Bayesian Forecasting of UEFA Champions League under alternative seeding schemes*

Submitted by:

Manas Mishra [†‡]

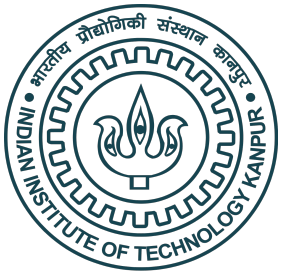
Ritwik Vashishtha [§†]

Shubha Sankar Banerjee [¶†]

Shresth Grover [‖†]

Supervised by:

Dr. Arnab Hazra [†]



Submitted on:

22^{nd} April, 2022

Abstract

Seeding rules plays an important role in determining the fate of teams competing in UEFA Champions League. In slight changes of seeding practices, weak teams (ranked according to UEFA coefficients) can stand to either benefit or lose out on chances in advancement in the tournament. We use probabilistic forecasting models to evaluate the effect of seeding. We review a Bayesian approach to account for the uncertainty involved with the parameters of the forecasting model. We perform a Monte Carlo simulation to estimate outcome probabilities by simulating the UEFA Champions League under alternative seeding regimes. We incorporate entropy in our analysis to evaluate the uncertainty in winning the tournament by all the teams under different seeding schemes.

Contents

[*] This report has been prepared towards the partial fulfillment of the requirements of the course *MTH535A: An Introduction to Bayesian Analysis.*

[†] Department of Mathematics & Statistics, Indian Institute of Kanpur, India.

[‡] 201340, M.Sc. Statistics (Final year).

[§] 201389, M.Sc. Statistics (Final year).

[¶] 201416, M.Sc. Statistics (Final year).

[‖] 190820, Btech Aerospace Engineering (3^{rd} year).

1 Introduction

Probabilistic models for sports forecasting are made to enhance the experience of fans, who anticipate the results of the games well before the actual game begins. In the economic sector, these forecasts are of direct utility to those associated with betting markets and also pundits and commentators who base their careers around their ability to forecast performances.

There can be a wide spectrum across which the motivation behind forecasting rests. One may want to model audience sizes and for that purpose must require a variable that measures the importance of that particular match, a metric which might be measured as the probability that the result of the current match will affect the outcome of the tournament. Such a modelling technique would also depend on the outcomes of matches which are yet to be played. Scarf and Shi (2008) measured the importance of the current match by incorporating in their simulation methodology a match result probabilistic forecasting model that is applied to matches that are yet to be played.

In this paper, we employ a model and use it to generate both a set of forecasts given that the state of the world as it is and a set of forecasts in some alternate setup, with the motivation to replicate the findings of Corona et al. (2019). We produce forecasts for various outcomes in the UEFA Champions League (e.g. the probability that the club i will reach the round of 16 and the probability that club i will reach the final) under both the actual and alternative schemes (discussed in Section 2). The motivation behind making these comparisons is to analyze which clubs gained and which clubs lost the most from the changes made due to seeding system.

Corona et al. (2019) introduced the Bayesian approach to match level forecasting and that enabled them to calculate predictive probability distributions for the tournament outcome probabilities under both current and alternate seeding schemes.

Classical techniques of forecasting include the plug-in approach where we get point estimates for individual match probabilities which are in turn used to simulate the outcome of the tournament. More than often, these match-level probabilities are themselves subject to uncertainties, which remain unaccounted for. Ignoring the issue would not affect the efficacy of the approach if the purpose is to simply generate each teams

chance of winning the league, but our objective is to compare probabilities of success under alternative seeding schemes. The procedure by Corona et al. (2019), makes it possible to obtain the posterior distribution of the model parameters and then estimate the predictive distribution of the probability that a given team will win the tournament. If the distribution of a team's probability of winning the tournament under one seeding system is different from that under another, it is impossible that the team will have same chance of winning under both systems. However, if the distributions overlap significantly, it could be that the true underlying probability of winning the tournament is essentially the same under both seeding systems.

