Brazil or Germany - who will win the trophy? Prediction of the FIFA World Cup 2014 based on team-specific regularized Poisson regression

Andreas Groll * Gunther Schauberger † Gerhard Tutz ‡

July 2, 2014

Abstract In this article an approach for the analysis and prediction of soccer match results is proposed. It is based on a regularized Poisson regression model that includes various potentially influential covariates describing the national teams' success in previous FIFA World Cups. Additionally, similar to Bradley-Terry-Luce models, differences of team-specific effects of the competing teams are included. It is discussed that within the generalized linear model (GLM) framework the team-specific effects can either be incorporated in the form of fixed or random effects. In order to achieve variable selection and shrinkage, we use tailored Lasso approaches. Based on the three preceding FIFA World Cups, two competing models for the prediction of the FIFA World Cup 2014 are fitted and investigated.
Keywords Football, FIFA World Cup 2014, Sports tournaments, Generalized linear model, Lasso, Variable selection.

1 Introduction

In the last few years various approaches to the statistical modeling of major soccer events have been proposed, among them the Union of European Football Associations (UEFA) Champions League (CL; Karlis and

*Department of Mathematics, Workgroup Financial Mathematics, Ludwig-Maximilians-University Munich, Theresienstr. 39, 80333 Munich, Germany, *groll@math.lmu.de*

†Department of Statistics, Seminar for Applied Stochastic, Ludwig-Maximilians-University Munich, Akademiestr. 1, 80799 Munich, Germany, *gunther.schauberger@stat.uni-muenchen.de*

‡Department of Statistics, Seminar for Applied Stochastic, Ludwig-Maximilians-University Munich, Akademiestr. 1, 80799 Munich, Germany, *gerhard.tutz@stat.uni-muenchen.de*

Ntzoufras, 2003, Leitner et al., 2011, Eugster et al., 2011), the European football championship (EURO; Leitner et al., 2008, Leitner et al., 2010a, Zeileis et al., 2012, Groll and Abedieh, 2013) or the Fédération Internationale de Football Association (FIFA) World Cup (Leitner et al., 2010b, Stoy et al., 2010, Dyte and Clarke, 2000). In particular, the current FIFA World Cup 2014 in Brazil is accompanied by various publications trying to predict the tournament winner, see, e.g., Zeileis et al. (2014), Goldman-Sachs Global Investment Research (2014), Silver (2014) and Lloyd's (2014).

In general, statistical approaches to the modeling of soccer data can be divided into two major categories: the first ones are based on the easily available source of "prospective" information contained in bookmakers' odds, compare Leitner et al. (2008), Leitner et al. (2010b), Zeileis et al. (2012) and Zeileis et al. (2014). They already correctly predicted the finals of the EURO 2008 as well as Spain as the 2010 FIFA World Champion and as the 2012 EURO Champion. The winning probabilities for each team were obtained simply by aggregating winning odds from several online bookmakers. Based on these winning probabilities, by inverse tournament simulation team-specific abilities can be computed by paired comparison models. Using this technique the effects of the tournament draw are stripped. Next, pairwise probabilities for each possible game at the corresponding tournament can be predicted and, finally, the whole tournament can be simulated. Using this approach, Zeileis et al. (2014) predict the host Brazil to win the FIFA World Cup 2014 with a probability of 22.5%, followed by Argentina (15.8%) and Germany (13.4%).

It should be noted that this method will always predict the team that has the lowest (average) bookmaker odds as the tournament winner and, hence, is implicitly assuming that all available information is covered by the bookmakers expertise. This is not unrealistic, as one can indeed expect bookmakers to use sophisticated models when setting up their odds, as they have strong economic incentives to rate the team strengths of soccer teams correctly. Nevertheless, from time to time a clear underdog wins a major tournament, as, for example, Greece at the EURO 2004[1]. Although any statistical model will have serious difficulties to correctly predict such an unexpected event, it would be desirable to at least draw conclusions from such an event with regard to future tournaments. This is only possible if models are used, which incorporate covariates of the competing teams, while methods based solely on bookmakers' odds (or solely on the market

[1]The German state betting agency ODDSET ranked Greece on place twelve among the favorites for the EURO 2004 with odds of 45.00 (usually, in statistics odds represent the ratio of the probability that an event will happen to the probability that it will not happen; however, European bookmakers specify the gross ratio, which represents the ratio of paid amount to stake. So putting 1 Euro on Greece as the EURO 2004 champion would have given back 45 Euro. Thus, European odds can be directly transformed into probabilities by taking the inverse and adjusting for the bookmakers' margins).

value instead) have no chance to account for such information. Hence, our goal is to determine additional relevant influence variables that may provide further information regarding the teams' abilities.

This task leads to the second category of approaches that are based on regression models. A simple standard linear regression approach was used by Stoy et al. (2010) to analyze the success of national teams at FIFA World Cups. The success of a team at a World Cup is measured by a defined point scale that is supposed to be normally distributed. Beside some sport-specific covariates also political-economic, socio-geographic as well as some religious and psychological influence variables are considered. Based on this model, a prediction for the FIFA World Cup 2010 was obtained.

In contrast to Stoy et al. (2010), most of the regression approaches directly model the number of goals scored in single soccer matches, assuming that the number of goals scored by each team follows a Poisson distribution model, see, e.g., Maher (1982), Lee (1997), Dyte and Clarke (2000), Rue and Salvesen (2000) and Goldman-Sachs Global Investment Research (2014). For example, Dyte and Clarke (2000) predict the distribution of scores in international soccer matches, treating each team's goals scored as conditionally independent Poisson variables depending on two influence variables, the FIFA world ranking of each team and the match venue. Poisson regression is used to estimate parameters for the model and based on these parameters the matches played during the 1998 FIFA World Cup can be simulated.

Similarly, Goldman-Sachs Global Investment Research (2014) set up a regression model based on the entire history of mandatory international football matches—i.e., no friendlies—since 1960, ending up with about 14,000 observations. The dependent variable is the number of goals scored by each side in each match, assuming that the number of goals scored by a particular side in a particular match follows a Poisson distribution. They incorporate six explanatory covariates: the difference in the Elo rankings[2] between the two teams, the average number of goals scored/received by the competing teams over the last ten/five mandatory international games, a dummy variable indicating whether the regarding match was a World Cup match, a dummy variable indicating whether the considered team played in its home country, a team-specific dummy variable indicating whether the considered team played on its home continent. Finally, based on the estimated regression parameters, a probability distribution for the outcome of each game is obtained and a Monte Carlo simulation with 100,000 draws is used to generate the probabilities of teams reaching particular stages of the tournament, up to winning the championship. The forecast tournament

[2]The Elo ranking is a composite measure of national football teams' success, which is based on the entire historical track record and which, in contrast to the FIFA ranking, is available for the entire history of international football matches (see Elo, 2008).

winner at the FIFA World Cup 2014 is Brazil with a rather high winning probability of 48.5%, followed by Argentina (14.1%) and Germany (11.4%).

