Analyzing group stages of football tournaments: using the bivariate Poisson model for goal difference

Yuan He

Abstract

This paper focuses on group stage matches and ranking rules of football tournaments. Data comes from 960 games played in the UEFA Champions League group stages. After initial analysis, the games are modeled using the bivariate Poisson model. The response is the goal difference between the home and away teams. Simulations based on the model are then used to discuss some ranking rules.

1 Introduction

Football (or in American English, soccer) has always been one of the most popular sports worldwide and data-driven analysis of football games becomes increasingly popular in the recent years. Among football analytics, statistical predictive modeling is the jewel in the crown for its capability of predicting game outcomes. Accurate predictions, on the one hand, will allow football-related organizations to improve their performance and thus maximize the profit; on the other hand, it helps individuals in betting.

However, it is proved hard to produce highly accurate predictions because so many factors can possibly influence the outcome of a football game: team strengths, referee, crowds, weather, injuries and randomness itself. Despite the difficulties, statisticians have come up with various models over the years. The fundamental paper by Lee (1997) proposed to use the Poisson distribution to model the number of goals scored by each team. Brillinger (2006, 2009) modeled the Norwegian Premier League and the Chinese Super League, respectively, using ordinal-valued (win-draw-loss) responses along with the Poisson model. Karlis and Ntzoufras (2003) extended the bivariate Poisson model to incorporate some time-dependent effects. Their 2008 paper used the goal difference, instead of goals scored by each team, as the response. By modeling the goal difference

using the Skellam's distribution (or Poisson difference distribution), Karlis and Ntzoufras claimed that the effect of correlation between the two competing teams was removed. This paper will use the same model (based on Skellam's distribution) to predict outcomes of group stage games of football tournaments.

Among various tournaments in European football, the UEFA Champions League (ECL) never fails to be the spotlight since late 1990s. Top teams from major European leagues qualify for this tournament and the group stage consists of 32 teams (ever since 1999-2000 season) divided into 8 groups. The 32 teams will be firstly divided into 4 seeding pots according to their UEFA club coefficients (calculated based on the club's performance in recent European competitions), and then each group will have 4 teams, with one team from each seeding pot. In other words, each group from the ECL group stage will have one tier-1 team, one tier-2 team, one tier-3 team and one tier-4 team. Each team in the group will play against each one of the other teams, twice, home and away. So there will be a total of 12 games in each group.

In this paper, match results were recorded for each of the ECL group stages from season 2005/06 to season 2014/15 (a duration of ten years). The model for goal difference was fitted on these results and parameters of the model were estimated using the data. Simulations based on the model were then used to discuss some of the ranking rules of ECL group stages.

The structure of this paper is as follows: Section 2 describes the raw data and provides some exploratory analysis. Section 3 describes the model for goal difference, along with methods for parameter estimation and prediction algorithms. Section 4 assesses the model and produces simulations to discuss ranking rules. Finally section 5 discusses possible problems and future developments.

2 Data Description & Exploratory Analysis

2.1 Data Description

Data was collected from Wikipedia entries of each season's (from 05/06 season to 14/15

season) ECL group stage. Game results were manually entered into a spreadsheet, and then imported into R for visualization and analysis (it was rather a pleasure, not torture, for a football fan to record and savor the results). A glimpse of the raw data can be shown in Figure 1:

Year	Group	Home_name	Home_seed	Away_name	Away_seed	Home_score	Away_score	Score_diff
2015	A	Olympiacos	3	At.Madrid	1	3	2	1
2015	A	Juventus	2	Malmo	4	2	0	2
2015	A	Malmo	4	Olympiacos	3	2	0	2
2015	A	At.Madrid	1	Juventus	2	1	0	1
2015	A	At.Madrid	1	Malmo	4	5	0	5
2015	A	Olympiacos	3	Juventus	2	1	0	1
2015	A	Malmo	4	At.Madrid	1	0	2	-2
2015	A	Juventus	2	Olympiacos	3	3	2	1
2015	A	At.Madrid	1	Olympiacos	3	4	0	4
2015	A	Malmo	4	Juventus	2	0	2	-2
2015	A	Olympiacos	3	Malmo	4	4	2	2
2015	A	Juventus	2	At.Madrid	1	0	0	0
2015	B	Liverpool	3	Razgrad	4	2	1	1
2015	B	R.Madrid	1	Basel	2	5	1	4
2015	B	Basel	2	Liverpool	3	1	0	1

***Figure 1**: A glimpse of raw data.*

The raw data has 960 rows with each row representing the information of one game. The fields (columns) of raw data are pretty straightforward: "Year" and "Group" record the season and group label of each game; "Home_name" and "Home_seed" record the club name and seeding pot of the home team, same for "Away_name" and "Away_seed"; "Home_score" and "Away_score" are the goals scored by the home team and away team, respectively; Lastly, "Score_diff" is simply "Home_score" minus "Away_score" (a value of "2" means the home team win by 2 goals; "-1" means the home team lose by 1 goal and "0" indicates a draw).

