Modeling the Evolution of Beliefs in Abstract Geopolitical Forecasting with Hidden Markov Models

9.66 Self-Selected Project

Gatlen Culp (gculp@mit.edu)
Massachusetts Institute of Technology, Undergraduate

Abstract

We study belief evolution in geopolitical forecasting using survey forecasts from the Good Judgement Project (2011–2015). After filtering to non-voided binary questions and baseline (first) forecasts, the dataset contains 382 individual forecasting problems and 323,847 baseline probability reports. We first examine whether forecasts exhibit measurable cross-question dependency (via correlations) and outline a cognitive modeling approach in which a forecaster's latent belief state evolves over time and is imperfectly observed through sparse reports, formalized as a Hidden Markov Model. We conclude by proposing train/test split strategies and predictive baselines for comparing individual and collective models.

Keywords: forecasting; belief perseverance; hidden markov models; cognitive modeling

Link to Project Code on GitHub
https://github.com/GatlenCulp/966_cocosci_forecasting

Introduction

Forecasting – making informed predictions about the future[1] – is simultaneously extremely difficult and extremely valuable. Forecasting is used constantly in our personal lives and used professionally by consultants, politicians, analysts, entrepreneurs, and more. Forecasting geopolitical events and responses is critically important for international and domestic peace, prosperity, and stability – underestimating the likelihood is an error measured in human lives.[2]

From an evolutionary psychology perspective, forecasting had limited usefulness. Predicting potential personal emergencies (such as hunting accidents) is useful for rationing food preserves and improving fitness, but forecasting more abstract concepts (such as whether China will invade Taiwan or whether Serbia will be officially granted EU candidacy) would rarely improve fitness.

Forecasting is a skill that can be trained like any other, and because modern forecasting problems are far from our evolved niches, the ceiling for human capabilities is likely far beyond our innate ability and much is yet to be known, as evidenced by research volunteers outperforming senior professionals in the U.S. Intelligence Community (IC)[3] after just a year. (Tetlock & Gardner, 2015)

One of the most fundamental tools to improve any skill is to understand our current untrained ability to forecast – downsides, strengths, common mistakes, etc. By understanding our abilities, we can improve forecasting pedagogy and our intuitive understanding of the future.[4]

An important starting point is understanding not just how individuals update their beliefs in response to new information, but how individuals update their beliefs as they think more about the future. In particular, I decided to model the following: when people forecast whether or not an event will occur in the future, how does this affect their beliefs about other events they have yet to consider? And does the order in which they consider these events "stick", or become hard to change? If they believe a far-off event is likely to occur, might they be averse to revising their belief even as they later consider related events happening beforehand?[5]

In parallel, I formalize an individual's day-by-day forecasting behavior as a Hidden Markov Model: a hidden belief vector over questions evolves over time, and the observed forecasts are noisy, sparse reports of those beliefs. I then evaluate whether this stylized model captures forecasting behavior by fitting it on subsets of each individual's timeline and measuring predictive accuracy on held-out forecasts, comparing against simple baselines (e.g. the collective average) and a collective model fit across forecasters. This sets up an empir-

[1] Typically on the order of months to years

[2] In addition to individual forecasting, there is much room for improvement in group forecasting and communication. E.g. the 2011 Abbottabad raid of the U.S. into Pakistan was rife with miscommunications about the likelihood of Osama bin Laden's location within the country, straining U.S.-Pakistan relations and not completely eliminating al-Qaeda. (Tetlock & Gardner, 2015)

[3] Consisting of the CIA and FBI among others

[4] Other reasons why understanding human forecasting is interesting: (A) We may be able to build algorithms around it. For example, more efficiently sampling and de-biasing predictions about the future provided by LLMs. (B) Beyond improving forecasting, understanding how others perceive the future (e.g. citizens) may itself be valuable information in forecasting.

[5] For example, I have been talking with people about the consequences of advanced AI up to and including superintelligent AI before ChatGPT was released. I find that many people find it improbable, not from forward-chaining events from today but from some bias towards a future that they are familiar with, and backwards-chain from that future to discount the probability of a much crazier future (may be an availability heuristic)

ical testbed for belief perseverance (resistance to revising beliefs) in sequential, interdependent forecasting.

Background

Despite the extreme importance of geopolitical forecasting, little research has investigated how humans develop their intuition for the abstract and distant future.

