies—a problem that might be labelled "integration failure" (9/11 commission: 91-92).

[1] (Corresponding author) Dr. Goldstein is a Program Manager at IARPA, IARPA/MS2 Building, Office of the Director of National Intelligence (ODNI), Washington, DC, 20511. Although the analyses presented herein are unclassified, the ICPM data that produced them remains classified. However, the authors will make replication data available through appropriate channels to appropriately cleared individuals.

[2] Dr. Hartman is Principal Applied Psychologist, The MITRE Corporation, Department of Social and Behavioral Sciences.

[3] Mr. Comstock is currently Senior Financial Data & Reporting Analyst at the University of Virginia School of Medicine.

[4] Ms. Baumgarten is currently an unaffiliated scholar.

[5] Additional IC tasks include the provision of analysis concerning the current or historical status, capabilities, interrelations, and evolving dynamics of a diverse array of global security actors and challenges.

This latter strategic surprise prompted the Intelligence Reform and Terrorism Prevention Act of 2004, resulting in the creation of the Office of the Director of National Intelligence (ODNI), an organization chartered to focus on intelligence integration and sharing of data across IC agencies.

In spite of the community-wide desire to avoid strategic surprise, analytic culture in the IC has not historically emphasized predictive accuracy (Johnston 2005: 107-116). More consistent with this renewed attention to surprise avoidance, voices from both outside (cf. Friedman and Zeckhauser 2012; Friedman and Zeckhauser 2014) and inside the IC (Clapper 2014) have recently called for an increased IC emphasis on forecasting capabilities and the tracking of predictive accuracy. Within ODNI in particular, the founding of the IC Prediction Market (ICPM) and the Intelligence Advanced Research Projects Activity's (IARPA's) funding of the Aggregative Contingent Estimation (ACE) research and development (R&D) program, reflect two recent initiatives aimed at enhancing IC "anticipatory intelligence" capabilities. Both initiatives were inspired by the "wisdom of crowds" or "crowd wisdom" (CW) phenomenon, where statistical aggregation of diverse human judgments is used to derive an optimal forecast estimate for a given anticipatory intelligence question (Surowiecki 2005). From simple or weighted averaging of individual forecaster probability judgments to the derivation of forecasts via virtual currency "prediction markets," CW approaches to forecasting have proven highly accurate in domains ranging from U.S. election forecasting to professional sporting event forecasting. Moreover, from the standpoint of avoiding strategic surprise, CW methods would seem to be a powerful antidote to the concerns over stovepiping and integration failure, inasmuch as such methods encourage participation and interaction among a wide and diverse group of forecasters (in this case, intelligence analysts and other cleared employees)—without respect to rank, home agency, experience, or job function.

As a first step toward the goal of evaluating IC anticipatory intelligence practices and capabilities, this paper presents initial analyses of the accuracy of the ICPM, both in absolute terms and as it relates to ACE program-related forecasting methods. Sponsored by ODNI's IARPA, the ICPM exists on a classified network and is populated by a pool of voluntary participants, who self-select which forecasting questions they respond to and receive no material (e.g., financial or administrative) benefit from their participation. This pool of forecasters consists entirely of top secret-cleared government employees and contractors from a variety of agencies across the intelligence and defense communities.

2. About the ICPM

Questions for the ICPM come from classified and unclassified intelligence products and from the suggestions of users. These questions can be of binary,[6] multinomial/multiple choice,[7] ordered multinomial,[8] and conditional[9] forms. Primary factors driving the development and

[6] For instance, Will X happen before Date Y?

[7] For instance, Who will win election Z? Candidate A; Candidate B; or someone else

[8] For instance, When will event Q happen? Between dates D and E; Between dates E and F; not before date F.

[9] For instance, Will event T happen before date E, if country L takes action M beforehand; If country L does not take action M beforehand.

ultimate selection of ICPM questions[10] include geopolitical relevance[11] and passage of the "clairvoyance test" (Tetlock 2005: 243), which requires that a question be written in a sufficiently clear and operationally concrete fashion that its eventual outcome can be determined with zero-to-minimal controversy (e.g., "Will GDP increase by 3%+?" vs. "Will there be significant economic improvement?"). Formulating questions in such a way that ground truth can be resolved non-controversially allows forecasters to clearly understand what they are forecasting on (and what they are not).