2.2 Exploratory Analysis

We are particularly interested in the last column of raw data since the response of our model is goal difference. A histogram visualization of the goal differences of these 960

games is shown in Figure 2:

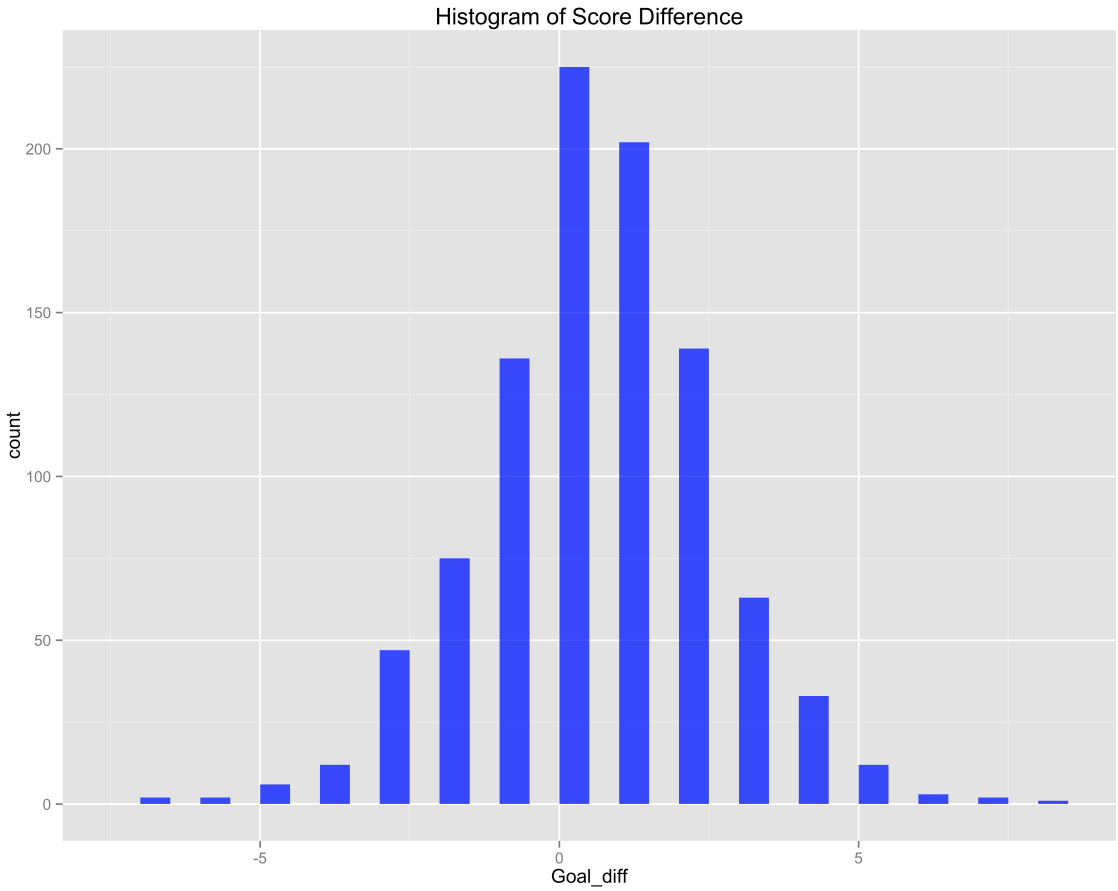

Figure 2*: Histogram of goal difference of 960 ECL group stage matches. Skewness towards the right indicates home advantage.*

It can be seen from the histogram that for most matches, the goal difference centers around 0 and lies between the interval of -5 to 5. The distribution is significantly skewed towards the right (positive direction), indicating the presence of home advantage. In fact, the average of goal difference of these 960 games is 0.385. In other words, disregarding all the tier difference, the home team scores 0.385 more goals in general, a pretty revealing number addressing the home-court advantages.

What about the seedings? It is expected that teams from higher seeding pots (or higher tiers) will perform better than teams from lower seeds. And in fact this is the truth, as revealed in Figure 3.

