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

Stay home and let the simulation play

Predicting Kreisliga football league outcomes with statistical simulations

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:	PhD Economics	M.Sc. Economics
Semester:	2nd	2nd
Graduation (est.):	–	Winter Term 2021

Contents

List of Figures

List of Tables

List of Abbreviations

1 Abstract

Publicly available data and public attention are contributing to the relevance and interest in forecasting football game results. We provide a short overview of the state of the literature and use data from the aborted German local men's league season 2019-20 to predict the season's outcome using three different statistical approaches. A measure of each team's strength is calculated from past plays and used as weight in the simulation. For the football league's organizers, using a prediction algorithm to find a season outcome might be fairer than either annulling the games played thus far or using the table as of now, ignoring the missing games. Research has shown, that measures like the Elo rating system are better predictors of teams' performance than for example league table points on their own. For this data set, we find that gains from using advanced methods are marginal when evaluating them with past seasons' data. Simulation results are evaluated using an out of sample error calculated from previous seasons.

2 Introduction

The Covid-19 epidemic forced sports leagues in Germany to suspend championships that were already in full swing. For example, the local men's league Recklinghausen class A1 finished around 150 games, before the rest were canceled starting from Sunday March 12, 2020, leaving around 90 games left unplayed until the last planned day of the tournament on Sunday May 24, 2020. There was a high probability that the games would not be made up for later, which proved to be true, a burning question for many players and fans was naturally: What would the outcome of the season have been? We use data on games already played from the website fussball.de to answer this question, drawing on established forecasting methods from the literature.

The league system in Germany implies that everyone plays a set against every opponent: Once on their home field and once on the opponents field for each league participant. This means it's easier to forecast when compared with the mode of tournaments the World Cup. There in the group stage, groups are determined by chance, a process known as "seeding". Groups then play can be sorted. But the World Cup then continues with single-elimination, or a knock-out stage, which introduces even more random path dependencies that are not needed for forecasting the Kreisliga. This implies that the part of the existing literature on forecasting results in the FIFA World Cup concerning the group stage remains highly relevant for the task at hand, since the game rules are otherwise identical.

In the next part, we give an overview of models used and evaluated for the purpose

of predicting football match outcomes. We introduce a small subset of models in more detail in the third part. The fourth part contains the results from calculating a simulation based on these for the local men's league Recklinghausen class A1. We also present some comparative statistics of the model performance and draw some conclusions in the last segment.

3 Literature

A natural starting point for forecasting match or season outcomes in football tournaments is using the FIFA points ranking method that is widely used 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 McHale and Davies (2007), especially because it does not update based on new information fast enough, and other methods have been proposed and evaluated. The benchmark study by Lasek et al. (2013) compares established and proposed rankings. They find that FIFA rankings perform slightly worse than alternative methods, especially a version of the Elo rating system originally proposed by Arpad Elo for the United States Chess Federation to rate competitive chess players that was adapted for football championships by the authors of the website EloRatings.net (2012).

Other studies show the effective prediction power of FIFA rankings, e.g. Suzuki and Ohmori (2008). Leitner et al. (2010) find that bookmakers odds are more predictive than FIFA rankings. In our case we don't expect betting markets to be deep enough to make this a feasible approach, although it would be an interesting reference point. We do however adopt their use of Spearman's rank correlation between simulated and real final tournament rankings to evaluate models' performance and complement it with Kendall's tau. Lasek et al. (2013) evaluate using rating points, which are less relevant for our use case than the absolute rankings, which determine whether a team advances, stays or drops out of a league.

We consider three models for our calculation: First, a benchmark model based on the table points of each team at the time when the league was aborted. Second, an Elo rating system, and third a simple model based on the Poisson distribution.

The benchmark model calculates the probability of winning a match by dividing a 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". This model does not include the possibility of a draw. The probability is not updated with after each simulated game, because this does not generate new information about a team's strength. Averaging

the results over enough simulations, this approach will converge to the current table ranking, so it is in fact just a weighted randomization of the current table.

