African Journal of Applied Statistics
Published Article : www.jafristatap.net
DOI: http://dx.doi.org/

AJAS / SPAS

ISSN 2316-0861

Prediction Model for the Africa Cup of Nations 2019 via Nested Poisson Regression

Lorenz A. Gilch*

University of Passau, Innstrasse 33, 94032 Passau, Germany

Article published in AJAS, www.jafristatap.net

Abstract. This article is devoted to the forecast of the Africa Cup of Nations 2019 football tournament. It is based on a Poisson regression model that includes the Elo points of the participating teams as covariates and incorporates differences of team-specific skills. The proposed model allows predictions in terms of probabilities in order to quantify the chances for each team to reach a certain stage of the tournament. Monte Carlo simulations are used to estimate the outcome of each single match of the tournament and hence to simulate the whole tournament itself. The model is fitted on all football games on neutral ground of the participating teams since 2010.

Résumé. Cet article est consacré aux prévisions du tournoi de football de la Coupe d'Afrique des Nations 2019. Il est basé sur un modèle de régression de Poisson qui inclut les points Elo des équipes participantes en tant que covariables et intègre les différences de compétences spécifiques des équipes. Le modèle proposé permet des prédictions en termes de probabilités afin de quantifier les chances pour chaque équipe d'atteindre une certaine phase du tournoi. Les simulations Monte Carlo sont utilisées pour estimer le résultat de chaque match du tournoi et donc pour simuler l'ensemble du tournoi. Le modèle est adapté sur tous les matchs de football sur terrain neutre des équipes participantes depuis 2010.

Key words: Africa Cup of Nations 2019; football; Poisson regression; Elo.
AMS 2010 Mathematics Subject Classification Objects : 97K80; 47N30.

Article published in: African Journal of Applied Statistics . Volume 6(1), pp 599-615.
Doi : http://dx.doi.org/10.16929/ajas/2019.599.233

---

*Lorenz A. Gilch: Lorenz.Gilch@uni-passau.de

1. Introduction

1.1. Problem formulation

Football is a typical low-scoring game and games are frequently decided through single events in the game. These events may be extraordinary individual performances, individual errors, injuries, refereeing errors or just lucky coincidences. Moreover, during a tournament there are most of the time teams and players that are in exceptional shape and have a strong influence on the outcome of the tournament. One consequence is that every now and then alleged underdogs win tournaments and reputed favorites drop out already in the group phase.

The above effects are notoriously difficult to forecast. Despite this fact, every team has its strengths and weaknesses (e.g., defense and attack) and most of the results reflect the qualities of the teams. In order to model the random effects and the "deterministic" drift forecasts should be given in terms of probabilities.

Among football experts and fans alike there is mostly a consensus on the top favorites, e.g. Senegal, Cameroon or Egypt, and more debate on possible underdogs. However, most of these predictions rely on subjective opinions and are not quantifiable. An additional difficulty is the complexity of the tournament, with billions of different outcomes, making it very difficult to obtain accurate guesses of the probabilities of certain events. In the particular case of the African championship it is still more unclear to estimate the strengths of the participating teams or even to determine the divergence of the teams' strengths, since many teams or players are not so well-known as the teams from Europe or South America. Hence, the focus of this article is not to make an exact forecast, which seems not reasonable due to many unpredictable events, but to make the discrepancy between the participating teams *quantifiable* and to measure the chances of each team. This approach is underlined by the fact that supporters of the participating teams typically study the tournament structure after the group draw in order to figure out whether their teams have a rather simple or hard way to the final. Hence, the aim is to quantify the difficulty for each team to proceed to the different stages of the tournament.

1.2. State of the art

We give some background on modelling football matches. A series of statistical models have been proposed in the literature for the prediction of football outcomes. They can be divided into two broad categories. The first one, the result-based model, models directly the probability of a game outcome (win/loss/draw), while the second one, the score-based model, focusses on the prediction of the exact match score. In this article the second approach is used since the match score is a non-neglecting, very important factor in the group phase of the championship and it also implies a model for the first one. In contrast to the FIFA World Cup, where the two best teams in each group of the preliminary round qualify for the round of 16, the situation becomes more difficult in the Africa Cup of Nations 2019, where also the four best third-placed teams in the group phase qualify for the round of 16. As we have seen in former World Cups before 1994 or during the European Championship 2016, in most cases the goal difference is the crucial criterion which decides whether a third-placed team moves on to the round of 16 or is eliminated in the preliminary round. This underlines the importance and necessity of estimating the exact score of each single match and not only the outcome (win/loss/draw).

There are several models for this purpose and most of them involve a Poisson model. The easiest model, Lee (1997), assumes independence of the goals scored by each team and that each score can be modeled by a Poisson regression model. Bivariate Poisson models were proposed earlier by Maher (1982) and extended by Dixon and Coles (1997) and Karlis and Ntzoufras (2003). A short overview on different Poisson models and related models like generalised Poisson models or zero-inflated models are given in Zeileis et al. (2008) and Chou and Steenhard (2011). Possible covariates for the above models may be divided into two major categories: those containing "prospective" informations and those containing "retrospective" informations. The first category contains other forecasts, especially bookmakers' odds, see e.g. Leitner et al. (2010), Zeileis et al. (2012) and references therein. This approach relies on the fact that bookmakers have a strong economic incentive to rate the result correctly and that they can be seen as experts in the matter of the forecast of sport events. However, their forecast models remain undisclosed and rely on information that is not publicly available. The second category contains only historical data and no other forecasts. Models based on the second category allow to explicitly model the influence of the covariates (in particular, attack/defense strength/weakness). Therefore, this approach is pursued using a Poisson regression model for the outcome of single matches.