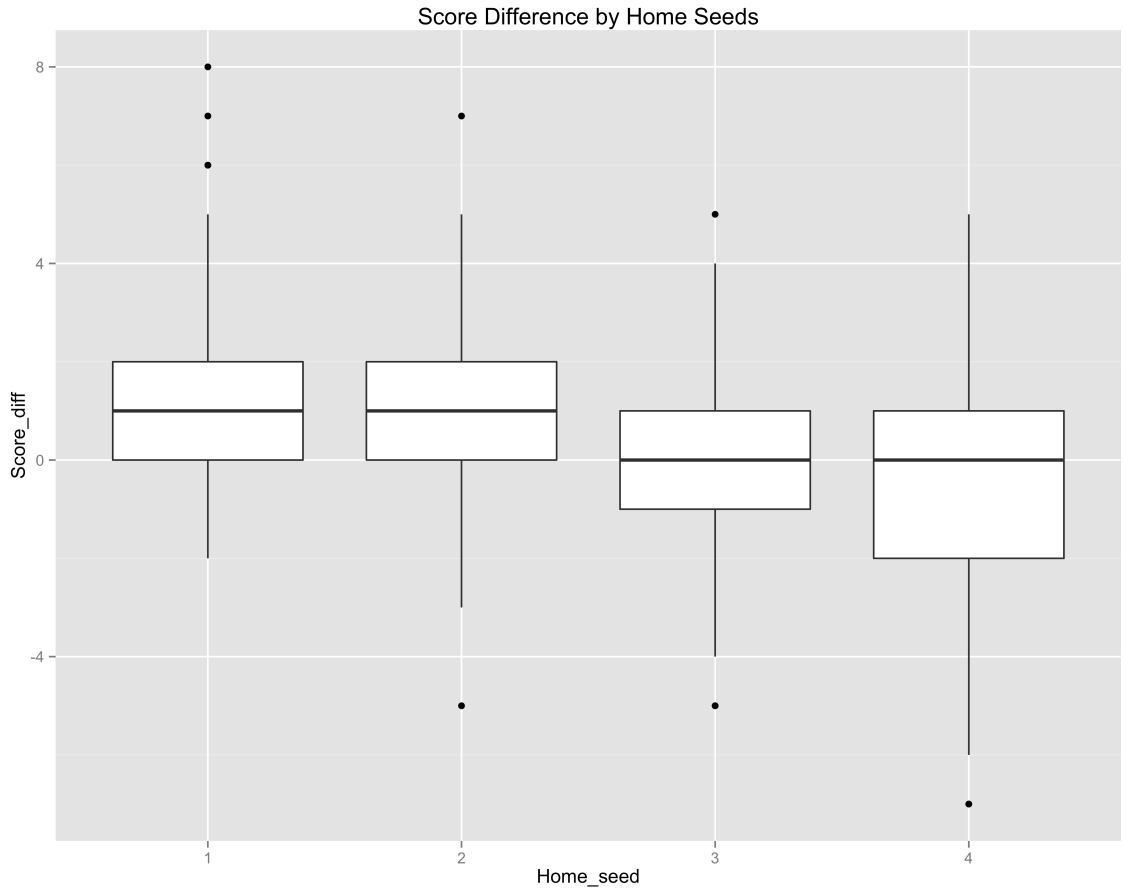

Figure 3*: Boxplots goal difference by home seeds/tiers. Teams from tier-1 and tier-2 have goal differences way above zero.*

Moreover, Table 1 gives the average goal difference for each pair of match-up. It can be seen that tier-1 teams have advantages for each possible pair of match-up, no matter home or away. The biggest average goal difference appears when tier-1 teams play tier-4 teams at home, where clubs from seeding pot 1 win by 2 goals on average. The closest match-up seems to be tier-3 teams playing tier-2 teams at home, which is reasonable considering the closeness of two tiers and the home advantages.

Away / Home	1	2	3	4
1	-	1.05	1.38	2.00
2	-0.08	-	0.56	1.18
3	-0.63	-0.01	-	0.85
4	-1.05	-0.77	0.11	-

***Table 1**: Pairwise average goal difference.*

3 The Bivariate Poisson Model of goal difference

3.1 Model Description

As Karlis and Ntzoufras (2008) proposed, consider two independent Poisson random variables X and Y and their difference Z = X – Y. Then Z is a discrete random variable defined on the set of {…, -2,-1,0,1,2,…}. In our case, Z is the random variable describing goal differences of individual games.

The distribution of Z (called the Skellam's distribution or Poisson difference distribution) was discussed by Skellam in 1946. We say that Z follows Skellam's distribution with parameters λ_1 and λ_2 (these are Poisson rates of X and Y, respectively) if X and Y are independent Poisson random variables with different means ($\lambda_1 \neq \lambda_2$). In notation, $Z \sim PD(\lambda_1, \lambda_2)$ with density function:

$$f_{PD}(z|\lambda_1,\lambda_2) = P(Z = z|\lambda_1,\lambda_2) = e^{-(\lambda_1+\lambda_2)} \left(\frac{\lambda_1}{\lambda_2}\right)^{z/2} I_{|z|}(2\sqrt{\lambda_1\lambda_2}) \qquad (3.1)$$

where $\lambda_1, \lambda_2 > 0$ and $I_r(x)$ is the modified Bessel function of order r.

For each individual game i, we have the goal difference Z_i as

$$Z_i = X_i - Y_i \sim PD(\lambda_{1i}, \lambda_{2i})$$

where i = 1,2,…,n and n is the number of games. For the Poisson parameters λ_{1i}, λ_{2i}, the bivariate Poisson model could be used (see Lee, 1997; Karlis and Ntzoufras, 2003):

$$\lambda_{1i} = exp\{\mu + H + A_{home\_i} + D_{away\_i}\} \qquad (3.2)$$

$$\lambda_{2i} = exp\{\mu + A_{away\_i} + D_{home\_i}\} \qquad (3.3)$$

where μ is a fixed constant, H is the home advantage factor, A_K and D_K are attacking

and defending coefficients of team K. For each game i, λ_{1i} depends on the attacking coefficient of the home team and the defending coefficient of the away team; the order is reversed for λ_{2i}.