The rest of the paper is arranged as follows: Section 2 describes the structure of the UEFA Champions League and the seeding schemes in place. Section 3 describes the measure of entropy which we employ to assess the degree of outcome uncertainty regarding which clubs will reach various stages of the competition as well as winning it. Section 4 sets our Bayesian Poisson model framework for probabilistic forecasting of individual match results. We use the estimates obtained from the previous section to simulate the tournament under different seeding regimes in Section 5. Finally we present our conclusions as a discussion in Section 6.

2 Seeding rules in the UEFA Champions League

We consider the seeding rules which was prevailing when the Corona et al. (2019) was published. 32 clubs participated in the annual edition of the tournament. The current holders of the Champions League trophy were reserved a place in the games. 21 places were for clubs from strong footballing leagues that qualified directly owing to their finishing positions in their respective domestic league tables. The final 10 places were for clubs which were successful in qualifying tournament held in the summer. It is to be noted that the finishing positions of the teams playing the qualifiers were important in deciding their potential for playing in the League.

The names of the 32 clubs that can take part are declared by the end of August. It is at this point, that the variation in the strength across the the playing clubs are analyzed and the competitive balance of the tournament is fixed. The eventual seeding plays an

important role in the outcome uncertainty of the tournament.

According to the taxonomy of tournament designs set out by Scarf and Yusof (2011), the Champions League is a "hybrid 1G-KO" competition. It means that there is a single "group stage" (played from September to December), followed by a series of knock-out or elimination rounds (from February to May), which culminate in a grand final (held in June).

At the start of the competition, the 32 clubs are split into eight groups (A to H). Each group is a mini double round-robin league such that each club plays each other both home and away. A win is rewarded with 3 points and a draw with 1 point. The top two clubs in each group proceed to the round of 16. From then on, the tournament is straight knock-out competition, with a fresh draw being conducted for each round. A constraint imposed for the draw of round of 16 is that top place holders from group stage play second place holders from the group stage.

The seeding is applied at the draw which allocates clubs between the eight initial groups. The clubs are allocated to pots 1, 2, 3 or 4 according to their seedings. This ensures that each group will include one team from each pot. Each group is also constrained to consist clubs from four difference countries (or domestic leagues).

The seeding in question is based on the "UEFA coefficients" of each club prior to the tournament. These ranking points are earned by wins and progression in the UEFA Champions League and in the UEFA Europa League over the preceding five seasons by the club itself and by the clubs from the respective domestic association (which yields lesser weights).

The method of allocation of clubs in pots mentioned earlier has changed over time. Prior to the 2015-16 season, the allocation of the clubs were done strictly according to the UEFA coefficients, i.e. the eight strongest clubs according to the UEFA rankings were put in pot 1. This further implied that these clubs could not play with each other in the group stage of the League. This is a case of classic seeding to ensure minimized probability of an early exit by the "best" competitors. The other pots were allocated in a similar fashion.

From the 2015-16 season. seeding arrangements were changed. Pot 1 now consisted of the reigning champion of the Champions League and the champions from

the seven strongest national league according to UEFA coefficients across Europe. FC Barcelona were coincidentally the champions of the Spanish League and the reigning Champions of the League, and as a result entitled to Pot 1. This made way for an extra place for the champions from the eighth-ranked league from Netherlands (according to UEFA coefficients), i.e. PSV Eindhoven. The remaining clubs were distributed according to their UEFA rankings.

The phenomenon we are interested in is that a minor change in the seeding rules set for allocation to pot 1 has apparently the potential to affect the degree of outcome uncertainty surrounding the tournament.

These changes can potentially have high impact on the games. We state the example cited by Corona et al. (2019). Real Madrid was the strongest club of all according to the UEFA coefficients however they failed to secure Pot 1 since they did not win the Spanish League. This might potentially reduce the chance of Chelsea progressing in the tournament, which is placed in Pot 1 (for winning the English Premier League). The club is no longer protected from the possibility of playing Real Madrid in the group stage since now they are in different pots. On the other hand, they might secure a rank in top two (in a group with Real Madrid) and thereby eliminate the possibility of playing against them in the round of 16. Weaker UEFA rated clubs like PSV Eindhoven would be the club to gain something from thus change in seeding. The new seeding reduces the chances of them playing against stronger clubs in group stage. Since the effects of the changes in seeding are difficult to acknowledge, we perform simulation analysis of the tournament to come to a conclusion.