The second model is based on a version of the Elo rating system published anonymously on the website EloRatings.net (2012). The algorithm was originally developed for ranking chess-players. As an "earned" rating system (Lasek et al. (2013)) a team's rating is iteratively updated according to the outcome of single matches and depending on the expected outcome with regard to the opponent's rating. This version was especially adapted for the use in ranking football teams. Glickman (1995) offers a comprehensive discussion of the Elo rating system.

The third model is a very simple implementation of a Poisson distribution that approximates a probability distribution of goals in each game with a fixed parameter to adjust for the home advantage. This approach follows the literature influenced by Maher (1982) and others. Generally, these models include different parameters to allow for team-specific strengths when playing home or away, and while defending or attacking. Parameters for e.g. random effects can be added, which we omit here for simplicity. For a general discussion see Karlis and Ntzoufras (2003). Many extensions of this model as well as model selection algorithms are possible.

For a more recent review of advances in the literature and a new approach based on the Weinbull distribution we refer to Boshnakov et al. (2017a). They use an evaluation based on calibration curves as well as the payoff from betting strategies and find that their model improves on previous models and can yield positive betting returns.

4 Data

For the simulation study we decided to use data from the local men's league Recklinghausen class A1 in Westphalia 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 Covid-19 pandemic, the association has decided to cancel all matches from March 15, 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 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, the game result can be reconstructed. The data record was then divided into completed and un-played games. The latter amount to 89 in this season, which were simulated with the methods in the following chapters.

For further analysis we decided to scrape the data of season 16/17, 17/18 and 18/19 as well to perform out-of-sample error (OOSE) test statistics. The latter will indicate the performance of the different methods.

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

5 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 by way of calculating a winning probability for each missing game. Specifically we calculate

- the points model,
- the EloRankings.net model,
- a Poisson model

The first model is just a simple baseline model that calculates the probability of a team A winning a game againsta team B using the formula

$$P(Awins) = \frac{tablepoints_A}{tablepoints_A + tablepoints_B}, \tag{1}$$

where *table points* corresponds to the number of games won at the current state of the season valued at three points plus the number of draws valued at one point. This value also governs the ranking and ultimate placement of the teams in the league. Two questions arise from this approach. Firstly, should the table points be updated after each simulated game? We argue no, because this would not include new information about the relative strength of the teams and just increase the variance of the result.

Secondly, the average over many simulated runs will converge to the initial table when the season was interrupted. This will defeat the purpose of running a simulaion in the first place, because it does not yield any new information, and we could have just used the table as it were. Using this way of simulation is however preferable to an unweighted coin toss, because that would unfairly favor below average teams.

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

$$R_n = R_0 + K \times (W - W_e). \tag{2}$$

Here, R_n is defined as the new rating as an update of R_0, which 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 $\frac{3}{4} \times K$ 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:

$$W_e = 1/(10^{(-dr/400)} + 1), \tag{3}$$

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)).

To simulate the outcome of the league with the points and Elo ranking method, we follow Correa et al. (2018) and draw 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.

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 from which

the game result is drawn is a $n \times n$ matrix where each cell indicates the probability 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 score 0 goals. The maximum number of goals n can be set high enough to cover all possible outcomes. We set it to 10 in our simulation. The poisson probability function of our model can be expressed as:

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \lambda > 0 \tag{4}$$

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

$$goals \sim f(home, team, opponent) \tag{5}$$

Where *goals* represents the number of goals scored by a team in a game, *home* is a dummy variable that equals 1 if the team plays on it's home pitch, and *team* and *opponent* represent dummies for each team respectively.

In summary, the coefficients of the model show that the club "Altendorf-Ulfkotte", is least likely to score a goal (low team estimate) and teams playing against them have the highest chance to score (high opponent estimate), with both estimates being highly significant. Since the club is in the last place in the current table, as mentioned in the Data section, this is the expected result. Conversely, we observe the opposite for the current table leader "VfL Ramsdorf". For the simulation the result is drawn from the Poisson distribution, and the score probabilities are based on the estimated parameters.

