University of Duisburg-Essen
Faculty of Business Administration and Economics
Chair of Economeics

"Stay home" and let the simulation play

Predicting the Outcome Kreisliga A Reklinghausen Season 2019-2020

Working Paper

Submitted to the Faculty of
Economics
at the
University of Duisburg-Essen

from:

Nils Paffen, David Schulze

Reviewer: Prof. Dr. Christoph Hanck

Deadline:

Name:	David Schulze	Nils Paffen
Matriculation Number:	–	3071594
E-Mail:	david.schulze@rgs-econ.de	nils.paffen@stud.uni-due.de
Study Path:	–	M.Sc. Economics
Semester:	5^{th}	2^{nd}
Graduation (est.):	–	Winter Term 2021

Contents

List of Figures

List of Tables

List of Abbreviations

1 Introduction

The Covid-19 epidemic forced sports leagues in Germany to suspend championships that were already in full swing. For example, the Kreisliga Herren Klasse 2 in Recklinghausen finished around 150 games, before the rest were canceled starting from Sunday 2020-03-12, leaving around 90 games left unplayed until the last planned day of the tournament on Sunday 2020-05-24. Given the distinct possibility that there won't be a chance to make up those games later, a burning question for many players and fans is naturally: What if they would have played those games? We attempt to use data on games already played from the website fussball.de to answer this question, drawing on established forecasting methods from the literature.

2 Literature

Some difficulties that have to be addressed in predicting outcomes of e.g. the FIFA World Cup, are not relevant to the Kreisliga. For example, the league system in Germany features two rounds per season. Each club plays each other club tiwce: Once in each round, and once on each club's home field. This reduces the uncertainty, when compared with the mode of the World Cup. There, in the group stage groups are determined by luck of the draw, a process known as "seeding". Groups then play a so called round-robin tournament, also known as all-play-all, were every every group member plays each other, which corresponds to the mode in which the Kreisliga plays. The World Cup then continues with single-elimination, or a knock-out stage, which introduces another even more random path dependencies. This implies that the existing literature on forecasting results in the FIFA World Cup group stage is highly relevant for the task at hand, since the game rules are identical.

A natural starting point for forecasting match or season outcomes in football tournaments is using the FIFA points ranking method that is used canonically to evaluate the strength of a team and updated after each game. For example, a recent study by Correa et al. (2018) uses FIFA points to forecast the results of the 2018 FIFA Men's World Cup. This approach has however generated criticism and other methods have been proposed and evaluated. The benchmark study by Laseck et al. (2013) compares established and proposed rankings. They find that FIFA rankings used to perform slightly worse than alternative methods, especially two versions of the Elo rating system originally proposed by Arpad Elo for the United States Chess Federation to rate competitive chess players. We will consider using the FIFA points method or these two alternative candidates.

The first candidate model is published by the anonymous website EloRatings.net (2012), and is especially adapted for the use in ranking football teams. The second one is the

FIFA Women's World Rankings, which has been in use since 2003. It is worth noting that the FIFA Men's World Rankings have been adapted in 2018 to also be based on a modification of the Elo ranking.

The website FuPa.net (2020) also publishes detailed statistics as well as a so called "Power-Ranking" to evaluate the "Formkurve", a rough measure of recent performance. Because it is only reflective of very recent games, it is not useful for long term forecasting. Another raw measure would be to calculate the probability of winning by dividing team's current points (victories are 3, draws are 2) by the total of their and their opponent's points, we can call this the "points model".

All these models just use weighted game results or point scores. Another approach would be to use more data, as in Berbée et al. (2015). While this concerns baseball, the principle is not dependent on the rules of the game, but it is based on the influence of individuals' and team's statistics on game outcomes. This is an application of the model proposed by Albert and Bennet (2007). They use the fact that wins and losses appear to be normally distributed in the long run, so chance and skill should both play a role. Team ability is then calculated as the deviation from the average winning probability over time. Very generally, available statistics are then evaluated to find those that matter the most for team performance, measured in runs per game, which would be goals in football. Those candidate factors are then weighted by regressing them on the team performance. The resulting index is then used to calculate a winning probability and this used then to simulate the season outcome by using weighted coin-flips.

