How accurate are forecasts on geopolitical events from human collectives? Evidence from a real-money prediction market

Oliver Strijbis
Franklin University Switzerland and University of Zurich
ostrijbis@fus.edu

Bernd Bucher
Franklin University Switzerland

Elisa Volpi
Franklin University Switzerland

Abstract

Accurate forecasts of geopolitical events are essential for security, foreign, and macroeconomic policy. Among human-based forecasting methods, predictions of collectives have established themselves as particularly accurate and useful. In particular, prediction polls and prediction markets have become well-studied and established methodologies. This article evaluates the discrimination and calibration of a prediction market on geopolitical events conducted in 2023 and 2024. It makes two contributions to the literature. First, it is the first article to provide evidence of the forecasting accuracy of a real-money prediction market on geopolitical events. Second, it provides one of the first comparisons of a prediction market's forecasting accuracy with those of prediction polls for geopolitical events. This way, it contributes to a still small but growing literature that tries to establish the conditions under which prediction polls or prediction markets generate more accurate forecasts.

Acknowledgments: For helpful comments, we would like to thank the participants of the Energizing Breakfast Seminar at Franklin University Switzerland, the CCEW SYMPOSIUM 2023 at Bundeswehr University Munich, and the Empirical Methods Workshop of the Annual Congress of the Swiss Political Science Association 2024. Special thanks for helpful comments go to Hanna Frank. We would also like the collaborators of the PREMIA project Marc Wildi and Carolin Strobl.

Introduction

On January 1, 2022, *The Economist* summarized the forecasts of the prediction markets for the coming year. One prediction that many observers in the West would have considered to be too high, concerned the probability that Russia would invade Ukraine. It was 43% – close to the probability of picking the right side in a coin toss. This is a high probability for the rare event of a war onset. Unfortunately, the high probability assigned to the scenario of the Russian invasion turned out to be realistic.

While a single forecast does not provide evidence about the accuracy of a forecasting method – we can impossibly know if the real probability (i.e. the "ground truth") was around 43% at the time – the fact that prediction markets gave it a higher probability than other forecasting methods might not be a coincidence. As it has been demonstrated before, prediction markets are a highly accurate forecasting method for some types of political events such as elections (Berg, Nelson, and Rietz 2003; Berg, Nelson, and Rietz 2008; Berg et al. 2008; Graefe 2017). However, it remains more questioned how accurate prediction markets are in forecasting geopolitics. So far, prediction markets for geopolitical events have almost exclusively been investigated in the context of a forecasting tournament (see the Intelligence Advanced Research Projects Activity - IARPA).[1] In the context of this tournament, prediction markets have proven to be less accurate than forecasts from optimized prediction polls. Hence, it remains an open question whether prediction markets are generally able to generate accurate forecasts on geopolitical events and more so than prediction polls.

This article describes the forecasting accuracy of a prediction market conducted in 2023 and 2024 and so doing provides the first analysis on the forecasting accuracy of a prediction markets for geopolitical events in another context than the IARPA-tournament (Atanasov et al. 2017; Dana et al. 2019; Powell et al. 2013; Goldstein et al. 2015). The article makes two contributions to the literature. First, we provide for the first time evidence of the forecasting accuracy of a real-money prediction market on geopolitical events. Second, we offer a comparison of its forecasting accuracy with those of prediction polls. So far, only very few articles have compared the forecasting accuracy between prediction polls and prediction markets. Hence, the article contributes to a still small but growing literature that tries to establish the conditions under which prediction polls or prediction markets generate more or less accurate forecasts.

[1] The only partial exception we are aware of are the "Saddam Securities" (Wolfers and Zitzewitz 2009). However, their analysis focuses only on one question (and three time points) and can therefore not be understood as an analysis of geopolitical events. Similarly, the many studies that focus on US presidential elections are excluded because they also focus only on one specific type of event.

The results from the evaluation of the prediction market on geopolitical events are encouraging. The prediction market was well-calibrated as the average probability forecast came close to the real base rate according to which the events finally occurred. At the same time, the forecasts were strongly discriminating the probabilities according to which the events were forecasted to happen or not. Finally, although the analysis is far from conclusive, the comparison with prediction polls also suggests that if prediction markets are designed and conducted properly, they might on average be more accurate than prediction polls. Our result contrasts with that of Atanasov et al. (2017) and Dana et al. (2019), who found prediction polls to outperform prediction markets when forecasting geopolitical events. Hence, in the conclusion we speculate about potential reasons for the diverging results and point to avenues of further research.

Prediction markets as aggregators

For most social scientists, making predictions is a way to test theories by examining how well they hold up against real-world events. The goal is to refine or revise these explanations based on evidence. However, forecasting is more about providing useful information to guide decision-making. Predictions are valuable when they accurately reflect what is likely to happen. This article evaluates the accuracy of one specific forecasting method that uses group input: prediction markets.

Prediction markets, as defined by Berg, Nelson, and Rietz (2003), are Internet-based financial markets designed to use the information contained in market prices to make predictions about certain future events. With the aim of bringing to light the best collective prediction, the participants buy and sell (henceforth 'trade') their expectations regarding specific events, called 'contracts,' that will occur in the future. The values of traded contracts depend directly on future events and therefore the prices of these contracts provide information about the outcome of these events.

The backbone of prediction markets (as for any market) is the efficient market hypothesis (EMH), which states that prices reflect all information (Fama 1970; Hayek 1945). We do not assume that this assumption perfectly holds on our prediction market on geopolitical events, nor any other market. Markets tend towards information inefficiencies (Grossman and Stiglitz 1980) and potential 'misvaluation' (Hirshleifer 2001) as a consequence of limited rationality (Simon 1978) among individuals[2]. We *do* assume, however, that the price mechanism is

[2] A specific advantage of prediction markets relative to other markets is that since the market's time horizon is typically rather short, even in the case of inefficient markets, the probability of 'speculation bubbles' is small.

efficient enough in aggregating even highly decentralized information to be superior to most other aggregation mechanisms such as the calculation of some type of average of individual forecasts. Also, we assume that participants in prediction markets arrive at their expectations independently and this reduces systematic bias through avoidance of interpersonal dynamics such as "groupthink" (Janis 1972) and reliance on "opinion leaders" (Lazarsfeld, Berelson, and Gaudet 1948) that might bias aggregation.

If the prediction market aims to forecast the likelihood of an event, a probability market needs to be applied[3]. In probability markets, the final values of the contracts are defined as 100 if the event materializes and 0 if it does not. Assuming risk neutral utility maximizers, this translates into probabilities. The reason is that these economically rational participants trade shares based on the probabilities they attach to events. For example, suppose a rational trader on the market thinks that the likelihood of an outcome is 60% while the share's current price is at 50. In that case, she buys as many shares until the price is at 60 because, until this value, the expected value (EV) for her return on investment is above 0[4].

Varying accuracy of prediction markets

Prediction markets have proven themselves in practice specifically in the forecasting of elections in the USA (Berg, Nelson, and Rietz 2003; Berg, Nelson, and Rietz 2008; Berg et al. 2008). They have also been widely used to forecast election outcomes in Europe, where they have outperformed other forecasting models based on polls, expert panels, and economic indicators (Graefe 2017) [5]. Finally, prediction markets have proven to be quite accurate even for hard to forecast events such as the outcome of replication studies (Gordon et al. 2021) or migration movements (Morgenstern and Strijbis 2024).

[3] An alternative to probability markets are 'index markets', which can directly generate point estimates (Snowberg, Wolfers, and Zitzewitz 2013). Those markets, however, have the disadvantage that they do not come with information on the uncertainty of the forecast and are particularly prone to bias introduced by market scoring rules (Arnesen and Strijbis 2015; Arnesen and Strijbis 2015).

[4] This only holds true if the trader has enough capital to conduct the trades and no other contracts deviate (in the eyes of the trader) more from the probability of the event taking place.

[5] It is sometimes argued that prediction markets have a comparative advantage for forecasts in the «long run» (Berg, Nelson, and Rietz 2008). However, this is only true to the degree that the traders are not strongly discounting for payout of trades with regards to events in the distant future relative to those in the near future. Otherwise, markets for events in the distant future suffer from inefficient markets (Page and Clemen 2013). In order to limit this negative effect of the discount rate we only offered 'contracts' that would settle on the same date in the not so near future (eight monthts).