3 Entropy Measure

Our focus is to measure the outcome uncertainty from the perspective of the point just prior to the draw for allocating the 32 competing clubs between the eight groups which compromise the first stage of the tournament. It is to be noted that the seeding is implemented at this stage. This potentially influences the events for the rest of the season.

I. (1997) first used Entropy, a measure of unpredictability of information, to measure the apparent "improvement" in uncertainty over time in Major League Baseball. We are interested in measuring the uncertainty over several outcomes, such as which club will win the tournament and which clubs will emerge from the group stage to take places in the round of 16.

Let $p_{j0} = P(V_j|\mathscr{E}_0)$ be the probability of victory for club j conditional on pre-tournament information concerning the 32 clubs in the competition.

The entropy is defined as follows:

$$e_0 = -\sum_{j=1}^{32} p_{j0} \log_2 p_{j0} \tag{1}$$

Minimum entropy or maximum information occurs when some $p_{j0} = 1$ while others are zero. In that case e_0. At the other extreme, the maximum entropy is when all probabilities p_{j0} are equal to $1/32$ so that there is a maximum outcome uncertainty. It is to be noted that entropy is calculated before the start of the tournament. Our interest in the entropy measure is the comparison of the tournament outcome under different seeding policies.

4 Bayesian Poisson model

For simulating a tournament, we first require a viable model which can be used for probabilistic forecasting of individual matches which might take place within the tournament. We will focus on Poisson regression model, which yields the estimated probability for each team, for scoring no goals, one goal, two goals and so on. Combining these probabilities enables us to produce point estimates of winning probabilities for each side. This itself is a feature of Poisson regression, that we can incorporate the goals (and not only the match outcomes), that we use it in our setup. Thus if the clubs are tied in points, the tie break rules include a comparison of goal differences. Thus, the model underlying the simulation must generate the probability of each possible scoreline in a match, as well as win-draw-lose probabilities.

Application and variations of Poisson model has its roots across many papers in

football forecasting literature. We will however focus on the Maher model, following Scarf and Yusof (2011) since our focus is on the tournament simulation rather than on the match-level forecasting models themselves.

We apply a Poisson regression model (BP) for the goals scored by each team in a match as follows:

$$Y_{t,k} \sim Poisson(\lambda_{T,k})$$

where, $Y_{t,k}$ represents the number of goals scored by team T in a match at time k and

$$\log(\lambda_{T,k}) = \beta_{A_T} x_{A_T,k} + \beta_{A_0} x_{A_0,k} + \beta_{H_T} x_{H_T,k} + \beta_{Aw_T} x_{Aw_T,k} + \beta_{F_T} x_{F_T,k} \tag{2}$$

where,

- $x_{A_T,k}$ is the strength of team T.
- $x_{A_0,k}$ indicates the strength of the opposing club.
- $x_{H_T,k}$ indicates that team T is playing at home.
- $x_{Aw_T,k}$ indicates that the team T is playing away.
- $x_{F_T,k}$ indicates that the match is the final of the whole competition (the only match played at a neutral ground).

The parameters β_{A_T}, β_{A_0}, β_{H_T}, β_{Aw_T} and β_{F_T} are coefficients that express the relationship between explanatory variables and $\lambda_{T,k}$.

We use the UEFA points (or coefficients) at the start of each years competitions as a measure of strength of the respective clubs.

It could be true strength of a club would depend on its form which might vary over time. Given that UEFA Champions league is played over a span of 8 months and the simulations are from the perspective of before the start of the competition. Therefore, a teams form in August, when the regular league season starts, might not be a relevant measure to predicting results of matches in April. Thus we use the UEFA coefficients, which indicate a long run behaviour, obtained on past performances. It is

further helpful as a universal metric since it is difficult to assess the relative strength of clubs on their propensities to score or concede goals across national leagues.