At this point, we also want to mention two other, completely different prediction approaches, which cannot be classified into one of the two aforementioned major categories of statistical approaches for modeling soccer data. The first one was proposed by Silver (2014) and is based on the so-called Soccer Power Index (SPI). The SPI is a rating system, which uses historical data on both the international and club level to predict the outcome of a match. The algorithm uses several years of data, taking into account goals scored and allowed, quality of the lineup fielded, and the location of the match. In addition, the index weights recent matches more heavily, and also takes into account the importance of the match – e.g., World Cup matches count much more than friendly matches. Based on the SPI, Silver (2014) forecasts again Brazil as the tournament winner at the FIFA World Cup 2014, also with a rather large winning probability of 45.2%, followed by Argentina (12.8%) and Germany (11.9%).

The other alternative approach is from a more economical perspective: Lloyd's (2014) use players wages and endorsement incomes together with a collection of additional indicators to construct an economic model, which estimates players incomes until retirement. These projections form the basis for assessing insurable values by players age, playing position and nationality. As Germany and Spain are associated with the largest estimated insured values, according to this approach they turn out to be the top favorites for winning the current World Cup.

The approach that we propose here is based on a model similar to Goldman-Sachs Global Investment Research (2014). We focus on FIFA World Cups and use a Poisson model for the number of goals scored by national teams in the single matches of the tournaments. Several potential influence variables are considered and, additionally, team-specific effects are included, either in the form of fixed or random effects, resulting in a flexible generalized linear (mixed) model, in short GL(M)M. The 192 matches of the FIFA World Cups 2002-2010 serve as basis for our analysis[3], where each match occurs twice in the data set in the form of two different rows, one for each team. Each row contains the differences between the covariates corresponding to the team whose goals are considered and those of its opponent. Incorporating a method for the selection of relevant predictors, we obtain a regularized solution for our model.

[3]Though this represents a quite small basis of data, we abstain from using earlier FIFA World Cups, as one of our main objectives is to analyze the explanatory power of bookmakers' odds together with many additional, potentially influential covariates. Unfortunately, the possibility of betting on the overall cup winner before the start of the tournament is quite novel. For example, the German state betting agency ODDSET offered the bet for the first time at the FIFA World Cup 2002.

The variable selection is based on L_1-penalization techniques and two different methods are used modeling either fixed or random team-specific effects. For the model with fixed team-specific effects the `grplasso` function from the corresponding R-package (see Meier et al., 2008) can be used, while for the model with random team-specific effects a fitting approach is used, which works by combining gradient ascent optimization with the Fisher scoring algorithm and is presented in detail in Groll and Tutz (2014). It is implemented in the `glmmLasso` function of the corresponding R-package (Groll, 2014; publicly available via CRAN, see http://www.r-project.org).

Finally, we compare the results of both approaches in order to determine a final model, which is then used to predict the current FIFA World Cup 2014. It should be noted that in contrast to other team sports, such as basketball, ice-hockey or handball, in soccer pure chance plays an important role. A major reason for this is that, compared to other sports, in soccer fewer points (goals) are scored and thus singular game situations can have a tremendous effect on the outcome of the match. One consequence is that every now and then alleged (and unpredictable) underdogs win tournaments[4]. Nevertheless, it can be interesting to investigate the relationship and dependency structure between different potentially influential covariates and the success of soccer teams (in our case in terms of the number of goals they score). Besides, we hope to get more insight into which covariates are already covered by bookmakers' odds and which covariates may give some additional useful information.

The rest of the article is structured as follows. In Section 2, we introduce the team-specific Poisson model for the number of goals. A list of several possible influence variables that will be considered in our regression analysis and the data are presented in Section 3. Next, a final model is determined in Section 4, which is then used to predict the FIFA World Cup 2014. Note that all computations have been performed by use of the statistical software R (R Core Team, 2014).

2 Model and estimation

The underlying model of our analysis concentrates on the number of goals a team scores against a specific opponent. For every team, specific attack and defense parameters are considered. Furthermore, the covariates of both teams, which might have an influence on the number of scored goals, are considered in the form of differences.

[4]There are countless examples in history for such events, throughout all competitions. We want to mention only some of the most famous ones: Germany's first World Cup success in Switzerland 1954, known as the "miracle from Bern"; Greece's victory at the EURO 2004 (compare Footnote 1); FC Porto's triumph in the UEFA CL season 2003/2004.

Let for n teams $y_{ijk}, i, j \in \{1, \ldots, n\}$, $i \neq j$, denote the number of goals scored by team i when playing team j at World Cup k. The considered model has the form:

$$\begin{aligned} y_{ijk}|\boldsymbol{x}_{ik}, \boldsymbol{x}_{jk} &\sim Pois(\lambda_{ijk}) \\ log(\lambda_{ijk}) &= \beta_0 + (\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk})^\intercal \boldsymbol{\beta} + att_i - def_j. \end{aligned}$$

It is assumed that the number of goals that team i scores follows a Poisson distribution with given team-specific parameters and covariates. In addition, the two observations of one match are assumed to be independent, given the team-specific parameters and covariates.

The linear predictor consists of the attacking parameter att_i of the team i and the defending parameter def_j of its opponent j. The covariates of team i at World Cup k are collected in a vector $\boldsymbol{x}_{ik} = (x_{ik1}, \ldots, x_{ikp})^\intercal$ of length p. Note that the covariates of each team can vary over different World Cups (but not within a tournament). Each covariate is incorporated as the difference between the respective covariates of both teams. The covariate effects are collected in the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ and β_0 represents the intercept.

Generally, the estimation of the covariate effects will be obtained by regularized estimation approaches. The idea is to first set up a model with a rather large number of possibly influential variables and then to regularize the effect of the single covariates. This regularization aims at diminishing the variance of the parameter estimates and, hence, to provide lower prediction error than the unregularized maximum likelihood estimator. The basic idea of regularization is to maximize a penalized version of the log-likelihood $l(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\intercal$ represents a general parameter vector. More precisely, one maximizes the penalized log-likelihood

$$l_p(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - \lambda J(\boldsymbol{\alpha}), \tag{1}$$

where λ represents a tuning parameter, which is used to control the strength of the penalization. In practice, this tuning parameter has to be chosen either by suitable criteria for model selection, like AIC (Akaike, 1973) or BIC (Schwarz, 1978) or by cross-validation. The penalty term $J(\boldsymbol{\alpha})$ can have many different shapes. Hoerl and Kennard (1970) suggested the so-called Ridge penalty

$$J(\boldsymbol{\alpha}) = \sum_{i=1}^{p} \alpha_i^2,$$

where the sum over the squares of all parameters in the model is penalized. The Ridge penalty has the feature to shrink the respective parameter estimates towards zero. After all, Ridge cannot set estimates to zero exactly and, therefore, can not perform variable selection. In our analysis, we will use a penalty based on the absolute values of the parameters instead of the squared values resulting in a so-called Lasso penalty. The Lasso estimator was originally proposed by Tibshirani (1996) and uses the penalty

$$J(\boldsymbol{\alpha}) = \sum_{i=1}^{p} |\alpha_i|.$$

In contrast to the Ridge penalty, Lasso can provide parameter estimates, which are exactly zero and, therefore, enforces variable selection.