Notwithstanding its unique participant population and subject matter interests, the ICPM is in all other respects a garden variety prediction market. Participants are granted 5,000 symbolic (non-monetary) points upon registration and earn additional points through "buying" shares in events that happen and "selling" shares in events that don't happen. The market is run by a Logarithmic Market Scoring Rule (LMSR; cf. Hanson 2003) wherein participants can cash in and out of positions at any time according to the market's current price. The current price, interpretable as a collective probability judgment, dynamically updates as a function of the "action" on either side of a question, similar to a point spread on a sporting event. As one pushes the aggregate probability away from the starting price[12] (50 for a binary question)[13] towards the probabilistic extremes (e.g. 0 or 100), it takes more and more points to push the probability a comparable distance. As with most prediction markets, users can view the community probability before making a trade and can also share the rationales for their trades via comment threads.

As ICPM "traders" answer more questions correctly, they gain points, and the degree to which they can push the community probability on any given question is increased. In this sense, while individual users make trades (as opposed to making direct probability forecasts), their trades should be at least a partial reflection of the probability they ascribe to an event's occurrence at any given point in time. The LMSR market maker interacts with all users making trades and distills disparate user actions into a single (aggregate) collective probability, reflected in the current market "price" for a given possible outcome. All results presented in this paper concern the behavior and accuracy of this aggregate CW probability.

3. Evaluating and Benchmarking the ICPM

[10] Each question also contained definitional terms that clearly and empirically unpacked the terms contained within each question. For instance, a question about a substantial lethal confrontation would have separate, empirically grounded definitions for "substantial" and "lethal confrontation".

[11] The group of questions was also screened for a priori plausibility—while crowdsourcing methods have been hypothesized to show promise for accurately forecasting events, many argue that such methods are not suitable for forecasting rare events due to the timeframes needed to resolve very rare-event questions (Taleb and Tetlock 2013) and other human biases that inhibit effective forecasting of long-shot events (Sobel and Raines 2003).

[12] The starting price varies based on question type. The opening prices for binary questions are set at 50%, while those for multi-option questions are set at 1/N (*100), where N corresponds to the number of answer options in a question.

[13] The strategy of setting opening probabilities at 50% for a binary question may be inefficient for prediction markets. Some have proposed that batch auctions might eliminate this inefficiency in prediction markets (Hou 2015) or in global equity markets more broadly (Budish et al. 2013).

The overarching goals of this paper are to (1) assess the absolute forecast accuracy of a CW platform whose participants are IC analysts with access to classified information *and* to (2) compare the accuracy of that classified system to a set of platforms that have access to only unclassified, open-source information, using a common set of *unclassified* geopolitical forecasting questions concurrently launched on both platforms. Despite the value of such an evaluative exercise, it is important to highlight a fundamental limitation: because classified questions are by definition inaccessible to those operating outside a classified environment, this analysis is limited in its ability to inform conclusions about the *overall* forecasting value or capabilities of the IC at large or the ICPM particularly. One might logically expect IC analysts to have informational advantages over the general public in forecasting on classified forecast questions (excluded from this analysis), it is less obvious that IC analysts have a unique relative informational advantage in forecasting against unclassified questions (the focus of this analysis). Nevertheless, the current effort provides one important indicator of the ICPM's forecasting capabilities.

Consistent with the paper's aims, the below analyses assess the performance of the ICPM for *unclassified* forecast questions in absolute terms and compare the performance to external benchmarks—forecasting platforms developed and/or managed by the Good Judgment Project (GJP) (cf. Mellers et al. 2014) as part of the IARPA ACE Program. Analysts who forecast on the GJP platforms are amateur forecasters recruited from the general public for the purposes of participation in the ACE program and were randomly assigned into the various conditions for purposes of GJP's experimental tests of interventions and conditions that might maximize forecast accuracy, one of which was a prediction market.[14] Both the GJP Prediction Market analyzed herein and the ICPM employ the Cultivate (formerly Inkling Markets) prediction market interface (Siegel 2009). In addition to these market-to-market comparisons, ICPM Accuracy is compared to GJP's unweighted linear opinion pool (ULinOP) platform, wherein probabilistic CW forecasts are derived by a simple averaging of individual forecasters direct probability judgments, as opposed to the prediction market method of deriving probabilities via trading activity. Finally, we compare ICPM accuracy to that of GJP's single most accurate CW method for the set of questions being analyzed[15]—a method called "All Surveys Logit."[16] All Surveys Logit takes the most recent forecasts from a selection of individuals in GJP's survey elicitation condition, weights them based on a forecaster's historical accuracy,

[14] GJP participants, unlike ICPM participants, were paid a small honorarium for their active participation on unclassified forecasting platforms.

