Trends in Computational and Applied Mathematics

Trends in Computational and Applied Mathematics, **23**, N. 1 (2022), 175-192
Sociedade Brasileira de Matemática Aplicada e Computacional
Online version ISSN 2676-0029
www.scielo.br/tcam
doi: 10.5540/tcam.2022.023.01.00175

What If the Forecaster Knew? Assessing Forecast Reliability via Simulation

M. INÁCIO[1*], R. IZBICKI[2], D. LOPES[3], M. A. DINIZ[4],
L. E. SALASAR[5] and J. C. P. FERREIRA[6]

Received on March 25, 2021 / Accepted on August 10, 2021

ABSTRACT. The FIFA Men's World Cup (FWC) is the most important football (soccer) competition, attracting worldwide attention. A popular practice among football fans in Brazil is to organize contests in which each participant informs guesses on the final score of each match. The participants are then ranked according to some scoring rule. Inspired by these contests, we created a website to hold an online contest, in which participants were asked for their probabilities on the outcomes of upcoming matches of the FWC. After each round of the tournament, the ranking of all participants based on a proper scoring rule was published. In this article we estimate, by means of simulations, the ability of the best forecasters of our contest, considering that their good performances could be due to randomness. We also study the performance of some methods to aggregate individual forecasts, in order to study if some sort of wisdom of crowds (WOC) phenomenon was verified in the contest.

Keywords: forecast, risk, soccer contest.

1 INTRODUCTION

A probabilistic forecast is the attachment of a probability to a random (i. e. as yet unknown) event. The quality of such a kind of forecast can be assessed by a scoring rule, which quantifies, after the event is verified or not, how "close" the prediction was from the observed phenomenon. A

*Corresponding author: Marco Inacio – E-mail: m@marcoinacio.com

[1] University of São Carlos, Rod. Washington Luiz, s/n, Monjolinho, São Carlos-SP, Brazil – E-mail: m@marcoinacio.com https://orcid.org/0000-0002-6865-5404

[2] University of São Carlos, Rod. Washington Luiz, s/n, Monjolinho, São Carlos-SP, Brazil – E-mail: rafaelizbicki@gmail.com https://orcid.org/0000-0003-0379-9690

[3] University of São Carlos, Rod. Washington Luiz, s/n, Monjolinho, São Carlos-SP, Brazil – E-mail: lopes@ufscar.br

[4] University of São Carlos, Rod. Washington Luiz, s/n, Monjolinho, São Carlos-SP, Brazil – E-mail: marcio.alves.diniz@gmail.com https://orcid.org/0000-0002-8239-4263

[5] University of São Carlos, Rod. Washington Luiz, s/n, Monjolinho, São Carlos-SP, Brazil – E-mail: luis.salasar@gmail.com https://orcid.org/0000-0003-4715-8633

[6] University of São Carlos, Rod. Washington Luiz, s/n, Monjolinho, São Carlos-SP, Brazil – E-mail: jcpoloniato@gmail.com https://orcid.org/0000-0002-0029-9327

tentative way to measure the ability of several forecasters is by means of forecasting tournaments, in which each forecaster or participant informs her forecasts to a set of unknown events. After the mentioned phenomena are settled, the forecasts are scored and the forecasters ranked according to their overall score.

However, given the limited number of phenomena included in these tournaments, it is possible that the obtained score of any given forecaster does not reflect her true prediction abilities [21]. This limitation is particularly relevant for sports events, whose predictability is sometimes difficult to assess [1].

Our objective was to estimate the true ability of a group of forecasters that participated in a tournament to forecast the results of the 64 matchups of the 2018 FIFA Men's World Cup (FWC). We also estimated the robustness of different methods to aggregate forecasts and compared them to the performance of the individual participants. The estimates were obtained by simulating fictitious tournaments in which the probabilities of the results of each match were assumed equal to the forecasts of some individual participant (one of the best four). The results showed that a tournament with only 64 events (matchups) does not provide enough evidence to measure the ability of the best forecasters. In fact, the simulations revealed that, for our data, one would need a tournament with at least 500 events in order to distinguish the ability of the best forecasters with a probability larger than 90%. The aggregation methods also required the same magnitude of simulated events to allow a clear assessment of their predictive ability.

The following section briefly describes the forecast tournament and the scoring rule we used to evaluate the submitted forecasts. After a summary of the data set, composed by all the submitted forecasts for all the matches of the FWC, we describe the aggregation methods we have considered in our analysis and comment on their performance when compared to the individual participants of the tournament. Section 3 brings the main results, detailing how the simulations were used to estimate the ability of each forecaster and aggregation methods and Section 4 closes with some remarks. The appendices bring more details on the scoring rule we used in our contest (Appendix A) and mathematical proofs of results mentioned in the text (Appendix B).

2 METHODS AND DATA

In 2018, between June 14 and July 15, we promoted an online tournament where participants or forecasters would submit their probabilities on the matches of the FWC, played in Russia during that period. To give more incentive to participants, we announced that they would compete against two "mathematical models" whose forecasts were publicly available on websites.

The participants of the contest had to access the website `fifaexperts.com` where, after registering, they were able to inform their probabilistic forecasts for the results of all the scheduled matches of the FWC. More specifically, they had to provide, for each match, a vector $P = (P_1, P_2, P_3)$, where P_1 denotes the probability of victory of the first team, P_2 the probability of victory of the second team and P_3 the probability of a draw. The website enforced the constraints $P_1 + P_2 + P_3 = 1$ and $P_i \geq 0$, $i = 1,2,3$.

After the end of each match the reported forecasts were numerically ranked by a scoring rule, which quantitatively measures how “close” the forecast was from the match outcome. We adopted the Brier score [2], which is the squared Euclidean distance between the forecast and the outcome of the event, the football/soccer match in our case. Applying a suitable linear transformation we obtained a standardized score between 0 (the worst possible forecast) and 100 (the best score, i.e. the one associated with a forecast that assigns probability one to the observed outcome of the match). Appendix 4 brings the mathematical definition and main properties of the Brier score. The mentioned forecasts provided by mathematical models were available at the following websites.