While prediction markets generally allow for accurate forecasts, forecasting accuracy can vary in important ways even for the same type of events that are to be forecasted (Strijbis and Arnesen 2019). A much-discussed problem of prediction markets, for example, is 'thin trading' (Pennock and Sami 2007). Thin trading results from a lack of matching buy and sell offers. Applied to our case, this situation would occur if the participant preferred a certain trade given their expectations on geopolitical events but doesn't get the opportunity to follow their plans due to the market situation. As a result, no market price is generated, and therefore no forecast can be derived. Large samples of participants can prevent the probability for thin trading. However, automated price makers may ensure infinite liquidity (Hanson 2003; Hanson 2007; Othman et al. 2013) using an algorithm that offers a new price for the expectations on the likelihood of an event after each transaction and prevents a situation where the participant is not willing to accept the price. In our application, we use the Logarithmic Market Scoring Rule (LMSR) developed by Hanson (2003) known for least systemic bias introduction in forecasts (Dudík et al. 2017) [6].

The 'wisdom of crowds' literature indicates that larger samples reduce prediction errors (Galton 1907; Surowiecki 2005). Dudík et al. (2017) also show with simulations for prediction markets with market scoring rules that under certain conditions the discrepancy between market clearing prices and ground truth goes to zero as the population of traders increases. However, the extent to which the wisdom of crowds mechanism plays a role in prediction markets and how many participants need to participate in a prediction market to arrive at accurate forecasts is still an unresolved issue. While the wisdom of the crowd theory suggests that the number of participants should be considerable, some argue that market efficiency is achieved already with a handful of participants. For instance, Christiansen (2007) reported in a case study that prediction markets with more than 16 participants were well-calibrated and McHugh & Jackson (2012) found varying the number of participants in the prediction market has a minimal impact on its accuracy (also Gordon et al. 2021).

A significant factor that may influence prediction market accuracy is financial incentivization. Markets encourage participation and the honest disclosure of information through performance-based rewards. The simplest approach is to allow participants to invest real money. However,

[6] Due to the LMSR's logarithmic function, it becomes increasingly expensive to push the price further down from the midpoint towards the minimum and increasingly expensive to push the price further up from the midpoint towards the maximum. In a context of cash constraints, this could lead to the overpricing of contracts for which the expected final price is low and the underpricing of contracts for which the final price is expected to be high (Arnesen & Strijbis 2015).

real-money markets are prohibited by law in many countries or may be deemed ethically inappropriate. An alternative is to use play money. Somewhat counterintuitively, it has been demonstrated that markets using play money also remain accurate (Pennock et al. 2001). Play money might work sufficiently well as incentives because social esteem can be an important motivator (Qiu and Kumar 2017). However, albeit also prediction markets without financial incentives yield informative forecasts, research on the *relative* performance of play-money and real-money markets is limited and inconclusive (Graefe 2017). Although one study (Servan-Schreiber et al. 2004) reported no difference in accuracy between the two market types, two other studies (Rosenbloom and Notz 2006; Deimer and Poblete 2010) concluded that real-money markets are more accurate than play-money markets.

Prediction markets and prediction polls

An interesting alternative to prediction markets are prediction polls. Prediction polls are surveys among a group of individuals where they are asked about the probability they attach to certain events. The forecasts of the polls are then (weighted) averages of the individual forecasts of the crowd. Prediction polls are an almost direct translation of the theory of the wisdom of the crowds into the realm of forecasting. Similar as with prediction markets, prediction polls have become known among a larger public in the context of election forecasting. In citizen forecasting models, survey responses on the question who the citizen expects to win the election are aggregated to the level of prediction. These citizen forecasting models have often been proved to be highly accurate (see the reviews in Graefe 2014; Stegmaier, Jokinsky, and Lewis-Beck 2023; Lewis-Beck and Stegmaier 2016; Leiter et al. 2018).

Prediction polls for geopolitical events have become famous through the work of Philipp Tetlock. Tetlock has shown their relevance in two volumes. In the *Expert Political Judgement* (Tetlock 2009), he showed that forecasts from a large number of laypersons outperformed those of experts. However, in this book he rather emphasized the limitations of expert judgement than the accuracy of the crowd forecast. This is different in the second volume, *Superforecasting* (Tetlock and Gardner 2015), where he and his co-author emphasize the accuracy of the forecasts of prediction polls with laypersons that are particularly good in forecasting.

In *Superforecasting*, Tetlock and Gardner (2015, 209) include a comparison of prediction poll forecasts with prediction market forecasts and concluded that forecasts of prediction polls of forecasting teams outperform prediction market forecasts. The evidence stems from prediction polls and prediction markets conducted in the context of the IARPA-tournament. Astanasov et al. (2017; also see Dana et al. 2019) specifically present results from a comparison of prediction

polls and double auction prediction markets of the Good Judgement Project. Here more than 2,400 participants made forecasts on 261 events in a geopolitical prediction tournament. Importantly, forecasters were randomly assigned to either prediction markets or prediction polls and their forecasting accuracy was subsequently compared. As Astanasov et al. (2017) underscore, in both seasons of the tournament, prices from the prediction market were more accurate than the simple mean of forecasts from prediction polls. Only if they further processed the forecasts of the prediction polls did they outperform the markets.

The forecasts from prediction polls that were optimized according to sophisticated weighting procedures also outperformed two additional prediction markets that have been conducted in the tournament: The prediction market DAGGRE applied a combinatorial prediction market with Hanson's logarithmic market scoring rule (LMSR) (Powell et al. 2013). The market could count on more than 3,000 forecasters and was conducted over the course of 20 months. Furthermore, the Intelligence Community Prediction Market (ICPM), which also applied the logarithmic market scoring rule (LMSR), involved intelligence analysts (Goldstein et al. 2015), as well as more than 4,300 traders,.

Why did the optimized prediction polls outperform the prediction markets? The reason for this result is probably twofold: First, the organizers invested great effort with regards to the forecasting team, for which they e.g. organized meetings. Hence, the motivation the forecasting team members were probably higher than those of the individual forecasters assigned to the prediction market. Consequently, we do not know if the prediction market forecast would not also have been better if the prediction market participants were motivated in a similar way. Second, the prediction market might have been conducted in a suboptimal way. For instance, the prediction market set-up did not provide financial incentives. While the literature does not conclusively demonstrate that real money really improves forecasting accuracy (see above), the role of incentives cannot be discarded as a potential explanation for the performance of the prediction market.

A real money prediction market on geopolitical events

In the following, we describe the application of a real money prediction market to the forecasting of geopolitical events in 2023 and the first two quarters of 2024. By geopolitical events, we refer to political occurrences or developments with significant impact on international relations, often influencing or reshaping global power dynamics. These events can include political conflicts, changes in governments of powerful countries, wars, economic sanctions, trade negotiations, or alliances between nations. They might also include domestic

political events within powerful countries such as the US if these events are considered to have far-reaching effects on regional stability, global economies, or international diplomacy.[7]

We conducted three different cycles – each for a period of approximately five months. A forecasting cycle starts when (most) questions are uploaded to the market environment and the participants are informed that they can start trading. The cycle ends when all open questions are resolved and hence the market as such is resolved. During the prediction market cycle, the organizer can choose to add new questions and contracts. Additionally, any questions with outcomes known before the cycle ends are resolved during the cycle.

Table 1: Events and participation in three prediction market cycles

Cycle	Geopolitical events	Contract-sets (questions)	Contracts (answers)	Participants	Trades
Spring 2023	20	19	100	334	7797
Fall 2023	13	17	76	297	4246
Spring 2024	22	17	72	281	4331
Total	55	53	248	912	16374

Our prediction market on geopolitical events covered 55 geopolitical events, the forecasts of which we periodically published on our university's website (see the Supplemental Material for the list of events).[8] In order to cover the 55 geopolitical events, we uploaded 53 contract-sets with 248 contracts on the prediction market (see Table 1). Each contract-set can be understood as a question for which the probabilities of different answers (the contracts)sum up to 100% probability. All questions referred to specific events with the answers describing different scenarios. For instance, we asked about the probability for violent action and/or a sea blockade by China against Taiwan by July 31, 2023. We provided four different contracts on which the participants could bet: a) There would be violent action, but no maritime blockade; b) There would be a maritime blockade, but no violent action; c) There would be both violent action and

[7] It is true that "geopolitics" is rather elusive concept, the meaning of which has changed over time (Nickel 2024).