[15] Throughout the ACE forecasting tournament, GJP submitted forecasts for at least 10 unique (experimental) methods, per question, per day.

[16] The TGJ best method was determined retrospectively. The Good Judgment Project submitted 20 forecasts (each derived from different methods) for each open question, every day, during the scored period, to MITRE; MITRE and the government then scored the accuracy of these methods. While one may object to ex-post selection of best method, it serves as a "tough" test for the ICPM's accuracy, because it provides a slight advantage to the Good Judgment Project in assessing accuracy. On the logic of case selection more generally and of case "toughness" specifically, see Eckstein 1975. Another objection would hold that allowing 20 different forecast methods all but guarantees at least one will be successful. We reply to this objection by noting that several other GJP methods were of similar accuracy (<2% difference in accuracy) to the "best" one; so prospective identification of any one of several excellent methods would yield similar conclusions.

expertise, and psychometric profile, and then extremizes the aggregate forecast (towards 1 or 0) using an optimized extremization coefficient (Good Judgment Project 2014: 154).[17]

4. Metrics

Our analysis considers two primary metrics: (1) the Brier score, a quadratic scoring rule for numerical probability judgments; and (2) a measure of "directional" accuracy that is concerned with whether a forecasting system assigns a plurality of its probability to the correct outcome.

The Brier Score

Originally developed for use in evaluating weather forecasts, the Brier score (Brier 1950) is a quadratic scoring rule (an honesty-incentivizing reward or penalty function) that measures the accuracy of probabilistic judgments.[18] In this analysis, we calculated Daily Brier Scores (DBSs) for each active day of each forecast question analyzed, treating the market's daily probabilities as of Noon ET as the daily forecast. Generally speaking, for binary and unordered multi-option forecasting questions,[19] a DBS is computed by squaring the difference between the actual outcome and the forecasted value, and taking the sum of all of these squared errors:

where *p* is the forecasted probability falling between 0 and 1; *o* is the resolved outcome (1 for occurrence and 0 for non-occurrence; K is the total number of answer options (two for a binary question) with k corresponding to option (answer choice) from 1 to K.

To characterize the accuracy of forecasts over the multi-day "lifespan" of a question, we computed a *Mean* Daily Brier (MDB), wherein Daily Brier Scores are averaged over all days during which the question was open and the outcome unresolved:

where d corresponds to a given day of a given question's lifespan out of the D total number of days that constitute that lifespan.

To characterize the overall accuracy of a forecasting platform (system, method) over the set of forecast problems analyzed, we calculated a mean of question-specific MDBs over all problems, the Mean Mean Daily Brier (MMDB):

where n corresponds to the nth forecast problem, with N the total number of forecasting questions. As a "mean of means," the MMDB has the effect of according equal weight to each

[17] On the logit transformation for forecasts, see Satopää et al. 2014a. On GJP aggregation algorithms designed to combine and enhance the accuracy of individual probability judgments more generally, see Satopää et al. 2014b; Baron et al. 2014).

[18] The Brier score is a "strictly proper" scoring rule—that is, one that is not gameable, and one for which a forecaster's (or forecasting system's) best score is obtained by a forecaster reporting their "true" probabilities—without hedging or attempting to strategically modulate their beliefs.

[19] The individual forecasting question is the unit of analysis for this paper.

forecasting question regardless of its duration. This approach can be contrasted with an approach that simply averages over all days of all questions giving equal weight to each question-day, which would tend to overweight longer-duration questions relative to shorter-duration questions. Note that the DB, MDB, and MMDB all follow the same scale, ranging from 0 to 2, where lower values indicate *higher* accuracy.