Running the simulation for each method repeatedly should indicate the distribution and expected average of outcomes, after the averages converge. Correa et al. (2018) execute 200,000 runs, but because the league in question is less complex than the World Cup they analyze, especially because there are no eliminiation 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).

Table 1: Regression output of the Poisson model

	control	goals team	opponent
	(1)	(2)	(3)
Constant	0.752*** (0.228)		
homey	0.241*** (0.076)		
Adler Weseke II		−1.047*** (0.249)	0.595*** (0.218)
BVH Dorsten		−0.289 (0.199)	0.051 (0.250)
FC RW Dorsten		−0.877*** (0.232)	0.178 (0.237)
Fenerbahce I. Marl		−0.564*** (0.206)	0.109 (0.244)
SC Marl-Hamm		−0.145 (0.192)	0.507** (0.226)
SC Reken II		−0.405** (0.206)	0.697*** (0.220)
SV Altendorf-Ulfkotte		−1.252*** (0.277)	1.089*** (0.205)
SV Lembeck		−0.216 (0.196)	0.356 (0.230)
SV Schermbeck II		−0.167 (0.186)	−0.267 (0.272)
TSV Raesfeld		0.021 (0.179)	−0.085 (0.258)
TuS 05 Sinsen II		−0.902*** (0.241)	0.581*** (0.219)
TuS Gahlen		−0.266 (0.191)	−0.812*** (0.315)
TuS Velen		−0.409** (0.202)	0.280 (0.233)
VfL Ramsdorf		0.072 (0.177)	−0.435 (0.283)
Westfalia Gemen II		−0.559*** (0.210)	0.591*** (0.220)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

While the independent Poisson distributions already allowed for a better fit and to model the outcome of draws, Boshnakov et al. (2017b) used a Weibull count model to improve even on the bivariate Poisson model, allowing them even to outperform betting market in selected bets.

6 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. Then we draw from these three probabilities the game result, team home wins, team away wins or draw. We repeat this procedure for all games and evaluate the results with 3 points for the winning team, 1 point for both teams in case of a draw and 0 points for the losing team.

In the poissonmodel we calculate for each match the goal probabilities of both teams as a probability matrix based on the model estimation as described in predictive models part.

All simulations were repeated until the rate of change of the point average was 1% or less. Aggregation to this point occurred after about 2580 for the elo model and after about 1980 for the poisson model.

Table 2: Simulated Final Score Table

	Poisson Distribution Model		Elo Rating Model		Points Model	
rank	club_name_poisson	score_poisson	club_name_elo	score_elo	club_name_points	score_points
1	VfL Ramsdorf	63.46088	VfL Ramsdorf	64.34676	VfL Ramsdorf	72.19792
2	TuS Gahlen	57.54770	TuS Gahlen	58.74429	TuS Gahlen	65.04059
3	SV Schermbeck II	56.68854	SV Schermbeck II	57.74274	SV Schermbeck II	63.51325
4	Fenerbahce I. Marl	52.45432	Fenerbahce I. Marl	53.52383	Fenerbahce I. Marl	58.06307
5	1. SC BW Wulfen	49.74609	1. SC BW Wulfen	50.88454	1. SC BW Wulfen	57.22107
6	TSV Raesfeld	49.58253	TSV Raesfeld	50.64084	TSV Raesfeld	55.75243
7	TuS Velen	40.56790	TuS Velen	41.79078	TuS Velen	45.37873
8	SC Marl-Hamm	39.67895	SC Marl-Hamm	40.90701	SV Lembeck	44.14961
9	BVH Dorsten	38.56991	BVH Dorsten	39.89500	SC Marl-Hamm	43.87890
10	SV Lembeck	38.49319	SV Lembeck	39.56838	BVH Dorsten	42.64140
11	FC RW Dorsten	34.61080	FC RW Dorsten	35.82449	FC RW Dorsten	37.59242
12	Westfalia Gemen II	31.46037	Westfalia Gemen II	32.40101	Westfalia Gemen II	32.82422
13	SC Reken II	27.79455	SC Reken II	29.11275	SC Reken II	27.60483
14	TuS 05 Sinsen II	22.68854	TuS 05 Sinsen II	23.86207	TuS 05 Sinsen II	21.27809
15	Adler Weseke II	21.69107	Adler Weseke II	22.59512	Adler Weseke II	18.88125
16	SV Altendorf-Ulfkotte	18.52751	SV Altendorf-Ulfkotte	19.50446	SV Altendorf-Ulfkotte	13.98222