Since the Africa Cup of Nations 2019 is a more complex tournament, involving for instance effects such as group draws, e.g. see Deutsch (2011), and dependences of the different matches, Monte-Carlo simulations are used to forecast the whole course of the tournament. For a more detailed summary on statistical modeling of major international football events, see Groll et al. (2015) and references therein.

Different similar models based on Poisson regression of increasing complexity (including discussion, goodness of fit and comparing them in terms of scoring functions) were analysed and used in Gilch and Müller (2018) for the prediction of the FIFA World Cup 2018. Among the models therein, in this article we will make use of the most promising Poisson model and omit further comparison and validation of different (similar) models. The model under consideration will not only use for estimating the teams' chances to win the Africa Cup but also to answer questions like how the possible qualification of third-ranked teams in the group phase affects the chances of the top favourites. Moreover, since the tournament structure of the Africa Cup of Nations 2019 has changed in this edition to 24 participating teams, a comparison with previous editions of this tournament seems to be quite difficult due to the heavy influence of possible qualifiers for the round of 16 as third-ranked teams.

Finally, let me say some words on the data available for feeding our regression model. These days a lot of data on possible covariates for forecast models is available. Groll et al. (2015) performed a variable selection on various covariates and found that the three most significant retrospective covariates are the FIFA ranking followed by the number of Champions league and Euro league players of a team. In this article the Elo ranking (see `http://en.wikipedia.org/wiki/World_Football_Elo_Ratings`) is preferably considered instead of the FIFA ranking (which is a simplified Elo ranking since July 2018), since the calculation of the FIFA ranking changed over time and the Elo ranking is more widely used in football forecast models. See also Gásques and Royuela (2016) for a discussion on this topic and a justification of the Elo ranking. At the time of this analysis, the composition and the line ups of the teams have not been announced and hence the two other covariates are not available. This is one of the reasons that the model under consideration is solely based on the Elo points and matches of the participating teams on neutral ground since 2010. The obtained results show that, despite the simplicity of the model, the model under consideration shows a good fit, and the obtained forecast is conclusive and gives *quantitative insights* in each team's chances. In particular, we quantify the chances of each team to proceed to a specific phase of the tournament, which allows also to compare the challenge for each team to proceed to the final.

1.3. Questions under consideration

The aim is to perform many simulations of the whole tournament and to give a summary of the chances of the participating teams to reach certain stages of the Africa Cup of Nations 2019. This is done by simulating the *exact score* of each single match, and – following the tournament structure – to simulate the whole tournament itself. In order to simulate the score of the match between team A and team B as G_A:G_B, where G_A (resp. G_B) is the number of goals scored by team A (resp. by team B), we use a regression model, which allows to simulate the number of scored goals as concrete realisations of some random variables.

The problem which arises at this point is that not only a single match is forecasted but the course of the whole tournament. Hence, the uncertainty of the results of the single matches lead to a even much higher uncertainty of the outcome of the whole tournament. Even the most probable tournament outcome (i.e., exact forecast how far the participating teams will come in the tournament) has a probability very close to zero to be actually realized. Hence, deviations of the *true* tournament outcome from the model's most probable one are not only possible, but most likely. However, simulations of the tournament yield estimates of the probabilities for each team to reach different stages of the tournament and allow to make the different team's chances *quantifiable*. This means that we want to describe if, e.g., some teams have almost the same chance to reach the semifinals or to win the Africa Cup or if one team has much higher chances to win than other teams. Since we measure these chances in terms of probabilities we can describe the difference between two team's chances to reach different stages of the tournament. In particular, we are interested to give quantitative insights into the following questions:

1. How are the probabilities that a team will win its group or will be eliminated in the group stage?
2. Which team has the best chances to become new African champion?
3. What is the effect of the fact that the four best third-ranked teams in the group phase qualify for the round of 16? How does it affect the chances of the top favourites?

As we will see, the model under consideration in this article favors Senegal (followed by Nigeria) to win the Africa Cup of Nations 2019.

2. The model

2.1. Involved data

The model used in this article was proposed in Gilch and Müller (2018) (together with several similar bi-variate Poisson models) as *Nested Poisson Regression* and is based on the World Football Elo ratings of the teams. It is based on the Elo rating system, see Elo (1978), but includes modifications to take various football-specific variables (like home advantage, goal difference, etc.) into account. The Elo ranking is published by the website `eloratings.net`, from which all historic match data was retrieved.