Because employing the traditional Brier score on ordinal questions would treat any probability assigned to an incorrect bin as equally wrong, the scoring procedure for ordinal questions employs an adjustment to the Brier score as described in Jose, Nau, and Winkler (2009) that gives partial credit to closer to correct (i.e. less wrong) forecasts.[20]

Directional Accuracy

Despite offering the advantages of a strictly proper forecast scoring rule, one concern with the Brier score is that it overemphasizes fine gradations of probability. For example, if one assigns a p=0.8 to the correct outcome, the resulting Brier score is four (4) times as large (. 08) as the score one would obtain if one had assigned p=0.9 to that outcome (.02). The difference in absolute Brier scores for these two forecasts is low, but the relative difference (e.g., as reflected on a percentage basis) appears substantial. Critics of the Brier score's use for this purpose may note that humans have considerable difficulty making meaningful distinctions among such small gradations of probability. For example, even the best poker players in the world are unlikely to be able to distinguish between more than about 20 degrees of probabilistic occurrence (i.e. the ability to distinguish a 50-50 shot from a 55-45 shot (Tetlock 2014: 37)), and most mortals are far less precise than this. As such, a scoring rule that views p=0.9 as "four times as accurate" as p=0.8 could be faulted for exaggerating the practical difference between two forecasts.

To address this concern, we also consider a measure of directional accuracy (DA). Rather than providing higher rewards for incremental increases in accuracy, DA metrics make a binary judgment as to whether a forecasting system identified the "right" outcome by way of assigning the single largest portion of its probability to that (ultimately) correct outcome. For instance, for a binary (yes/no) question that resolves "yes," a DA metric would look at each daily probability and determine whether it was "correct" by being greater than .5 or "incorrect" by being less than or equal to .5.[21] For a forecast question that was active for multiple days (as virtually all ICPM/ACE questions are), the DA metric then takes the percentage of a question's total active days during which the forecasting system identified the correct outcome in this directional sense. This results in a percentage of days directionally accurate (PDDA). From there, we take each question's PDDA and average them across all questions to get a mean percentage of days directionally accurate (MPDDA), in a manner somewhat analogous to the MMDB.[22]

[20] In this vein, a forecast occurs for an ordinal bin that is simply the correct or incorrect bin (1 or 0), but the weight to the forecast for a given ordinal bin is determined by how close to the "correct" ordinal bin a given bin lies. The procedure for implementing the Jose et al., ordinal scoring rule is to break the original answer bins (A-B-C-D-E) into a set of binary categories (A-BCDE; AB-CDE; ABC-DE; ABCD-E); apply the Brier scoring procedure to each binary pair; and then calculate an average Brier score across the binary pair scores (Jose, Nau, & Winkler, 2009).

[21] Or, for a multi-option question, a probability is directionally accurate if it exceeds the probability of all other answer options.

[22] Higher MPDDA is good (more accurate).

5. Analysis

The accuracy figures (Brier and DA) in this analysis were calculated for the GJP Prediction Market (GJPPM) and the ICPM using noon prices from each platform for each day from August 1 2013 through May 9 2014[23]; during which 139 questions were posted jointly on both platforms. We turn first to the Brier score, focusing on the analysis of MMDBs.

Brier Score Results

Figure 5.1 and Table 1 present the MMDBs and associated (bootstrap-based) 95% confidence intervals for the ICPM and the various GJP platforms (Figure 3.1 and Table 1), along with tests of the null hypothesis. These results indicate that the GJP platform MMDBs are, on average, equal to those of the ICPM (Table 1).[24] Collectively, the results indicate that the ICPM is significantly more accurate than the GJP ULinOP, is of comparable statistical accuracy[25] to the GJPPM, and is significantly less accurate than the GJP most accurate method ("All Surveys Logit").