The team-specific attack and defense parameters can be modeled in two different ways, namely either by fixed or random effects. This distinction leads us to two different estimation procedures.

Model 1. If the team-specific ability parameters att_i and def_j are considered as fixed effects, they are coded by dummy variables within the design matrix. From this perspective, the attack (and, analogously, the defense) variables are seen as categorical covariates with as many categories as there are teams[5]. One assigns -1 to the dummy variables associated with att_i, if the goals of team i are considered, and 0 otherwise. Similarly, one assigns -1 to the dummy variables associated with def_j, if team j is the opponent, and 0 otherwise. An extract of the corresponding design matrix is given in Table 2.

Since the attack and defense abilities are considered to be covariates, the corresponding parameters are regularized similar to the parameters of the covariates. Commonly, for categorical covariates a so-called Group Lasso penalty (Yuan and Lin, 2006) is used so that all dummy variables corresponding to either the attack or the defense abilities are treated as two groups of variables. The Group Lasso penalizes the L_2-norm of the respective parameter vector $\boldsymbol{att} = (att_1, \ldots, att_n)^\intercal$ or, equivalently, $\boldsymbol{def} = (def_1, \ldots, def_n)^\intercal$. Hence, the single parameters within the groups of attack or defense abilities are shrunk towards each other and, if shrunk to zero, the whole parameter group is excluded from the model The penalty term for this model is given by

$$J(\alpha) = \sum_{i=1}^{p} |\beta_i| + \sqrt{\sum_{i=1}^{n} att_i^2} + \sqrt{\sum_{i=1}^{n} def_i^2}.$$

[5]Usually, for reasons of identifiability, categorical predictors with k factor levels are coded by $k-1$ dummies. However, the regularization approach (with $\lambda > 0$) provides unique estimates despite the issues of identifiability.

The model can easily be fitted by use of the `grplasso` function from the corresponding R-package (see Meier et al., 2008).

Model 2. Alternatively, the team-specific effects can be estimated as random effects. Then, the attack and the defense parameter of team i are assumed to be normally distributed

$$(att_i, def_i) \sim N(\mathbf{0}, \mathbf{\Sigma}), \tag{2}$$

where $\mathbf{\Sigma}$ is an unknown 2×2 covariance matrix.

In this case, the ability parameters are automatically regularized by the assumption of a distribution and only the covariate effects $\boldsymbol{\beta}$ are explicitly penalized by using

$$J(\alpha) = \sum_{i=1}^{p} |\beta_i|.$$

When integrating out the random effects from the corresponding log-likelihood function, the Laplace approximation proposed in Breslow and Clayton (1993) results in a penalized quasi-likelihood, see, for example, Fahrmeir and Tutz (2001) and Tutz (2012).

For this model, a fitting approach is used that works by combining gradient ascent optimization with the Fisher scoring algorithm and is presented in detail in Groll and Tutz (2014). It is implemented in the `glmmLasso` function of the corresponding R-package (Groll, 2014) and is publicly available via CRAN, see http://www.r-project.org.

In general, Model 2 can be seen as an extension of the model used in Groll and Abedieh (2013) with a more realistic random effects structure, now considering random effects of both competitors.

3 Data

In addition to the team-specific dummy variables for attack and defense, the model introduced above uses additional covariates. For each participating team, the covariates are observed either for the year of the respective World Cup (e.g. GDP per capita) or shortly before the start of the World Cup (e.g. FIFA ranking). Therefore, the covariates of a team vary from one World Cup to another and, hence, the model allows for a prediction of a new World Cup based on the current covariate realizations. In the following, we give a brief description of the covariates that are used.

Economic Factors:

GDP per capita. The gross domestic product (GDP) per capita represents the economic strength of a country. To account for the

general increase of the GDP, a ratio of the GDP per capita of the respective country and the worldwide average GDP per capita is used. The GDP data were collected from is the website of the United Nations Statistics Division (`http://unstats.un.org/unsd/snaama/dnllist.asp`).

Population. The population size of a country may have an influence on the success of a national team as small countries will have a smaller amount of players to choose from. The population size is used as a ratio with the respective global population to account for the general growth of the world population. The data source is the website of the world bank (`http://data.worldbank.org/indicator/SP.POP.TOTL`).

Sportive factors:

ODDSET odds. Bookmakers' odds on the probability to win a World Cup already entail a great amount of covariates and information about the respective team and, therefore, can be assumed to be a good predictor for the success of a country. The odds were provided by the German state betting agency ODDSET. The bookmakers' odds are converted into winning probabilities by taking the inverse of the odds followed by elimination of the bookmakers' margin. Hence, the variable reflects the probabilities of ODDSET for each team to win the respective World Cup.

FIFA ranking. The FIFA ranking provides a ranking system for all national teams measuring the performance of the team over the last four years. The exact formula for the calculation of the FIFA points and all rankings since implementation of the FIFA ranking system can be found at the official FIFA website (`http://de.fifa.com/worldranking/index.html`). Since the calculation formula of the FIFA points changed after the World Cup 2006, the rankings according to FIFA points are used instead of the points.

Home advantage:

Host. The host of the World Cup could have an advantage over its opponents because of the stronger support of the crowd in the stadium. Therefore, a dummy variable for the respective host of the World Cup is included.

Continent. Before the World Cup 2014, many discussions revolved around the climatic conditions in Brazil and who would deal best with these conditions. One could assume that teams from the same continent as the host of the World Cup (including the host

itself) may have advantages over teams from other continents, as they might better get along with the climatic and cultural circumstances. A dummy variable for the continent of the World Cup host is included.

Factors describing the team's structure:

The following variables are thought to describe the structure of the teams. Each variable was observed with the final squad of 23 players nominated for the respective World Cup.

(Second) maximum number of teammates. If many players from one club play together in a national team, this could lead to an improved performance of the team as the teammates know each other better. Therefore, both the maximum and the second maximum number of teammates from the same club are counted and included as covariates.

Average age. The average age of all 23 players is observed to include possible differences between rather old and rather young teams.