For model interpretability, sum to zero constraints are made to A_K and D_K. In other words:

$$\sum_{k=1}^{K} A_k = \sum_{k=1}^{K} D_k = 0 \quad (3.4)$$

where K is the number of teams in the tournaments. In our case $K = 4$ because for a general ECL group, there will be 4 teams competing in it.

All the parameters in this model are easy to interpret and carry certain meanings. H reflects home advantage and is the expected goal difference if both teams have the same defense and offense abilities. A_K and D_K measure a team's attacking and defending abilities comparing to average performance. It can be easily seen that positive A_K indicates above-average offense and larger A_K means better offense. On the other hand, positive D_K indicates below-average defense and larger D_K means poorer defense.

3.2 Parameter Inference

The set of parameters for the ECL group stages can be listed as follows:

$$\boldsymbol{\Theta} = (\mu, H, A_2, A_3, A_4, D_2, D_3, D_4)$$

Note that according to (3.4) we can calculate A_1 and D_1 easily with the zero sum constraints.

To estimate the parameters using training data (i.e. the 960 games of ECL group stage), we used the maximum likelihood estimate (MLE). More specifically, individual probabilities could be calculated by substituting the values in $\boldsymbol{\Theta}$ into the equations (3.1), (3.2) and (3.3). Then the negative log likelihood (which is the summation of negative log terms of individual probabilities) was minimized to solve for the parameters. In notations, if we denote the individual likelihood in (3.1) as $L(z_i|\boldsymbol{\theta})$, then the negative log likelihood is defined as:

$$NLL = -\sum_{i=1}^{n} \log\left(L(z_i|\boldsymbol{\theta})\right)$$

where n is the number of games. Also the parameters can be estimated using:

$$\widehat{\boldsymbol{\theta}} = argmin_{\theta}(NLL)$$

Confidence intervals of estimated parameters were obtained using the estimated standard errors from the resulting Hessian matrix of MLE. Table 2 provides a list of estimated values for each parameter:

Parameter	μ	H	A_1	A_2	A_3	A_4	D_1	D_2	D_3	D_4
Value	0.195	0.262	0.196	0.126	-0.14	-0.18	-0.37	0.027	0.045	0.300

Table 2*: list of estimated parameter values.*

It can be seen from table 2 that tier-1 teams are the best both offensively and defensively. In fact, the defensive performance of tier-1 teams is so good that D_1 is the only defending coefficient that falls below zero, which stands for sub-par defensive performance. Figure 4 & 5 further visualizes the offensive and defensive performance of each seeding pot by showing the confidence interval and the observed values of average goals scored (conceded) per game.

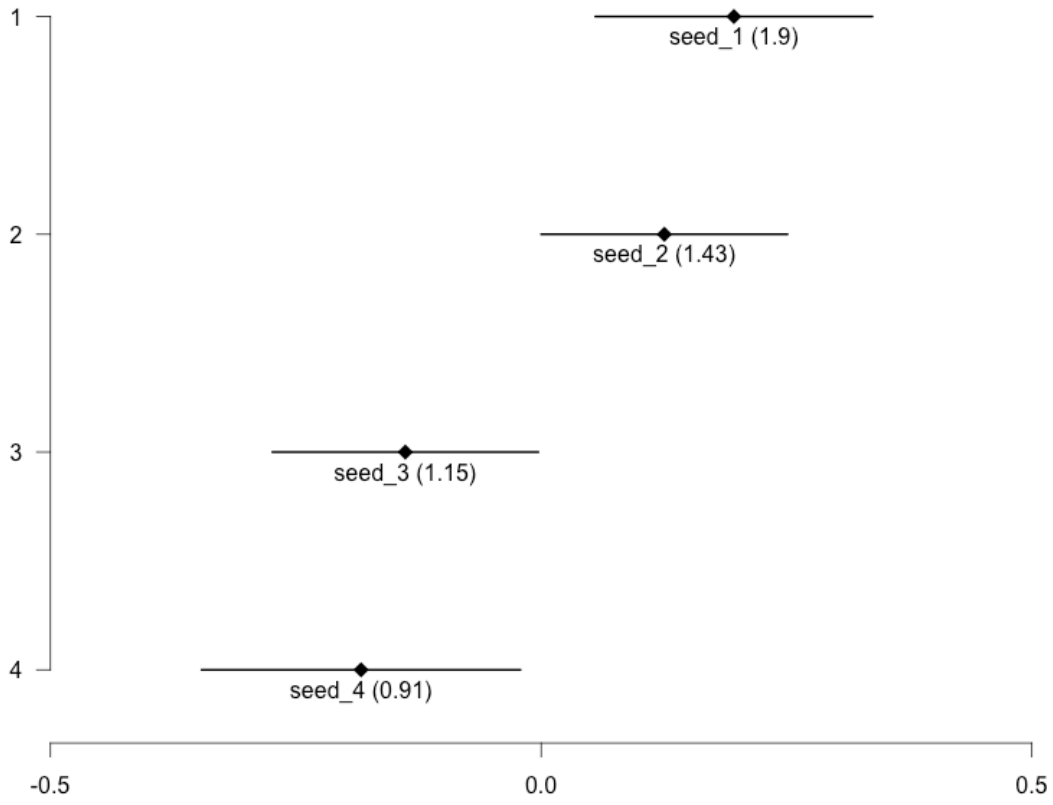

Figure 4*: 95% confidence interval for attacking coefficients (A_i). Numbers within parenthesis are observed values of average goals scored per game.*