[8] The forecasts were published as one out of two indicators of the Franklin Political Risk and Opportunity Index (FRISKOP). FRISKOP is conducted by a team of professors at Franklin University Switzerland including the authors of this paper. The index and its methodology can be consulted here: https://www.fus.edu/research/FRISKOP. In the spring 2024 edition, we also included 13 contract-sets (questions) on the outcome of the European elections in Spain. These forecasts were published by the newspaper *El Periódico* and proved to be highly accurate – outperforming all polls. However, since this election cannot be categorized as a geopolitical event as defined above, we excluded these events from all analyses.

a maritime blockade; And d) there would be none of the two. Importantly, these events were also further defined to clarify what was meant by violent action and maritime blockade[9].

We selected the geopolitical events based on three considerations. First, we have included events that – according to an expert panel – would have a strong impact on the global economy. This was done because the forecasts were used for an index of geopolitical risks (see footnote 6). Second, the selection of events was inspired by what some practitioners in the reinsurance and defense sector thought was particularly relevant to them. Finally, in order to compare the forecasting accuracy of the prediction market with prediction polls, we have chosen some questions based on the availability of identical or very similar questions on the prediction poll platform Good Judgement Open (GJO) and Metaculus.

On January 11th, the first set of participants were tagged to the market and invited to participate. As some invitees had previously been recruited to participate prediction markets on elections in Germany, Spain or Switzerland, they were already familiar with trading on the prediction market, and only received a short description with information on the incentives and date of closure. More specifically, they were informed that they would receive 10 € of starting capital and that the market would close on July 31, 2023. For the subsequent two prediction markets in fall 2023 and spring 2024 the same participants were invited. We recruited another 160 participants before or during the fall 2023 market and another 18 in spring 2024. The former were overwhelmingly political science students from the University of Zurich while the latter were students from Franklin University Switzerland.

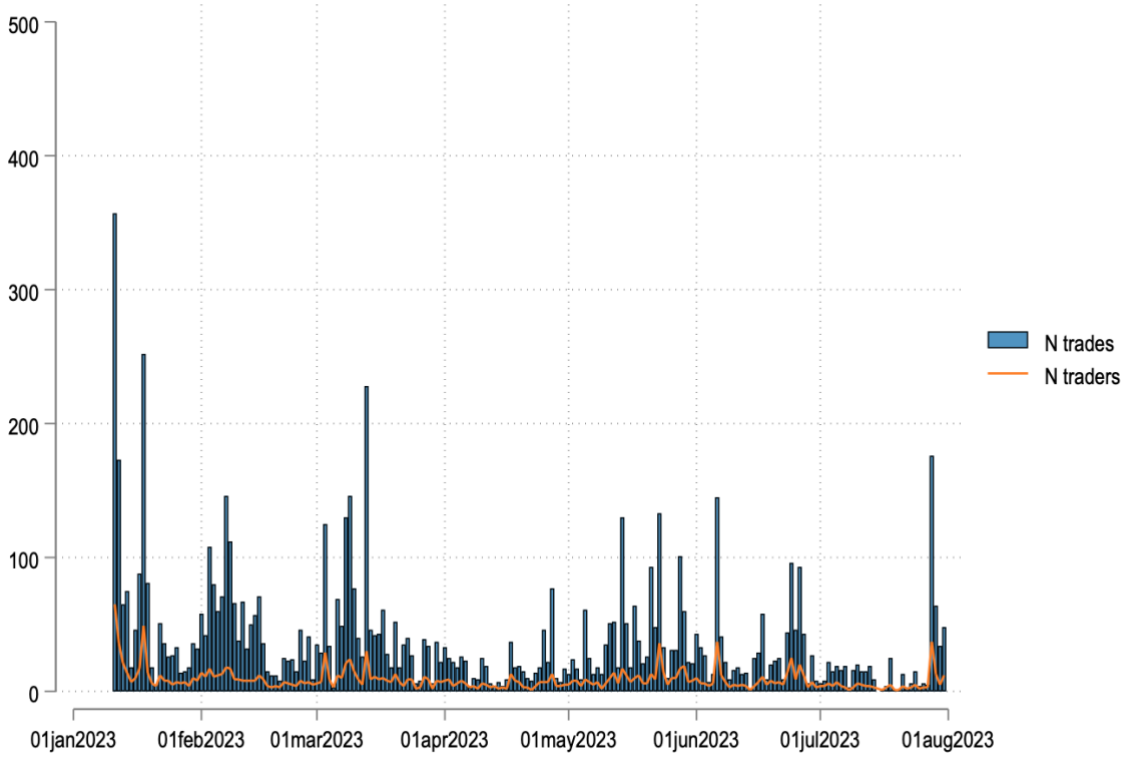
In each market between 281 and 334 participants traded at least once. In each of the prediction market cycles between 4246 and 7797 trades were made in total (see Table 1). Figure 1 shows the number of participants that conducted at least one trade per day and the total number of transactions performed. The trade activity peaked at the beginning of each project, but trades were conducted throughout the entire period with peaks at different points in time. The peaks took place when important events – like the Wagner mutiny in Russia – had important effects on the expectations of the participants or when we sent an Email to the participants to remind them about the market. In the spring 2023 edition the number of trades and active traders was quite well-spread over the whole cycle. In the fall 2024 version the activity peaked with the two

[9] In this specific case, violent action by China was defined as "[a]ny action by the Chinese military that results in the deaths of at least 25 Taiwanese military personnel and/or civilians." Seablockade was defined as "a situation in which China attempts to block Taiwan from accessing the sea or from conducting trade or other activities through the sea by using military or other means involving blockading ports, intercepting and inspecting ships, or using naval or air forces to patrol and enforce the blockade."

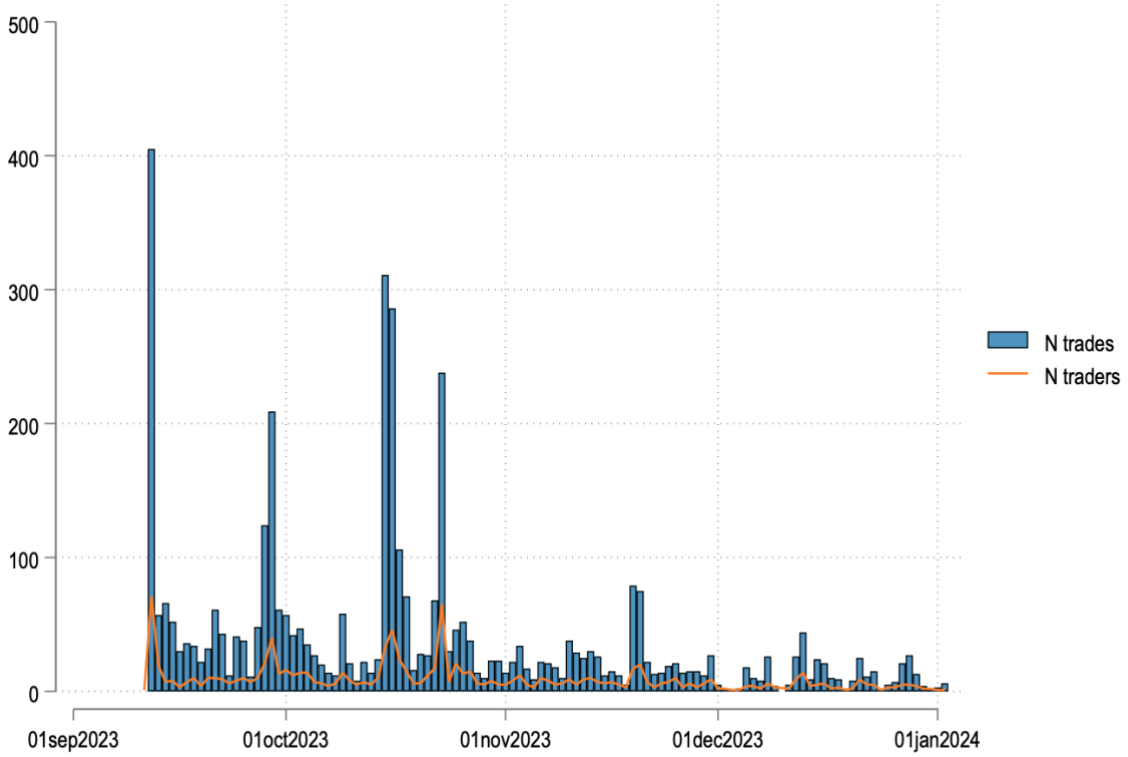
rounds of elections in Poland – a main topic of this market – but strongly decreased towards the end of the cycle. We cannot rule out that the low participation towards the end had negative effects on the forecasting accuracy. In the spring 2023 section trading on the geopolitical events was least frequent in April and May. This is because in this period the questions on the European elections in Spain were placed more prominently than the geopolitical events. However, after the European elections participation increased again.