Figure 5.1

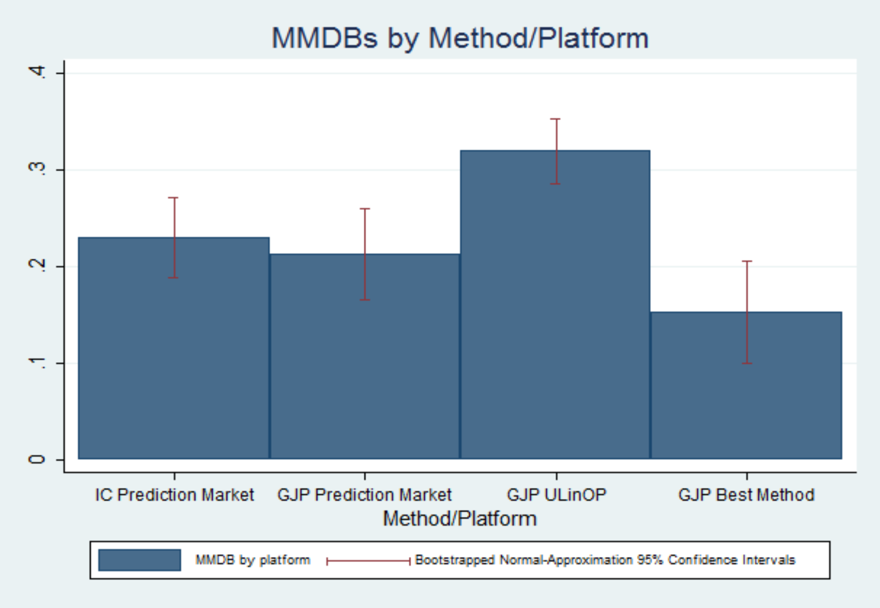

Table 1

[23] This comparison constitutes only a subset of the four-year duration of the IARPA ACE program, which ran from late-2011 through mid-2015.

[24] Recall that for MMDB, based on Brier Scores, lower scores are better (more accurate).

[25] The phrases "comparable statistical accuracy" or "statistically indistinguishable," as used in this paper, should be understood as indicating that no statistically significant difference exists (at the p=.05 level of confidence). These phrases should not be interpreted as implying anything regarding the sufficiency (or lack thereof) of the statistical power of the hypothesis tests underpinning these statements.

Platform	**Mean of Mean Daily Brier (MMDB)**	**Bootstrapped 95% confidence interval for MDB; p(z) statistical significance for comparison to non-ICPM platforms**
IC Prediction Market	.23	(.19, .27)
Good Judgment Project Prediction Market	.21	(.17, .26)
Good Judgment Project ULinOP	.32	(.29, .35)***
Good Judgment Project best method	.15	(.10, .21)***

***p<.001

DA Results

As noted previously, DA is generally a blunter tool than the Brier score, because it collapses probabilities into binary states, treating all "directionally" correct (or incorrect) probability judgments as equivalent. For instance, as long as it assigns a higher probability to the correct outcome than to any other competing outcome, a forecast is considered directionally accurate, regardless of whether that probability was hedged toward an "ignorance prior"[26] (e.g., assigning p=0.501 to the correct outcome for a binary yes-no question) or extremized toward complete certainty (e.g., assigning p=0.999 to the correct answer. As such, one would expect this blunter tool to suppress the apparent variation in accuracy across the various platforms analyzed. This expectation is indeed borne out. As depicted in Figure 5.2 and Table 2, the ICPM is not statistically better than any GJP method, as the confidence intervals for the differences in MPDDAs of the ICPM vs. the GJP Prediction Market and ULinOP both straddle zero. The ICPM is, however, statistically worse than GJP's best method in terms of directional accuracy.

(Text Continues)

Figure 5.2

[26] The "ignorance prior" would be 50% for a binary question.