First, we present the formula for the Elo ratings, which uses the typical form as described in `http://en.wikipedia.org/wiki/World_Football_Elo_Ratings`: let $\mathrm{Elo}_{\mathrm{before}}$ be the Elo points of a team before a match; then the Elo points $\mathrm{Elo}_{\mathrm{after}}$ after the match against an opponent with Elo points $\mathrm{Elo}_{\mathrm{Opp}}$ is calculated via the following formula:

$$\mathrm{Elo}_{\mathrm{after}} = \mathrm{Elo}_{\mathrm{before}} + K \cdot G \cdot (W - W_e),$$

where

- K is a weight index regarding the tournament of the match (World Cup matches have weight 60, while continental tournaments matches have weight 50)
- G is a number from the index of goal differences calculated as follows:

$$G = \begin{cases} 1, & \text{if the match is a draw or won by one goal} \\ \frac{3}{2}, & \text{if the match is won by two goals} \\ \frac{11+N}{8}, & \text{where } N \text{ is the goal difference otherwise} \end{cases}$$

- W is the result of the match: 1 for a win, 0.5 for a draw, and 0 for a defeat.
- W_e is the expected outcome of the match calculated as follows:

$$W_e = \frac{1}{10^{-\frac{D}{400}} + 1},$$

  where $D = \mathrm{Elo}_{\mathrm{before}} - \mathrm{Elo}_{\mathrm{Opp}}$ is the difference of the Elo points of both teams.

The Elo ratings as they were on 12 april 2019 for the top 5 participating nations (in this rating) are as follows:

Senegal	Nigeria	Morocco	Tunisia	Ghana
1764	1717	1706	1642	1634

The forecast of the outcome of a match between teams A and B is modelled as

$$G_A \ : \ G_B,$$

where G_A (resp. G_B) is the number of goals scored by team A (resp. B). The model is based on a Poisson regression model, where we assume (G_A, G_B) to be a bivariate Poisson distributed random variable; see (Gilch and Müller, 2018, Section 8) for a discussion on other underlying distributions for G_A and G_B. The distribution of (G_A, G_B) will depend on the current Elo ranking Elo_A of team A and Elo ranking Elo_B of team B. The model is fitted using all matches of Africa Cup of Nations 2019 participating teams on *neutral* playground between 1.1.2010 and 12.04.2019. Matches, where one team plays at home, have usually a drift towards the home team's chances, which we want to eliminate. In average, we have for each team 29 matches from the past and for the top teams even more. In the following subsection we explain the model for forecasting a single match, which in turn is used for simulating the whole tournament and determining the likelihood of the success for each participant.

2.2. Nested Poisson regression

We now present a *dependent* Poisson regression approach which will be the base for the whole simulation. The number of goals G_A, G_B respectively, shall be a Poisson-distributed random variable with rate $\lambda_{A|B}$, $\lambda_{B|A}$ respectively. As we will see one of the rates (that is, the rate of the weaker team) will depend on the concrete realisation of the other random variable (that is, the simulated number of scored goals of the stronger team).

In the following we will always assume that A has *higher* Elo score than B. This assumption can be justified, since usually the better team dominates the weaker team's tactics. Moreover the number of goals the stronger team scores has an impact on the number of goals of the weaker team. For example, if team A scores 5 goals it is more likely that B scores also 1 or 2 goals, because the defense of team A lacks in concentration due to the expected victory. If the stronger team A scores only 1 goal, it is more likely that B scores no or just one goal, since team A focusses more on the defense and secures the victory.

The Poisson rates $\lambda_{A|B}$ and $\lambda_{B|A}$ are now determined as follows:

1. In the first step we model the number of goals $\tilde{G}_A$ scored by team A only in dependence of the opponent's Elo score $\text{Elo} = \text{Elo}_B$. The random variable $\tilde{G}_A$ is modeled as a Poisson distribution with parameter μ_A. The parameter μ_A as a function of the Elo rating Elo_B of the opponent B is given as
$$\log \mu_A(\text{Elo}_B) = \alpha_0 + \alpha_1 \cdot \text{Elo}_B, \qquad (1)$$
where α_0 and α_1 are obtained via Poisson regression.
2. Teams of similar Elo scores may have different strengths in attack and defense. To take this effect into account we model the number of goals team B receives against a team of Elo score $\text{Elo} = \text{Elo}_A$ using a Poisson distribution with parameter ν_B. The parameter ν_B as a function of the Elo rating Elo_B is given as
$$\log \nu_B(\text{Elo}_B) = \beta_0 + \beta_1 \cdot \text{Elo}_B, \qquad (2)$$
where the parameters β_0 and β_1 are obtained via Poisson regression.
3. Team A shall in average score $\mu_A\big(\text{Elo}_B\big)$ goals against team B, but team B shall have $\nu_B\big(\text{Elo}_A\big)$ goals against. As these two values rarely coincides we model the numbers of goals G_A as a Poisson distribution with parameter
$$\lambda_{A|B} = \frac{\mu_A\big(\text{Elo}_B\big) + \nu_B\big(\text{Elo}_A\big)}{2}.$$

4. The number of goals G_B scored by B is assumed to depend on the Elo score $E_A = \text{Elo}_A$ and additionally on the outcome of G_A. More precisely, G_B is modeled as a Poisson distribution with parameter $\lambda_B(E_A, G_A)$ satisfying

$$\log \lambda_B(E_A, G_A) = \gamma_0 + \gamma_1 \cdot E_A + \gamma_2 \cdot G_A. \tag{3}$$

   The parameters $\gamma_0, \gamma_1, \gamma_2$ are obtained by Poisson regression. Hence,

$$\lambda_{B|A} = \lambda_B(E_A, G_A).$$

5. The result of the match A versus B is simulated by realizing G_A first and then realizing G_B in dependence of the realization of G_A.