Figure 1: Number of trades and trading participants

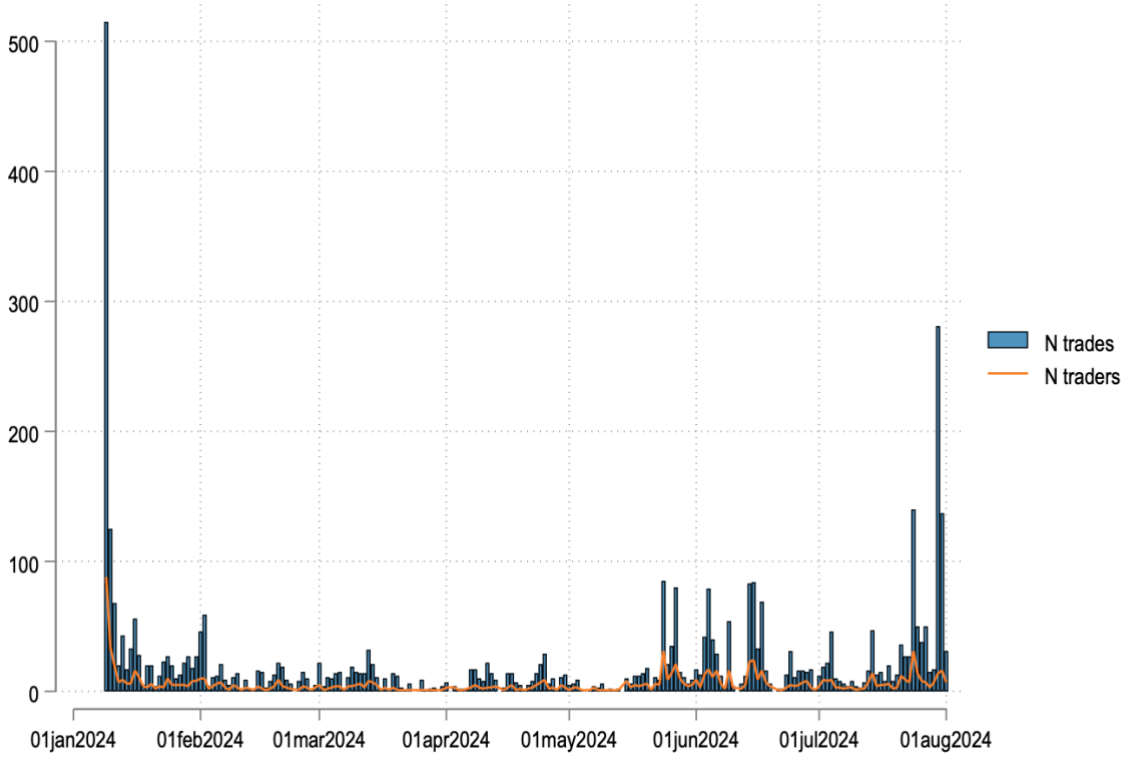
a) Spring 2023 prediction market



b) Fall 2023 prediction market



c) Spring 2024 prediction market



Prediction markets work along a monetary logic and financial incentives are consequently assumed to work towards participants revealing their true expectations. As mentioned above, the participants received 10 € starting capital with which they could trade for each of the three markets. In the spring 2023 edition, the best trader won 116.03 €, while the mean payout was 10.88 €. In fall 2023 the equivalents were 133.53 € and 11.90 €. And in the spring 2024 market the best forecaster gained 197.96 € with the average payout being 11.39 €.

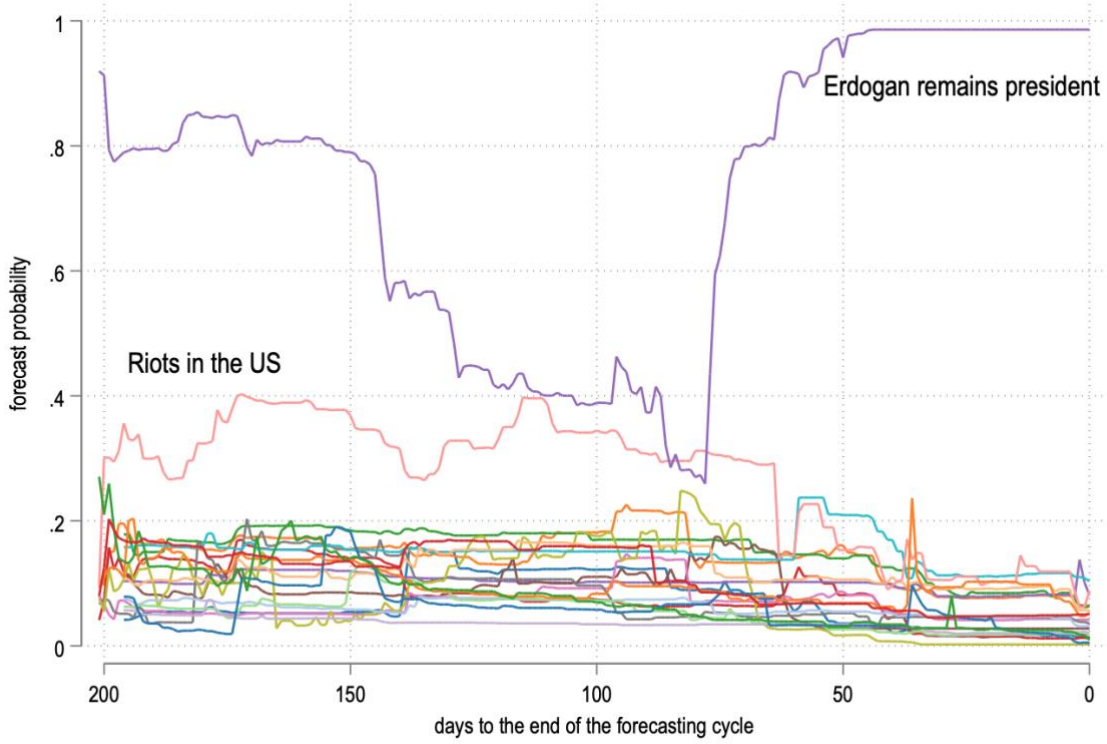
The participants were probably not only motivated by the financial incentives. We can also speculate that participants were motivated by social esteem since previous research has shown that on prediction markets without financial incentives this can be an important motivator (Qiu and Kumar 2017). This is also why our prediction market provided a ranking which made each participant's performance visible to the other participants. Additional motivations might have been curiosity and/or an interest in supporting research. Based on our design, it is impossible to know whether the prediction market would have been even more successful if stronger financial incentives could have been provided. This said, our mix of incentives allowed recruiting enough actively trading participants and to create market efficiency that resulted in informed forecasts over six months.