Number of Champions League (Europa League) players. The European club leagues are valuated to be the best leagues in the world. Therefore, the competitions from teams between the best European teams, namely the UEFA Champions League and the UEFA Europa League (previously UEFA Cup) can be seen as the most prestigious and valuable competitions on club level. As a measurement of the success of the players on club level, the number of players in the semi finals (taking place only weeks before the respective World Cup) of these competitions are counted.

Number of players abroad. Finally, the national teams strongly differ in the numbers of players playing in a league of the respective country and players from leagues of other countries. For each team, the number of players playing in clubs abroad (in the season previous to the respective World Cup) are counted.

Factors describing the team's coach:

Also covariates of the coach of the national team may have an influence on the performance of the team. Therefore, the *age* of the coach and the duration of the *tenure* of the coach are observed. Furthermore, a dummy variable is included, if the coach has the same *nationality* as his team or not.

Note that the differences of the three binary variables *host*, *continent* and *nationality* lead to new categorical variables with the three factor levels -1,

0 and +1. Each of these categorical variables is represented by two new dummy variables with -1 as the reference category.

At this point we also want to mention that at the FIFA World Cup 2014 two teams participate, which have not participated in any of the World Cups from 2002-2010, namely Bosnia and Herzegovina and Colombia. In order to obtain nonetheless reasonable estimates for the team-specific effects of both teams, which can then be used for the prediction of the FIFA World Cup 2014, we collect all teams that have only participated once in the tournaments between 2002 and 2014 in a group called NEWCOMERS. This concerns the following 12 teams: Angola, China, Czech Republic, Ireland, New Zealand, North Korea, Senegal, Slovakia, Togo, Trinidad & Tobago, Turkey, Ukraine.

As already mentioned, in the model specifications of Model 1 and 2 from Section 2 all covariates are considered in the form of differences. For example, in the first match of the FIFA World Cup 2002 in Japan and South Korea, where France played against Senegal (which is among the group of NEWCOMERS in our sample), the French team had an *average age* of 28.30 years, was on first place in the current *FIFA ranking* and had a winning probability given by the *ODDSET odds* of 15%, while Senegal's team had an *average age* of 24.30 years, was on 42th place in the current *FIFA ranking* and had a winning probability of 1%. Hence, when the goals of France are considered, this results in the following differences for the metric covariates: $age = 28.30 - 24.30 = 4.00$, $rank = 1 - 42 = -41$, $odd = 0.15 - 0.01 = 0.14$. For the categorical variable $host \in \{-1, 0, 1\}$ we get $host = 0 - 0 = 0$, resulting in the dummies $host0 = 1$ and $host1 = 0$, as the factor level -1 was chosen as the reference category. An extract of the design matrix part, which corresponds to the covariates is presented in Table 1. The matrix resulting from the encoding of the team-specific effects is illustrated in Table 2.

goals	team	opponent	age	rank	odds	host0	host1	...
0	France	Newcomer	4.00	-41	0.14	1	0	...
1	Newcomer	France	-4.00	41	-0.14	1	0	...
1	Uruguay	Denmark	-2.10	4	-0.00	1	0	...
2	Denmark	Uruguay	2.10	-4	0.00	1	0	...
1	Denmark	Newcomer	3.10	-22	0.01	1	0	...
1	Newcomer	Denmark	-3.10	22	-0.01	1	0	...
0	France	Uruguay	3.00	-23	0.14	1	0	...
0	Uruguay	France	-3.00	23	-0.14	1	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Table 1: Extract of the design matrix part which corresponds to the covariates.

FRA.att	FRA.def	NEW.att	NEW.def	URU.att	URU.def	DEN.att	DEN.def
1	0	0	-1	0	0	0	0
0	-1	1	0	0	0	0	0
0	0	0	0	1	0	0	-1
0	0	0	0	0	-1	1	0
0	0	0	-1	0	0	1	0
0	0	1	0	0	0	0	-1
1	0	0	0	0	-1	0	0
0	-1	0	0	1	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Table 2: Encoding of the team specific-effects

4 Results and prediction of the FIFA World Cup 2014

In this section, first we compare the fit of Model 1 and Model 2 from Section 2, in order to select a final model, which is then used for the prediction of the FIFA World Cup 2014.

4.1 Comparison of team-specific Poisson models

As pointed out in Section 2 we use Lasso-type penalization approaches to fit Model 1 and Model 2, with the major difference that in Model 1 the team-specific effects are treated as fixed effects and hence, are also penalized directly by the L_1-penalty term, while in Model 2 they are treated as random effects and, hence, are penalized implicitly by the restrictions imposed by the corresponding normal distribution assumption. The crucial step is now to determine the optimal value of the tuning parameter λ from Equation (1). Note that different levels of sparseness are obtained depending on the selection of the optimal tuning parameter λ. In general, information criteria such as Akaike's information criterion (AIC, see Akaike, 1973) or the Bayesian information criterion (BIC, see Schwarz, 1978), also known as Schwarz's information criterion, could be used, but as our main focus is on achieving good prediction results in order to be able to provide a realistic forecast of the FIFA World Cup 2014, we decided to use 10-fold cross validation (CV) based on the conventional Poisson deviance score [6]. The corresponding 10-fold CV results are illustrated in Figure 1 and 2. There, also the coefficient paths for the (scaled) covariates are shown along the penalty parameter λ. Note that in order to correctly apply the Lasso algorithms, all covariates were scaled to have mean 0 and variance 1. In

[6] As two observations corresponding to the goals of the same match belong together, we do not exclude single observations from the training data, but single matches.

Table 3, the fixed effects estimates for the (unscaled) covariates are shown for both models.

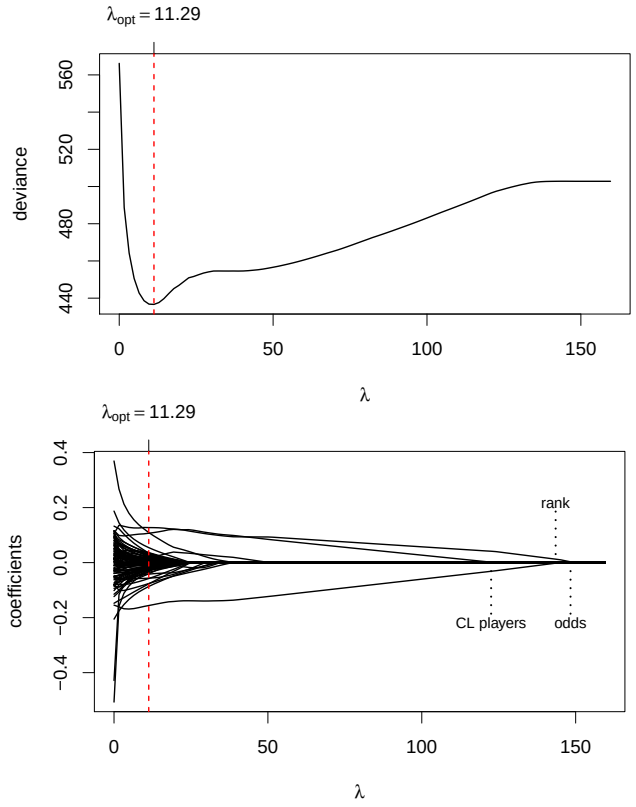

Figure 1: Deviance for 10-fold CV (top) together with coefficient paths (bottom) vs. the penalty parameter λ for Model 1; the optimal value of the penalty parameter λ is shown by the vertical lines.