Previous work by the **Good Judgement Project (GJP)**[6] in collaboration with the U.S. Intelligence Community collected 888,328 geopolitical forecasts on questions such as *"Will Daniel Ortega win another term as President of Nicaragua during the late 2011 elections?"* and identified practices, personality traits, and institutional procedures to improve systemic forecasting ability even beyond experts in the CIA, FBI, and more. (Friedman et al., 2018; Tetlock & Gardner, 2015) While this work explored how experts may improve, it said little about how everyday people make their predictions.

In this paper, I use the dataset of geopolitical forecasts provided by the Good Judgement Project (Friedman et al., 2018).

Previous work by (Lieder et al., 2012) attempted to model the psychological concept of anchoring – weighting the first piece of evidence more than subsequent pieces – by using MCMC with a limited number of iterations, and demonstrated that the perceived distribution was different from the steady-state, demonstrating a probability "burn-in" of visited states.

The Good Judgement Project Dataset

Before introducing any forecasting model, I first describe the subset of the GJP dataset I analyze and show that it contains a measurable dependency structure worth modeling.

The GJP has collected forecasts via both prediction markets and volunteer surveys. For the sake of this research, I limited myself to all four years of the survey forecasts from September 2011 to May 2015.

Forecasting Problems (IFPs) and Filtering

Each question in the dataset is referred to as an **Individual Forecasting Problem (IFP)**. For two IFP examples, see Table 1.[7]

Table 1: Example IFPs from the GJP dataset (simplified)

`ifp_id`	`short_title`	`q_text`	`options`	dates
1004-0	UN-GA recognize Palestine	Will the United Nations General Assembly recognize a Palestinian state by 30 September 2011?	(a) Yes (b) No	2011-09-01 to 2011-09-30
1007-0	confrontation in S. or E. China Sea	Will there be a lethal confrontation involving government forces in the South China Sea or East China Sea by 31 December 2011?	(a) Yes, by 15 October 2011 (b) Yes, between 16 Oct and 31 Dec (c) No	2011-09-01 to 2011-12-11

To guarantee the validity of my results, I filtered any IFPs labeled `voided`. And for model simplification purposes, later explained in *Forecasting Models*, I limited my analysis to boolean (yes/no) questions.[8] This reduced the number of IFPs from 617 to 382. Figure 4 displays the remaining questions on a timeline.[9]

Individual Forecasts (Baseline Survey Priors)

Each of the volunteers in the GJP study were assigned a unique and anonymous ID and could make forecasts on IFPs via an online platform during each IFP's window shown in Figure 4. The forecasters would rate the probability of each option as a value between 0 and 1.

Because users can update their forecasts on the same IFP, I decided to consider only the first forecast, which I call the "baseline" forecast. Because I had filtered for only boolean IFPs, I was able to reduce the forecasts of an agent i on any IFP to a single value, $\Pr[\text{answer}_i = \texttt{"(a) yes"}]$. Unless otherwise noted, consider all forecasts below to be baseline forecasts.[10]

After filtering all survey forecasts to include only the baseline and the restricted IFP set, the number of forecasts in this dataset was reduced to $323,847$.

An example of five forecasts can be seen in Table 2.[11]

[6] *Superforecasting*, written by administrators of GJP, was the main inspiration behind this research project.

[7] To fit this table, a number of columns were removed/combined. An important column is `q_desc`, a 1-2 paragraph-long description containing more information and formal resolution criteria. `q_status` notes whether the IFP was `closed` or `voided`.

[8] Extension to multiple-choice questions using a different state-space is possible, but unnecessary for analysis.

[9] Originally an interactive graph. May be too hard to make out in the paper, but generally useful to understand the structure.

[10] Only after I had done most of my project had I realized that some of the questions had important conditionals inside of the `outcomes` field instead of in the question text. E.g. "If Parti Quebecois does not hold a majority of seats in the Quebec provincial legislature beforehand : (a) Yes, (b) No". These conditionals certainly have a large influence on the volunteers' responses which I did not consider in my project. However, because these conditionals $\Pr(\text{yes} \mid \text{condition})$ are positively correlated with $\Pr(\text{yes})$, the results here are not entirely invalid.

[11] As with Table 1, a number of columns were omitted, but the ones displayed are the focus of this study.