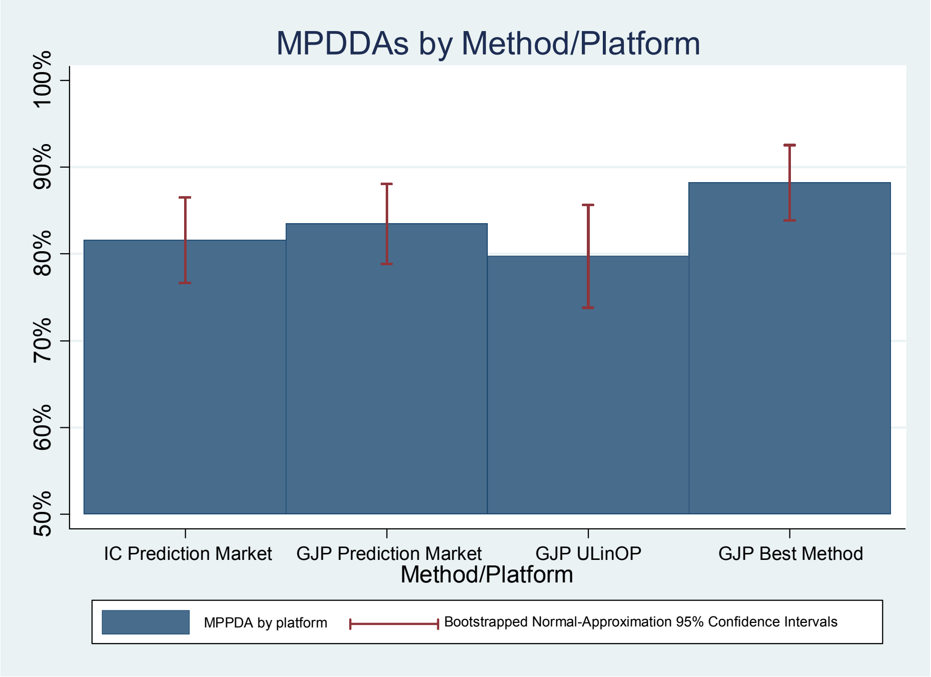

Table 2

Platform	Mean Percentage of Days Directionally accurate (MPDDA)	Absolute Difference in platform MPDDA vs. ICPM; 95% bootstrapped confidence interval of individual platform MPDDA	Bootstrapped 95% confidence interval for absolute difference between MPDDAs; p(z) statistical significance for comparison to non-ICPM platforms (two-tailed)
IC Prediction Market	81.58	N/A; (76.63, 86.54)	N/A
Good Judgment Project Prediction Market	83.45	-1.87; (78.83, 88.06)	(-5.00, 1.27)
Good Judgment Project ULinOP	79.74	1.85; (73.82, 85.66)	(-2.62, 6.31)
Good Judgment Project best method	88.20	-6.62***; (83.87, 92.53)	(-9.83, -3.40)***

*P<.05; **p<.01; ***p<.001

6. Discussion

The evidence reviewed here suggests that the ICPM is statistically superior to the GJP ULinOP, statistically inferior to GJP's most accurate method, and statistically indistinguishable from the GJP Prediction Market. While we must be cautious in the inferences we draw from these comparisons (they employ only about 9 months' worth of forecast data), the results are consistent with previous comparative work on optimal CW methods. Namely, those conclusions are: (1) prediction markets will outperform standard unweighted opinion pools[27]; (2) optimized/weighted opinion pools will outperform prediction markets.[28]

Why Didn't the ICPM Beat the GJP Prediction Market?

One may be tempted to draw overgeneralized inferences from the ICPM's failure to outperform the GJP prediction market. After all, IC analysts have access to classified information where GJP analysts do not, meaning IC analysts have access to more information and more types of information.[29] However, recall again that this analysis considers only comparative performance on *unclassified* questions. This distinction will be a critical one to the extent that we expect classified information to be most helpful or diagnostic in forecasting against classified questions and least uniquely helpful on unclassified questions.[30] Thus, it is not obvious that IC analysts should have a unique advantage over GJP analysts generally and especially in the case of GJP analysts augmented by novel interventions and innovative algorithmic methods.

Indeed, we cannot rule out the possibility that cleared analysts' access to classified information may even limit their forecast accuracy in some contexts. For instance, Travers et al. (2014) identified the "secrecy heuristic"[31] as a hypothetical tendency of cleared analysts

[27] This conclusion was reached by the Good Judgment Project in 2012—following the first year of the IARPA ACE program--see Ungar et al. 2012. Although for a different take—that unweighted opinion pools and prediction markets are of roughly comparable accuracy—see Chen et al. 2005.

[28] Again, see Ungar et al. 2012.

[29] While this line of thinking is common, once one has a reasonable amount of necessary information to render a judgment, having more information is not necessarily beneficial to intelligence analysis (cf. Heuer 1999: 51-64).