For a better understanding, we give an example and consider the match Senegal vs. Ivory Coast: Senegal has 1764 Elo points while Ivory Coast has 1612 points. Against a team of Elo score 1612 Senegal is assumed to score in average

$$\mu_{\text{Senegal}}(1612) = \exp(2.73 - 0.00145 \cdot 1612) = 1.48$$

goals, while Ivory Coast receives against a team of Elo score 1764 in average

$$\mu_{\text{Ivory Coast}}(1764) = \exp(-4.0158 + 0.00243 \cdot 1764) = 1.31$$

goals. Hence, the number of goals, which Senegal will score against Ivory Coast, will be modelled as a Poisson distributed random variable with rate

$$\lambda_{\text{Senegal}|\text{Ivory Coast}} = \frac{1.48 + 1.31}{2} = 1.395.$$

The average number of goals, which Ivory Coast scores against a team of Elo score 1764 provided that G_A goals against are received, is modelled by a Poisson random variable with rate

$$\lambda_{\text{Ivory Coast}|\text{Senegal}} = \exp(1.431 - 0.000728 \cdot 1764 + 0.137 \cdot G_A);$$

e.g., if $G_A = 1$ then $\lambda_{\text{Ivory Coast}|\text{Senegal}} = 1.33$.

As an important remark, let me mention that the presented dependent approach may also be justified through the definition of conditional probabilities:

$$\mathbb{P}[G_A = i, G_B = j] = \mathbb{P}[G_A = i] \cdot \mathbb{P}[G_B = j \mid G_A = i] \quad \forall i, j \in \mathbb{N}_0.$$

Finally, let me note that classical bi-variate Poisson regression allows to estimate the parameters of a bi-variate Poisson distributed random vector in dependence of the covariates. Simulation of a bi-variate Poisson distributed random variable is then performed via realisation of *three independent* Poisson distributed random variables (the third one controls the correlation); see e.g. Kawamura (1973) for further details. In our case we do *not* model the outcome (G_A, G_B) as a bi-variate Poisson random vector, but in the sense of a *nested* model, where first G_A is realised via a Poisson distribution and *thereafter* G_B is realised in dependence of the *concrete* outcome of G_A. For comparision of our nested model in contrast to similar Poisson models, we refer once again to Gilch and Müller (2018). In the following subsections we present some regression plots and will test the goodness of fit. All calculations were preformed with `R` (version 3.3.1).

2.3. Regression plots

As two examples of interest, we sketch in Figure 1 the results of the regression in (1) for the number of goals scored by Senegal and Cameroon. The dots show the observed data (i.e, number of scored goals on the y-axis in dependence of the opponent's strength on the x-axis) and the line is the estimated mean μ_A depending on the opponent's Elo strength.

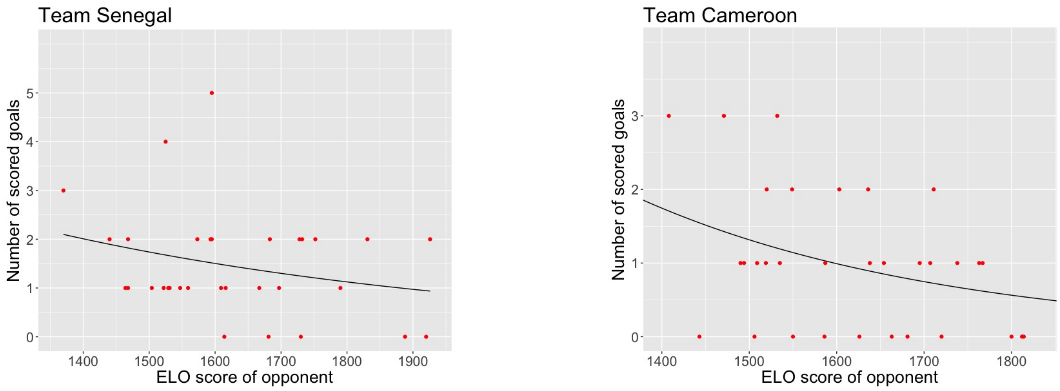

Fig. 1. Plots for the number of goals scored by Senegal and Cameroon in regression (1).

Analogously, Figure 2 sketches the regression in (2) for the (unconditioned) number of goals against of Nigeria and Egypt in dependence of the opponent's Elo ranking. The dots show the observed data (i.e., the number of goals against in the matches from the past) and the line is the estimated mean ν_B for the number of goals against.