It is seen that for Model 1, where the team-specific effects are treated as fixed effects and, hence, are directly penalized by the L_1-penalty in the same way as the covariate effects, in addition to the team-specific effects also most of the covariates are selected at the optimal value for λ. The strong explanatory power of the bookmakers' odds is reflected by the fact that this is the first covariate to enter the model. Next, the *FIFA ranking* and the *number of CL players* are included. In the final model all covariates except for the *host* dummy, the *number of players abroad* and the *maximum number of teammates* are included, which indicates that in fact there is additional information provided by other covariates, which is not yet covered by the odds.

Meanwhile, for Model 2, where the team-specific effects are treated as random effects, not a single covariate is selected and the model contains

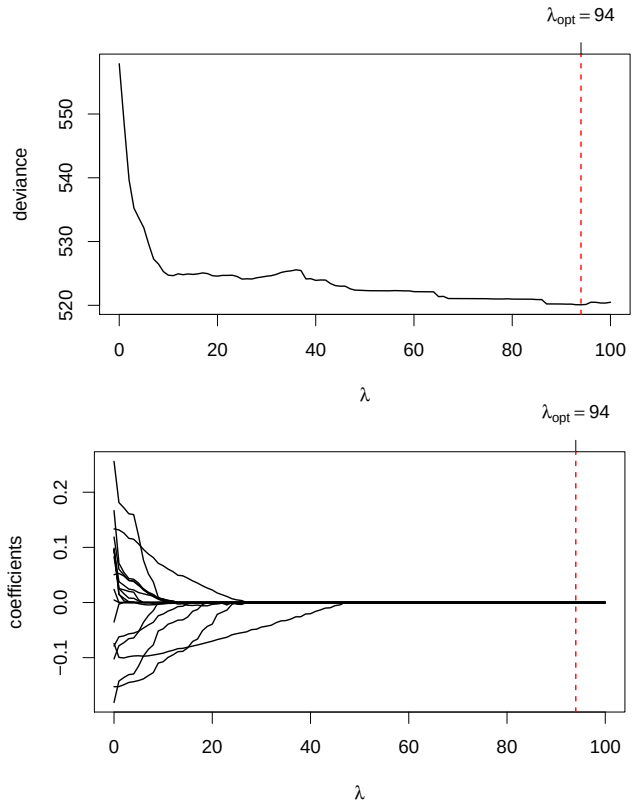

Figure 2: Deviance for 10-fold CV (top) together with coefficient paths (bottom) vs. the penalty parameter λ for Model 2; the optimal value of the penalty parameter λ is shown by the vertical lines.

only random effects for each team's attacking and defending abilities ($\hat{\sigma}^2_{att} = 1.145, \hat{\sigma}^2_{def} = 0.387, \widehat{cov}(att, def) = 0.008$). This seems surprising at first glance. In particular, the first covariates to enter in Model 2 are *population* and *continent* and, therefore, not the covariates one would expect from intuition (or from the results of Model 1). But, the team-specific effects of Model 1 and Model 2 have to be interpreted completely different. Model 2, at a strong level of penalization, is completely dominated by the random effects. Furthermore, as was to be expected, we found high correlations between the team-specific effects and several covariates. For example, rather strong teams like Brazil or Germany always had low values for the *FIFA ranking* or high values for the *ODDSET odds* or for the *number of CL players*. It is well known that estimates of random effects models suffer strongly if the random effects are correlated with covariates, see, for example, Verbeke et al. (2001), Neuhaus and McCulloch (2006). The bias is strongly reduced in fixed effects models (Tutz and Oelker, 2014).

	Model 1	Model 2
	0.051	-0.174
CL.players	0.034	0
UEFA.players	0.008	0
age.Coach	-0.004	0
tenure.Coach	-0.036	0
legionnaires	0	0
max.teammates	0	0
sec.max.teammates	-0.035	0
age	-0.014	0
rank	-0.006	0
GDP	0.035	0
odds	1.907	0
population	-2.722	0
continent0	0.071	0
continent1	-0.006	0
nationality.coach0	0.003	0
nationality.coach1	-0.001	0
host0	0	0
host1	0	0

Table 3: Estimates of the covariate effects for Model 1 and Model 2.

This phenomenon is also reflected in the different ways Model 1 and Model 2 build up their coefficients. At the point of the highest penalization, Model 1 starts with nothing but an intercept, whereas Model 2 starts with (almost unpenalized) team-specific random effects. With decreasing level of penalization, either covariate effects or team-specific effects can enter Model 1, depending on which effects contribute the largest part of information. In contrast, Model 2 can only be improved by covariate effects, if these contain additional information not yet covered by the random effects. Nevertheless, the goodness-of-fit criterion presented in the next paragraph shows that both models lead to an adequate fit (at least "in sample").

Goodness-of-fit. In order to assess the performance of our models, we use a possible goodness-of-fit criterion. In addition to the 32 odds corresponding to all possible tournament winners, which are fixed before the start of the tournament, also the "three-way" odds[7] were provided from the German state betting agency ODDSET for all 192 games of the FIFA World Cups 2002-2010. By taking the three quantities $\tilde{p}_i = 1/\text{odds}_i, i \in$

[7]Three-way odds consider only the tendency of a match with the possible results *victory of team 1*, *draw* or *defeat of team 1* and are usually fixed some days before the corresponding match takes place.