While there is not as much individual player data in football on the level of our analysis, the team average is still something that can be calculated. Also, more information on the games, such as location at home or away, could be candidate factors for performance. For example, one team might benefit more from playing at home than it hurts by being away, receiving a net benefit from location, or the other way around. While these data are certainly interesting, we must be aware of the danger of overfitting the model.

Finally the literature contains references to using betting markets as a benchmark, as they perform relatively well as predictors. For the Kreisliga however, we don't expect betting markets to be deep enough to reach this level of accuracy. Otherwise, they would also be an interesting reference point.

3 Data

For the simulation study we decided to use data from the amateur football league of the Kreisliga A Kreis Recklinghausen class A1 in Westphalia, further called Kreisliga A1, for the 2019/2020 season. 16 clubs will play against each other on a total of 30 match days in one home and one away game each. Due to the well known Corona crisis the association has decided to cancel all matches from 15.03.2020. After all, 20 matchdays have already been played until this point in time which corresponds to a database of 158 matches. As the first half of the season had already been completed, each team had already played at least once against each team in the table. The extraction of real data from websites using scraping scripts can be complicated, as website operators have an interest in protecting their data from such automated queries. "Fussball.de" is a website of the DFB (German Football Association) which acts as a collection point for match results and news, especially in the amateur sector. The match results of the website itself cannot be directly read out. They are masked ,so they are made unreadable when viewing the HTML file and are only evaluated afterwards using a javascript and transferred to the CSS of the site. The site also offers a match report, which graphically represents a temporal course of the match. This is broken down in the HTML code, in contrast to the match results, unmasked, and shows the course of the match in text form. With the help of regular expression operations, such as pattern matching, the game result can be reconstructed. The data record was then divided into completed and unplayed games. The latter amount to 89 in this season, which were simulated with the methods of the following chapters.

Give a short overview of the actual standing (ranking table after matchday 20)

(David: Describe calculating the values needed for the models.)

4 Predictive Models

To predict the outcome of the cancelled games, we calculate the candidate rankings and use them to simulate the end of the 2019/2020 season. Specifically we calculate the EloRankings.net Ranking, the FIFA Women's Ranking, the points model and a version of the Total Team Average incorporating additional information. We discard the classic FIFA Men's Ranking, since it was shown to perform worse than the other candidates and was discarded by the FIFA Men's World Rankings as well in 2018.

Our second model is based on the rating algorithm of The anonymous site operator formulates the rating, representative for the strength of a team, as follows :

$$R_n = R_0 + K \times (W - W_e). \tag{1}$$

Here R_n is defined as the new rating and R_o is the old rating. The weighting factor for each match is defined by the type of tournament in which the match takes place and also controls for friendly matches, which is given to the lowest weight of 20. While matches in world championships and other major international tournaments are given weights between 40 and 60, the rest falls into the category "all other tournaments" which are given a weighting factor of 30. Following this example, we also set K to 30 for matches already played in the Kreisliga A. The weighting factor K is adjusted again based on the goal difference of the result. Thus, K is increased by $\frac{K}{2}$ if the match was won with two goals, by $dfrac34 \times K$ if the match was won with three goals and by $\frac{3}{4} + \frac{(N-3)}{8} \times K$, where N defines the goal difference of the match if the match was won with four or more goals. W is the result of the match. 0 for a loss, 0.5 for a draw and 1 for a win. W_e is the probability of winning defined by the following formula :

$$W_e = 1/(10^{(-dr/400)} + 1) \tag{2}$$

where dr is defined as the rating difference and the home team receives a bonus of 100 points. This bonus is considered to be a psychological advantage resulting from the fact that the game is played in the home stadium(see, e.g., Pollard (2008)).

Our third model uses the poisson distribution to simulate the match result with the probability of a goal in every minute of a match. The probability matrix where the game result is drawn from is a $n \times n$ matrix where each cell indicates the proability of that specific match result. While the rows indicates the goals of the home team, the column indicates the goals of the away team. For example the cell of the first row and in the first column indicates the likelihood that the both teams socre 0 goals. The poisson probability function of our model can be expressed as :

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \lambda > 0 \tag{3}$$

where the lambda represents the average number of goals First we estimate the following model from the matches already played:

$$goals = home + team + opponent. \tag{4}$$

Where goals represents the number of goals scored by a team in a game, home presents a dummy variable that is 1 for the home team and 0 for the away team, team representing the home team, and opponent, team representing the opponent team.