- Chance de Gol [`http://www.chancedegol.com.br/`]. This website uses a bivariate Poisson regression model for the final score of the matches considering offensive and defensive factors of each team as explanatory variable. See [9] for more details.
- Previsão Esportiva [`www.previsaoesportiva.com.br`]. The statistical model of this website is similar to that of “Chance de Gol”, the only difference being the inclusion of expert information as explanatory variable. See [9, 16] for more details.

During the FWC we were informed that at least two other participants reported forecasts of statistical models:

- Esportes em números. The statistical model used by [4] (`http://www.fgv.br/emap/copa-2018/`), which is based on the model proposed by [19] and estimates the inherent offensive and defensive strengths of each team in a Poisson model; and
- Groll et al. [14], which is a new hybrid model for the score of international football matches. It combines random forests and Poisson ranking methods. While the random forest is based on covariates such as economic and sportive factors of each country, the ranking method estimates ability parameters on historical match data that adequately reflect the current strength of the teams.
- We have also included as participant the forecasts provided by the website FiveThirtyEight [`https://fivethirtyeight.com`].

It is important to remark that our results are based on the Brier score. We could have used different (proper) scoring rules to rank the submitted forecasts, as those mentioned by [18], but since their interpretation is not as intuitive as the Brier score, this could confuse some participants.

2.1 Dataset description

At the end of the FWC, there were 511 registered participants in the contest, though not all submitted forecasts for the 64 matches. Table 1 summarizes the number of participants according to the number of submitted forecasts, recalling that the group stage had 48 matches, the round of 16 had eight, the quarter-finals had four, the semi-finals two and the finals, two. The analyses presented in the sequel regards only the 57 participants that submitted forecasts to all matches of the FWC.

The number of submitted forecasts for each matchday presented a decreasing trend. After the group phase, several participants did not submit forecasts for the final matches, probably because some lost interest due to their poor performance.

Table 1: Number of participants by number of forecasts.

Participants	Submitted forecasts
57	64
104	$\geqslant 60$
139	$\geqslant 56$
217	$\geqslant 48$

2.2 Aggregation methods and their performance

The fact that combining forecasts often leads to good results has been named *Wisdom of the Crowds* (WOC) [3, 6, 26], which has been applied to several fields including cosmology [11, 17], medicine [23], natural language processing [24], and computer vision [28]. See [26] and references therein for other examples.

Although it is common to aggregate forecasts by unweighted averages [10, 20], many other methods are based on assigning different weights to each forecast [13]. These weights can be computed by evaluating how much opinions from different forecasters differ among each other [3, 8], or by evaluating the past performance of each forecaster [5, 27]. Other approaches take into account additional information about each forecaster that can be correlated to their performance [15, 23, 29]. See [12, 22] and references therein for a review of some approaches.

In order to assess if the wisdom of crowds was a phenomenon observed in our contest, we compared the forecasts of all participants to the following methods of aggregating their individual forecasts.

- Top-n. Arithmetic average of the forecasts made by the top-n participants (i.e., the n participants with best score) before a specific match. We considered $n = 1, 5, 10, 20$.
- Local wisdom. Average of all forecasts submitted for a given match.
- Global wisdom. Betting odds were collected from 18 online betting websites and the respective outcome probabilities were calculated using basic normalization [25]. The forecasts of the best three websites, according to their scores up to that point, were then averaged and reported as one forecast.

- CWM. The *Contribution Weighted Model* proposed by [3]. In our context, in the $(N+1)$-th match, this approach assigns to each forecaster j a contribution factor of $C_j := \sum_{i=1}^{N}(S_i - S_i^{-j})/N$, where S_i is the score of the above-mentioned local wisdom strategy for the i-th match, and S_i^{-j} is the score of the local wisdom strategy for the same match *removing forecaster j*. The aggregated forecast is given by the weighted average of the forecasts with positive C_j's, using the latter (normalized) constants as the weights.

- ISP-η. This is an *individual sequence prediction* (ISP) approach [5]. This approach is based on a weighted average of the forecasts given by each participant. More precisely, the forecast for match t is

$$\widehat{p}_t := \frac{\sum_{j=1}^{k} w_{j,t} f_{j,t}}{\sum_{j=1}^{k} w_{j,t}},$$

where $w_{j,t} \geq 0$ is the weight given to participant j for that match. The weights are taken to be $w_{j,1} \propto 1$ and, for $t > 1$, $w_{j,t} \propto \exp(\eta R_{j,t-1})$ where $R_{j,t-1}$ is the regret for the j-th participant up to match $t-1$, defined as

$$R_{j,t-1} = \sum_{t_0=1}^{t-1} \left[l(\widehat{p}_{t_0}, y_{t_0}) - l(f_{j,t_0}, y_{t_0}) \right],$$

where f_{j,t_0} is the forecast of the j-th participant for match t_0 and l is a loss function (in our case, the negative value of the score defined in Equation (A.1) in Appendix 4). η must be chosen by the participant; we have considered four values 0.001, 0.01, 0.1 and 1.

As a baseline for comparisons, we have also included the following simple strategies.

- Uniform. It randomly chooses a point on the simplex, that is, the forecast is a uniformly distributed vector over the 2-simplex: a Dirichlet distribution with parameter vector $(1,1,1)$.

- Edges. It randomly chooses a point at one of the edges of the simplex, meaning that the probability of one of the results (victory of team 1, victory of team 2 or draw) is set to 0 and the other probabilities are randomly drawn from the remaining possible values.

- Vertices. It randomly picks one of the vertices $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$ of the simplex, i.e., randomly selects a forecast that gives total certainty to one of the possible outcomes.

- Maximin. It assigns equal probabilities for every possible result (i.e., the forecast is $(1/3,1/3,1/3)$ for all the matches, called above naive or conservative strategy). This is the non-randomized maximin strategy.