We are focused on using the Poisson regression model in a Bayesian setup since Bayesian methods take the parameter uncertainty directly into account.

Before proceeding with the methodology, we need to define the prior distribution of the regression coefficients $\beta = (\beta_{A_T}, \beta_{A_0}, \beta_{H_T}, \beta_{Aw_T}, \beta_{F_T})'$ in Eq (2). Corona et al. (2019) used improper, uniform prior distribution $f(\beta) \approx 1$, as recommended by Martín et al. (2011). For convenience purposes of implementation, we use $Uniform(-2,2)$ prior distribution for each component of β.

Since exact calculation of the posterior distribution is impossible, we can generate an approximate Monte Carlo sample of values from the posterior distribution using Markov Chain Monte Carlo (MCMC) methods. As per Sherlock et al. (2010), we use random walk Metropolis algorithm to generate successive elements of β from their conditional posterior distributions.

After calculating the Monte Carlo samples, we obtain the win, draw and loss probabilities, $p_{W,k}, p_{D,k}$ and $p_{L,k}$ respectively by following the procedure used by Corona et al. (2017). Thus the probabilities for each game during the competition for team T against team O becomes:

$$p_{W,k} = \sum_{i_T=0}^{\infty} \sum_{i_0=1}^{i_T-1} P(y_{T,k} = i_T) P(y_{O,k} = i_O) \tag{3}$$

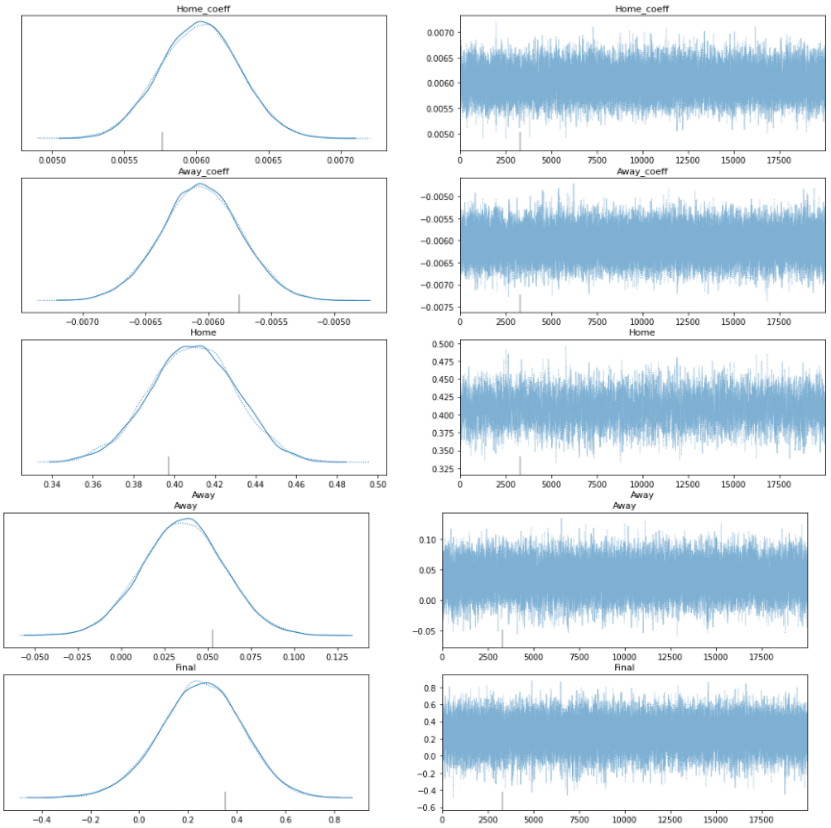
$$p_{D,k} = \sum_{i_T=0}^{\infty} P(y_{T,k} = i_T) P(y_{O,k} = i_T) \tag{4}$$

$$p_{L,k} = \sum_{i_O=0}^{\infty} \sum_{i_T=1}^{i_O-1} P(y_{T,k} = i_T) P(y_{O,k} = i_O) \tag{5}$$