Evaluating the forecasting accuracy

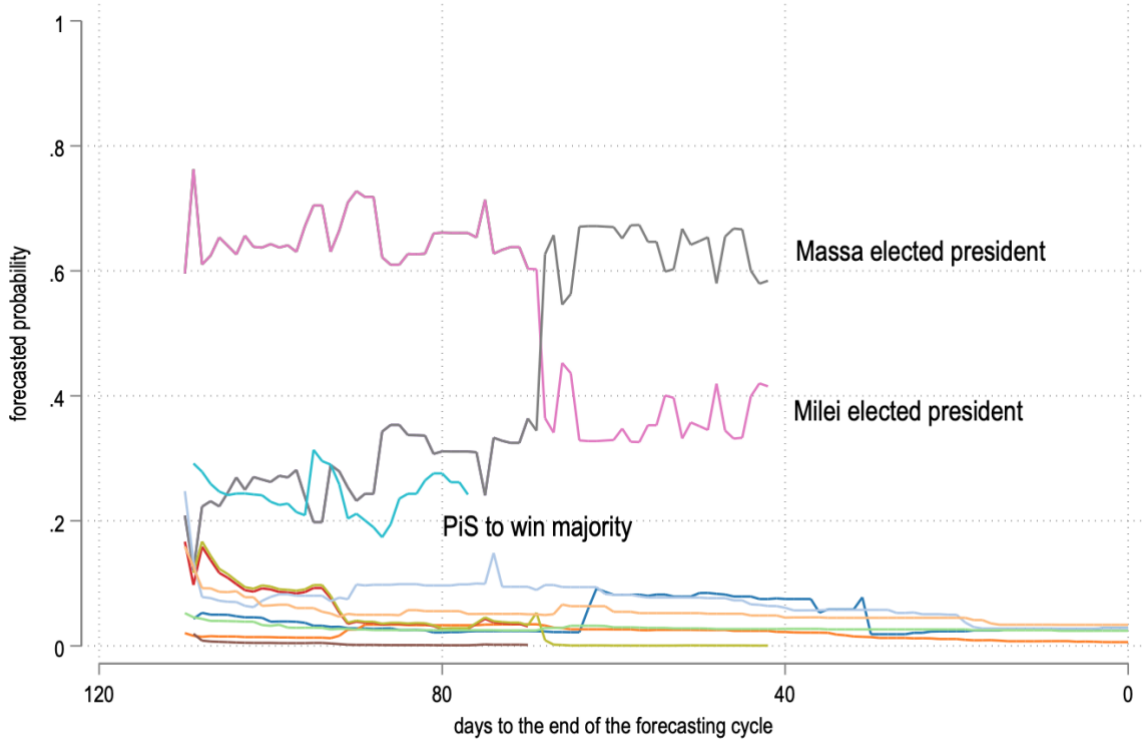
All forecasts on the prediction market were about the likelihood that an event would take place. Hence, the forecasts consisted of probability estimates regarding the likelihood according to which these events would occur. Since the real probability for events to happen before the time point for which the forecast is made is unknown, we can only investigate the accuracy of the prediction market method in the aggregate. For binary events, forecasts are typically considered to be accurate when they are good in discrimination, i.e. give specific events probabilities that are close to 0% or 100%, and are well calibrated, i.e. assign types of events the probability of happening as they do on average (Tetlock 2009). Hence, we will investigate the forecasting accuracy of the prediction market on both dimensions. Finally, we will also compare its forecasting accuracy relative to 25 very similar or identical forecasts from prediction polls.

Figure 2: Predicted probabilities of prediction market forecasts over time

a) Spring 2023 prediction market



b) Fall 2023 prediction market



c) Spring 2024 prediction market

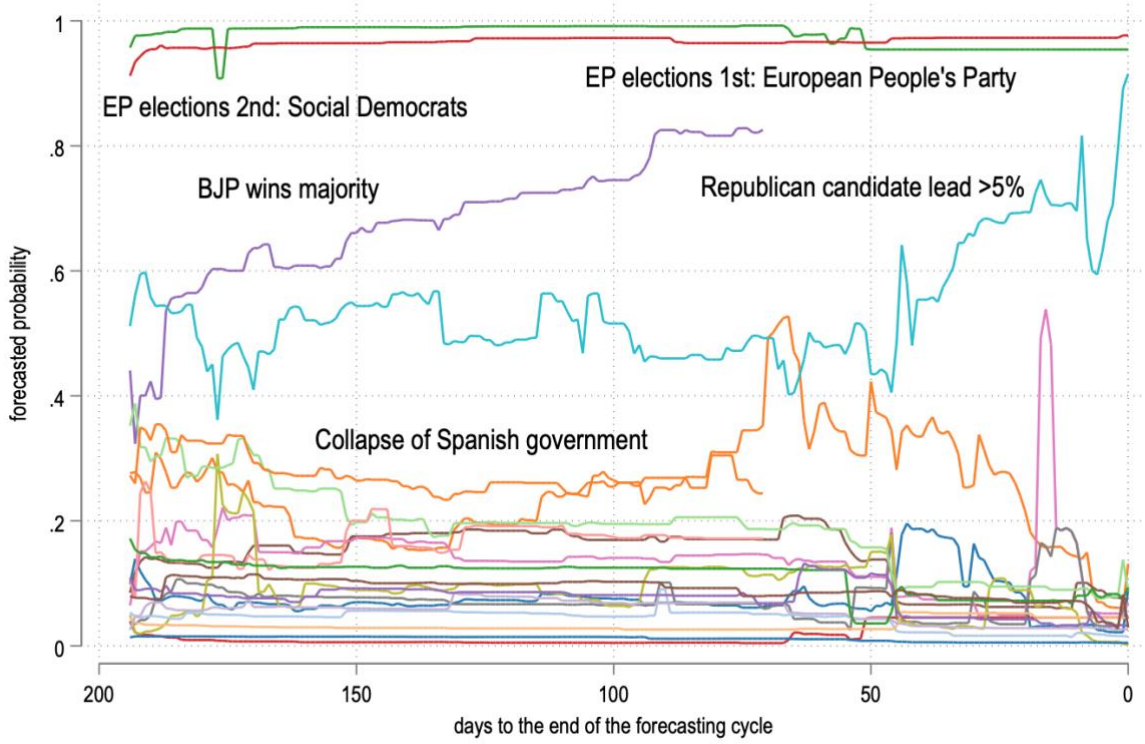


Figure 2 describes the forecasts of the 55 events over time divided by the three forecasting cycles. The Y-axis describes the probability as revealed in the market price on the prediction market, while the X-axis shows the time point at which the prediction was made relative to the end of the forecasting cycle. For most of the questions the timeline was the last day of the forecasting cycle (July or December 31). In the cases of elections (Argentina, EU, India, Mexico, and Poland), the event was decided with the election result. This explains why a few forecasts end before the end of the forecasting cycle.

Most forecasts have attached a very low probability to an event happening and have typically remained very low over time. This is because the forecasts focused mostly on "political risks", i.e. rather unlikely events that would have a high economic impact if they would take place. Second, there was a smaller group of forecasts on events for which their probability to take place was estimated to be very high throughout the whole period. These events often consisted of scenarios according to which the status quo would remain unchanged. Importantly, for both types of events the forecasts would in most cases rather slowly trend towards 0% and 100% respectively. This indicates that the prediction market adjusted to the fact that the shorter the time horizon of the forecast, the smaller the probability that the event would (not) happen.

Finally, there is a small group of forecasts for which the estimated probability varied between 20% and 80%. These forecasts do not strongly discriminate, which is problematic if they are not well calibrated. In other words, these are only good forecasts if – on average – these kinds of events did happen only between 20% and 80% of the time.