The proofs of the mathematical results concerning these strategies are in Appendix 4.

Table 2 shows the ranking and final scores obtained for the FWC for all aggregation strategies and statistical models mentioned above. With exception of Previsão esportiva, statistical models had a good performance, in particular, Esportes em números and Groll et al. which obtained the top two scores.

The best aggregation strategy was Global wisdom, which is based on bets from external sources and the best aggregation strategy that only used bets made by other participants was CWM. The other aggregation strategies yielded poor results, the worst one being Top-1, which is intuitive. It is also interesting to note that the performance of ISP heavily depends on the tuning parameter η. For this application, taking $\eta = 0.01$ gave the best results. Still, the method was not among the top-10 best forecasters.

Table 2: Performance of the aggregation strategies as well as some participants.

Forecaster	Total Score	Average score/game	Rank
Esportes em números	4650	72.7	1.0
Groll et al.	4644	72.6	2.0
Global wisdom	4634	72.4	3.5
FiveThirtyEight	4634	72.4	3.5
Chance de gol	4611	72.0	5.0
CWM	4601	71.9	6.0
ISP-0.01	4569	71.4	11.0
ISP-0.001	4567	71.4	12.5
Local wisdom	4567	71.4	12.5
Top-20	4553	71.1	17.0
Top-10	4549	71.1	18.5
Top-5	4525	70.7	23.0
ISP-0.1	4492	70.2	31.0
Previsão esportiva	4450	69.5	37.0
ISP-1	4440	69.4	39.0
Top-1	4438	69.3	40.0
Maximin	4267	66.7	59.0
Uniform	3733	58.3	69.0
Edges	3200	50.0	70.0
Vertices	2133	33.3	71.0

3 MAIN RESULTS: ESTIMATING FORECASTERS' ABILITIES

Table 2 shows that the total scores obtained by the best forecasters are very close to each other, especially considering that they are based only on 64 matches. This raises the question of whether it is possible to say that the winner is indeed the best forecaster, or if was only a matter of luck. In this section we evaluate this question using simulations. Here we only consider participants that submitted forecasts to our tournament; a similar analysis for aggregated strategies is done in Section 3.1.

In order to make our simulations realistic, we first used the probabilities assigned by the winner of the contest (Esportes em números) as if they were the generating probabilities of the match outcomes. Using these probabilities, several independent tournaments were simulated and the rankings of each participant evaluated for each of these simulated tournaments. The idea is to check how many times the participant that is the best by construction (because it is used to generate the outcomes) is indeed the winner of the contest.

The details are as follows: for each simulation, we generated the outcomes of the 64 matches independently according to the probabilities submitted by the actual winner of the contest; we then computed the total score of each participant using these simulated outcomes and ranked the participants according to their simulated total scores. At the end of 100,000 simulations, we calculated the proportion of times a given participant ended up in a given rank.

Figure 1 shows a heat map of estimated probabilities that a participant with a given rank (horizontal axis) in our observed contest would end up in a given rank in the simulated tournaments (vertical axis). The darker the pixel the higher the corresponding probability is; the dashed diagonal line is the equality line. As a reference, all the probability values in the represented data matrix add up to one either by row or by column.

The figure shows that there is some association between the final rank in the actual contest and the rank of the same participant in the simulations (most of the shades in the graph are concentrated along the equality line). However, this association is not very strong, as indicated by the lightness of those shades, with most of the probabilities around the equality line being below 10%. Also, the pixel in the lower left corner is not as dark as one would expect: the actual winner ended up in the first place only in 28% of the simulations. It is also worthwhile noting that the participant with the lowest final ranking has the darkest pixel in Figure 1 as a combination of her assertiveness and some distance of her forecasts to those of the other forecasters. Table 3 presents a summary of the results and corroborates our conclusion that the ranking of the contest does not clearly define who is indeed the best forecaster.

Table 3: Percentages of simulations a given participant ended up in a certain rank, average and standard deviation of the ranks in the simulations. The forecasts of 57 participants for each of the 64 matches in the FWC are fixed as the ones informed in the actual contest, but the match outcomes are simulated according to the forecasts of the actual winner (Esportes em números).

	1st	2nd	3rd	4th	Other ranks	Average Rank	SD of Rank
Actual 1st place	28.2%	17.7%	11.4%	8.2%	34.5%	4.7	4.8
Actual 2nd place	5.1%	6.1%	6.1%	5.8%	76.9%	11.5	8.0
Actual 3rd place	0.8%	2.0%	3.4%	5.1%	88.7%	9.5	4.2
Actual 4th place	2.5%	6.4%	9.8%	11.7%	69.6%	7.1	4.1

In order to check the robustness of this conclusion, we repeated the simulations using probabilities submitted by other participants to generate the match outcomes. In this scenario, for each

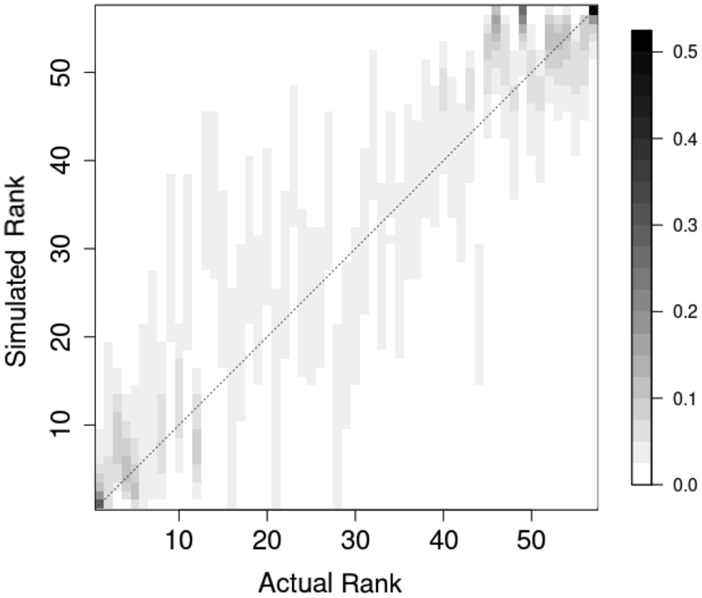


Figure 1: Probability that a participant who ended up in a given rank (horizontal axis) during the contest with 64 forecasts would end up in some other rank (vertical axis) in simulated contests.

of the top four participants of the actual contest, we generated 100,000 tournaments using their forecasts as probabilities to simulate the matches and then calculated the proportion of times this participant ended up in a given rank.

