

Andreas Groll*, Gunther Schaubberger and Gerhard Tutz

Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA World Cup 2014

Abstract: In this article an approach for the analysis and prediction of international 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, within the generalized linear model (GLM) framework, also differences of team-specific effects are incorporated. In order to achieve variable selection and shrinkage, we use tailored Lasso approaches. Based on preceding FIFA World Cups, two models for the prediction of the FIFA World Cup 2014 are fitted and investigated. Based on the model estimates, the FIFA World Cup 2014 is simulated repeatedly and winning probabilities are obtained for all teams. Both models favor the actual FIFA World Champion Germany.

Keywords: FIFA World Cup 2014; football; LASSO; prediction; sports tournaments; variable selection.

DOI 10.1515/jqas-2014-0051

1 Introduction

In the last few years various approaches to the statistical modeling of major international soccer events have been proposed, among them the Union of European Football Associations (UEFA) Champions League (CL; Karlis and Ntzoufras 2011; Eugster, Gertheiss, and Kaiser 2011), the European football championship (EURO; Leitner, Zeileis, and Hornik 2010a; Zeileis, Leitner, and Hornik 2012; Groll and Abedieh 2013) or the Fédération Internationale de Football Association (FIFA) World Cup (Leitner, Zeileis,

and Hornik 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, Leitner, and Hornik (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. (2010a) and their follow-up papers. 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) predicted 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. Although the bookmakers' models certainly contain covariate information of the competing teams, at least indirectly, an alternative approach is to explicitly model the influence of covariates on the success of soccer teams.

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)

*Corresponding author: **Andreas Groll**, Department of Mathematics, Ludwig-Maximilians-University, Theresienstr. 39, 80333 Munich, e-mail: andreas.groll@math.lmu.de

Gunther Schaubberger and Gerhard Tutz: Department of Statistics, Ludwig-Maximilians-University, Munich, Bavaria, Germany

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), Dixon and Coles (1997), Dyte and Clarke (2000), Rue and Salvesen (2000) and Karlis and Ntzoufras (2003). 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 rankings¹ 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 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 other prediction approaches, which cannot be classified into one of the two aforementioned major categories of statistical approaches for modeling soccer data. For example, Dobson and Goddard (2011) or Forrest and Simmons (2000) use discrete choice models for the modeling of match outcomes. Concerning the prediction of the FIFA World Cup 2014, an approach proposed by Silver (2014) 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: the London insurance market Lloyd's (2014) uses 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.

In the approach that we propose here we focus on international soccer tournaments, here applied to FIFA World Cups, and use a Poisson model for the number of goals scored by competing teams in the single matches of the tournaments. Several potential influence variables are considered and, additionally, team-specific effects are included in the form of fixed effects, resulting in a flexible generalized linear model (GLM). Incorporating a method for the selection of relevant predictors, we obtain a regularized solution for our model. The variable selection is based on suitable L_1 -penalization techniques and is performed with the `grplasso` function from the corresponding R-package (see Meier, Van de Geer, and Bühlmann 2008). As an application, the approach is used to fit data from previous FIFA World Cups and finally, based on the obtained estimates, the FIFA World Cup 2014 is predicted.

¹ 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).

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. 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; FC Porto's triumph in the UEFA CL season 2003/2004. 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).

The rest of the article is structured as follows. In Section 2, we introduce the team-specific Poisson model for the number of goals. Section 3 entails a description of the data for the application to the FIFA World Cup, including a list of possible influence variables. Furthermore, the model is fitted to the data and 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

Our proposed model 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.

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

$$y_{ijk} \mid \mathbf{x}_{ik}, \mathbf{x}_{jk} \sim \text{Pois}(\lambda_{ijk})$$

$$\log(\lambda_{ijk}) = \beta_0 + (\mathbf{x}_{ik} - \mathbf{x}_{jk})^\top \boldsymbol{\beta} + \text{att}_i - \text{def}_j. \quad (1)$$

It is assumed that the number of goals that team i scores follows a Poisson distribution with given team-specific parameters and covariates of both teams. 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 tournament k are collected in a vector $\mathbf{x}_{ik} = (x_{ik1}, \dots, x_{ikp})^\top$ of length p . In the following, we assume that the covariates of each team can vary over different tournaments (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 $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ and β_0 represents the intercept.