where i_T and i_O are the numbers of goals scored for teams T and O respectively, where $P(y_{T,k} = i_T)$ and $P(y_{O,k} = i_0)$ represent the corresponding Possion probabilities of goals scored (with $\lambda_{T,k}$ and $\lambda_{O,k}$ as means) by the two teams in the games. Similar to Scarf and Yusof (2011), we assume independence between the goals scored by the two teams. We estimated the model by employing a data set containing the results of every Champions League match played between seasons 2002–03 and 2014–15. This is the period

since the current 1G-KO structure of the tournament was put in place. The parameter values β in Eq. (2) were sampled from the posterior distribution using 10,000 MCMC iterations, with 5,000 iterations to burn in the chain and thinning to reduce autocorrelation.

We ran two chains of metropolis samplers and plotted their traces. We can clearly see from the following plot that the chains are well mixed denoting that the sampler is exploring the space of the posterior distribution of each parameter efficiently. From the samples obtained from the posterior distribution, we estimated the density function using Kernel density estimation and get an idea about how each parameter is distributed.



We report the mean and standard deviation of the parameter from the samples obtained from the above samples obtained.

Parameter	mean	s.d.
β_{A_T}	0.0060	0.0003
β_{A_O}	-0.0061	0.0003
β_{H_T}	0.4091	0.0215
β_{Aw_T}	0.0359	0.0239
β_{F_T}	0.2525	0.1698

Table 1: Mean and standard deviation of Posterior samples

This gives us an idea in the forecasting of individual matches. For example, The UEFA coefficient for Real Madrid is 172 and that for FC Barcelona is 165. The expected number of goals Real Madrid scores in a match with FC Barcelona at Santiago Bernabéu (home ground for Real Madrid) would then be:

$$\lambda_{\text{Real Madrid}} = \exp\{0.0060(172) - 0.0061(165) + 0.4091\} \approx 1.54$$

and that for FC Barcelona (for whom it would be an away match),

$$\lambda_{\text{Barcelona}} = \exp\{0.0060(165) - 0.0061(172) + 0.0359\} \approx 0.976$$

Thereby, to the nearest integer, the most likely result would be 2-1 to Real Madrid.

However, we are not interested in match wise forecasting but on an overall affect of the seeding regimes. This motivates us to use the estimates obtained above and thereby the probabilities of winning the different stages of the league to simulate the tournament which is being done in the next section.

5 Monte Carlo Simulation

We carried out three separate simulations of the 2019– 20 Champions League. Two of these related to the old and new seeding regimes described in Section 2 above. The third simulation assumed the use of a completely random draw for determining the

allocation of the 32 clubs to the eight initial groups. We term these three possible seeding systems traditional (denoted by O), new (denoted by N) and random (denoted by R).

In our case, the simulation includes simulations of both the group draw and the draw for the Round of 16 (and indeed the draws for the quarter- and semi-finals), as well as simulations of the evolution of winners and losers as the tournament progresses. Similarly, to mimic the tournament closely, we apply the same tie-break rules as those set down by UEFA for determining which club proceeds to the next stage if two clubs are tied on points in a group. In the subsequent knock-out rounds, ties up to and including the semi-final are two-legged. In this case, where two teams are drawing after two legs, 30 minutes' extra time is played, and if the teams are still tied after that extra time, a penalty shoot-out is used to decide the result. In this case, we determine the winner of the tie by a random process because of the small number of observations of extra time plus penalties in the data set.

In the context of the simulation of tournaments, an innovation in how we proceed is the use of MCMC estimation (described in the previous section) instead of classical estimation. This means that we obtain not only an estimation of the model parameters and their standard deviations, but also an estimation of their posterior distribution. This gives us a much fuller picture of the different values each parameter can take.We estimate the probability of a team winning the tournament, or surviving to the next round, using a Monte Carlo simulation as follows:

- Through MCMC we obtain 5000 samples for each parameter.
- For each parameter we simulate the tournament 1000 times and calculate the probability of winning and qualifying from he group stage.
- We repeat the step 2 for each 4999 values of the parameters obtained in step 1