Table 2: Example baseline survey forecast priors from the GJP dataset. Each row is a user's first forecast on an IFP for a given answer option.

ifp_id	user_id	answer_option	confidence	timestamp
1340-0	6324	a	0.10	2014-02-04 10:42:04
1453-0	23336	a	0.15	2014-10-25 03:43:54
1468-6	9015	b	0.25	2015-01-14 12:32:09
1221-6	9219	d	0.97	2014-03-30 17:58:13
1129-2	2827	b	0.95	2012-08-13 11:16:23

In Figure 1, there is an example of a user's complete forecasting profile depicted as a timeline. This user was in the top 80% most active forecasters after applying the filters mentioned above.

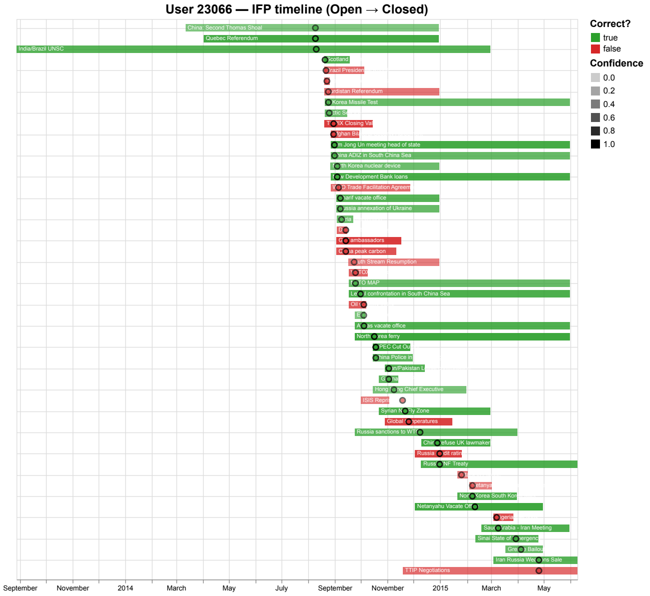


Figure 1: Zoomed-in example of user 23066′s forecasting timeline. Only IFPs the user submitted a forecast on are shown. The dot on each bar is the forecast timestamp; green/red indicates correct/incorrect; opacity scales with confidence.

Forecast Correlations

As a preliminary check before doing any dependency-based modeling, I searched for highly correlated (positive or negative) baseline forecasts across IFPs. If forecasts were close to independent across questions, then any modeling based on cross-IFP structure would be poorly motivated. In Figure 2, you can see the top 10 correlations per IFP.

At first glance, these events seem to have little to do with one another and future research may be interested in exploring the chain of logic or personality/cultural biases by which individuals make these correlated forecasts – for example, the two most correlated forecasts involve: "Will India and/or Brazil become a permanent member of the U.N. Security Council before 1 March 2015?" and "Will a referendum on Quebec's affiliation with Canada be held before 31 December 2014?". This strong correlation may indicate some underlying personal bias towards or against "cooperation".[12]

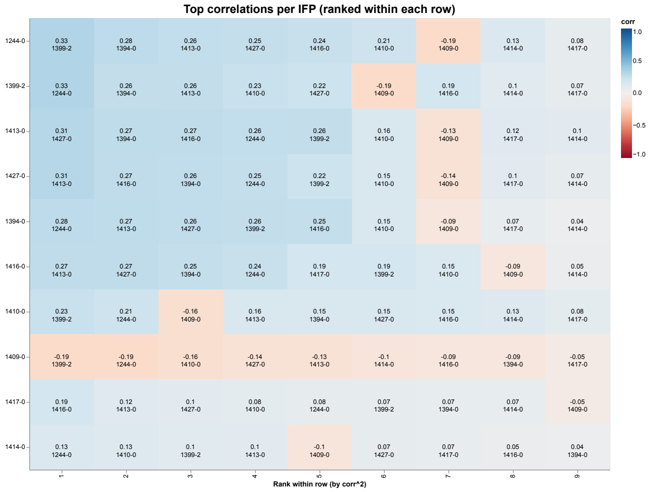


Figure 2: Top correlations per IFP (ranked within each row)

The top 5 pairs, as displayed below in Table 3, coincidentally have reasonably high positive correlations, ranging from 0.266 to 0.325. This weakly supports using the dependency structure between an individual's forecasts.[13]

Table 3: Top 5 correlated IFP pairs in the filtered survey forecasts.[14]

ifp_id_a	short_title_a	ifp_id_b	short_title_b	corr
1244-0	India/Brazil UNSC	1399-2	Quebec Referendum	0.325
1413-0	Kurdistan Referendum	1427-0	Russia annexation of Ukraine	0.305
1244-0	India/Brazil UNSC	1394-0	China: Second Thomas Shoal	0.277
1394-0	China: Second Thomas Shoal	1413-0	Kurdistan Referendum	0.270
1413-0	Kurdistan Referendum	1416-0	North Korea nuclear device	0.266

[12] Principal Component Analysis across forecasters may be an interesting direction.