Each row of Table 4 corresponds to a different set of simulations and shows the distribution of the rank of a given participant when her own forecasts were the probabilities generating the fictitious matches. The participant who held second place in the actual contest (Groll et al.) had a similar pattern when she knew the truth compared to the actual winner (Esportes em números), with an even higher probability of taking the first place. On the other hand, the actual third and fourth places (Global wisdom and FiveThirtyEight, respectively) had higher probabilities of being in second, third or fourth places than being in the first place when each of them knew the truth. They all had the highest probabilities of ending up in each of the top ranks among participants, but the distributions of their ranks during simulations are concentrated on higher values, as indicated by their respective mean and standard deviation.

The lack of stability in the rankings is in part due to the small number of matches. In order to test if with a larger number of matches a participant who knows the truth would necessarily show her superiority, we performed a bootstrap-like simulation: we analyzed the behavior of the forecasts in a situation with *n* matches by taking a sample with replacement of size *n* from the original 64 matches. Then, for a given number *n* of matches and a given participant who knows the truth, we ran 10,000 simulations such that, in each simulation, we sampled *n* matches with replacement,

Table 4: Percentage of simulations a given participant ended up in a certain rank, average and standard deviation of the ranks in the simulations, when her respective forecasts were true. The forecasts of 57 participants for each one of the 64 matches in the FWC are fixed as the ones informed in the actual contest, but the match outcomes are simulated according to the forecasts of the participant in each row.

	1st	2nd	3rd	4th	Other ranks	Average Rank	SD of Rank
Actual 1st place	28.2%	17.7%	11.4%	8.2%	34.5%	4.7	4.8
Actual 2nd place	37.4%	16.2%	9.7%	6.8%	29.9%	4.3	4.8
Actual 3rd place	5.8%	10.4%	12.6%	12.9%	58.3%	5.8	3.5
Actual 4th place	9.1%	13.0%	13.2%	12.2%	52.5%	5.6	3.9

generated n outcomes independently according to the truth (if the same match was selected more than once, it possibly had different simulated outcomes), calculated the total score and ranked all the participants according to their total score.

Figure 2 shows the estimated probability of winning the contest (vertical axis) for the actual winner (Esportes em números, solid line) and the actual third place (Global wisdom, dashed line) when knowing the truth versus the number of matches (horizontal axis) during this simulated fictional contest. The actual third place was chosen due to her very small chance of winning when knowing the truth with only 64 matches (5.8%, as shown in Table 4). As expected, both lines increase and get closer to one as the number of matches increases. The best forecaster of our tournament needed a contest with approximately 576 matches in order to have a probability of 95% of winning when knowing the truth, while the actual third place would need approximately 1,024 matches in order to have at least the same probability of 95%.

Considering that our simulations show that hundreds of matches would be necessary to tell two very good forecasters apart, this kind of comparison in real life seems doable only for forecasting models, since human forecasters would probably find the task of providing hundreds of forecasts a very burdensome task.

3.1 Evaluating aggregation strategies

The results discussed in Section 2.2 show that aggregation strategies were superior to most of the individual participants including the statistical models and the simpler strategies. However one could ask, in the same spirit of the last section: how could one tell that these strategies were really the best ones and not just "got lucky" about the 64 matches of this particular tournament?

To answer this question, similarly to the procedure described above, we simulated 100,000 fictitious tournaments using the forecasts of a given participant as the generating mechanism for the match outcomes. Then, we compared the performance of the aggregation strategies Top-5, Local

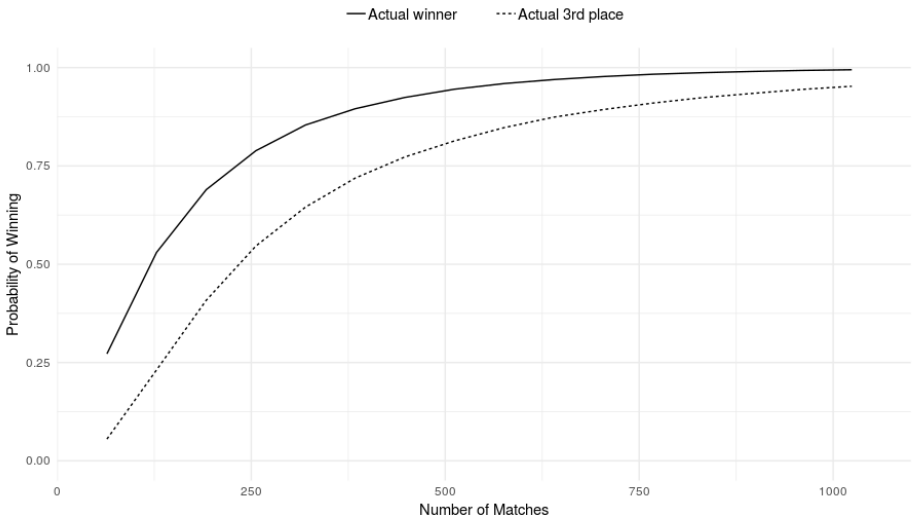


Figure 2: Probability that the actual winner (solid line) and the actual third place (dashed line) win the contest with when their respective forecasts were true versus the number of matches in the contest.

Wisdom and CWM separately to all the 57 participants (i.e., strategies are not compared among them when computing simulated ranks).