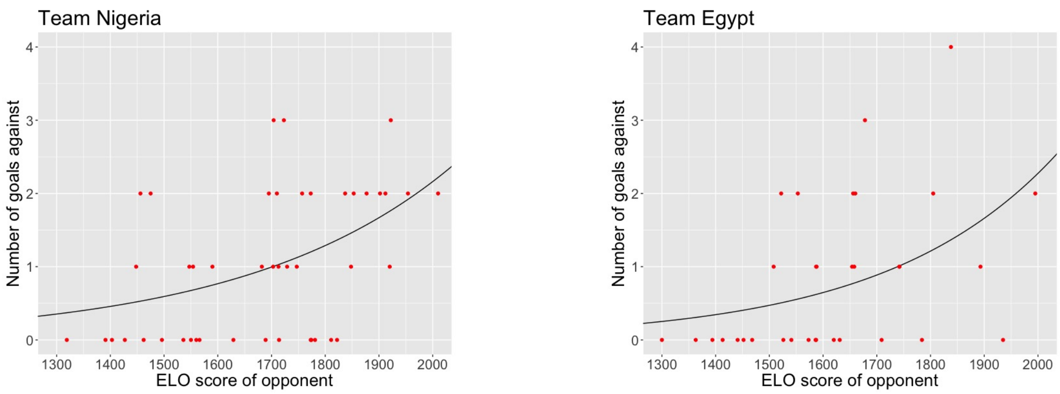

Fig. 2. Plots for the number of goals against for Nigeria and Egypt in regression (2).

2.4. Goodness of fit tests

We check goodness of fit of the Poisson regressions in (1) and (2) for all participating teams. For each team $\mathbf{T}$ we calculate the following χ^2-statistic from the list of matches from the past:

$$\chi_{\mathbf{T}} = \sum_{i=1}^{n_{\mathbf{T}}} \frac{(x_i - \hat{\mu}_i)^2}{\hat{\mu}_i},$$

where $n_{\mathbf{T}}$ is the number of matches of team $\mathbf{T}$, x_i is the number of scored goals of team $\mathbf{T}$ in match i and $\hat{\mu}_i$ is the estimated Poisson regression mean in dependence of the opponent's historical Elo points.

We observe that most of the teams have a very good fit, except Namibia with a p-value of 0.048. In average, we have a p-value of 0.476. In Table 1 the p-values for some of the top teams are given. Similarly, we can calculate a χ^2-statistic for each team which measures the goodness of fit for the regression in (2) which models the number of goals against. Here, we get an average p-value of 0.67; see Table 2. Finally, we test the goodness of fit for the regression in (3) which models the number of goals against of the weaker team in dependence of the number of goals which are scored by the stronger team. We obtain an average p-value of 0.33; see Table 3. As a conclusion, the p-values suggest good fits.

2.5. Deviance analysis

We calculate the null and residual deviances for each team for the regressions in (1), (2) and (3). Tables 4, 5 and 6 show the deviance values and the p-values for the residual deviance for some of the top teams. Most of the p-values are not low, except for Nigeria. We remark that the level of significance of the covariates is also of fluctuating quality, but it is still reasonable in many cases.

3. Africa Cup of Nations 2019 Simulations

Finally, we come to the simulation of the Africa Cup of Nations 2019, which allows us to answer the questions formulated in Section 1.3. We simulate each single match of the Africa Cup of Nations 2019 according to the model presented in Section 2, which in turn allows us to simulate the whole Africa Cup tournament. After each simulated match we update the Elo ranking according to the simulation results. This honours teams, which are in a good shape during a tournament and perform maybe better than expected. Overall, we perform 100.000 simulations of the whole tournament, where we reset the Elo ranking at the beginning of each single tournament simulation. All simulations were preformed with `R` (version 3.3.1).

3.1. Single Matches

As the basic element of our simulation is the simulation of single matches, we visualise how to quantify the outcomes of single matches. Group C starts with the match between Senegal and Tanzania. According to our model we have the probabilities presented in Figure 3 for the result of this match: the most probable score is a $2-0$ victory of Senegal, but a $3-0$ or $1-0$ win would also be among the most probable scores.

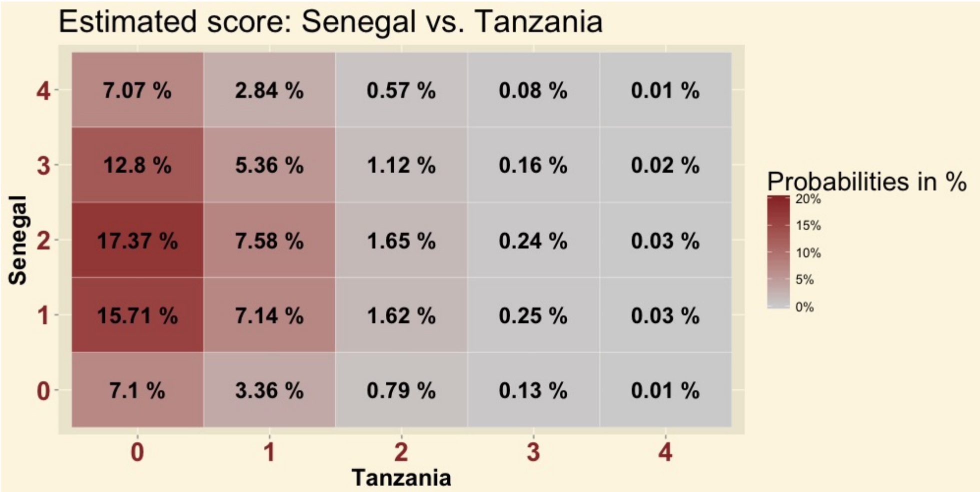

Fig. 3. Probabilities for the score of the match Senegal vs. Tanzania in Group C.