If the linear predictor of the model is re-formulated, it can be denoted by

$$\begin{aligned} \eta_{ijk} &= \beta_0 + (\mathbf{x}_{ik} - \mathbf{x}_{jk})^\top \boldsymbol{\beta} + \text{att}_i - \text{def}_j \\ &= \beta_0 + \mathbf{x}_{ik}^\top \boldsymbol{\beta} + \text{att}_i - \mathbf{x}_{jk}^\top \boldsymbol{\beta} + \text{def}_j \\ &= \beta_0 + \gamma_i - \delta_j. \end{aligned}$$

Here, $\gamma_i = \mathbf{x}_{ik}^\top \boldsymbol{\beta} + \text{att}_i$ and $\delta_j = \mathbf{x}_{jk}^\top \boldsymbol{\beta} + \text{def}_j$ represent the total attack ability of team i and defense ability of team j , respectively. Hence, att_i and def_j act as additional parameters covering ability differences that are not covered by the covariate effects yet.

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 concept of regularization is to maximize a penalized version of the log-likelihood $l(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)^\top$ represents a general parameter vector. More precisely, one maximizes the penalized log-likelihood

$$l_p(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - \lambda J(\boldsymbol{\alpha}), \quad (2)$$

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 or by cross-validation. Model selection criteria are usually based on a compromise between the model fit (e.g., in terms of the likelihood) and the complexity of the model, like AIC (Akaike 1973) or BIC (Schwarz 1978). 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 least absolute shrinkage and selection operator (LASSO) penalty. The LASSO estimator was originally proposed by Tibshirani (1996) and uses the penalty

$$J(\alpha) = \sum_{i=1}^p |\alpha_i|. \quad (3)$$

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

The team-specific ability parameters att_i and def_j are considered as fixed effects and 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.² 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.

In the following, both team-specific effects corresponding to one team are treated as a group. Hence, the original LASSO penalty from Equation (3) has to be modified appropriately according to the so-called Group LASSO penalty proposed by Yuan and Lin (2006). The Group LASSO penalizes the L_2 -norm of the respective parameter vectors $(att_i, def_i)^\top, \dots, (att_n, def_n)^\top$. Hence, the parameters of attack and defense abilities of single teams are simultaneously shrunk towards zero and, if shrunk exactly to zero, excluded from the model. Besides, the covariate effects β are penalized using the ordinary LASSO penalty from Equation (3). Altogether, the penalty term for Model (1) is given by

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

The prefactor $\sqrt{2}$ controls for the group sizes of the groups of team-specific parameters, compare Yuan and Lin (2006) or Meier et al. (2008). Another advantage of penalization is the way correlated predictors are treated.

If two predictors are highly correlated, the parameter estimates are stabilized by the penalization. The chosen LASSO penalty tends to include only one of the predictors and only includes the second predictor if it entails additional information for the response variable. Therefore, if several variables possibly contain information on the strength of teams they can be used simultaneously. The most informative variable is chosen automatically by the penalty term. The model can easily be fitted by use of the `grplasso` function from the corresponding R-package (see Meier et al. 2008).

Note that, alternatively, similar to the model used in Groll and Abedieh (2013) the team-specific effects could be estimated as random effects. Then, the attack and the defense parameter of team i are assumed to be multivariate normally distributed. In this case, the ability parameters are automatically regularized by the assumption of a distribution and only the covariate effects β are explicitly penalized by using LASSO. The algorithm `glmLasso` proposed in Groll and Tutz (2014) can be used to fit this model. However, this results in a model more focused on team-specific effects than covariate effects due to the different, namely lower, penalization of the random team-specific effects. Therefore, this modeling approach is not pursued in the following.

3 Application

In the following, the proposed model is applied to data from previous FIFA World Cups and is then used to predict the FIFA World Cup 2014 in Brazil.

3.1 Data

In this section, we give a brief description of the used 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.

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

² Usually, for reasons of identifiability, categorical predictors with k factor levels are coded by $k - 1$ dummies. However, the regularization approach introduced in the following (with $\lambda > 0$) provides unique estimates despite the issues of identifiability.

worldwide average GDP per capita is used. The GDP data were collected from 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 national team. 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.³

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.⁴

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 valued 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.

Unfortunately, the covariate *ODDSET odds* is not available before the FIFA World Cup 2002. But as this covariate can be assumed to contain already a lot of expertise and information about an upcoming World Cup, we decided to perform a separate analysis for the FIFA World Cup data from 2002 to 2010 (from now on denoted by WC2002),

³ 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.