The comparison was based on the proportion of simulations each strategy ended up in a given rank under two different simulation scenarios for the 64 matches: (i) match outcomes simulated from the forecasts of the actual winner (Esportes em números) and (ii) match outcomes simulated from the forecasts of the actual third place (Global wisdom). Tables 5 and 6 present the results for the scenarios (i) and (ii), respectively.

Notice that the proportions presented in Tables 5 and 6 are much smaller than those presented in Tables 3 and 4, which is expected since it is difficult for the aggregation strategies to identify the best forecasters using only 64 matches. For both scenarios, Top-5 is the riskier strategy, with higher chance of reaching top ranks, but also with higher variability, as indicated by the greater average and large standard deviation of the final rank. Local Wisdom strategy is the most conservative one, with moderate values of average and standard deviation of final ranks.

In order to study the effect of increasing the number of forecasts in the contest on the performance of the three aggregation strategies, we created fictitious tournaments of n matches by sampling with replacement from the 64 matches in the FWC using their corresponding forecasts submitted to the contest, similarly to the procedure described in Section 3.

For n varying from 1 to 1,024, we simulated 30,000 fictitious tournaments from the probabilities submitted by the best forecaster of the contest and then computed the average final rank for each

Table 5: Percentages of simulations the selected aggregation strategies ended up in a given rank relative to the 57 participants, including the respective average and standard deviation of the final ranks. The match outcomes were simulated from the forecasts of the actual first place participant.

	1st	2nd	3rd	4th	Other ranks	Average Rank	SD of Rank
Top-5	1.2%	1.9%	2.5%	2.7%	91.7%	15.2	7.7
Local Wisdom	0.2%	0.9%	2.1%	3.8%	93.0%	10.1	3.9
CWM	0.2%	0.8%	1.2%	1.6%	96.2%	13.6	4.6

Table 6: Percentages of simulations the selected aggregation strategies ended up in a given rank relative to the 57 participants, including the respective average and standard deviation of the final ranks. The match outcomes were simulated from the forecasts of the actual third place participant.

	1st	2nd	3rd	4th	Other ranks	Average Rank	SD of Rank
Top-5	2.1%	3.2%	3.7%	3.8%	87.1%	13.6	7.9
Local Wisdom	0.4%	1.8%	3.9%	6.8%	87.1%	8.3	3.5
CWM	0.5%	1.5%	2.3%	2.8%	92.9%	12.5	5.1

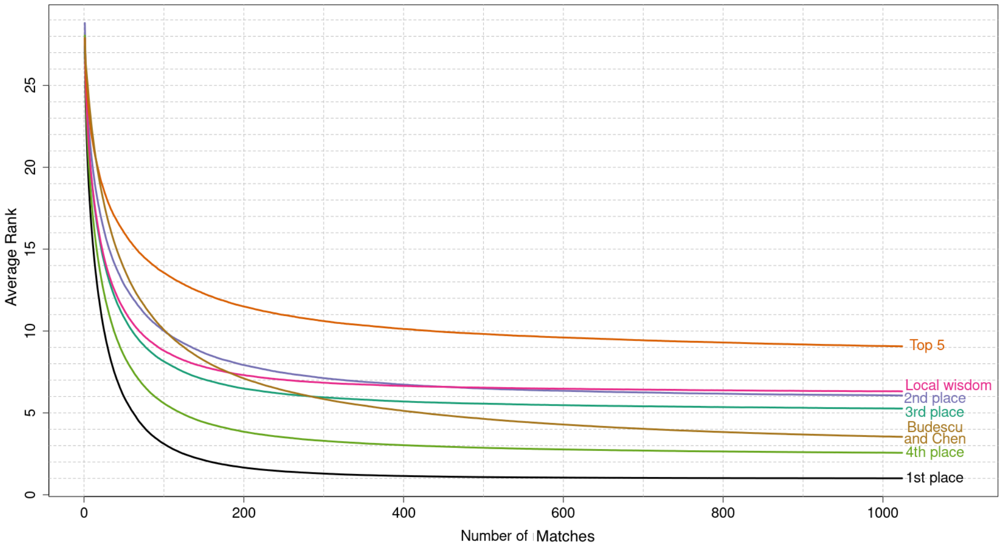


Figure 3: Average final rank of a given participant (or strategy) in the simulations when forecasts of the actual winner were true versus the number of matches in the contest with 57 participants.

one of the actual top-4 participants and for the three WOC strategies (once again, we compared each strategy individually to the 57 participants). Results are shown in Figure 3.

All the strategies and top participants improved their average ranks as we increased the number of forecasts in the contest, but the most noticeable improvement was accomplished by the strategy CWM. The intuitive reason for the good performance of this method is that its weighted average can correctly identify the best forecasters, such as the actual winner, and at the same time disregard the misinformed participants, those with negative contribution factors. Learning about which participants bring relevant information and which do not is the main advantage of the CWM method, but it may take some hundreds of forecasts for its prevalence among strategies to appear.

4 FINAL REMARKS

The probabilistic previsions submitted to our website to forecast the matches of the 2018 FIFA Men's World Cup revealed some interesting characteristics of the behaviour of such contests and of the performance of WOC strategies, considering different ways of opinion aggregation.

Regarding the aggregation (WOC) strategies, the best ones (Global wisdom and CWM) had an outstanding overall performance of the same level of the best forecasters, which adopted some sort of statistical model or algorithm. However, this was not true for all of them, such as ISP-0.1, which ranked 31. An important remark about the strategies Global wisdom and CWM is that one should know all the submitted forecasts of an upcoming match to provide a forecast for the same match, which would not be feasible for a regular user of our website. Moreover, the fact that CWM yields good results heavily depends on the fact that there are good predictors among the forecasts that were submitted.

Finally, our simulations revealed that a tournament with 64 matches, like the FWC, is not sufficient to identify the best forecasters since the observed performance of any given participant might be due to randomness. Thus, longer tournaments such as the English Premier League, which has 380 matches every season, may yield further insights on which aggregation strategies work better. However, conducting an open contest of the same sort we did for the FWC would require incentive mechanisms to avoid individual participants of leaving the contest before its end.