3.2. Group Forecast

Among football experts and fans a first natural question after the group draw is to ask how likely it is that the different teams survive the group stage and move on to the round of 16. Since the individual teams' strength and weaknesses are rather hard to quantify in the sense of tight facts, one of our main aims is to quantify the chances for each participating team to proceed to the round of 16. With our model we are able to quantify the chances in terms of probabilities how the teams will end up in the group stage. In the following tables 7-12 we present these probabilities obtained from our simulation, where we give the probabilities of winning the group, becoming runner-up, to qualify as one the best third-placed teams or to be eliminated in the group stage. In Group D, the toughest group of all, a head-to-head fight between Morocco, Ivory Coast and South Africa is expected with slight advantage for the team from Ivory Coast.

3.3. Playoff Round Forecasts

Finally, according to our simulations we summarise the probabilities for each team to win the tournament, to reach certain stages of the tournament or to qualify for the round of last 16. The result is presented in Table 13. E.g., Senegal will at least reach the quarterfinals with a probability of 67, 70%, while Ghana has a 17% chance to reach the final. The regression model favors Senegal, followed by Nigeria, Ivory Coast and Egypt, to become new football champion of Africa.

3.4. Simulation without third-placed qualifiers

One important and often asked question is whether the current tournament structure, which allows third-placed teams in the preliminary round still to qualify for the round of

16, is reasonable or not. In particular, it is the question whether this structure is good or bad for the top teams and to quantify this factor. Hence, the simulation was adapted in the sense that third-placed teams in the group stage are definitely eliminated, while the winners of those groups, which are intended to play against a third-ranked team in the round of 16, move directly to the quarter finals. This leads to the results in Table 14: it shows that the top teams have now slightly higher chances to win the tournament.

In order to compare the uncertainty of both models we evaluate the entropy for the teams' probabilities to win the cup. Recall that the *entropy* of a probability vector $p := (p_1, \dots, p_n) \in (0,1)^n$, $n \in \mathbb{N}$, with $\sum_{i=1}^{n} p_i = 1$ is defined as

$$h(p) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i,$$

where we set $0 \cdot \log_2 := 0$. The entropy $h(p)$ describes the amount of uncertainty contained in the probability vector: the higher $h(p)$, the more uncertain (i.e., harder to guess) is the outcome of a p-distributed random variable. In other words, the entropy can be interpreted as the average number of questions with answers "yes/no" which are needed in order to guess the outcome. In our setting let p be the winning probabilities (according to our simulations) for the teams, that is, p_i is the probability that team i wins the tournament. In the case of the current tournament structure, where also third-ranked teams may move to the quarterfinals, the entropy is 3.7093, while in the case when third-ranked teams are definitely eliminated the entropy goes down to 3.6365. Hence, there is a measurable decrease of uncertainty.

In Table 15 we compare the probabilities of reaching different stages in the case of the adapted tournament (third-ranked teams are definitely eliminated) versus the real tournament structure, which still allows third-ranked teams to qualify for the round of 16. As one can see, the differences are rather marginal. However, the top favourite teams would profit from the adapted setting slightly. Moreover, many teams have a chance of 10% or more to qualify for the round of 16 as one of the best four third-ranked teams. Thus, the chances to win the African championship remain more or less the same, making it neither harder nor easier for top ranked teams to win.

4. Discussion on Related Models

In this section we want to give some quick discussion about the used Poisson models and related models. Of course, the Poisson models we used are not the only natural candidates for modeling football matches. Multiplicative mixtures may lead to overdispersion. Thus, it is desirable to use models having a variance function which is flexible enough to deal with overdispersion and underdispersion. One natural model for this is the *generalised Poisson model*, which was suggested by Consul (1989). We omit the details but remark that this distribution has an additional parameter φ which allows to model the variance as λ/φ^2; for more details on generalised Poisson regression we refer to Stekeler (2004) and Erhardt (2006). Estimations of φ by generalised Poisson regression lead to the observation that φ is close to 1 for the most important teams; compare with Gilch and Müller (2018). Therefore, no additional gain is given by the use of the generalised Poisson model.

Another related candidate for the simulation of football matches is given by the *negative binomial distribution*, where also another parameter comes into play to allow a better fit. However, the same observations as in the case of the generalised Poisson model can be made, that is, the estimates of the additional parameter lead to a model which is almost just a simple Poisson model. We refer to Joe and Zhu (2005) for a detailed comparison of generalized Poisson distribution and negative Binomial distribution.

For further discussion on adaptions and different models, we refer once again to the discussion section in Gilch and Müller (2018).