[13] To compare against generally temporally-similar data, see Figure 5 in the appendix, showing the correlation coefficient matrix for the first 10 IFPs from the dataset. The values here range from −0.21 to 0.28.

[14] Given more time, I would have also liked to examine which IFPs had the most variance and how much the variance of one IFP could explain that of another. The figure here does not tell us much about the differences between individuals.

Modeling Forecasts

Now that we have a better understanding of the data and reason to believe that forecasts are correlated across IFPs, we can go about modeling (and later simulating) how forecasters form and update their predictions. The first step is choosing a representation for the discrete time series of an individual's forecasts. Because we are interested in representing the internal beliefs of the human forecaster and not just representing the forecasts themselves, one promising formalization of this data is as a Hidden Markov Model (HMM) as depicted in Figure 3.

For a single forecaster, I write the hidden belief state at time t as $\mathbf{Z}_t$, where t is a discrete day index measured from the start of the dataset (so $t = 1$ is the day of the first GJP datapoint). Let $\mathcal{I}$ be the set of boolean IFPs after filtering. Then $\mathbf{Z}_t$ is a vector in $[0, 1]^{|\mathcal{I}|}$, Where for $i \in \mathcal{I}$, $Z_{t,i}$ is the forecaster's subjective probability (on day t) that IFP i will resolve to "Yes".

Because the forecaster's mind cannot be read directly, the belief state must be inferred from the forecasts they submitted. I therefore represent the forecast at time t as $\boldsymbol{X}_t$, a vector with the same indexing as $\mathbf{Z}_t$. When a forecaster submits a forecast on day t for IFP i, the value $X_{t,i}$ represents the reported probability that the forecaster assigns to "Yes". Let lowercase variables represent the realized values of their uppercase counterparts. When they do **not** submit a forecast for i on day t, that coordinate is unobserved, which I encode as $x_{t,i} = -1$.

For simplicity, I assume a stationary HMM (both the belief transitions and the reporting emissions are time-independent), with **transition** distribution $p_{\mathbf{Z}_{t+1}|\mathbf{Z}_t}(\cdot \mid \cdot)$ and **emission** distribution $p_{\mathbf{X}_t|\mathbf{Z}_t}(\cdot \mid \cdot)$ (see the Notation Reference). I also do not attempt to model exogenous information (e.g. reading the news) which may be represented (albeit at a lower resolution) in the hidden belief transitions.

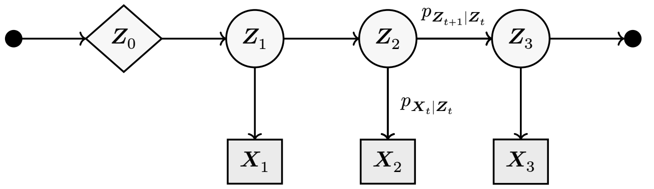


Figure 3: Hidden Markov Model. The initial dot represents that the individual's background (provided in the GJP dataset) may be used to help configure their priors $\mathbf{Z}_0$, which we do not do here.

We can additionally assume that forecasters are simply report their hidden beliefs from a tight normal distribution ($\sigma = 0.05$), clamped by the probability range $[0, 1]$ and with a mean around their actual hidden belief and the forecasts they decide to make at time t are independent of their internal beliefs (e.g. we ignore that individuals may be more likely to make forecasts on IFPs they have greater certainty on). Mathematically:

$$p_{\mathbf{X}_t|\mathbf{Z}_t}\left(x_{t,i} \mid z_{t,i}\right) = \begin{cases} 1 \text{ if } x_{t,i} = -1 \\ 1 \text{ if } x_{t,i} = z_{t,i} \\ 0 \text{ otherwise} \end{cases}$$

The last thing to do now would be to learn the transitions probabilities $p_{\mathbf{Z}_{t+1}|\mathbf{Z}_t}$ and the prior (initial belief-state) $\mathbf{Z}_0$. Which can be done using the Baum-Welch (Forward-Backward) Algorithm. (Devijver, 1985) We can start a uniform prior and identical transition probabilities:

$$\mathbf{Z}_0 \leftarrow [1/2, \cdots, 1/2]^{\mathsf{T}},$$

$$p_{\mathbf{Z}_{t+1}|\mathbf{Z}_t}(z_{t+1} \mid z_t) \leftarrow \begin{cases} 1 \text{ if } z_{t+1} = z_t \\ 0 \text{ otherwise} \end{cases}$$

For the individual, the emissions distribution will simply be learned to be a normal distribution with the mean of the observation.