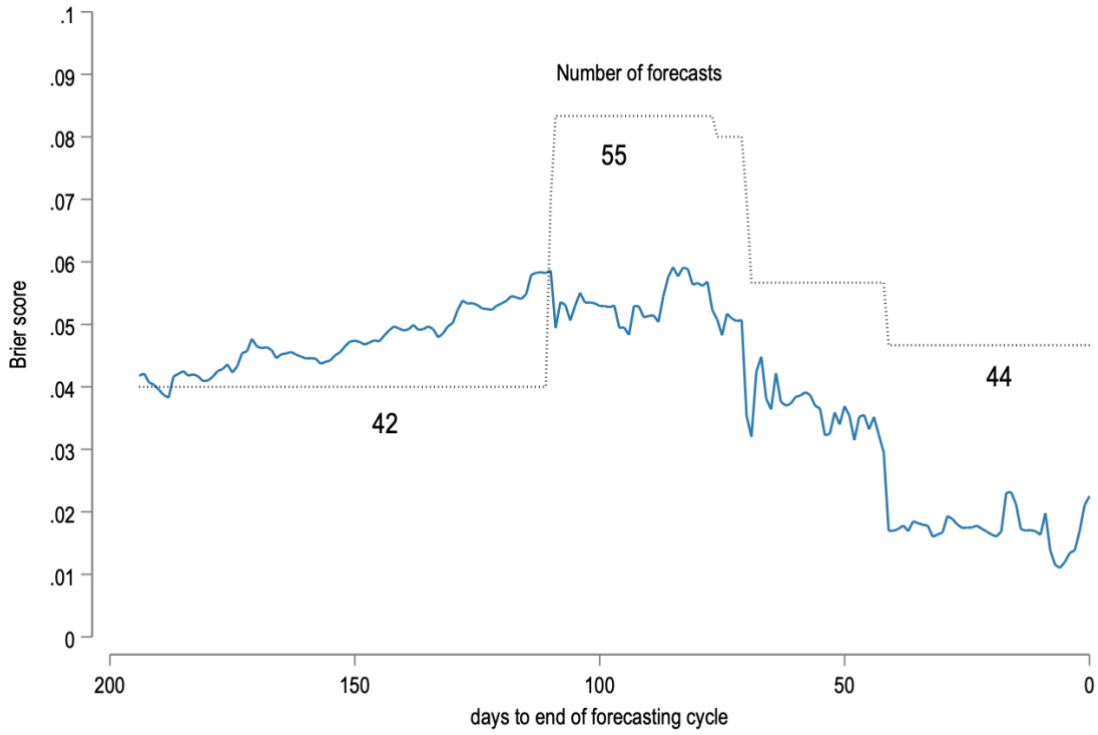
Figure 3 shows the overall accuracy of the forecasts based on the Brier score (Brier 1950). The Brier Score is a metric used to evaluate the accuracy of probabilistic predictions. It measures the mean squared difference between the predicted probability assigned to possible outcomes and the actual outcome (coded as 0 or 1). The score is commonly used in binary classification but can also be extended to multi-class classification problems. For binary classification, the Brier score is calculated as:

$$Brier\ Score = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2$$

where *N* is the number of predictors, *f* the predicted probability for the positive class, and *o* the actual outcome (1 if the event occurred, 0 otherwise).

Figure 3 shows generally high accuracy with Brier score of between 0.04 and 0.06. Surprisingly, the Brier score does not improve over the first 130 days but rather worsens. To an important extent this is because the forecast regarding the future of the Turkish president Erdogan went in the oppositive direction of the outcome (the probability of Erdogan losing the election was considered to increase, while he eventually remained in office) and the same was true for the probability of the House of Representatives to increase the debt ceiling (which the ultimately did). Also, in the period with 55 forecasts around 110 to 70 days before the end of the forecasting cycles the dataset includes a few elections. The outcome of these elections had far higher base rates than most of the other events. Apparently, it was more difficult for the forecasters to predict the outcome of these events, which is why the forecasting accuracy only improved once the outcomes of these events were clear and they consequently no longer were part of the dataset. This might point to the fact that elections are a different type of political event than most of the other events in our sample. Once these elections took place and the debt ceiling was raised, the forecasts very much aligned with the eventual outcomes and the Brier score consequently approximated 0.01.

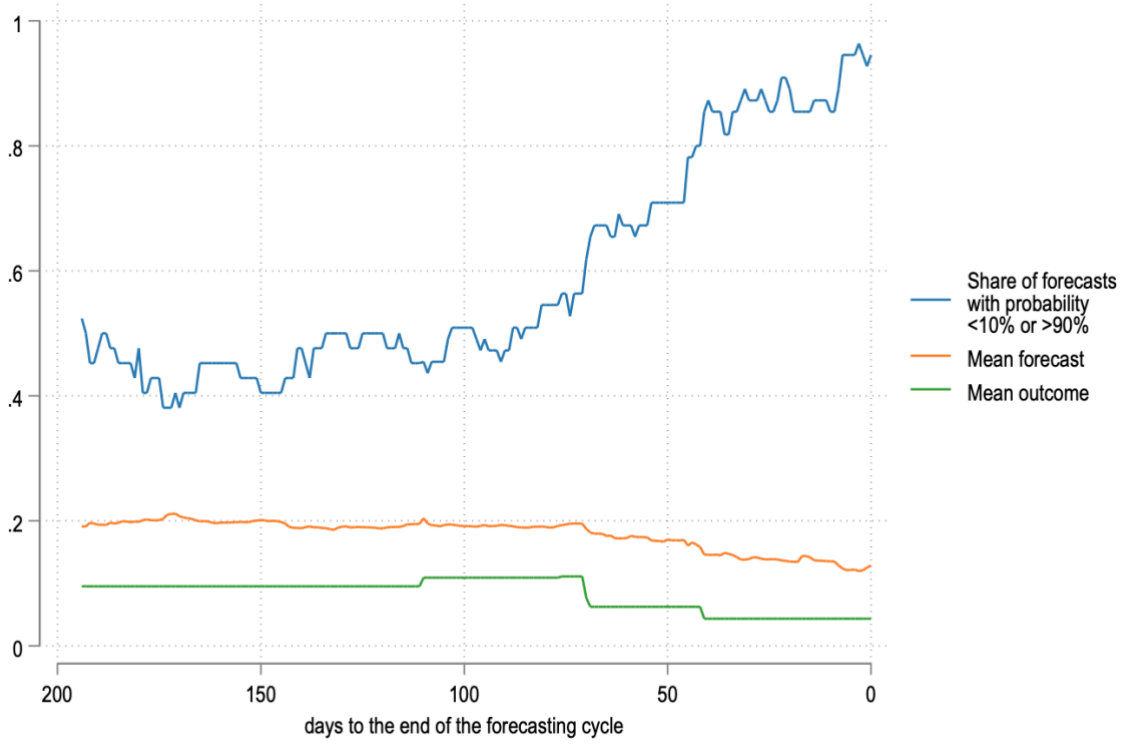
Figure 3: Brier score for 55 forecasts on geopolitical events



While Figure 3 describes the overall accuracy of the forecasts, Figure 4 informs with easily interpretable indicators how it did in terms of calibration and discrimination. In order to assess whether the forecasts discriminated strongly, we use a simple measure by calculating the share of events for which a higher (smaller) probability than 90% was forecasted. As the Figure shows, this was initially the case for about half of the cases. Only in roughly the middle of the forecasting period did discrimination increase and achieve a value above 90% at the end of the forecasting cycle.

To get a better sense of biases in calibration of the forecast, Figure 4 also shows the mean probability forecast and the base rate of the outcomes (i.e. the mean outcome). It shows that on average the forecasters overestimated the probability of events taking place by around 10%. Combining this with the evidence of a low Brier score (Figure 3), it shows that while the prediction market was good in attaching higher probabilities to more likely events than to less likely events, it systematically overestimated the base rate of the events.

Figure 3: Degree of discrimination and calibration of prediction market forecasts



In the final step of our analysis, we compare forecasts of the prediction markets with the two most established prediction polls – the Good Judgement Open (GJO) and Metaculus. Both prediction polls do not report simple forecasting averages, but averages that are weighted on the historical accuracy of the individual forecasters. Especially the GJO method has been developed by many of the proponents of public opinion polls cited in the third section of this paper. Table 2 shows forecasts for the 25 events for which we can compare the forecasts. As time points for comparison, we have chosen the first date after the launch of the prediction market at which we could easily access the forecast from the prediction polls. In order to not bias the comparison in favor of prediction markets, we chose a time point that was not only shortly after the launch of the question, but also when many forecasters on the prediction polls revealed their expectations. Somewhat arbitrarily, we have chosen January 31st for the 2023 cycle, September 15th for the fall 2023 market, and January 15th for the spring 2024 market. In some cases, like the forecast on the debt ceiling by Metaculus or the riots in the US by GJO, forecasts were only available at a later point in time.