"'{r , results = "asis"}#Does not work right now library(tidyverse) require(stargazer) library(knitr) opts_chunk$set(echo = TRUE)

database_mr <- readr::read_rds(stringr::str_c(here::here() , "data", "database_match_results_1920.r
sep = "/"))

home_goals_avg <- database_mr$goals_team_home %>% mean()

away_goals_avg <- database_mr$goals_team_away %>% mean()

poisson_model <- bind_rows(tibble(goals = database_mr$goals_team_home, team = database_mr$club_name_home, opponent=database_mr$club_name_away, home = 1), tibble(goals = database_mr$goals_team_away, team=database_mr$club_name_away, opponent = database_mr$club_name_home, home=0)) %>%

glm(goals ~ home + team +opponent, family=poisson(link=log),data=.) half <- list(coefficients =list(poisson_model$coefficients[1:17]))

stargazer(half, title = "These are awesome results!", notes="stargazer html", no.space = T) "'

In summary, the coefficients of the model show that the club "Altendorf-Ulfkotte", both as home and away club, has a strong negative and a strong positive influence, both highly significant, on goals, i.e. the number of goals. Since the club is in the last place in the current table, as mentioned in the Data section, we expected that it would be easier to score goals if they played against the team on the last place of the seasons soccer league table rather than the team on the first place. On the other hand it will be harder for this last placed team to score goals even if they are the home team.

Following Correa et al. (2018), we then run the simulation by drawing the results of each game from a binomial distribution. For each game and team, the probability of winning is dividing the ranking points awarded each team by their and their competitors sum of points.

Running this simulation repeatedly should indicate the distribution and expected average of outcomes. Correa et al. (2018) execute 200,000 runs, but because of the relatively low complexity of the Kreisliga's format compared to the World Cup, especially because there are no eleminiation rounds, we expect to need less repetitions.

A few alternatives have been developed for forecasting football games. The potential of using independent Poisson distributions to match the empirical distribution of goals scored by a team has been improved on by introducing correlation between the teams playing against one another in a bivariate Poisson distribution Karlis and Ntzoufras (2003).

While the independent Poisson distributions already allowed for a better fit and to model the outcome of draws, Boshnakov et al. (**boshnakov2016**) used a Weibull count model to improve even on the bivariate Poisson model, allowing them even to outperform betting market in selected bets.

5 Results

For the simulation study using the elo rating, as explained in the predictive models chapter,we used the average of all matches played in the current season resulting in a tie for the probability of a draw. Half of the percentage points are deducted from the home team's winning probability and half from the away team's winning probability.

6 Conclusion

References

Albert, J., & Bennett, J. (2007). *Curve ball: Baseball, statistics, and the role of chance in the game*. Springer Science & Business Media.

Correa, M., Barrera-Causil, C., & Marmolejo-Ramos, F. (2018). The next winner of the 2018 fifa world cup will be...: An illustration of the use of statistical simulation to make a prediction in a complex tournament. *Chilean Journal of Statistics (ChJS)*, *9*(1).

Fupa.net [Accessed: 2020-04-16]. (2020).

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(3), 381–393. https://doi.org/10.1111/1467-9884.00366

Lasek, J., Szlávik, Z., & Bhulai, S. (2013). The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recognition*, *1*(1), 27–46.

Lobeck, M., Schulze, D., & Berbée, P. (2015). Can he hit? measuring and prognosis of offensiveperformance by baseball teams and individual teammembers. *Seminar Paper, B.Sc. VWL HU Berlin.*

Pollard, R. (2008). Home advantage in football: A current review of an unsolved puzzle. *The Open Sports Sciences Journal*, *1*. https://doi.org/10.2174/1875399X00801010012

The world football elo rating system [Accessed: 2012-03-03]. (2012).