After developing an HMM model from the forwards-backwards algorithm for an individual, we can then also develop another HMM algorithm for the collective set of forecasts to compare against, treating each individual forecast as if they come from the same, unknown distribution.[15]

Evaluating Models

There are a few ways we can evaluate this model using the GJP dataset, by combining (A) all of the train / test splits below with any of the (B) evaluation metrics also found below. The "*dataset*" referenced here is the *set of forecasts made by the individual* that the model was fit to. Each of these should be repeated across all users and the distribution of results should be analyzed.

(A) Train/Test Split Strategies

For any of the train/test splits below, we fit our model with the forwards-backwards algorithm only using the training data and evaluate it based on the witheld testing data.

1. **Randomized Split**: Select a random x% of the available forecasts to include in training, the rest become part of the testing set. This is meant to evaluate overall quality of our model
2. **Temporal Cut-Off Split**: Select a time τ as the stopping point. Everything before is used for training, everything after is used for testing. This is meant to evaluate our stationary-transition condition – being able to successfully forecast a long and uninterrupted sequence of events would be positive evidence that a stationary-transition assumption is reasonable.
3. **Time-Interval Split**: Select a start and end time such that $1 < \tau_{\text{start}} < \tau_{\text{end}} < |\text{dataset}|$. This is meant to contrast against the temporal cut-off split – just how much prediction accuracy is gained by knowing forecasts in the far future? If similarly sized train/test splits have higher accuracy in this split than the cut-off split and the stationary-transition assumption appears false, then this is positive evidence for a predictably evolving belief transition function – i.e. evidence as time progresses, the

[15] We could also use the correlation coefficient matrix as a transition function and renormalize after each step.

way beliefs “interact” in your head to form new ideas and beliefs mutates slowly and continuously.

(B) Evaluation Metrics

Both of these metrics should be used and compared with one another. After a model is fit using the training data, the following are measured on the testing data. **Accuracy** here is defined to be the average squared error across model-predictions. E.g. if the model predicts a user’s belief state of ifp i and j to be 0.25 at time t but they had actually reported 0.75 and 0.5 at time t respectively, then the **error** of that prediction is defined to be the euclidean distance between the predicted and true forecasts, in this instance: $\sqrt{(0.75 - 0.25)^2 + (0.5 - 0.25)^2}$.

1. **Collective-Average Prediction Accuracy**: As a baseline statistic – what is the accuracy of predicting that the forecaster will make a forecast on every IFP equal to the average over the collective forecasts of all users?
2. **Collective-Model Prediction Accuracy**: If we were to use the HMM that fit to the collective predictions across all users – what is the accuracy?
3. **Individual-Model Prediction Accuracy**: If we were then to use the individually-fitted model, how much better does it perform than the collective-model? If the accuracy is much better, this is positive evidence for there being strong differences between how individuals make forecasts.

Conclusion

In this paper, I introduce the importance of further developing the field of computational cognitive science in understanding how humans develop their beliefs about the future and forecast geopolitical events as a starting point for developing better forecasting algorithms and practices. I identify and outline a promising dataset by the Good Judgement Project to further develop this research. And I design a simple Hidden Markov Model algorithm, parameter-fitting process, evaluation procedure, and methods of analysis that may be used to understand this topic in depth.

While I did not have the time to implement my procedure and collect data due to spending a considerable amount of time parsing, filtering, and understanding the GJP dataset that I had plan to fit and test my model on, I believe this project is in line with the spirit of computational modeling and analysis that was focused on in the majority of the 9.66 Computational Cognitive Science class.

In my original project I had intended to simulate MCMC with limited iterations on the tree for each respondent, representing “thinking about events most closely related to the one they are forecasting” in the style of (Lieder et al., 2012). Then analyze the effect of this sampling method against forward chaining from the present and through each intermediate event.