⁴ The FIFA ranking was introduced in August 1993.

including the odds. But as this results in a quite small data basis, another analysis will be performed on a data set including the World Cups from 1994 to 2010, excluding the covariate *ODDSET odds* (from now on denoted by WC1994).

Note that the differences of the three binary variables *host*, *continent* and *nationality*, which originally have been encoded with $\{0, 1\}$, lead to new categorical variables with the three factor levels $-1, 0$ and $+1$. For each of these new categorical covariates we use dummy encoding with -1 as the reference category and, hence, obtain two columns per covariate in the design matrix, e.g., *host0* and *host1*, corresponding to the factor levels 0 and 1 , respectively. The dummy variables corresponding to one categorical covariate are treated as groups and, hence, are also penalized by a Group LASSO penalty, similar to the attack and defense ability parameters.

It should be noted that at the FIFA World Cup 2014 the national team of Bosnia and Herzegovina participated for the very first time. Therefore, for this team no estimates of its team-specific effects are available. Analogously, the national team of Colombia participating also at the FIFA World Cup 2014 did not participate in any of the FIFA World Cups from 2002 to 2010. In order to obtain nonetheless reasonable estimates for the team-specific effects of such 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 from the respective data basis in a group called *newcomers*. Therefore, these teams share the same team-specific ability parameters. Exemplarily, for the WC2002 data 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 specification of Model (1) 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: $age = 28.30 - 24.30 = 4.00$, $rank = 1 - 42 = -41$, $odds = 0.15 - 0.01 = 0.14$. For the categorical variable *host* $\in \{-1, 0, 1\}$ we get $host = 0 - 0 = 0$, which results in the entries $host0 = 1$ and $host1 = 0$ in the two columns of the design matrix corresponding to the dummy encoding, 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.

3.2 Estimation results

In this section, we present the fit of Model (1) from Section 2 on the basis of both data sets, i.e., the FIFA World Cups 1994–2010 and 2002–2010, which is then used for the prediction of the FIFA World Cup 2014.

As pointed out in Section 2 we use LASSO-type penalization approaches to fit the Model (1). The crucial step is now to determine the optimal value of the tuning parameter λ from Equation (2). 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

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

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 2: Encoding of the team specific-effects.

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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

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.⁵ The corresponding 10-fold CV results are illustrated in Figure 1, exemplarily for the WC1994 data. Additionally, in Figure 2 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 (both binary and continuous) were scaled to have mean 0 and variance 1. Besides, Figure 3 illustrates the coefficient paths of the team-specific attack and defense parameters. In Table 3, the fixed effects estimates for the (scaled) covariates are shown for both data sets.

The optimal tuning parameter λ , which minimizes the deviance shown in Figure 1, leads to a model with 10 (out of possibly 17) regression coefficients different from zero for the WC1994 data set. The paths illustrated in Figure 2 show that the first covariate to be selected is

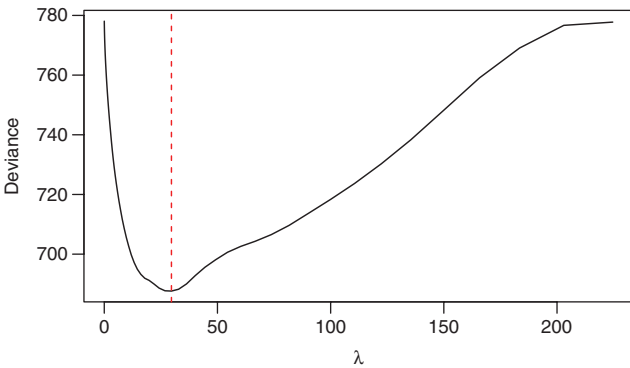


Figure 1: Deviance for 10-fold CV for Model (1), exemplarily for the FIFA World Cup data 1994–2010; the optimal value of the penalty parameter λ is shown by the vertical line.

⁵ 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.

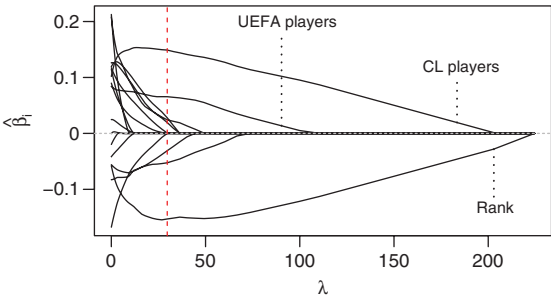


Figure 2: Coefficient paths of the covariate effects vs. the penalty parameter λ , exemplarily for the FIFA World Cup data 1994–2010; the optimal value of the penalty parameter λ is shown by the vertical line.