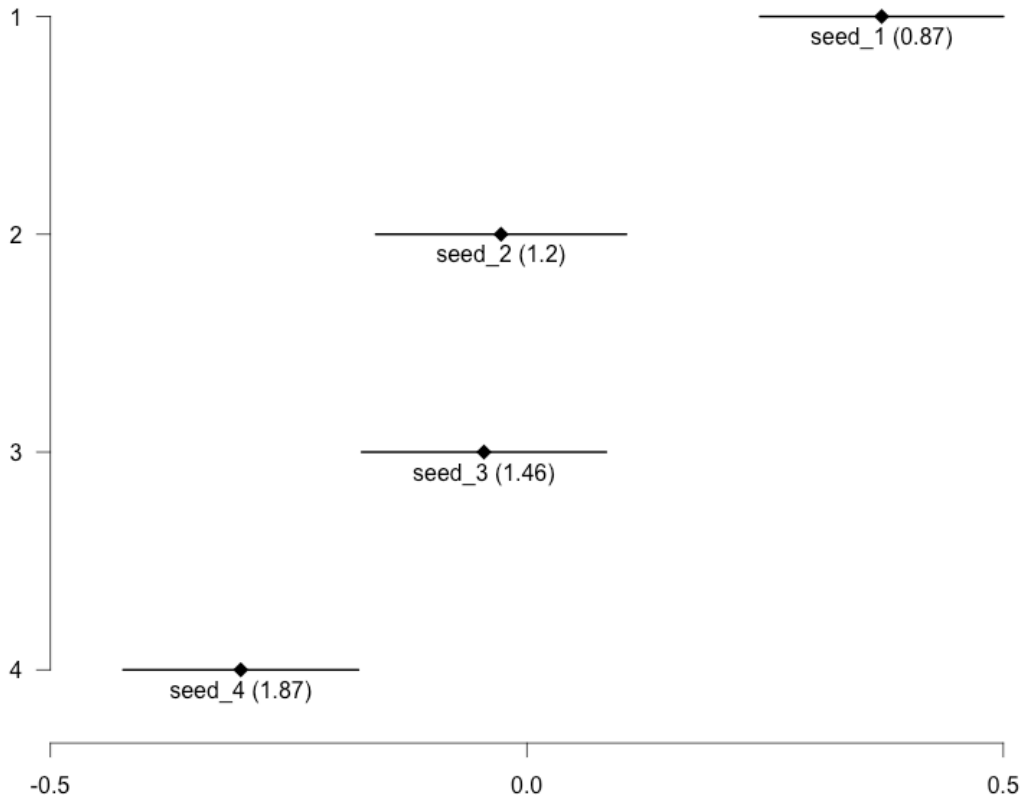

***Figure 5**: 95% confidence interval for defending coefficients (-D_i). Numbers within parenthesis are observed values of average goals conceded per game.*

3.3 Predicting future games

The following mechanism was used to predict future group stage games:

- Calculate λ_1^{pred} and λ_2^{pred} based on the estimated parameters (pick the corresponding A_i and D_i according to the seeding pots of home and away teams), see (3.2) and (3.3).
- Generate x^{pred} and y^{pred} from Poisson distributions with parameters λ_1^{pred} and λ_2^{pred}, respectively.
- $z^{pred} = x^{pred} - y^{pred}$ will be the generated prediction of the goal difference.

The generation process in the second step introduces randomness. Therefore the outcomes of predicted games will fluctuate slightly, even though the home and away seeding pots remain the same.

With the estimated coefficients and the predicting mechanism, we can simulate outcomes of each individual match-up in a general group. As a result, simulating ranking tables and deciding qualifying teams become approachable.

4 Application: Discussing rules of ECL group stage

4.1 Model Assessment

The bivariate Poisson model works as follows:

- For a given match in an ECL group, pick the home team seeding pot and the away team seeding pot (from {1, 2, 3, 4}).
- Find the corresponding attacking and defending coefficients according to the seeding pots.
- Generate predicted result from the model, based on learned parameter.

The model can be assessed by simulating match-ups of certain pairs and comparing the simulated results with observed results.

Firstly, the scenario of tier-3 teams playing tier-2 teams at home was simulated. According to Table 1, this match-up has the closest goal difference. Figure 6 visualizes the distribution of simulated results of 80 games (there were 80 games observed for each pair of match-up), comparing to the observed distribution. It could be seen that the model performed well, producing projected distributions very similar to the observed distribution in this case.