Appendix

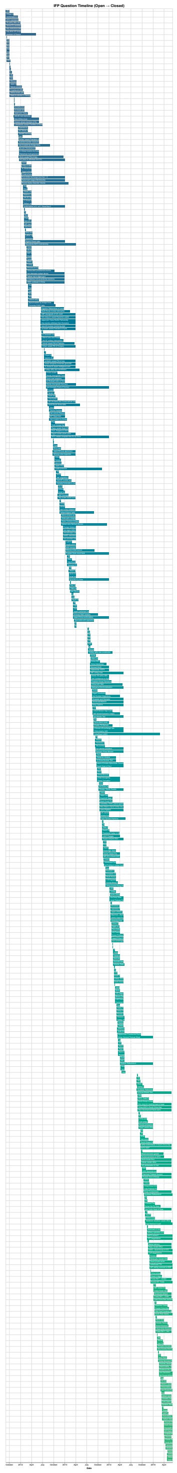
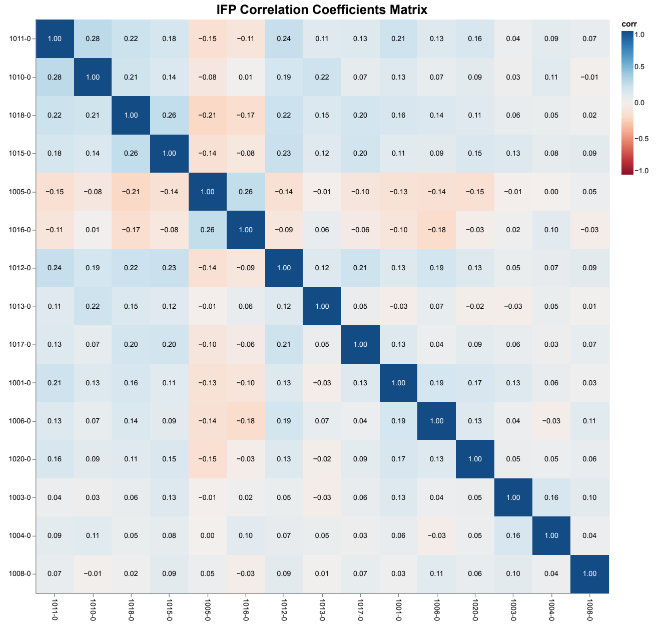


Figure 4: Timeline of all IFPs. The x-axis is time. Each horizontal bar represents an IFP: the left end is when it opened for forecasts and the right end is when it closed, either by deadline (e.g. “Will the city of Kiev be bombed by Russia by Nov 11th, 2012?”) or by resolution (e.g. if Russia bombed Kiev on Oct 11th, 2012).



	1011-0	1010-0	1018-0	1015-0	1005-0	1016-0	1012-0	1013-0	1017-0	1001-0	1006-0	1020-0	1003-0	1004-0	1008-0
1011-0	1.00	0.28	0.22	0.18	−0.15	−0.11	0.24	0.11	0.13	0.21	0.13	0.16	0.04	0.09	0.07
1010-0	0.28	1.00	0.21	0.14	−0.08	0.01	0.19	0.22	0.07	0.13	0.07	0.09	0.03	0.11	−0.01
1018-0	0.22	0.21	1.00	0.26	−0.21	−0.17	0.22	0.15	0.20	0.16	0.14	0.11	0.06	0.05	0.02
1015-0	0.18	0.14	0.26	1.00	−0.14	−0.08	0.23	0.12	0.20	0.11	0.09	0.15	0.13	0.08	0.09
1005-0	−0.15	−0.08	−0.21	−0.14	1.00	0.26	−0.14	−0.01	−0.10	−0.13	−0.14	−0.15	−0.01	0.00	0.05
1016-0	−0.11	0.01	−0.17	−0.08	0.26	1.00	−0.09	0.06	−0.06	−0.10	−0.18	−0.03	0.02	0.10	−0.03
1012-0	0.24	0.19	0.22	0.23	−0.14	−0.09	1.00	0.12	0.21	0.13	0.19	0.13	0.05	0.07	0.09
1013-0	0.11	0.22	0.15	0.12	−0.01	0.06	0.12	1.00	0.05	−0.03	0.07	−0.02	−0.03	0.05	0.01
1017-0	0.13	0.07	0.20	0.20	−0.10	−0.06	0.21	0.05	1.00	0.13	0.04	0.09	0.06	0.03	0.07
1001-0	0.21	0.13	0.16	0.11	−0.13	−0.10	0.13	−0.03	0.13	1.00	0.19	0.17	0.13	0.06	0.03
1006-0	0.13	0.07	0.14	0.09	−0.14	−0.18	0.19	0.07	0.04	0.19	1.00	0.13	0.04	−0.03	0.11
1020-0	0.16	0.09	0.11	0.15	−0.15	−0.03	0.13	−0.02	0.09	0.17	0.13	1.00	0.05	0.05	0.06
1003-0	0.04	0.03	0.06	0.13	−0.01	0.02	0.05	−0.03	0.06	0.13	0.04	0.05	1.00	0.16	0.10
1004-0	0.09	0.11	0.05	0.08	0.00	0.10	0.07	0.05	0.03	0.06	−0.03	0.05	0.16	1.00	0.04
1008-0	0.07	−0.01	0.02	0.09	0.05	−0.03	0.09	0.01	0.07	0.03	0.11	0.06	0.10	0.04	1.00