5. Conclusion

A team-specific Poisson regression model for the number of goals in football matches facing each other in international tournament matches has been used for quantifying the chances of the teams participating in the Africa Cup of Nations 2019. They all include the Elo points of the teams as covariates and use all matches of the teams since 2010 as underlying data.The fitted model was used for Monte-Carlo simulations of the Africa Cup of Nations 2019. According to this simulation, Senegal (followed by Nigeria) turns out to be the top favorite for winning the title. Besides, for every team probabilities of reaching the different stages of the cup are calculated.

A major part of the statistical novelty of the presented work lies in the construction of the nested regression model. This model outperforms previous studied models, that use (inflated) bivariate Poisson regression, when tested on the previous FIFA World Cups 2010, 2014 and 2018; see the technical report Gilch and Müller (2018).

References

Chou, N.-T. and Steenhard, D. (2011). Bivariate count data regression models – a SAS® macro program. *SAS Global Forum 2011 - Paper 355-2011*, pages 1–10.

Consul, P. (1989). *Generalized Poisson distributions : properties and applications*. Statistics, textbooks and monographs v. 99. New York, M. Dekker.

Deutsch, R. C. (2011). Looking back at South Africa: Analyzing and reviewing the 2010 FIFA world cup. *CHANCE*, 24(2):15–23.

Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 46(2):265–280.

Elo, A. E. (1978). The rating of chessplayers, past and present. *Arco Pub.*, New York.

Erhardt, V. (2006). Verallgemeinerte Poisson und Nullenueberschuss – Regressionsmodelle mit regressiertem Erwartungswert, Dispersions- und Nullenueberschuss-Parameter und eine Anwendung zur Patentmodellierung. *Master's thesis, Technical University of Munich.*

Gásques, R. and Royuela, V. (2016). The determinants of international football success: A panel data analysis of the elo rating*. *Social Science Quarterly*, 97(2):125–141.

Gilch, L. A. and Müller, S. (2018). On Elo based prediction models for the FIFA Worldcup 2018. *Technical Report, Number MIP-1801, Department of Informatics and Mathematics, University of Passau, Germany.*

Groll, A., Schauberger, G., and Tutz, G. (2015). Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA World Cup 2014. *Journal of Quantitative Analysis in Sports*, 11(2):97–115.

Joe, H. and Zhu, R. (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. *Biometrical Journal*, 47(2):219–229.

Karlis, D. and 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.

Kawamura, K. (1973). The structure of bivariate Poisson distribution. *Kodai Math. Sem. Rep.*, 25(2):246–256.

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

Leitner, C., Zeileis, A., and 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.

Stekeler, D. (2004). Verallgemeinerte Poissonregression und daraus abgeleitete zero-inflated und zero-hurdle Regressionsmodelle. *Master's thesis, Technical University of Munich.*

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software, Articles*, 27(8):1–25.

Zeileis, A., Leitner, C., and Hornik, K. (2012). History repeating: Spain beats Germany in the EURO 2012 final. *Working Papers in Economics and Statistics 2012-09, University of Innsbruck, Innsbruck.*

Team	Senegal	Nigeria	Egypt	Ivory Coast	South Africa
p-value	0.74	0.10	0.60	0.94	0.72

Table 1. Goodness of fit test for the Poisson regression in (1) for some of the top teams.

Team	Senegal	Nigeria	Egypt	Ivory Coast	South Africa
p-value	0.99	0.79	0.38	0.51	0.76

Table 2. Goodness of fit test for the Poisson regression in (2) for some of the top teams.

Team	Senegal	Nigeria	Egypt	Ivory Coast	South Africa
p-value	0.99	0.38	0.27	0.78	0.74

Table 3. Goodness of fit test for the Poisson regression in (3) for some of the top teams.

Team	Null deviance	Residual deviance	p-value
Senegal	28.14	26.34	0.66
Nigeria	71.36	66.39	0.03
Egypt	43.94	38.15	0.29
Cote d'Ivoire	47.15	46.8	0.71
South Africa	12.0	10.49	0.65

Table 4. Deviance analysis for some top teams in regression (1)

Team	Null deviance	Residual deviance	p-value
Senegal	19.41	19.21	0.94
Nigeria	58.50	45.35	0.87
Egypt	49.63	38.09	0.29
Cote d'Ivoire	69.97	59.61	0.25
South Africa	12.14	11.92	0.53

Table 5. Deviance analysis for some top teams in regression (2)

Team	Null deviance	Residual deviance	p-value
Senegal	28.1	24.8	0.69
Nigeria	71.4	62.1	0.05
Egypt	43.94	37.98	0.25
Cote d'Ivoire	47.15	45.45	0.73
South Africa	12.01	10.36	0.58

Table 6. Deviance analysis for some top teams in regression (3)

Team	1st	2nd	Qualified as Third	Preliminary Round
Egypt	51.00	28.30	11.30	9.50
DR of Congo	32.00	31.80	15.60	20.70
Uganda	4.70	14.10	16.00	65.10
Zimbabwe	12.40	25.80	21.20	40.60