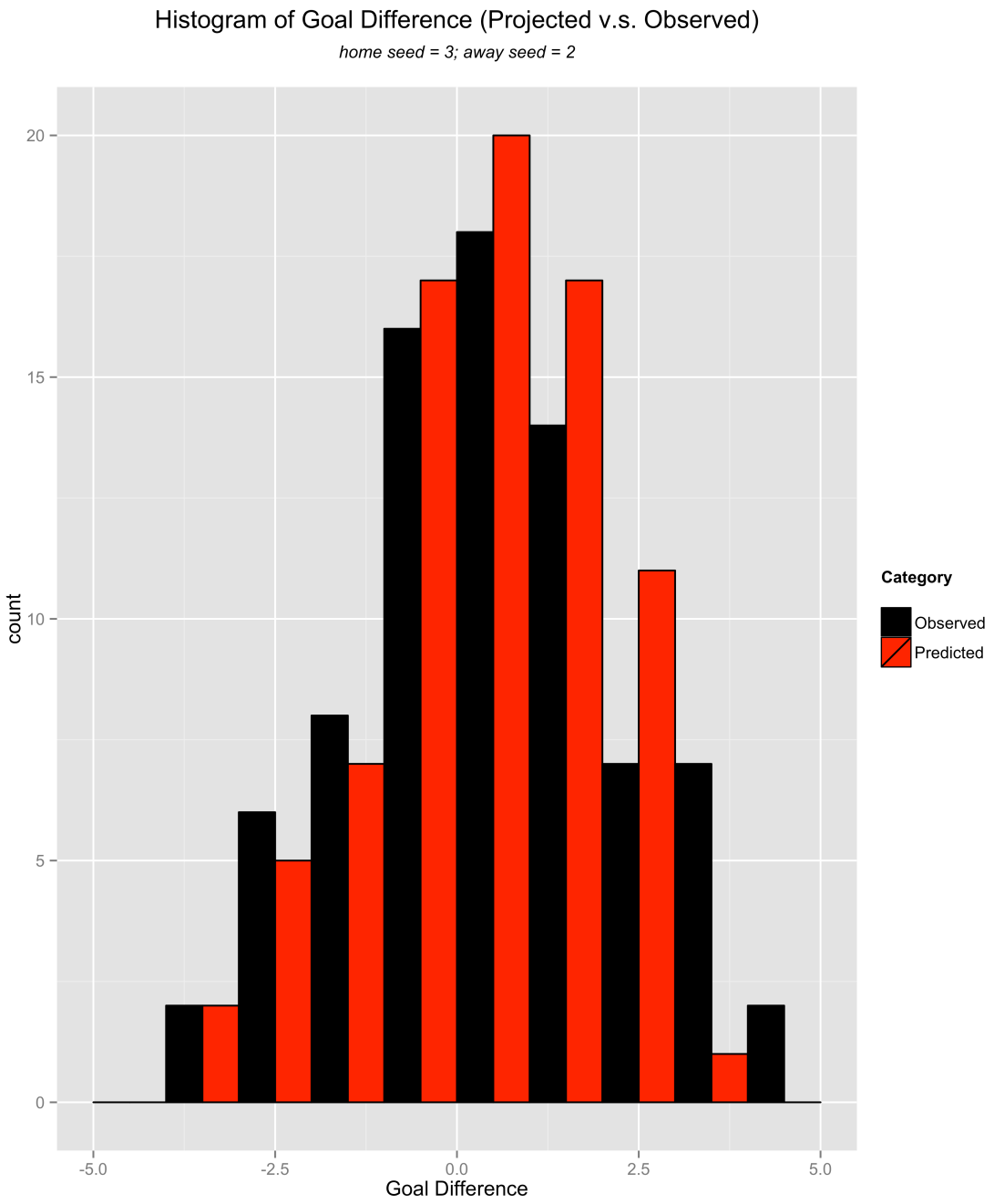

***Figure 6**: Distribution of simulated 3-2 matchups. Black bars are observed results and red bars are estimated results.*

Secondly, the scenario of tier-1 teams playing tier-4 teams at home was simulated. This is the match-up with biggest goal difference, according to Table 1. Figure 7

visualizes the simulated distribution against the observed distribution.