Acknowledgments

Marco Inácio is grateful for the financial support of CAPES: this study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Rafael Izbicki is grateful for the financial support of FAPESP (2019/11321-9) and CNPq (309607/2020-5).

REFERENCES

[1] E. Ben-Naim, F. Vazquez & S. Redner. Parity and Predictability of Competitions. *Journal of Quantitative Analysis in Sports*, **2**(4) (2006), Article 1.

[2] G.W. Brier. Verification of Forecasts Expressed In Terms Of Probability. *Monthly Weather Review*, **78**(1) (1950), 1–3.

[3] D.V. Budescu & E. Chen. Identifying Expertise to Extract the Wisdom of Crowds. *Management Science*, **61**(2) (2014), 267–280.

[4] P.C.P. Carvalho, M.A. Silva & A.P. Carneiro. Previsões para os jogos da Copa do Mundo de 2018. Unpublished manuscript (2018).

[5] N. Cesa-Bianchi & G. Lugosi. "Prediction, learning, and games". Cambridge University Press (2006).

[6] C.P. Davis-Stober, D.V. Budescu, J. Dana & S.B. Broomell. When Is a Crowd Wise? *Decision*, **1**(2) (2014), 79.

[7] A.P. Dawid & M. Musio. Theory and applications of proper scoring rules. *Metron*, **72**(2) (2014), 169–183.

[8] A.P. Dawid & A.M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **28**(1) (1979), 20–28.

[9] M.A. Diniz, R. Izbicki, D. Lopes & L.E. Salasar. Comparing probabilistic predictive models applied to football. *Journal of the Operational Research Society*, **70**(5) (2019), 770–782.

[10] L.G. Esteves, R. Izbicki & R.B. Stern. Teaching decision theory proof strategies using a crowdsourcing problem. *The American Statistician*, **71**(4) (2017), 336–343.

[11] P.E. Freeman, R. Izbicki, A.B. Lee, J.A. Newman, C.J. Conselice, A.M. Koekemoer, J.M. Lotz & M. Mozena. New image statistics for detecting disturbed galaxy morphologies at high redshift. *Monthly Notices of the Royal Astronomical Society*, **434**(1) (2013), 282–295.

[12] B. Frénay & M. Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, **25**(5) (2013), 845–869.

[13] C. Genest & J.V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, **1**(1) (1986), 114–135.

[14] A. Groll, C. Ley, G. Schauberger & H. Van Eetvelde. A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*, **15**(4) (2019), 271–287.

[15] R. Izbicki & R.B. Stern. Learning with many experts: model selection and sparsity. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **6**(6) (2013), 565–577.

[16] A.J. Lee. Modeling scores in the Premier League: is Manchester United really the best? *Chance*, **10**(1) (1997), 15–19.

[17] C.J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M.J. Raddick, R.C. Nichol, A. Szalay, D. Andreescu *et al.* Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, **389**(3) (2008), 1179–1189.

[18] R.L. Machete. Contrasting Probabilistic Scoring Rules. *Journal of Statistical Planning and Inference*, **143**(10) (2013), 1781–1790.

[19] M.J. Maher. Modelling Association Football Scores. *Statistica Neerlandica*, **36**(3) (1982), 109–118.

[20] S. Makridakis & R.L. Winkler. Averages of forecasts: Some empirical results. *Management Science*, **29**(9) (1983), 987–996.

[21] E. Merkle, M. Steyvers, B. Mellers & P. Tetlock. Item Response Models of Probability Judgments: Application to a Geopolitical Forecasting Tournament. *Decision*, **3**(1) (2016), 1–19.

[22] H. Olsson & J. Loveday. A Comparison of Small Crowd Selection Methods. In "Proceedings of the 37th Annual Meeting of the Cognitive Science Society. Austin, TX". Cognitive Science Society (2015), p. 1769–1774.

[23] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni & L. Moy. Learning from crowds. *Journal of Machine Learning Research*, **11**(Apr) (2010), 1297–1322.

[24] R. Snow, B. O'Connor, D. Jurafsky & A.Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In "Proceedings of the Conference on Empirical Methods in Natural Language Processing". Association for Computational Linguistics (2008), p. 254–263.

[25] E. Štrumbelj. On determining probability forecasts from betting odds. *International Journal of Forecasting*, **30**(4) (2014), 934–943.

[26] J. Surowiecki. "The Wisdom of Crowds: Why the Many Are Smarter Than The Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations.". Doubleday & Co, New York, NY (2004).

[27] V. Vovk & F. Zhdanov. Prediction with expert advice for the Brier game. *Journal of Machine Learning Research*, **10**(Nov) (2009), 2445–2471.

[28] P. Welinder & P. Perona. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In "2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops". IEEE (2010), p. 25–32.

[29] Y. Yan, R. Rosales, G. Fung & J. Dy. Modeling multiple annotator expertise in the semi-supervised learning scenario. *arXiv preprint arXiv:1203.3529*, (2012).

APPENDICES

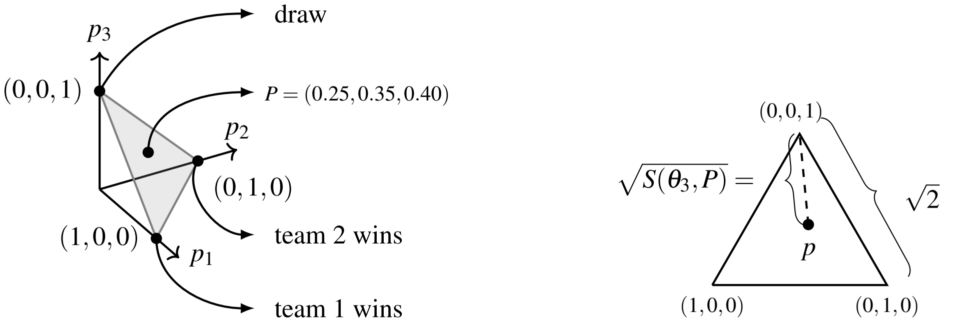
A - The scoring rule

As mentioned above, the participants of our contest were ranked according to their ability to make probabilistic forecasts about the outcome θ of a football match, where θ is a random variable taking values in $\Theta = \{\theta_1, \theta_2, \theta_3\}$, with θ_1 standing for a victory of the first team, θ_2 a victory of the second team and θ_3 a draw. The probabilistic forecast for θ is represented by the vector $P = (P_1, P_2, P_3)$ of probabilities for each possible outcome θ_1, θ_2 and θ_3.