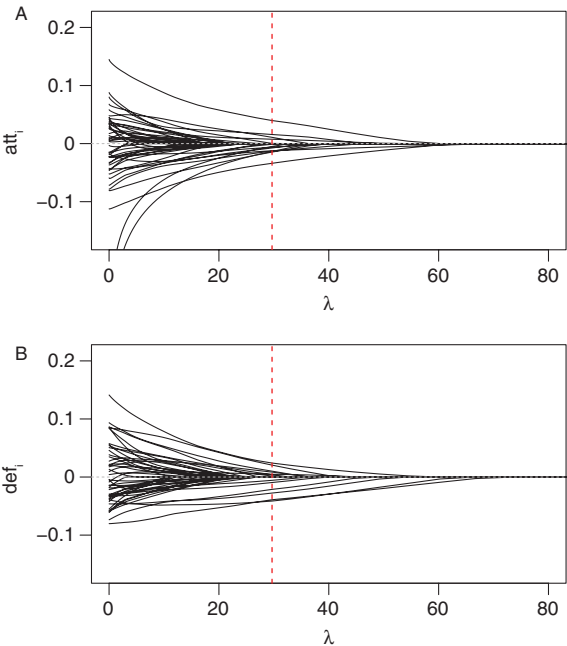


Figure 3: Coefficient paths of the team-specific attack (A) and defense effects (B) vs. the penalty parameter λ , exemplarily for the FIFA World Cup data 1994–2010; the optimal value of the penalty parameter λ is shown by the vertical lines.

Table 3: Estimates of the covariate effects for the FIFA World Cups 1994–2010 and 2002–2010.

	WC 1994–2010	WC 2002–2010
CL players	0.149	0.075
UEFA players	0.066	0
Age Coach	0	−0.017
Tenure Coach	0	−0.071
Legionaires	0	0
Max. # teammates	0	0
Sec. max # teammates	−0.053	0
Age	0	0
Rank	−0.153	−0.167
GDP	0.024	0.042
Odds	−	0.113
Population	−0.031	−0.060
Continent0	0.001	0.010
Continent1	0.000	−0.003
Nation Coach0	0	0
Nation Coach1	0	0
Host0	0.019	0
Host1	0.028	0

the *FIFA rank*, followed by the *number of CL players* and *number of UEFA players* (when the penalty parameter λ decreases). Together with the fact that the estimated effects of these three covariates also exhibit the highest absolute values, this indicates that the three covariates offer the highest explanatory power among all regarded covariates. The estimated coefficients show the intuitively expected effects: better, i.e., lower, FIFA ranks and more players that have been successful with their clubs in the UEFA Champions or Europa League have positive effects on the number of goals scored. It is also worth mentioning that at the optimal tuning parameter, for several teams the ability estimates are still zero, compare Figure 3.

In general, similar graphs are obtained for the smaller WC2002 data, which includes the *ODDSET odds* as a covariate. The major difference is that the *ODDSET odds* are the first variable to enter the model, followed by the *FIFA rank*. This confirms the supposition that the bookmakers' odds cover already a lot of information and, hence, provide strong explanatory power in the context of the success of soccer teams. Again, also the *number of CL players*, the third covariate that enters the model, seems to play an important role.

The model including the odds is sparser with only 9 out of 18 regression coefficients different from zero. A possible explanation is that the *ODDSET odds* already include a lot of information from other covariates, as for example the host effect, which has been found in the WC1994 data.