7 OOSE Test Statistics

Making predictions of events that might never happen can fairly criticized by a simple question. How do you know that your results reflect reality as good as possible? Following George E. P. Box who is known for his qoute “All models are wrong” which is often extenden by “but some are usefull” we want to show that our models cover the latter. The out-of-sample error test statistic is one way to achieve this. One simply divides a dataset a small test data set and a larger training data set. For the season 16/17, 17/18 and 18/19 we decided to split the dataset at the same point where the COVID-19 pandemic forced the

Following Leitner et al. (2010), we evaluate the models' performance using the rank correlation between their predicted and the real ranking tables for the three past years' seasons (2016, 2017 and 2018). To increase the relevance for our use case, we use as much training data as was available for this year's aborted season (2019-20). We find that the Elo ranking system improves on the baseline model, which in turn performs better than the simple Poisson model. The fact that the points model achieves a 1.00 correlation in the 2017-18 season however makes these results doubtful, since the points model converges to the table as it was a the point of interruption. A perfect correlation with the final table can thus only occur if there is no change in the ranking after that date.

Generally, the high correlation between the predicted and the actual table outcomes leads us to believe that adopting the results from each method would provide a fair improvement over anulling the 2019-20 season.

method	spearmans_rho	kendalls_tau
elo ranking	0.98	0.93
points	0.98	0.92
poisson	0.96	0.89

Table 3: Average rank correlation coefficients for simulation and actual data

8 Conclusion

The decision to quit all games later than 08th of march because of the pandemic was not revised while the infection rates relaxed during may and june in Germany. Combined with the unforseeble future of the COVID-19 situation we see a more fair and balanced decision making process by integrating statistical learning techniques, such as those, shown in this paper.

TO DO:

method	season	spearmans_rho	kendalls_tau
elo ranking	1617	0.96	0.88
elo ranking	1718	0.98	0.94
elo ranking	1819	0.99	0.96
points	1617	0.97	0.90
points	1718	0.96	0.88
points	1819	1.00	0.97
poisson	1617	0.94	0.85
poisson	1718	0.96	0.88
poisson	1819	0.99	0.93

Table 4: Rank correlation coefficients for simulation and actual data

-add caveats: what other factors play a role? -goal distribution, qq-plot -include reference to Blog by David Sheehan: https://dashee87.github.io/data%20science/football/r/predicting-football-results-with-statistical-modelling/ -rerun oose

References

Boshnakov, G., Kharrat, T., & McHale, I. G. (2017a). A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, *33*(2), 458–466.

Boshnakov, G., Kharrat, T., & McHale, I. G. (2017b). A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, *33*(2), 458–466. https://doi.org/https://doi.org/10.1016/j.ijforecast.2016.11.006

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).

Glickman, M. E. (1995). A comprehensive guide to chess ratings. *American Chess Journal*, *3*(1), 59–102.

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.

Leitner, C., Zeileis, A., & Hornik, K. (2010). Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the EURO 2008. *International Journal of Forecasting*, *26*(3), 471–481.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, *36*(3), 109–118.

McHale, I., & Davies, S. (2007). Statistical analysis of the effectiveness of the FIFA world rankings. In J. Albert & R. Koning (Eds.), *Statistical Thinking in Sports* (pp. 77–90). Chapman & Hall - CRC.

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

Suzuki, K., & Ohmori, K. (2008). Effectiveness of FIFA/Coca-Cola World Ranking in predicting the results of FIFA World Cup finals. *Football Science*, *5*, 18–25.

The World Football Elo Rating System [Accessed: 2012-03-03]. (2012).