The comparisons are not always perfect – as the wording of the questions and the time horizons of the forecast of geopolitical events were not always identical. The differences in the time horizons, however, do not bias systematically in favor of one or the other method: In four cases the difference in the time horizon was slightly in favor of the prediction polls and in five cases

in favor of the prediction market. The comparison is also not perfect because some events are correlated with each other. The correlation is most obvious for the elections in Argentina where the election outcome for the different candidates were directly related to each other.

While the comparison is imperfect, it nevertheless allows a prudent assessment of the forecasting accuracy of the prediction market relative to prediction polls. Table 2 shows that both the prediction market and the prediction polls made more or less accurate forecasts depending on the question (R = 0.86). On average, however, the prediction market was clearly more accurate than the prediction polls. This is demonstrated by the Brier score that is only 0.046 for the prediction market but 0.073 for the prediction polls. Regarding the fact that the comparison is imperfect, we should not read too much into this finding. But the analysis does suggest that if prediction markets are set up accurately such as in our case, they are at least if not more accurate in forecasting geopolitical events than prediction polls.

Conclusion

To our knowledge, this article provides the first assessment of the forecasting accuracy of a real money prediction market on geopolitical events. Hence, this article contributes to recent studies that compare the performance of prediction markets without monetary incentives and prediction polls. The evaluation has shown that real money prediction markets can provide forecasts on geopolitical events that do well both in terms of discrimination and calibration. This has been shown when describing the forecasting accuracy of the prediction market applied to 55 geopolitical events and the comparison with 25 events for which forecasts of prediction polls were also available.

While the presented evidence supports the argument that prediction markets constitute one of the most if not the most accurate forecasting method to predict rare and to an important extent idiosyncratic events, the analysis presented in this article is not entirely conclusive. Subsequent analyses should complement the data presented here and, in this way, increase the number of cases analyzed. This is important because among the events analyzed here many had base rates that were rather extreme. And those that were not extreme were often directly related to elections, which might be considered a very specific kind of political event. It therefore might be a problem that we mix the analysis of elections and other events and in the future – with a larger sample – systematic differences between election forecast and other geopolitical forecasts should be taken into account.

Hence, it remains still somewhat unclear how well the prediction market does for events that have base rates closer to 50% probability. Second, in this article we were only able to compare

the prediction market forecasts with prediction polls regarding 25 events. This is obviously too few to develop strong general statements about their respective accuracy. Hence, subsequent applications of our prediction market should specifically include events which are also run by prediction polls, and the resulting data should be added to what has been presented here.

Table 2: Prediction market and prediction poll forecasts compared

Event (probability)	Date of forecast	Timeline poll	Timeline market	Outcome	Market forecast	Absolute error	Poll forecast	Absolute error	Source poll
Riots in the US	31.01.23	15.04.23	31.07.23	0	0.29	0.29	0.19	0.19	Metaculus
Khamenei either flee Iran or cease to be its supreme leader	31.01.23	15.04.23	31.07.23	0	0.16	0.16	0.04	0.04	GJO
Russia and Ukraine announce a ceasefire	31.01.23	05.05.23	31.07.23	0	0.12	0.12	0.11	0.11	GJO
Erdogan reelected[1]	31.01.23	28.05.23	31.07.23	1	0.85	0.15	0.68	0.32	Metaculus
US debt ceiling raised	31.01.23	31.08.23	31.07.23	0	0.12	0.12	0.03	0.03	GJO
Giorgia Meloni cease to be the prime minister	31.01.23	01.10.23	31.07.23	0	0.17	0.17	0.21	0.21	GJO
Lethal confrontation between China and Taiwan	31.01.23	01.10.23	31.07.23	0	0.05	0.05	0.13	0.13	GJO
Russia detonate a nuclear device in Ukraine	31.01.23	01.10.23	31.07.23	0	0.16	0.16	0.10	0.10	GJO
Vladimir Putin cease to be the president	31.01.23	01.10.23	31.07.23	0	0.12	0.12	0.20	0.20	GJO
US default	26.04.23	31.07.23	31.07.23	0	0.08	0.08	0.09	0.09	Metaculus
Majority of the Law and Justice party (PiS)	15.09.23	15.10.23	15.10.23	0	0.25	0.25	0.13	0.13	GJO
President of Argentina: Milei	15.09.23	22.10.23	22.10.23	1	0.65	0.35	0.57	0.43	GJO
President of Argentina: Bullrich	15.09.23	22.10.23	22.10.23	0	0.12	0.12	0.27	0.27	GJO
President of Argentina: Massa	15.09.23	22.10.23	22.10.23	0	0.22	0.22	0.26	0.26	GJO
President of Argentina: Others	15.09.23	22.10.23	22.10.23	0	0.01	0.01	0.01	0.01	GJO
Armed conflict between China and Taiwan	15.09.23	31.12.23	31.12.23	0	0.05	0.05	0.05	0.05	Metaculus
US to ban TikTok	15.09.23	31.12.23	31.12.23	0	0.02	0.02	0.05	0.05	GJO
Cease-fire or peace agreement between Russian and Ukraine	15.09.23	31.12.23	31.12.23	0	0.07	0.07	0.03	0.03	Metaculus
Vladimir Putin cease to be the president	15.09.23	31.12.23	31.12.23	0	0.09	0.09	0.05	0.05	Metaculus
Will Russia use nuclear weapons against Ukraine[2]	15.09.23	31.12.23	31.12.23	0	0.04	0.04	0.01	0.01	Metaculus
President of Argentina: Milei	25.10.23	19.11.23	19.11.23	1	0.45	0.55	0.23	0.77	GJO
President of Argentina: Massa	25.10.23	19.11.23	19.11.23	0	0.55	0.55	0.67	0.67	GJO
Largest political group after EP elections: EPP	15.01.24	09.06.24	09.06.24	1	0.98	0.02	0.75	0.25	Metaculus
Largest political group after EP elections: S&D	15.01.24	09.06.24	09.06.24	0	0.02	0.02	0.17	0.17	Metaculus
Mean absolute error (MAE)						0.16		0.19	

Brier score	0.05	0.07

Notes: [1]In the prediction market the question was whether Erdogan would remain president. This explains the different timeline; [2]In the Metaculus forecast this was also made dependent on the US giving Russia fighter aircraft, which made it less likely.

References

Arnesen, Sveinung, and Oliver Strijbis. 2015. "Accuracy and Bias in European Prediction Markets." *Italian Journal of Applied Statistics* 2 (25). Citeseer: 123–138.

Atanasov, Pavel, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip E. Tetlock, Lyle Ungar, and Barbara Mellers. 2017. "Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls." *Management Science* 63 (3). INFORMS: 691–706. doi:10.1287/mnsc.2015.2374.

Berg, Joyce, Robert Forsythe, Forrest Nelson, and Thomas Rietz. 2008. "Results from a Dozen Years of Election Futures Markets Research." In *Handbook of Experimental Economics Results*, edited by Charles R. Plott and Vernon L. Smith, 1:742–751. Elsevier. doi:10.1016/S1574-0722(07)00080-7.

Berg, Joyce, Forrest Nelson, and Thomas Rietz. 2003. "Accuracy and Forecast Standard Error of Prediction Markets." *Tippie College of Business Administration, University of Iowa*.

Berg, Joyce, Forrest Nelson, and Thomas Rietz. 2008. "Prediction Market Accuracy in the Long Run." *International Journal of Forecasting*, US Presidential Election Forecasting, 24 (2): 285–300. doi:10.1016/j.ijforecast.2008.03.007.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

Christiansen, Jed D. 2007. "Prediction Markets: Practical Experiments in Small Markets and Behaviours Observed." *The Journal of Prediction Markets* 1 (1): 17–41. doi:10.5750/jpm.v1i1.418.

Dana, Jason, Pavel Atanasov, Philip E. Tetlock, and Barbara Mellers. 2019. "Are Markets More Accurate than Polls? The Surprising Informational Value of 'Just Asking.'" *Judgment and Decision Making* 14 (2). Cambridge University Press: 135–147. doi:10.1017/S1930297500003375.