[30] Some may counter that many times what causes a question to be classified is its end date (for instance, an intelligence product, based on classified information, may hint at an unclassified event to come by date X). In such a case, it is the suggestion that an unclassified event may occur before some date (if such a report is based on classified sources) that makes a question classified, rather than the intrinsic subject matter of a classified question itself (which is very often ipso facto unclassified).

[31] In the Travers et al. study, Amazon Mechanical Turk participants were confronted with declassified information that they believed to be classified (compared with a control group presented with the same document that they believed to be unclassified); the results indicated that participants rated apparently classified documents as higher quality than other participants rated an identical but apparently unclassified document.

to illogically overweight the importance or credibility of (apparently) classified information,[32] a tendency that could be at play even for unclassified forecast questions. In other words, to the extent that intelligence analysts with access to classified information have the same finite amount of time to read information relevant to a forecasting question as analysts without access to classified information, classified and unclassified documents are zero-sum competitors for the attention of the cleared analysts. And to the extent that classified information is not ipso facto more informative than open source information for any given question, ICPM analysts may spend too much time considering classified information in forecasting on unclassified questions. Such interpretations are, given the data analyzed in this study, of course speculative, but they seem sufficiently plausible as to challenge unreflective assumptions about the unconditional value or relevance of classified information in unclassified forecasting.

Yet another possible explanation (or contributing cause) of the ICPM's inability to beat the GJP prediction market may lie in the fact that GJP participants were directly compensated for their participation while ICPM participants were not[33], though no study has directly examined the effects of **compensation for participation[34] on accuracy** in crowdsourcing platforms.[35] Because direct compensation for participation may increase an individual's propensity to update their positions/forecasts, it is conceivable that compensation for participation may have conferred an advantage to GJP.

Conclusion

As important as the conclusions discussed are the ones not discussed. This study did not examine, for instance, the differences in accuracy across systems by region or by subject matter, which may be fruitfully analyzed in a follow-up study. This means that we cannot conclude that the ICPM and GJP systems were of equal accuracy across all geopolitical regions or question subject matter (either across the domain of the N=139 questions in this study or beyond that domain). The failure to examine these questions does not render them any less important; they are simply outside the scope and space of the present analysis.

Further, although the GJP best method was statistically superior to the ICPM, it is important to contextualize this finding. The GJP best method was afforded the fruits of a large research program devoted to enhancing the accuracy of crowdsourced forecasting systems, whereas the ICPM represents forecasts made on commercial-off-the-shelf technology, without the benefit of the use of training materials, teaming protocols, or aggregation algorithms, all of which have been shown to improve the accuracy of forecasts. While the ICPM cannot yet lay

[32] The Travers et al. study did not use actual classified information, but rather the experimenters re-marked declassified information to create in participants the impression that documents were classified.

[33] ICPM participants are, however, indirectly compensated for participation because all are either government employees or government-funded contractors and they participate on the ICPM while at work.

[34] That is to say, the effect of paying people simply for providing forecasts or trades.

[35] Some have, however, analyzed the effect of play-money (points) as opposed to real-dollars on prediction market accuracy (cf. Servan-Schreiber et al. 2004)—and found that no significant difference in accuracy among the two types of markets exists.

claim to being more accurate than its unclassified counterpart, future ICPM incorporation of technologies validated during the ACE program may allow its accuracy to increase over time.

Acknowledgments

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract 2015-14120200002-002 and other vehicles. The views and conclusions contained 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, ODNI, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon.

Works Cited

Baron, Jonathan, et al. "Two reasons to make aggregated probability forecasts more extreme." Decision Analysis 11.2 (2014): 133-145.

Brier, Glenn W. "Verification of forecasts expressed in terms of probability." Monthly weather review 78.1 (1950): 1-3.

Budish, Eric B., Peter Cramton, and John J. Shim. "The high-frequency trading arms race: Frequent batch auctions as a market design response." Fama-Miller Working Paper (2013): 14-03.

Chen, Yiling, et al. "Information markets vs. opinion pools: An empirical comparison." Proceedings of the 6th ACM conference on Electronic commerce. ACM, 2005.