Our scoring rule was based on the Brier score. To formally define it, notice that the forecasts $P = (P_1, P_2, P_3)$ lie in the 2-simplex, i.e., $P \in \Delta_2 = \{(p_1, p_2, p_3) : p_1 + p_2 + p_3 = 1,\ p_1, p_2, p_3 \geq 0\}$. The Brier score $S(\theta, P)$ is given by

$$S(\theta, P) = \sum_{i=1}^{3} \mathbb{I}(\theta = \theta_i)(1 - P_i)^2 + \sum_{i=1}^{3} \mathbb{I}(\theta \neq \theta_i) P_i^2$$

where $\mathbb{I}$ is the indicator function, i.e. it equals 1 if the argument of the function is verified and zero otherwise. Figure 4a illustrates the forecast $P = (0.25, 0.35, 0.40)$ represented on the 2-simplex.



(a) Δ_2 simplex: gray surface

(b) Brier score of P for a tied match

Figure 4: Δ_2 simplex viewed on three **(a)** and two **(b)** dimensions.

When $P = (1,0,0)$ the forecaster puts all the probability at one vertex of the simplex, believing that $\theta = \theta_1$ for sure. In this case, if team 1 in fact wins the match, the Brier score is zero, its minimum possible value. However, if the match is won by team two ($\theta = \theta_2$) or ends tied ($\theta = \theta_3$), the Brier score is 2, its maximum possible value.

Therefore, for the Brier score, the smaller the scores, the better are the forecasts, which could be counterintuitive for most of the participants. For this reason, we adopted a linear function of the Brier score as the scoring rule for our constest, namely

$$S^*(\theta, P) = 100 - 50 \cdot S(\theta, P) \tag{A.1}$$

that is bounded between zero (worst score) and 100 (best score). Since the World Cup had 64 matches, the perfect score would be 6400 points.

An important feature of the Brier score and our $S^*(\theta, P)$ is that they are *proper* [7], meaning that, in order to maximize the expected value of their score, the forecasters should inform their true probabilities.

B - Maximin and expected scores of strategies

In this appendix we prove that: (1.) the naive strategy is the (pure) maximin strategy; (2.) the expected scores (per match) of the strategies Uniform, Vertices and Edges are 58.33, 33.33 and 50, respectively.

1. The naive strategy $P^* = (1/3, 1/3, 1/3)$ is the non-randomized maximin strategy.

Proof:

A strategy P^* is maximin when

$$\inf_{\theta\in\Theta}[S^*(\theta,P^*)] \geq \inf_{\theta\in\Theta}[S^*(\theta,P)] \tag{B.1}$$

for all $P \in \Delta_2$, that is, P^* maximizes the minimum score given by $S^*(\theta,P) = 100-50\cdot S(\theta,P)$. Notice that, P^* satisfies (B.1) if, and only if,

$$\sup_{\theta\in\Theta} S(\theta,P) \geq \sup_{\theta\in\Theta} S(\theta,P^*) \tag{B.2}$$

for all $P \in \Delta_2$.

Thus, the naive strategy $P^* = (1/3, 1/3, 1/3)$ is maximin if

$$\sup_{\theta\in\Theta} S(\theta,P) \geq \sup_{\theta\in\Theta} S(\theta,P^*) = \frac{2}{3} \tag{B.3}$$

for all $P \in \Delta_2$. Observe that proving (B.3) is equivalent to prove that for any $P = (P_1, P_2, P_3) \in \Delta_2$ we have

$$S(\theta_i,P) \geq \frac{2}{3} \text{ for some } i = 1,2,3. \tag{B.4}$$

Associating the possible outcomes $\theta_1, \theta_2, \theta_3$ respectively to the points $O_1 = (1,0,0)$, $O_2 = (0,1,0)$, $O_3 = (0,0,1) \in \Delta_2$, we see that $S(\theta_i,P)$ is the squared Euclidean distance between the points O_i and P, denoted by $\|\overrightarrow{PO_i}\|^2$. Then, in order to prove (B.4), we shall prove that $\|\overrightarrow{PO_i}\|^2 \geq 2/3$ for at least one $i = 1,2,3$.

Since $P_1 + P_2 + P_3 = 1$, then $P_i \leq 1/3$ for some i. For instance, we can assume that $P_3 \leq 1/3$. Then, it is possible to show that

$$\langle\overrightarrow{P^*P}, \overrightarrow{P^*O_3}\rangle = P_3 - \frac{1}{3} \leq 0, \tag{B.5}$$

where $\langle\cdot,\cdot\rangle$ is the dot product of the Euclidean space. The inequality (B.5) is equivalent to say that the angle between the vectors $\overrightarrow{P^*P}$ and $\overrightarrow{P^*O_3}$ is greater or equal to $\pi/2$. See Figure 5 for an illustration of the mentioned vectors on the simplex.

From the polarization identity (law of cosines), it follows that

$$\|\overrightarrow{PO_3}\|^2 = \|\overrightarrow{P^*P}\|^2 + \|\overrightarrow{P^*O_3}\|^2 - 2\langle\overrightarrow{P^*P}, \overrightarrow{P^*O_3}\rangle \geq \|\overrightarrow{P^*O_3}\|^2,$$

where the last inequality follows from (B.5). Then,

$$S(\theta_3,P) = \|\overrightarrow{PO_3}\|^2 \geq \|\overrightarrow{P^*O_3}\|^2 = S(\theta_3,P^*) = \frac{2}{3},$$

which proves (B.4). Therefore, $P^* = (1/3, 1/3, 1/3)$ is a maximin strategy.