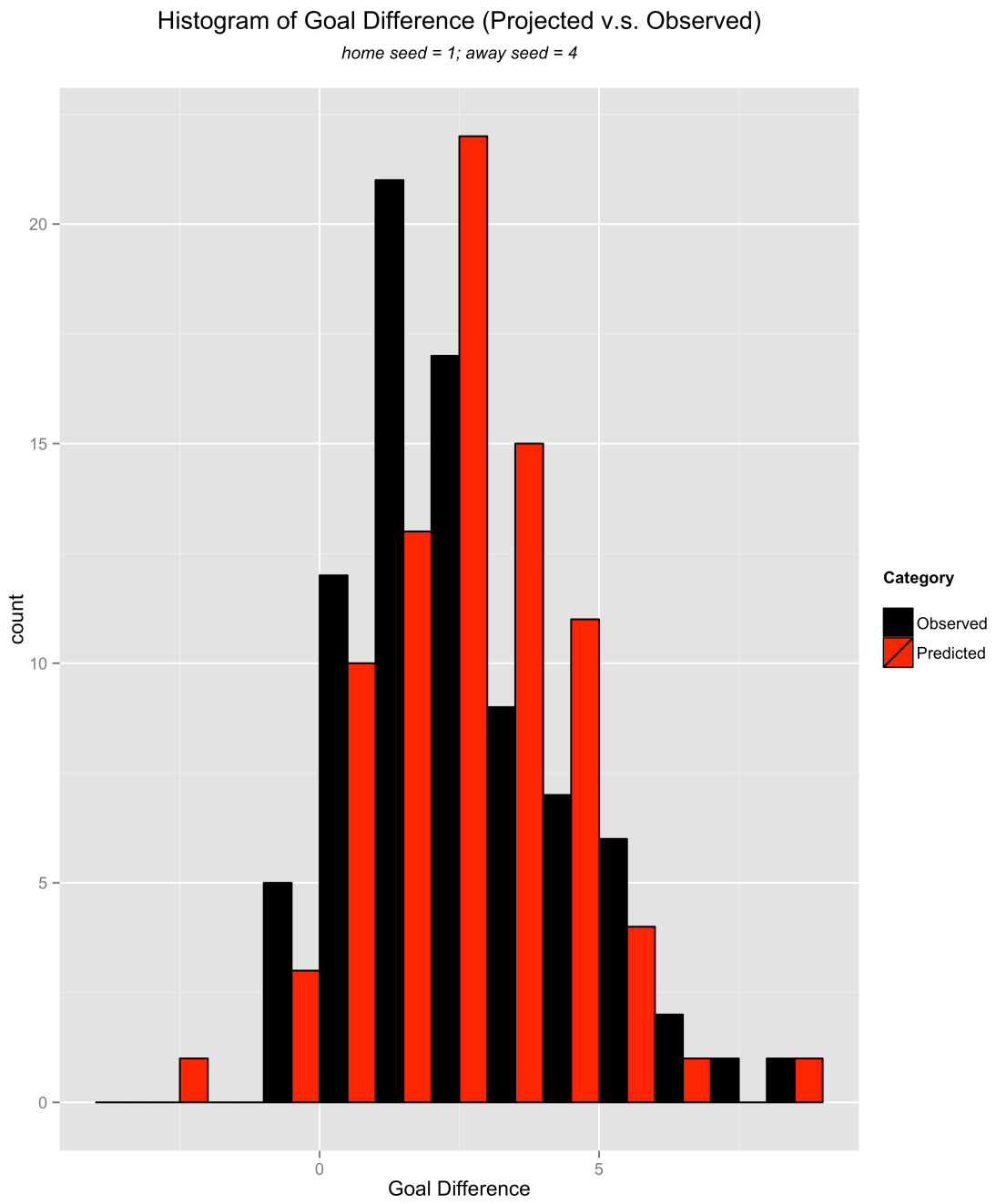

***Figure** 7: Distribution of simulated 1-4 matchups. Black bars are observed results and red bars are estimated results.*

The above histogram illustrate that the projected outcomes tend to have higher (more

positive) values comparing to the observed outcomes in this case, which indicates the model tends to overestimate the goal difference, even in the upper tail of observed values.

Other pairs of match-ups were also simulated and evaluated, and the overall performance of the model is acceptable.

4.2 Some background of group stage ranking rules

How to determine the ranking of a team in a given group? Various ranking rules have been developed in different football tournaments. But the universal priority rule will be the number of points achieved by a team. Generally speaking, a team gets 3 points for a win (in some rare cases, 2 points), 1 point for a draw and 0 point for a loss. And teams with higher points rank higher than teams with lower points. In a group of four teams, the two teams with higher ranks will **qualify** for the next rounds.

However, it is not unusual for two teams to have the same number of points. To break such a tie and determine which team ranks higher, the rule of **tiebreakers** will be applied. Different tiebreakers are being used in different tournaments, with the popular examples of goal difference, total number of goals and match-up history of teams in question. In real life, a number of tiebreakers usually apply altogether with a priority emphasis. In ECL, for example, ties are determined first by match-up history (with an emphasis on results and away goals), then by goal difference or total number of goals if the match-up history fails to tell the two teams apart.

The next two sections will talk about advantages and disadvantages of certain ranking rules, based on simulations using the Skellam model.

4.3 Discussion: number of points per win

The question is: should a win worth 3 points or 2 points? To answer this question, we simulated the general group stage 10,000 times with a win worth 2 points (rule #1), and another 10,000 times with a win worth 3 points (rule #2). With the main tiebreaker being goal difference, outcomes of these groups were compared, with regard to different rules.

First of all, we noticed that it resulted in a tie 41.55% of the times in groups with rule #1. As a comparison, only 32.1% of the times does a tie occur in groups with rule #2.

As for qualifications, we counted the qualifying ratio (QR = #qualify/#participation) of teams from each tier for both rules. Table 3 summarizes the resulted ratios. It can be seen that the results are very much similar, despite the difference in rules.

To summarize, both rules provide similar qualifying results. However, rule #2 results in less ties, thus less disputes. Therefore, 3 points for a win might be preferred for it makes the rankings more distinguishable.

Team Tier	1	2	3	4
QR (rule #1)	90.64%	63.84%	32.87%	12.65%
QR (rule #2)	90.40%	63.44%	33.47%	12.59%

Table 3*: Qualifying ratios of team tiers, with rule #1 and rule #2 being applied respectively.*

4.4 Discussion: choice of tiebreakers

Consider the following two tiebreaker rules:

Rule A: Teams with higher goal difference will be ranked higher;

Rule B: Teams that have advantages in match-up history with other teams in question will be ranked higher.