3.3 Goodness-of-fit

It is well-known that the scores of both competing teams in a soccer match are correlated. Several approaches to handle the correlation have been proposed in the literature. For example, in an unregularized setting McHale and Scarf (2006, 2011) model the dependence by using bivariate discrete distributions and by specifying a suitable family of dependence copulas. One of the first works investigating the topic of dependency between scores of competing soccer team is the fundamental article of Dixon and Coles (1997). They have shown that the joint distribution of the scores of both teams can not be well represented by the product of two independent marginal Poisson distributions of the home and away teams. They suggest to use an additional term to adjust for certain under- and overrepresented match results. After all, these findings are based on the marginal distributions and only hold for models where the predictors of both scores are uncorrelated. However, the model proposed by Dixon and Coles (1997) includes team-specific attack and defense ability parameters and then uses independent poisson distributions for the numbers of goals scored. Therefore, the linear predictor for the number of goals of a specific team depends both on parameters of the team itself and its competitor. When fitting such a model to our World Cup data it turned out that the estimates of the attack and defense abilities of the teams are positively correlated. Therefore, although independent Poisson distributions are used for the scores in one match, the linear predictors and, accordingly, the predicted outcomes are (negatively) correlated. This holds both for the model of Dixon and Coles (1997) and, even more, for our proposed model where the linear predictors additionally entail covariates of both teams. To check if this phenomenon represents the actual correlations between the scores in one match in an appropriate manner, we compared the correlations between the real outcomes and the predictions from our model, exemplarily for the WC1994 data. We measured the correlation between 10,000 predictions for every match from the data set and compared it to the actual correlation between the scores in these matches. While we found a rank correlation (Spearman) of $\rho_{data} = -0.0882$ for the real outcomes, the predictions from our model have a rank correlation of $\rho_{model} = -0.0908$. The correlations according to Bravais-Pearson show similar results, $r_{data} = -0.1387$ and $r_{model} = -0.0968$. Alternatively, one can also investigate the residuals of the fitted model. If the model is representing the correlation structure in the data appropriately, the residuals belonging to the same match should be uncorrelated. For the WC1994 data we found correlations (accompanied

by 95% bootstrap confidence intervals) according to Bravais-Pearson of 0.0198 (CI: $[-0.0867; 0.1283]$) for the deviance residuals and of 0.0062 (CI: $[-0.0977; 0.1141]$) for the Pearson residuals, respectively. In general, the point estimates show that the actual residuals of our model are uncorrelated. Still, due to the rather low number of observations, we obtain rather wide confidence intervals. Altogether, the correlations within the linear predictors for both teams competing in a match seem to fully account for the correlation between the scores of those teams and there is no need for further adjustment.

In a second step, we examined the actual distributions of the numbers of goals and compared them to the following (conditional) probabilities predicted by our model: separately for each plausible score from 0 to 5 goals we compared the observed proportion of the score in the data set with the probabilities for this score predicted by the model on only those observations showing this score. Figure 4 shows the corresponding boxplots, both using the WC1994 data (upper plot) and the WC2002 data including the odds (lower plot). The boxplots represent the probabilities of the respective scores predicted by our model, conditioned on those observations, whose actual number of goals equate to those scores. The red lines represent the

relative frequencies of the respective scores in the data set. Note that if no statistical model is available the relative frequencies would serve as a natural, simple basis for the sampling of scores. So every statistical model should be able to compete with these relative frequencies in the sense that it should produce conditional predicted probabilities for each score exceeding these frequencies as far as possible. It can be seen that the model shows a good prediction performance regarding the number of goals. For example, for those 191 observations with an actual number of goals of zero, we observed a median of the conditional predicted probabilities of 36.1%, while the proportion in the data set for scores of zero was only 31.0%. In general, for all scores, the predicted conditional probabilities exceed the relative frequencies in the majority of cases. With respect to this criterion, the model for the data set including the odds (World Cups 2002–2010) performs slightly better than the model on the WC1994 data.

Another important aspect when modeling soccer matches based on (Poisson distributed) scores is a possible underestimation of draws, see e.g., Dixon and Coles (1997) and Karlis and Ntzoufras (2003). For the actual match outcome (i.e., win of team A, draw or win of team B) we performed an analysis similar to the different number of goals shown above. Separately for all three possible match outcomes we compared the relative frequencies of the outcome to the predicted probabilities of the respective true match outcome, conditioned on only those matches showing this outcome, see Figure 5. Interestingly, the first-mentioned teams win more often than the second-mentioned teams. This is probably a consequence of the FIFA arrangement of the matches in the group stage and the round of sixteen. Hence, it seems reasonable to distinguish between wins of the “home” and “away” teams. Although draws are generally predicted less well than wins of one of the teams, we found no systematic underestimation of draws. Again, the performance on the WC2002 data is slightly better.