$I := \{1, 2, 3\}$ and by normalizing with $c := \sum_{i \in I} \tilde{p}_i$ in order to adjust for the bookmakers' margins, the odds can be directly transformed into probabilities using $\hat{p}_i = \tilde{p}_i / c$[8]. On the other hand, let G_i denote the random variables representing the number of goals scored by team i in a certain match and G_j the goals of its opponent, respectively. Then, we can compute the same probabilities by approximating $\hat{p}_1 = P(G_i > G_j), \hat{p}_2 = P(G_i = G_j)$ and $\hat{p}_3 = P(G_i < G_j)$ for each of the 192 matches using the corresponding Poisson distributions $G_i \sim Poisson(\hat{\lambda}_i)$, $G_j \sim Poisson(\hat{\lambda}_j)$, where the estimates $\hat{\lambda}_i$ and $\hat{\lambda}_j$[9] are obtained by our regression models. Hence, we can provide a goodness-of-fit criterion by comparing the values of the log-likelihood of the 192 matches for the ODDSET odds with those obtained for our regression models. For $\omega_j \in I, j = 1, \ldots, 192$, the likelihood is given by the product $L_{\text{three-way}} := \prod_{j=1}^{192} \hat{p}_{1j}^{\delta_{1\omega_j}} \hat{p}_{2j}^{\delta_{2\omega_j}} \hat{p}_{3j}^{\delta_{3\omega_j}}$, with δ_{ij} denoting Kronecker's delta. Based on this log-likelihood, we can compute a corresponding deviance-type score $D_{\text{three-way}} = -2 \log(L_{\text{three-way}})$. The deviance results corresponding to Model 1 and 2 and for the ODDSET odds are found in Table 4. In general, the regression models should be able to produce lower deviance scores compared to the deviance score corresponding to the ODDSET odds, indicating a better fit to the realized "three-way" tendencies. If the fits obtained by our models would not even be able to beat the bookmakers' odds "in sample", the whole regression analysis would be useless. That would mean that one would achieve a better fit just by following the bookmakers' odds, which are publicly available shortly before the matches and thus are "out-of-sample". The results in Table 4 show that for both settings the fit obtained by our regression models clearly outperforms the deviance score corresponding to the ODDSET odds and, hence, both models seem reasonable.

`glmmLasso`		ODDSET odds
Model 1	Model 2	
338.880	333.814	365.472

Table 4: Deviance scores $D_{\text{three-way}}$ for Model 1 and Model 2 and the ODDSET odds.

At this point of our analysis, we finally have to choose between Model 1 and Model 2 for the prediction of the current FIFA World Cup 2014.

[8] The transformed probabilities only serve as an approximation, based on the assumption that the bookmakers' margins follow a discrete uniform distribution on the three possible match tendencies.

[9] For convenience we suppress the index k for both teams here, which indicates the corresponding World Cup, as well as the indices corresponding to the opponent. One should correctly write $\hat{\lambda}_{ijk}$ and $\hat{\lambda}_{jik}$, if team i and team j are facing each other at World Cup k.

The final decision was primarily based on the expected prediction accuracy of both models for a new World Cup. The deviances plotted in Figures 1 and 2 clearly showed better results for Model 1 ($\min(dev_{Mod1}) \approx 440, \min(dev_{Mod2}) \approx 520$). This is not surprising. For the model fit, it does not matter whether (in the case of high correlations between teams and covariates) an effect is included in a team-specific effect or in a covariate effect. However, for the prediction, especially the prediction of a new World Cup, one can assume that strong covariate effects are better than strong team-specific effects, which vary over time. If the strength of a team differs strongly from the respective strength from the previous World Cups, this can only be covered by current covariates of the respective team. Belgium, for example, did not perform very well in the previous World Cups covered by our sample. For the World Cup 2014, however, they are ranked among experts as a secret favorite of the tournament. This cannot be taken into account by Model 2 solely consisting of team-specific random effects. Model 1, however, includes covariates as, for example, the *FIFA ranking* and the bookmakers' expectations, which show an improvement of Belgium with respect to the previous tournaments and, therefore, improve Belgium's predicted chances for the current World Cup. Based on these considerations, Model 1 was chosen as the final model to predict the outcome of the World Cup 2014.

In Table 5 and Table 6, the corresponding estimates of the (unscaled) fixed team-specific attacking and defending effects are summarized. It is striking that compared to all other teams Germany and Brazil both have rather high attacking and defending abilities: Germany's attack is on 4th place, its defense is on 6th place; Brazil's attack is on 5th place, its defense on 10th place. Most other teams have either a rather bad attacking or defense parameter. Nevertheless, one has to keep in mind that these team-specific effects cannot be interpreted independently from the covariate effects. In particular, they rather have to be seen as remaining effects not yet covered by the covariates. In this context, also the parameters of Switzerland are interesting. Switzerland has a rather bad attack, but the best defense parameter among all the teams. This can be easily explained, as Switzerland has received only a single goal in its seven games at the World Cups 2006 and 2010, but on the other hand only scored five goals in these seven matches.

4.2 Probabilities for FIFA World Cup 2014 winner

In the following, we have used the estimates from Model 1 to simulate the tournament progress of the FIFA World Cup 10,000 times. Note here that one advantage in comparison to several other prediction approaches is that we are able to draw exact match outcomes for each match by drawing the

1.	RSA	0.210	11.	DEN	0.075	21.	JPN -0.020	31.	IRN -0.094
2.	URU	0.204	12.	USA	0.073	22.	MEX -0.023	32.	NED -0.100
3.	CIV	0.157	13.	ARG	0.071	23.	NEW -0.023	33.	CRO -0.126
4.	GER	0.153	14.	POR	0.068	24.	SWE -0.033	34.	SUI -0.173
5.	BRA	0.137	15.	ECU	0.054	25.	POL -0.053	35.	TUN -0.193
6.	BEL	0.123	16.	PAR	0.015	26.	SRB -0.059	36.	FRA -0.223
7.	KOR	0.123	17.	CHI	0.001	27.	SVN -0.061	37.	KSA -0.227
8.	RUS	0.114	18.	GHA	-0.007	28.	GRE -0.067	38.	CMR -0.235
9.	CRC	0.104	19.	ESP	-0.011	29.	ENG -0.078	39.	HON -0.256
10.	AUS	0.079	20.	ITA	-0.019	30.	NGA -0.090	40.	ALG -0.371

Table 5: Estimates of the (fixed) team-specific attacking effects att_i for Model 1.

1.	SUI	0.407	11.	CRO	0.082	21.	ESP 0.002	31.	DEN -0.143
2.	HON	0.324	12.	NED	0.077	22.	NGA 0.001	32.	POL -0.169
3.	ALG	0.265	13.	FRA	0.059	23.	ARG -0.006	33.	AUS -0.196
4.	PAR	0.209	14.	JPN	0.047	24.	URU -0.010	34.	SRB -0.211
5.	POR	0.162	15.	CHI	0.042	25.	RUS -0.045	35.	SVN -0.219
6.	GER	0.130	16.	SWE	0.017	26.	GRE -0.060	36.	BEL -0.264
7.	ECU	0.123	17.	KOR	0.016	27.	RSA -0.092	37.	TUN -0.287
8.	GHA	0.115	18.	MEX	0.014	28.	USA -0.126	38.	IRN -0.299
9.	ENG	0.094	19.	NEW	0.009	29.	CIV -0.135	39.	CRC -0.466
10.	BRA	0.089	20.	ITA	0.009	30.	CMR -0.141	40.	KSA -0.622

Table 6: Estimates of the (fixed) team-specific defending effects def_i for Model 1.

goals of both competing teams from the predicted Poisson distributions, i.e. $G_i \sim Poisson(\hat{\lambda}_i)$, $G_j \sim Poisson(\hat{\lambda}_j)$, with estimates $\hat{\lambda}_i$ and $\hat{\lambda}_j$ from Model 1. This allows us to precisely follow the official FIFA rules when determining the final group standings[10]. If a match in the knockout stage ended in a draw, it was simulated again until a winner was found.