The figure below shows the kernel density estimates of the posterior predictive distributions of the probability that each team qualifies under each of the three seeding systems: the traditional (O), the new (N) and a hypothetical random draw (R) for 16

teams out of the 32 teams. From the posterior distribution for Bayern we can see that Bayern which has the highest UEFA coefficient qualifies from the group stages with a very high probability. In case of Inter, we can see that they are a moderately good team and have done very well domestically in recent years but do not have the consistency to maintain a high enough coefficient in comparison to the likes of Bayern. They suffer from Old seeding rule but are helped by random and newer seeding rule as newer seeding technique keeps certain lower ranked teams in pot 1 which increases their chances of drawing a lower seeded team from pot 1 as compared to the old seeding rule. Finally, in case of RB Salzburg, we see that it gains most from random and new seeding rules as under the new seeding regime it qualifies in pot 1 as the winner of Austrian league and gets drawn against lower level teams from pots 2,3 and 4 which increases its chances of qualifying as compared to old seeding regime which places it in pot 3 or 4 which almost guarantees facing top teams in group stages.

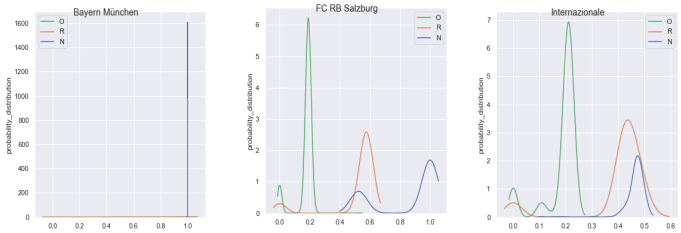


Figure 1: . Kernel density estimates of the posterior predictive density of the probability of qualifying from the group stage for Bayern,Inter and RB salzburg under the different seeding systems: O (green), N (blue) and R (orange)

From the plots below we see bimodal graphs which is due to the small sample size we had to take(because of computational constraints) while finding the distribution of probability of going to round of 16.However their interpretation is similar to that given for the three teams above.Rest of the plots are given in the link shared in the bibliography.

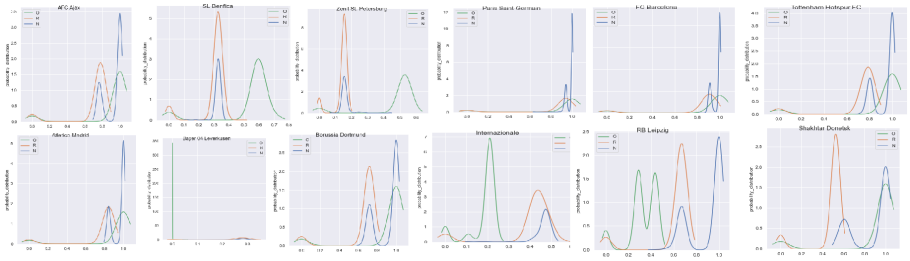


Figure 2: . Kernel density estimates of the posterior predictive density of the probability of qualifying from the group stage for 16 teams under the different seeding systems: O (green), N (blue) and R (orange)

Using a more formal approach, we can capture the overall effect on the uncertainty of outcome (in terms of which club will win the competition). Figure below shows the densities of the entropies for each of our three seeding regimes. The density associated with the random draw regime is more inclined to the right than those for either the traditional or new seeding systems, implying that failing to seed at all will increase the uncertainty as to which competitor will win the tournament. The density of the entropy under the new seeding system is also slightly to the right of that under the old system, again suggesting that it is likely that there is slightly higher uncertainty under the new system than the old. But perhaps entropy does not capture the full story about the competitive balance. The audience of the competition may be interested mainly in the very strongest clubs, which alone have a realistic chance of winning the tournament. It has been argued that the first (group) stage lacks interest in that there is too little risk of the most famous clubs failing to advance.