Table 7. Probabilities for Group A

Team	1st	2nd	Qualified as Third	Preliminary Round
Nigeria	53.90	26.90	10.90	8.40
Guinea	25.80	31.70	17.20	25.40
Madagascar	16.10	25.90	20.50	37.60
Burundi	4.30	15.60	17.20	62.90

Table 8. Probabilities for Group B

Team	1st	2nd	Qualified as Third	Preliminary Round
Senegal	54.40	27.80	10.80	7.10
Algeria	28.50	31.90	17.40	22.10
Kenya	12.30	24.80	21.20	41.70
Tanzania	4.80	15.50	16.70	63.10

Table 9. Probabilities for Group C

Team	1st	2nd	Qualified as Third	Preliminary Round
Morocco	29.40	27.10	17.50	26.00
Ivory Coast	33.60	28.80	16.70	20.90
South Africa	30.40	29.00	17.20	23.40
Namibia	6.60	15.10	17.40	60.90

Table 10. Probabilities for Group D

Team	1st	2nd	Qualified as Third	Preliminary Round
Tunisia	49.60	28.60	13.50	8.30
Mali	32.10	37.50	19.00	11.40
Mauritania	4.10	9.10	11.40	75.40
Angola	14.30	24.80	27.00	33.90

Table 11. Probabilities for Group E

Team	1st	2nd	Qualified as Third	Preliminary Round
Cameroon	38.80	42.60	11.90	6.80
Ghana	55.70	32.00	7.90	4.40
Benin	4.60	19.70	33.70	42.00
Guinea-Bissau	0.90	5.70	11.00	82.30

Table 12. Probabilities for Group F

Team	Champion	Final	Semifinal	Quarterfinal	Last16
Senegal	15.40	25.20	41.20	67.70	92.90
Nigeria	12.10	22.70	37.30	59.90	91.60
Ivory Coast	10.20	17.70	31.10	51.90	79.10
Egypt	10.10	19.20	34.60	56.60	90.60
Ghana	8.60	17.00	30.50	57.20	95.40
South Africa	8.40	15.50	28.50	48.80	76.50
Morocco	8.30	15.30	28.20	48.20	73.90
Tunisia	5.80	11.90	23.20	45.50	91.70
Algeria	5.10	10.30	21.40	43.30	77.80
Guinea	3.40	8.10	17.90	37.60	74.60
Cameroon	3.00	9.00	22.30	50.70	93.30
DR Congo	3.00	7.70	19.00	40.00	79.10
Mali	1.60	5.00	13.20	32.70	88.50
Madagascar	1.60	4.10	10.50	25.40	62.40
Kenya	1.10	3.10	9.10	23.90	58.40
Angola	1.00	2.80	8.00	22.10	66.10
Zimbabwe	0.40	1.80	7.40	22.80	59.50
Namibia	0.30	1.20	4.20	13.20	39.10
Uganda	0.10	0.50	2.60	10.30	34.90
Tanzania	0.10	0.50	2.60	10.10	36.90
Mauritania	0.10	0.40	1.50	5.90	24.40
Benin	0.10	0.60	3.40	15.10	58.00
Burundi	0.00	0.20	1.60	7.90	37.00
Guinea-Bissau	0.00	0.00	0.30	2.60	17.60

Table 13. Africa Cup of Nations 2019 simulation results for the teams' probabilities to proceed to a certain stage

Team	Champion	Final	1/2	1/4	Last16	1st	2nd	Pre.Round
Senegal	15.80	25.40	43.50	74.10	82.20	54.50	27.70	17.90
Nigeria	14.50	28.40	45.30	72.30	80.60	53.90	26.70	19.40
Egypt	11.70	22.60	41.10	67.90	79.30	50.50	28.70	20.70
Ivory Coast	9.90	17.00	30.40	51.50	62.50	33.50	29.00	37.50
South Africa	7.90	14.40	27.40	47.90	59.60	30.80	28.70	40.50
Ghana	7.80	16.00	28.40	53.40	87.70	55.60	32.10	12.30
Morocco	7.60	13.50	26.20	45.90	56.50	29.20	27.20	43.60
Algeria	5.10	10.20	21.90	45.90	60.10	28.50	31.70	39.80
Tunisia	4.70	9.60	19.00	39.00	77.70	49.30	28.50	22.20

Table 14. Adapted Africa Cup of Nations 2019 simulation results, where third-placed teams are definitely eliminated

Team	Champion	Final	Semifinal	Quarterfinal	Last16
Senegal	0.40	0.20	2.30	6.40	-10.70
Nigeria	2.40	5.70	8.00	12.40	-11.00
Egypt	1.60	3.40	6.50	11.30	-11.30
Ivory Coast	-0.30	-0.70	-0.70	-0.40	-16.60
South Africa	-0.50	-1.10	-1.10	-0.90	-16.90
Ghana	-0.80	-1.00	-2.10	-3.80	-7.70
Morocco	-0.70	-1.80	-2.00	-2.30	-17.40
Algeria	0.00	-0.10	0.50	2.60	-17.70
Tunisia	-1.10	-2.30	-4.20	-6.50	-14.00

Table 15. Difference of probabilities of adapted tournament simulation vs. real tournament structure