Based on these simulations, for each of the 32 participating teams probabilities to win the tournament are obtained. These are summarized in Table 7 together with the winning probabilities based on the ODDSET odds for comparison. In contrast to most other prediction approaches for the current World Cup clearly favoring Brazil, we get a neck-and-neck race between Germany and Brazil with slightly better chances for Germany. The major reason for this is that in the simulations with a high probability

[10]The final group standings are determined by (1) the number of points, (2) the goal difference and (3) the number of scored goals. If several teams coincide with respect to all of these three criteria, a separate chart is calculated based on the matches between the coinciding teams only. Here, again the final standing of the teams is determined following criteria (1)-(3). If still no distinct decision can be taken, the decision is taken by lot.

both Germany and Brazil finish their groups on the first place and then face each other in the semi final. In a direct duel, the model concedes Germany a wafer-thin advantage with winning odds of 50,4% against 49,6%. The favorites Germany and Brazil are followed by the teams of Spain, Belgium, Argentina and Portugal. Note that based on the 10,000 simulation runs also survival probabilities for each team and for each tournament stage can be obtained, but are skipped here for the sake of simplicity.

team	$\hat{p}_{\text{Lasso}}$	$\hat{p}_{\text{ODDSET}}$	team	$\hat{p}_{\text{Lasso}}$	$\hat{p}_{\text{ODDSET}}$
1. GER	0.2880	0.1420	17. GHA	0.0022	0.0071
2. BRA	0.2765	0.2028	18. KOR	0.0019	0.0024
3. ESP	0.0900	0.1092	19. ALG	0.0018	0.0071
4. BEL	0.0819	0.0592	20. ECU	0.0017	0.0071
5. ARG	0.0582	0.1420	21. USA	0.0016	0.0071
6. POR	0.0522	0.0237	22. MEX	0.0012	0.0071
7. SUI	0.0413	0.0071	23. JPN	0.0010	0.0047
8. CRO	0.0210	0.0071	24. BIH	0.0008	0.0047
9. ENG	0.0193	0.0355	25. GRE	0.0005	0.0071
10. FRA	0.0135	0.0355	26. RUS	0.0004	0.0118
11. NED	0.0129	0.0355	27. NGA	0.0004	0.0035
12. ITA	0.0094	0.0355	28. AUS	0.0003	0.0024
13. URU	0.0071	0.0284	29. HON	0.0002	0.0005
14. CHI	0.0063	0.0203	30. CRC	0	0.0071
15. COL	0.0052	0.0394	31. CMR	0	0.0024
16. CIV	0.0032	0.0071	32. IRN	0	0.0005

Table 7: Estimated winning probabilities for all 32 teams based on 10,000 simulation runs of the FIFA World Cup 2014 and based on the estimates of Model 1 together with winning probabilities based on the ODDSET odds.

4.3 Most probable tournament outcome

Finally, based on the 10,000 simulations, we also provide the most probable tournament outcome. Here, for each of the eight groups we selected the most probable final group standing, also regarding the order of the first two places, but without regarding the irrelevant order of the teams on place three and four. The results together with the corresponding probabilities are presented in Table 8.

Group A 44%	Group B 24%	Group C 16%	Group D 18%
1. BRA	1. ESP	1. COL	1. ENG
2. CRO	2. NED	2. CIV	2. ITA
MEX	CHI	JPN	URU
CMR	AUS	GRE	CRC

Group E 22%	Group F 36%	Group G 37%	Group H 26%
1. SUI	1. ARG	1. GER	1. BEL
2. FRA	2. BIH	2. POR	2. KOR
ECU	NGA	GHA	RUS
HON	IRN	USA	ALG

Table 8: Most probable final group standings together with the corresponding probabilities for the FIFA World Cup 2014 based on 10,000 simulation runs and on the estimates of Model 1.

It is obvious that there are large differences with respect to the groups' balances. While in Group A and Group G the model forecasts Brazil followed by Croatia as well as Germany followed by Portugal with rather high probabilities of 44% and 37%, respectively, other groups such as Group C, Group D and Group E seem to be quite close.

Based on the most probable group standings, we also provide the most probable course of the knockout stage, compare Figure 3. Finally, according to the most probable tournament course the German team will take

home the World Cup trophy. Although according to the model this reflects the most probable tournament outcome, it only has a very low overall probability of $3.991 \cdot 10^{-6}$ % (given as the product of all single probabilities of Table 8 and Figure 3). Hence, deviations of the true tournament outcome from the model's most probable one are not only possible, but very likely.

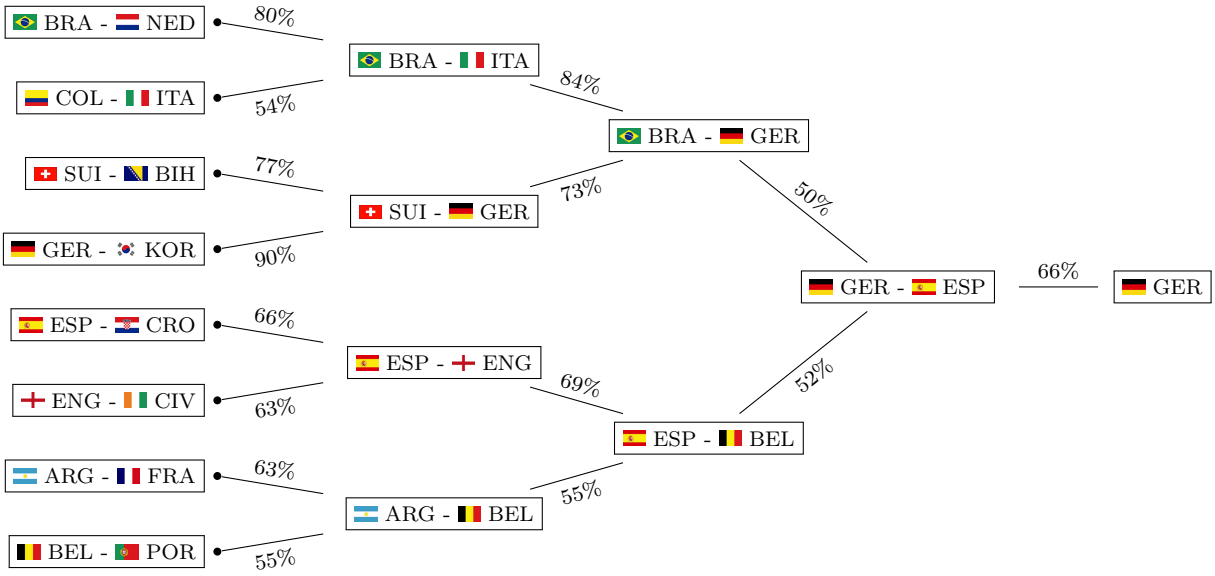

Figure 3: Most probable course of the knockout stage together with corresponding probabilities for the FIFA World Cup 2014 based on 10,000 simulation runs and on the estimates of Model 1.