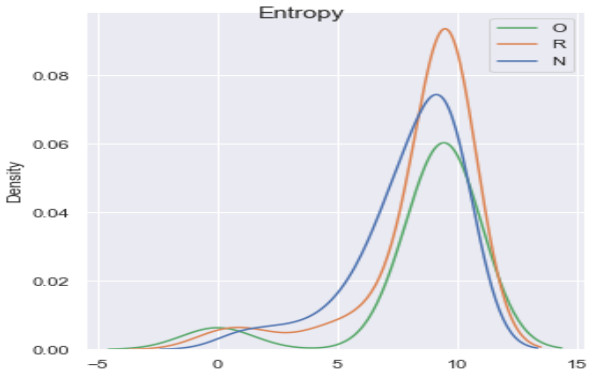


Figure 3: Predictive distributions of the entropy of the distribution of the probability of the competition winner: O (green), N (blue) and R (orange).

6 Discussion

In our project, we have successfully reproduced the results from the paper by Corona et al. (2019). The purpose of the paper was to forceast results of UEFA Champions Leauge for 2015-16 season and compare performance of different teams under alternative seeding rules. We use the model used in their paper, "Bayesian Poisson Model" but with a slightly different prior and observed similar results. Post estimation of parameters of our model, we used them to simulate the 2019-20 UEFA CL tournament and compare performances of different teams under various seeding rules. We observed that weak and moderately good teams benefited the most from the new and random seeding rule. In particular, European teams that play in some of the weaker leagues but are champions of their domestic league benefited from the new rule. This is true because due to the new seeding rule they will be assgined to pot 1 and will have low chances of facing strong teams in the group stage. However, earlier in the old seeding rule, they were kept in pot 2 or pot 3 as they have a low UEFA coefficient. These observations show that the seeding rules did indeed affect the chances of different teams qualifying from the group stage to the knock out rounds.

We also beyond the work done in paper and calculate the probability of winning the tournament for various teams. We found that Bayern Munchen had the highest mean posterior probability of winning the tournament (0.34). This alligns with the actual

result, where Bayern Munchen won the 2019-20 Champions League tournament.

7 Acknowledgements

We take this opportunity to heartily thank our supervisor Prof. Arnab Hazra for his valuable feedback and constant guidance on this project.

8 Contributions

- M. Mishra - Finding the paper, writing code for obtaining posterior distribution of parameters (50%) and preparing the presentation (40%).
- R. Vashistha - Writing code for for obtaining posterior distribution of parameters (50%), preparing report (10%) and presentation (40%).
- S. Grover - Finding data and writing the code for simulating the 2019-20 tournament.
- S. S. Banerjee - Provided theoretical inputs regarding Bayesian Poisson Model and its convergence, Preparing the report (90%) and the presentation (20%).

9 Supplementary Material

The interested reader is directed to Github which contains all the figures present here in the directory `images` and the corresponding codes to generate them in the `R` directory.

References

Corona, F., De Dios Tena Horrillo, J., and Wiper, M. P. (2017). On the importance of the probabilistic model in identifying the most decisive games in a tournament. *Journal of Quantitative Analysis in Sports*, 13(1):11–23. Cited By :1.

Corona, F., Forrest, D., Tena, J., and Wiper, M. (2019). Bayesian forecasting of uefa champions league under alternative seeding regimes. *International Journal of Forecasting*, 35(2):722–732.

I., H. (1997). The increasing competitive balance in major league baseball. *Review of Industrial Organization*, 12(3):373 – 387. Cited by: 66.

Martín, A., Cechich, A., and Rossi, G. (2011). Accessibility at early stages: Insights from the designer perspective. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*, W4A '11, New York, NY, USA. Association for Computing Machinery.

Scarf, P. A. and Shi, X. (2008). The importance of a match in a tournament. *Computers & Operations Research*, 35(7):2406–2418. Part Special Issue: Includes selected papers presented at the ECCO'04 European Conference on combinatorial Optimization.

Scarf, P. A. and Yusof, M. M. (2011). A numerical study of tournament structure and seeding policy for the soccer World Cup Finals. *Statistica Neerlandica*, 65(1):43–57.

Sherlock, C., Fearnhead, P., and Roberts, G. O. (2010). The Random Walk Metropolis: Linking Theory and Practice Through a Case Study. *Statistical Science*, 25(2):172 – 190.