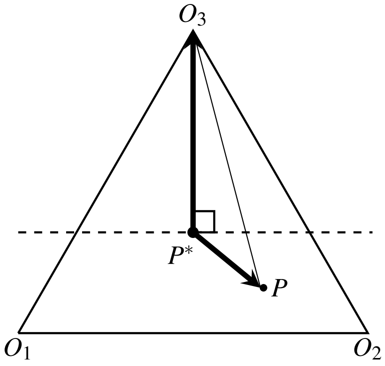


Figure 5: Illustration of vectors $\overrightarrow{P^*O_3}$, $\overrightarrow{PO_3}$ and $\overrightarrow{P^*P}$.

Figure 6 displays the surface of scores $\left(S^*(\theta_1,P),S^*(\theta_2,P),S^*(\theta_3,P)\right)$ with P varying on the simplex Δ_2 (yellow surface) and the level surface of the score vectors such that $\min\{S^*(\theta_1,P),S^*(\theta_2,P),S^*(\theta_3,P)\} = 200/3$ (blue surface), where the $200/3$ is the minimum score for the naive strategy P^*.

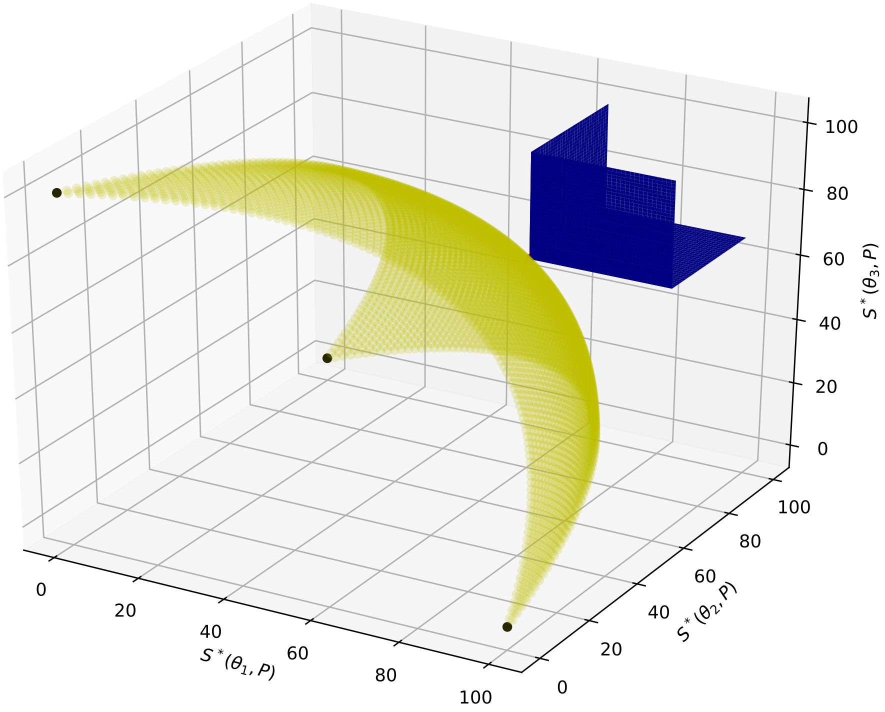


Figure 6: Score surface of each pure strategy P.

2. Expected scores of the mixed strategies: (i) **Uniform**: Dirichlet(1,1,1); (ii) **Vertices**: uniform on vertices; (iii) **Edges**: uniform of edges.

First we rewrite the expression for $S(\theta,P)$.

$$\begin{aligned} S(\theta,P) &= \sum_{i=1}^{3} \mathbb{I}(\theta=\theta_i)(1-P_i)^2 + \sum_{i=1}^{3} \mathbb{I}(\theta \neq \theta_i)P_i^2 \\ &= 1 + \sum_{i=1}^{3} P_i^2 - 2\sum_{i=1}^{3} P_i\, \mathbb{I}(\theta=\theta_i) \end{aligned}$$

Then, for $\theta \in \Theta = \{\theta_1,\theta_2,\theta_3\}$:

$$E[S(\theta,P)] = 1 + \sum_{i=1}^{3} E[P_i^2] - 2\sum_{i=1}^{3} E[P_i]\cdot \mathbb{I}(\theta=\theta_i)$$

(i) **Uniform**: Dirichlet (1,1,1)

$$E[S(\theta_1,P)] = E[S(\theta_2,P)] = E[S(\theta_3,P)] = 5/6 \Longrightarrow E[S(\theta,P)] = 5/6,$$

and therefore $E[S^*(\theta,P)] = 100 - 50\cdot E[S(\theta,P)] = 58.33$.

(ii) **Vertices** : each vertex with probability $1/3$

$$E[S(\theta_1,P)] = E[S(\theta_2,P)] = E[S(\theta_3,P)] = 4/3 \Longrightarrow E[S(\theta,P)] = 4/3,$$

and therefore $E[S^*(\theta,P)] = 100 - 50\cdot E[S(\theta,P)] = 33.33$. In fact, this strategy is the randomized maximin strategy, corresponding to a randomization of the three points highlighted on the utility surface displayed by Figure 6.

(iii) **Edges**: Uniform on the edges. For $i = 1,2,3$:

$$\begin{aligned} E[S(\theta_i,P)] =& \frac{1}{3}\left[\int_0^1 (0-u)^2 + (1-(1-u))^2\, du + \int_0^1 (0-(1-u))^2 + (1-u)^2\, du + \right.\\ & \left. \int_0^1 1 + (0-u)^2 + (0-(1-u))^2\, du \right] \quad \Rightarrow E[S(\theta,P)] = 1, \end{aligned}$$

and therefore $E[S^*(\theta,P)] = 100 - 50\cdot E[S(\theta,P)] = 50$.