Figure 5: Correlation coefficient matrix for the first 10 IFPs. Meant as a point of comparison between strongly correlated forecasts and those that may be correlated mainly due to temporal proximity.

Notation Reference

Conventions

X, Y: Uppercase denotes random variables; lowercase denotes realized values (e.g. X vs. x).

$\boldsymbol{x}, \boldsymbol{z}$: Bold denotes a vector / collection across IFPs at a time t.

t: Discrete time index; subscripts indicate time (e.g. z_t).

$p_{X|Y}(x \mid y)$: Conditional probability. We use stationary parameters unless explicitly time-indexed.

Variables and distributions

$\mathcal{I}$: Set of Individual Forecasting Problems (IFPs).

i: An individual IFP index, with i in $\mathcal{I}$.

Z: Latent belief state of a forecaster.

z: Realization of the latent belief state.

X: Observed report / forecast derived from the dataset.

x: Realization of an observed report.

$p_{\mathbf{Z}_{t+1}|\mathbf{Z}_t}$: Stationary transition distribution over belief states.

$p_{X_t|\mathbf{Z}_t}$: Stationary emission distribution from belief states to reports.

Notes for 9.66 Staff

Author Contributions: This project was my original idea, heavily inspired by the book *Superforecasting*. I am an MIT Undergraduate and received no outside assistance. This is not related to any of my other work and would not be done otherwise. However, it may include some interesting findings for future research or work I may do. (See Additional Note on Motivation)

AI Use: I did not use AI for any of the writing in this report other than for helping me generate some of the non-Python diagrams I used in explaining my methodology (I had completely described and understood the diagram I wanted, implementation was what I needed help with). I had also used to to last-minute generate the abstract and the notation reference.

As for the code, I used AI for debugging (improper API-usage) and refactoring (breaking up large functions). Because I decided to use unfamiliar data-manipulation and data-visualization libraries (`polars` and `altair` respectively), I used AI extensively to understand and describe the API calls needed to perform an operation. Nowhere in my code is there logic I personally did not personally design. I had used Claude 4.5 Sonnet/Opus or GPT-5.2.

Additional Note on Motivation: The LLM sampling mentioned in a footnote is a large component of my underlying motivation. I know the authors behind AI-2027 and they are interested in developing better geopolitical forecasting tools and have a small team working on something related. Forecasting and logically stepping through events is an extremely time consuming process and LLMs aren't great at doing this out of the box. I was originally interested in GenLM for my mini project as some way to tie LLMs to quantitative forecasts or parameters representing the modeler's assumptions. Link to my notes related project idea: https://gatlen.notion.site/automated-wargames?source=copy_link

References

Devijver, P. A. (1985). Baum's Forward-Backward Algorithm Revisited. *Pattern Recognition Letters*, *3*(6), 369–373. https://doi.org/10.1016/0167-8655(85)90023-6

Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E., & Zeckhauser, R. (2018). The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament. *International Studies Quarterly*, *62*(2), 410–422. https://doi.org/10.1093/isq/sqx078

Lieder, F., Griffiths, T., & Goodman, N. (2012). Burn-in, Bias, and the Rationality of Anchoring. *Advances in Neural Information Processing Systems*, *25*. https://papers.nips.cc/paper_files/paper/2012/hash/81e5f81db77c596492e6f1a5a792ed53-Abstract.html

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction* (First edition). Crown Publishers.

,