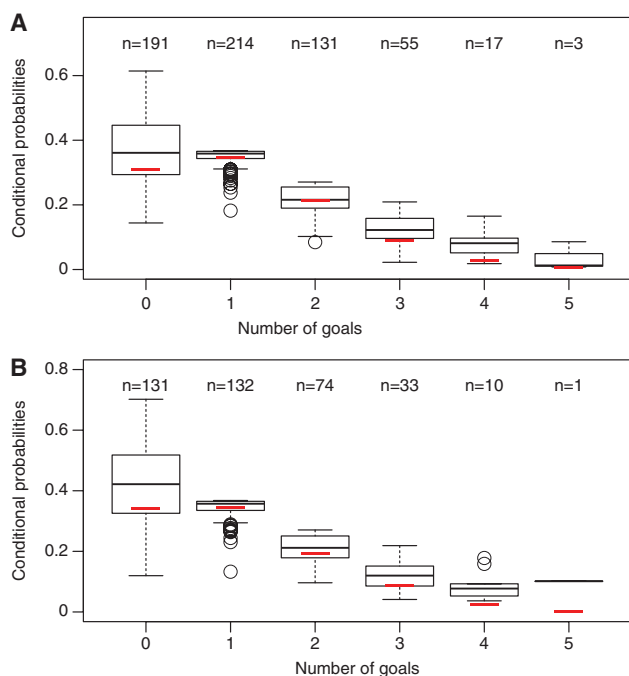


Figure 4: Conditional probabilities of the numbers of goals predicted by the model for the FIFA World Cup data 1994–2010 (A) and by the model for the 2002–2010 data set (B). Red lines represent the relative frequencies of the respective scores in the data set and the corresponding absolute frequencies are displayed on top of every boxplot.

3.4 Prediction power

In the following, we try to assess the performance with respect to prediction of our model. At <http://www.oddsportal.com/soccer/world/world-cup-2014/results/> “three-way” odds⁶ for all 64 matches of the FIFA World Cup 2014, averaged over 16 well-known bookmakers, are provided.

⁶ 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.

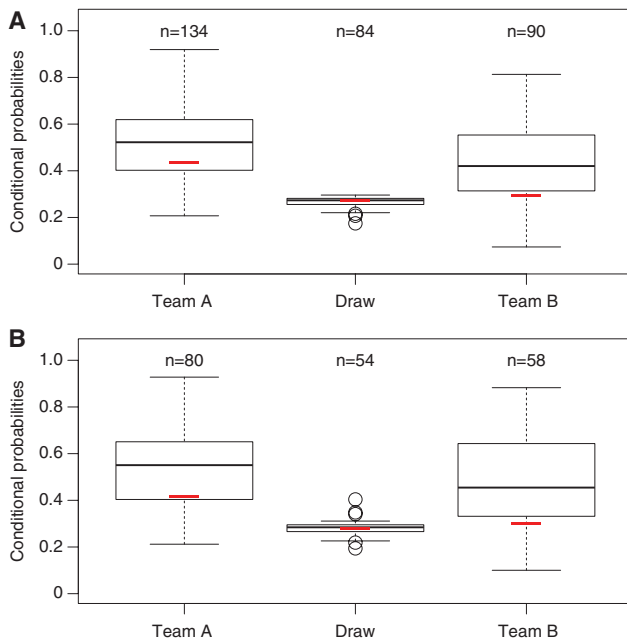


Figure 5: Conditional probabilities of the actual match outcome (i.e., win team A, draw or win of team B) predicted by the model for the WC1994 data (A) and by the model for the WC2002 data (B). Red lines represent the relative frequencies of the respective outcomes in the data set and the corresponding absolute frequencies are displayed on top of every boxplot.