Deimer, Sebastian, and Joaquin Poblete. 2010. "Real-Money vs. Play-Money; Forecasting Accuracy in Online Prediction Markets – Empirical Insights from iPredict." *The Journal of Prediction Markets* 4 (3): 21–58. doi:10.5750/jpm.v4i3.479.

Dudík, Miro, Sébastien Lahaie, Ryan Rogers, and Jennifer Wortman Vaughan. 2017. "A Decomposition of Forecast Error in Prediction Markets." In 31st Conference on Neural Information Processing Systems (NIPS 2017). https://www.microsoft.com/en-us/research/publication/a-decomposition-of-forecast-error-in-prediction-markets/.

Fama, Eugene F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance* 25 (2). [American Finance Association, Wiley]: 383–417. doi:10.2307/2325486.

Galton, Francis. 1907. "Vox Populi (the Wisdom of Crowds)." *Nature* 75 (7): 450–451.

Goldstein, Seth, Rob Hartman, Ethan Comstock, and Thalia Samash Baumgarten. 2015. "Assessing the Accuracy of Geopolitical Forecasts from the US Intelligence Community's Prediction Market." The MITRE Corporation. https://goodjudgment.com/wp-content/uploads/2020/11/Goldstein-et-al-GJP-vs-ICPM.pdf.

Gordon, Michael, Domenico Viganola, Anna Dreber, Magnus Johannesson, and Thomas Pfeiffer. 2021. "Predicting Replicability—Analysis of Survey and Prediction Market Data from Large-Scale Forecasting Projects." *PLOS ONE* 16 (4). Public Library of Science: e0248780. doi:10.1371/journal.pone.0248780.

Graefe, Andreas. 2014. "Accuracy of Vote Expectation Surveys in Forecasting Elections." *Public Opinion Quarterly* 78 (S1): 204–232. doi:10.1093/poq/nfu008.

Graefe, Andreas. 2017. "Political Markets." Edited by Kai Arzheimer, Jocelyn Evans, and Michael S. Lewis-Beck. *The Sage Handbook of Electoral Behavior. Los Angeles: Sage. FC.*

Grossman, Sanford J., and Joseph E. Stiglitz. 1980. "On the Impossibility of Informationally

Efficient Markets." *The American Economic Review* 70 (3). American Economic Association: 393–408.
Hanson, Robin. 2003. "Combinatorial Information Market Design." *Information Systems Frontiers* 5 (1): 107–119. doi:10.1023/A:1022058209073.
Hanson, Robin. 2007. "Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation." *The Journal of Prediction Markets* 1 (1): 3–15. doi:10.5750/jpm.v1i1.417.
Hayek, F. A. 1945. "The Use of Knowledge in Society." *The American Economic Review* 35 (4). American Economic Association: 519–530.
Hirshleifer, David. 2001. "Investor Psychology and Asset Pricing." *The Journal of Finance* 56 (4): 1533–1597. doi:https://doi.org/10.1111/0022-1082.00379.
Janis, Irving Lester. 1972. *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton, Mifflin.
Lazarsfeld, Paul F., Bernard Berelson, and Hazel Gaudet. 1948. *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. Columbia University Press.
Leiter, Debra, Andreas Murr, Ericka Rascón Ramírez, and Mary Stegmaier. 2018. "Social Networks and Citizen Election Forecasting: The More Friends the Better." *International Journal of Forecasting* 34 (2): 235–248. doi:10.1016/j.ijforecast.2017.11.006.
Lewis-Beck, Michael S., and Mary Stegmaier. 2016. "Election Forecasting: Scientific Approaches." In *Encyclopedia of Social Network Analysis and Mining*, edited by R Alhajj and J Rokne, 1–8. New York, New York, NY: Springer.
McHugh, Patrick, and Aaron L. Jackson. 2012. "Prediction Market Accuracy: The Impact of Size, Incentives, Context, and Interpretation." *The Journal of Prediction Markets* 6 (2): 22–46. doi:10.5750/jpm.v6i2.500.
Morgenstern, Sandra, and Oliver Strijbis. 2024. "Forecasting Migration Movements Using Prediction Markets." *Comparative Migration Studies* 12 (45): 1–8. doi:10.1186/s40878-024-00404-0.
Nickel, Carsten. 2024. "What Do We Talk about When We Talk about the 'Return' of Geopolitics?" *International Affairs* 100 (1): 221–239. doi:10.1093/ia/iiad295.
Othman, Abraham, David M. Pennock, Daniel M. Reeves, and Tuomas Sandholm. 2013. "A Practical Liquidity-Sensitive Automated Market Maker." *ACM Transactions on Economics and Computation* 1 (3): 14:1-14:25. doi:10.1145/2509413.2509414.
Page, Lionel, and Robert T. Clemen. 2013. "Do Prediction Markets Produce Well-Calibrated Probability Forecasts?" *The Economic Journal* 123 (568). Oxford University Press Oxford, UK: 491–513.
Pennock, David M., Steve Lawrence, C. Lee Giles, and Finn Årup Nielsen. 2001. "The Real Power of Artificial Markets." *Science* 291 (5506). American Association for the Advancement of Science: 987–988. doi:10.1126/science.291.5506.987.
Pennock, David M., and Rahul Sami. 2007. "Computational Aspects of Prediction Markets." In *Algorithmic Game Theory*, edited by Eva Tardos, Noam Nisan, Tim Roughgarden, and Vijay V. Vazirani, 651–676. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511800481.028.
Powell, Walter A., Robin Hanson, Kathryn B. Laskey, and Charles Twardy. 2013. "Combinatorial Prediction Markets: An Experimental Study." In *Scalable Uncertainty Management*, edited by Weiru Liu, V. S. Subrahmanian, and Jef Wijsen, 8078:283–296. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-40381-1_22.
Qiu, Liangfei, and Subodha Kumar. 2017. "Understanding Voluntary Knowledge Provision and Content Contribution Through a Social-Media-Based Prediction Market: A Field Experiment." *Information Systems Research* 28 (3). INFORMS: 529–546. doi:10.1287/isre.2016.0679.

Rosenbloom, E. S., and William Notz. 2006. “Statistical Tests of Real-Money versus Play-Money Prediction Markets.” *Electronic Markets* 16 (1). Routledge: 63–69. doi:10.1080/10196780500491303.

Servan-Schreiber, Emile, Justin Wolfers, David M. Pennock, and Brian Galebach. 2004. “Prediction Markets: Does Money Matter?” *Electronic Markets* 14 (3). Routledge: 243–251. doi:10.1080/1019678042000245254.

Simon, Herbert A. 1978. “Rationality as Process and as Product of Thought.” *The American Economic Review* 68 (2). American Economic Association: 1–16.

Snowberg, Erik, Justin Wolfers, and Eric Zitzewitz. 2013. “Chapter 11 - Prediction Markets for Economic Forecasting.” In *Handbook of Economic Forecasting*, edited by Graham Elliott and Allan Timmermann, 2:657–687. Handbook of Economic Forecasting. Elsevier. doi:10.1016/B978-0-444-53683-9.00011-6.

Stegmaier, Mary, Steven Jokinsky, and Michael S. Lewis-Beck. 2023. “The Evolution of Election Forecasting Models in the UK.” *Electoral Studies* 86 (December): 102694. doi:10.1016/j.electstud.2023.102694.

Strijbis, Oliver, and Sveinung Arnesen. 2019. “Explaining Variance in the Accuracy of Prediction Markets.” *International Journal of Forecasting*, Special Section: Supply Chain Forecasting, 35 (1): 408–419. doi:10.1016/j.ijforecast.2018.04.009.

Surowiecki, James. 2005. *The Wisdom of Crowds*. Knopf Doubleday Publishing Group.

Tetlock, Philip E. 2009. *Expert Political Judgment*. Princeton University Press.

Tetlock, Philip E., and Dan Gardner. 2015. *Superforecasting: The Art and Science of Prediction*. Random House.

Wolfers, Justin, and Eric Zitzewitz. 2009. “Using Markets to Inform Policy: The Case of the Iraq War.” *Economica* 76 (302): 225–250. doi:10.1111/j.1468-0335.2008.00750.x.