Clapper, James. *The National Intelligence Strategy of the United States of America (2014)*. 2014, at http://www.dni.gov/files/documents/2014_NIS_Publication.pdf

Eckstein, Harry. "Case Study and Theory in Political Science. Greenstein, E and N. Polsby, eds. Handbook of Political Science (Vol. 7)." (1975).

Friedman, Jeffrey A., and Richard Zeckhauser. "Assessing Uncertainty in Intelligence." Intelligence and National Security 27.6 (2012): 824-847.

Friedman, Jeffrey A., and Richard Zeckhauser. "Why Assessing Estimative Accuracy is Feasible and Desirable." Intelligence and National Security (2014): 1-23.

Good Judgment Project. “Exploring the Optimal Forecasting Frontier: How Much Room Is There to Improve Subjective Forecasting Accuracy?” (2014) Annual Report of the Good Judgment Project, UC-Berkeley, Prime Contractor, Submitted 2 Jun 2014.

Hanson, Robin. "Combinatorial information market design." Information Systems Frontiers 5.1 (2003): 107-119.

Hersh, Seymour. "The stovepipe." New Yorker 27.October (2003): pp-76.

Heuer, Richards J. *Psychology of intelligence analysis*. CIA, Center for the Study of Intelligence, 1999.

Hou, Yuan. Email to the author. January 29, 2015.

Johnston, Rob. *Analytic culture in the US intelligence community: An ethnographic study*. Central
Intelligence Agency Washington DC Center For Study Of Intelligence, 2005.

Jose, Victor Richmond R., Robert F. Nau, and Robert L. Winkler. "Sensitivity to distance and baseline
distributions in forecast evaluation." Management Science 55.4 (2009): 582-590.

Mellers, Barbara, et al. "Psychological strategies for winning a geopolitical forecasting tournament." Psychological science 25.5 (2014): 1106-1115.

Poi, Brian P. "From the Help Desk: Some Bootstrapping Techniques." Stata Journal 4 (2004): 312-328

Satopää, Ville A., et al. "Combining multiple probability predictions using a simple logit model." International Journal of Forecasting 30.2 (2014): 344-356.

Satopää, Ville A., et al. "Probability aggregation in time-series: Dynamic hierarchical modeling of sparse
expert beliefs." The Annals of Applied Statistics 8.2 (2014): 1256-1280.

Servan-Schreiber, Emile, et al. "Prediction markets: Does money matter?" Electronic markets 14.3
(2004): 243-251.

Siegel, Adam. "Inkling: One Prediction Market Platform Provider's Experience." The Journal of Prediction
Markets 3.1 (2009): 65.

Sobel, Russell S., and S. Travis Raines. "An examination of the empirical derivatives of the favourite-
longshot bias in racetrack betting." Applied Economics 35.4 (2003): 371-385.

Surowiecki, James. The wisdom of crowds. Anchor, 2005.

Taleb, Nassim Nicholas, and Philip E. Tetlock. "On the Difference between Binary Prediction and True
Exposure with Implications for Forecasting Tournaments and Decision Making Research." Available at SSRN 2284964 (2013).

Tetlock, Philip. *Expert political judgment: How good is it? How can we know?* Princeton University
Press, 2005.

Tetlock, P. (2014). “Forecasting Tournaments Test the Limits of Judgment and Stretch the Boundaries of Science.” Unpublished Manuscript, Good Judgment Project, University of Pennsylvania, Philadelphia, PA.

Tetlock, Philip E., and Barbara A. Mellers. "Intelligent management of intelligence agencies: beyond accountability ping-pong." American Psychologist 66.6 (2011): 542.

Tetlock, Philip, and Barbara Mellers. "Judging political judgment." Proceedings of the National Academy of Sciences 111.32 (2014): 11574-11575.

Travers, Mark, Leaf Van Boven, and Charles Judd. "The secrecy heuristic: Inferring quality from secrecy in foreign policy contexts." Political Psychology 35.1 (2014): 97-111.

Ungar, Lyle, et al. "The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions." (2012).

United States. *Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction*, Laurence H. Silberman, and Charles S. Robb. Commission on the intelligence capabilities of the United States regarding weapons of mass destruction. The Commission, 2004.

United States. *The 9/11 commission report: Final report of the national commission on terrorist attacks upon the United States*. Government Printing Office, 2011.