5 Concluding remarks

Two different team-specific generalized linear (mixed) Poisson models for the number of goals scored by national teams facing each other in FIFA World Cup matches are set up. The difference between the two models is the specification of the team-specific effects, either as fixed or as random effects. The FIFA World Cups 2002-2010 serve as the data basis for an analysis of the influence of several covariates on the success of national teams in terms of the number of goals they score in single matches. Procedures for variable selection based on an L_1-penalty, implemented in the R-packages `grplasso` and `glmmLasso`, are used and compared. An "in-sample" performance measure is applied that is based on the log-likelihood corresponding to the three-way tendencies of the considered matches.

With regard to the predictive performance of both models, the model with fixed team-specific effects was chosen as the final model. This model

was used for simulation of the FIFA World Cup 2014. According to this simulation, Germany and Brazil turned out to be the top favorites for winning the title, with a wafer-thin advantage for Germany. Besides, the most probable tournament outcome is provided.

In particular, the big differences between Model 1 and Model 2, especially with regard to the set of selected covariates are noteworthy. While Model 1 incorporates most covariates, for Model 2 not a single covariate was included and the model is solely consisting of random effects for each team's attacking and defending abilities. This can (at least partly) be explained by the high correlations that were found between the team-specific effects and some covariates such as the *FIFA ranking*, the *ODDSET odds* or the *number of CL players*. In contrast to Model 1, the effects for these covariates are (at least to some extent) included in the team-specific random effects and parameter estimates are biased. Apart from the major aim of this article, namely the prediction of a major soccer event, we found this to be an interesting and special data situation. In particular, in combination with the used regularization approaches, correlations between covariates and cluster-specific effects lead to interesting statistical challenges.

Acknowledgement

We are grateful to Falk Barth and Johann Summerer from the ODDSET-Team for providing us all necessary odds data and to Sven Grothues from the Transfermarkt.de-Team for providing us market values for the 2006 and 2010 FIFA World Cups. The article has strongly benefited from a methodical and statistical perspective by suggestions from Helmut Küchenhoff and Christian Groll. The insightful discussions with the hobby football experts Tim Frohwein and Johannes Stahl also helped a lot to improve the article.

References

Akaike, H. (1973). Information theory and the extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, 267–281.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association 88*, 9–25.

Dyte, D. and S. R. Clarke (2000). A ratings based Poisson model for World Cup soccer simulation. *Journal of the Operational Research Society 51 (8)*, 993–998.

Elo, A. E. (2008). *The Rating of Chess Players, Past and Present*. San Rafael: Ishi Press.

Eugster, M. J. A., J. Gertheiss, and S. Kaiser (2011). Having the second leg at home - advantage in the UEFA Champions League knockout phase? *Journal of Quantitative Analysis in Sports 7*(1).

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd ed.). New York: Springer-Verlag.

Goldman-Sachs Global Investment Research (2014). The world cup and economics 2014. `http://www.goldmansachs.com/our-thinking/outlook/world-cup-and-economics-2014-folder/world-cup-economics-report.pdf`.

Groll, A. (2014). *glmmLasso: Variable Selection for generalized linear mixed models by L_1-penalized estimation*. R package version 1.3.1.

Groll, A. and J. Abedieh (2013). Spain retains its title and sets a new record - generalized linear mixed models on European football championships. *Journal of Quantitative Analysis in Sports 9*(1), 51–66.

Groll, A. and G. Tutz (2014). Variable selection for generalized linear mixed models by L_1-penalized estimation. *Statistics and Computing 24*(2), 137–154.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Karlis, D. and I. Ntzoufras (2003). Analysis of sports data by using bivariate poisson models. *The Statistician 52*, 381–393.

Lee, A. J. (1997). Modeling scores in the premier league: is manchester united really the best? *Chance 10*, 15–19.

Leitner, C., A. Zeileis, and K. Hornik (2008). Who is going to win the EURO 2008? (A statistical investigation of bookmakers odds). Research report series, Department of Statistics and Mathematics, University of Vienna.

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

Leitner, C., A. Zeileis, and K. Hornik (2010b). Forecasting the winner of the FIFA World Cup 2010. Research report series, Department of Statistics and Mathematics, University of Vienna.

Leitner, C., A. Zeileis, and K. Hornik (2011). Bookmaker concensus and agreement for the UEFA Champions League 2008/09. *IMA Journal of Management Mathematics 22 (2)*, 183–194.

Lloyd's (2014). Fifa world cup: How much are those legs worth? `http://www.lloyds.com/news-and-insight/news-and-features/market-news/industry-news-2014/fifa-world-cup-how-much-are-those-leg-worth`.

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

Meier, L., S. Van de Geer, and P. Bühlmann (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society B 70*, 53–71.

Neuhaus, J. M. and C. E. McCulloch (2006). Separating between-and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(5), 859–872.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rue, H. and O. Salvesen (2000). Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician) 49*(3), 399–418.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Silver, N. (2014). Its Brazils World Cup to Lose. `http://fivethirtyeight.com/features/its-brazils-world-cup-to-lose/`.

Stoy, V., R. Frankenberger, D. Buhr, L. Haug, B. Springer, and J. Schmid (2010). Das Ganze ist mehr als die Summe seiner Lichtgestalten. Eine ganzheitliche Analyse der Erfolgschancen bei der Fußballweltmeisterschaft 2010. Working Paper 46, Eberhard Karls University, Tübingen, Germany.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58*, 267–288.

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge: Cambridge University Press.

Tutz, G. and M. Oelker (2014). Modeling clustered heterogeneity: Fixed effects, random effects and mixtures. Technical Report 156, Department of Statistics LMU Munich.

Verbeke, G., B. Spiessens, and E. Lesaffre (2001). Conditional linear mixed models. *The American Statistician 55*(1), 25–34.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B 68*, 49–67.

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

Zeileis, A., C. Leitner, and K. Hornik (2014). Home Victory for Brazil in the 2014 FIFA World Cup. Working paper, Faculty of Economics and Statistics, University of Innsbruck.