For rule B, ECL takes account of the number of away goals in match-up history. However, the bivariate Poisson model predicts the goal difference instead of goals scored by both team. Therefore, the simulations in this part will not consider away goals.

The question is: what are some characteristics of each rule? If only one rule will be picked as the main tiebreaker, which rule should we choose?

Generally speaking, a good tiebreaker should be able to break the tie effectively without undermining the fairness. To make quantitative comparisons, a total of 1000 general ECL groups were simulated. 339 ties were detected in the simulated pool (note that there might be more than one tie in each group). Both rule A and rule B were applied to each of the 339 ties, respectively, without the presence of other tiebreakers. Statistics regarding tie-breaking success rates and winning ratios of underdogs (i.e. teams from lower tiers) were recorded.

Out of the 339 ties of simulated ECL groups, rule A solved 89.97% of the ties while rule B solved 75.22% of the ties. Under conditions where rule A successfully broke a tie, teams from lower tiers won the duel **24.92%** of the times. The number, on the other hand, was **44.13%** with rule B being applied.

It can be seen that rule A breaks the tie more effectively while rule B arguably gives the underdogs some advantages. This is understandable because while rule A takes look at the overall performance in the entire group stage, rule B puts more emphasis on certain match-ups. The variance, or randomness, involved under rule B should be higher. For teams from lower tiers (usually with lower strengths), the variance works in their favor.

The choice between rule A and rule B can be tricky. While rule A seems more "fair" because top teams ought to qualify more often, people do enjoy the randomness in sports and stories of underdogs.

In our simulations, rule A and rule B were applied as the sole tiebreaker. In real life, however, multiple tiebreakers are always being applied together. With the consideration of away goals, rule B should be able to solve a tie easily more than 90% of the times. The same for rule A combined with the total number of goals. Simulations regarding a sequence of multiple tiebreakers are far more complex and are beyond the scope of this paper.

5 Concluding remarks

5.1 Problems of the model

There are potential fallbacks of the bivariate Poisson model for ECL group stage.

First of all, the model tends to overestimate the goal differences in certain conditions, as suggested in Figure 7. It might be a good idea to apply some penalty on λ^{pred} so that the Poisson process generates more reasonable results. However, the magnitude and form of the penalty term needs to be carefully studied and tuned.

Secondly, team tiers (seeding pots) are solely determined by the club coefficient, which may not be a true reflection of a team's strength at times. For example, clubs may get huge capital injection and suddenly become competent (examples like Chelsea,

Manchester City and Malaga). Clubs like these will still be grouped as a tier-4 team (for lack of performance in previous campaigns), despite the fact that their squad may be even better than a tier-1 team. An alternate way to determine team tiers is to find and compare the total market value of players from each team. In modern football, more money usually means more squad depth.

Lastly, (3.2) and (3.3) only capture limited factors that may influence a match's outcome. Other factors (moral difference, injuries, weather, to name a few) are not reflected in the formulae. It may be a good idea to try to incorporate additional factors into the model, yet some factors are hard to quantify and model complexity can be tricky to deal with.

5.2 Future developments

Section 3.2 utilized the frequentist's approach (MLE) to estimate model parameters. An alternative Bayesian approach could be used for model inference. For example, Karlis and Ntzoufras (2008) used Markov Chain Monte Carlo (MCMC) on the posterior distribution of parameters to solve for parameters. This approach may outperform MLE given informative priors.

In fact, a combination of MLE and MCMC can be used for inference with MLE serves as the "pilot round". Firstly, use MLE to get the rough distributions of each parameter; then apply MCMC with normal priors centered at the maximum likelihood estimates.

The bivariate Poisson model incorporates effects of home advantage and can thus be used for tournaments like European leagues and ECL where there are home and away games. However, some major tournaments (World Cup, for examples) have matches on neutral grounds without the distinction of home and away. To simulate these tournaments, one can remove the parameter H and reapply the model.

References

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

2 Karlis D, Ntzoufras I. Analysis of sports data using bivariate Poisson models. *The Statistician*, 52: 381–393 (2003)

3 Karlis D, Ntzoufras I. Bayesian modelling of football outcomes: using Skellam's distribution for the goal difference. *IMA J Manag Math*, 20(2): 133–145 (2009)

4 Brillinger D R. Modelling some Norwegian soccer data. In: *Advances in Statistical Modelling and Inference*. Nair V J, ed. New Jersey: World Scientific, 2006, 3–20

5 Dixon, M.J. and Coles, S.G. Modelling association football scores and inefficiencies in football betting market. *Applied Statistics*, 46, 265-280 (1997)

6 Brillinger D R. An analysis of Chinese Super League partial results. *Science in China Series A: Mathematics*, 52(6): 1139-1156 (2009)

7 2014-15 UEFA Champions League Group Stage. Wikipedia (2015). //en.wikipedia.org/wiki/2014-15_UEFA_Champions_League_group_stage