By taking the three quantities $\tilde{p}_r = 1 / \text{odds}_r$, $r \in \{1, 2, 3\}$ and by normalizing with $c := \sum_{r=1}^3 \tilde{p}_r$ in order to adjust for the bookmakers' margins, the odds can be directly transformed into probabilities using $\hat{p}_r = \tilde{p}_r / c$.⁷ On the other hand, let G_{ij} denote the random variables representing the number of goals scored by team i in a certain match against team j and G_{ji} the goals of its opponent, respectively. Then, we can compute the same probabilities by approximating $\hat{p}_1 = P(G_{ij} > G_{ji})$, $\hat{p}_2 = P(G_{ij} = G_{ji})$ and $\hat{p}_3 = P(G_{ij} < G_{ji})$ for each of the 64 matches of the FIFA World Cup 2014 using the corresponding Poisson distributions $G_{ij} \sim \text{Poisson}(\hat{\lambda}_{ij})$, $G_{ji} \sim \text{Poisson}(\hat{\lambda}_{ji})$, where the estimates $\hat{\lambda}_{ij}$ and $\hat{\lambda}_{ji}$ are obtained by our regression models. Based on these predicted probabilities, the average probability of a correct prediction of a FIFA World Cup 2014 match can be obtained. For the true match outcomes $\omega_m \in \{1, 2, 3\}$, $m = 1, \dots, 64$, it is given by $\bar{p}_{\text{three-way}} := \frac{1}{64} \sum_{m=1}^{64} \hat{p}_{1\omega_m}^{\delta_{1\omega_m}} \hat{p}_{2\omega_m}^{\delta_{2\omega_m}} \hat{p}_{3\omega_m}^{\delta_{3\omega_m}}$, with δ_m denoting Kronecker's delta. The quantity $\bar{p}_{\text{three-way}}$ serves as

⁷ 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.

Table 4: Average probability $\bar{p}_{\text{three-way}}$ of a correct prediction of a FIFA World Cup 2014 match for our model on both data sets and the bookmakers' odds.

WC1994	WC2002	Bookmakers' odds
40.15%	40.33%	41.45%

a useful performance measure for a comparison of the predictive power of the model and the bookmakers' odds and is shown for both data sets in Table 4. It is striking that the predictive power of our model compares well with the bookmakers' odds for both data sets, especially if one has in mind that the bookmakers odds are usually released just some days before the corresponding match takes place and, hence, are able to include the latest performance trends of both competing teams. In general, the out-of-sample prediction seems very satisfying to us, with slightly better results for the WC2002 data.

If one puts one's trust into the model and its predicted probabilities, it could serve as the basis of the following betting strategy: for every match one would bet on the three-way match outcome with the highest expected return, which can be calculated as the product of the model's predicted probability and the corresponding three-way odd offered by the bookmakers. We applied this strategy to the model results of both data sets, yielding a return of 33.52% for WC2002 and 19.28% for WC1994, when for all 64 matches equal-sized bets are placed. Again, this is a very satisfying result with an advantage for WC2002.

In Table 5, the corresponding estimates of the (unscaled) fixed team-specific attacking and defending effects are summarized, exemplarily for the WC2002 data. In contrast to the covariate effects from Table 3, we present the unscaled effects here, as this allows a direct comparison of both the attack and defense parameters of different teams. As already pointed out in Section 2, the full attack or defense abilities of team i are represented by the terms $\mathbf{x}_{ik}^\top \boldsymbol{\beta} + \text{att}_i$ and $\mathbf{x}_{ik}^\top \boldsymbol{\beta} + \text{def}_i$, respectively, and not only by the parameters att_i and def_i . Therefore, $\text{att}_i = \text{def}_i = 0$ simply indicates that for such teams no additional attack or defense effects are needed. In general, larger team-specific attack or defense parameters, respectively, increase the team's performance. 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 1st place, its defense is on 3rd place; Brazil's attack is on 2nd place, its defense on 5th place. 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

Table 5: (Unscaled) estimates of the team-specific attacking effects att_i and their exponentials $\exp(att_i)$ (left) and defending effects def_i and their exponentials $\exp(def_i)$ (right) for the WC2002 data.

Estimated attack parameters				Estimated defense parameters			
1.		GER	0.237 1.267	1.		SUI	0.599 1.821
2.		BRA	0.114 1.121	2.		ALG	0.205 1.227
3.		URU	0.101 1.106	3.		GER	0.181 1.199
4.		CRC	0.099 1.104	4.		HON	0.065 1.067
5.		RSA	0.060 1.062	5.		BRA	0.057 1.059
6.		BEL	0.042 1.043	6.		FRA	0.046 1.047
7.		POR	0.019 1.019	7.		POR	0.030 1.031
8.		ARG	0.000 1.000	8.		PAR	0.021 1.021
9.		AUS	0.000 1.000	9.		ARG	0.000 1.000
10.		CHI	0.000 1.000	10.		AUS	0.000 1.000
11.		CRO	0.000 1.000	11.		CHI	0.000 1.000
12.		DEN	0.000 1.000	12.		CRO	0.000 1.000
13.		ECU	0.000 1.000	13.		DEN	0.000 1.000
14.		ENG	0.000 1.000	14.		ECU	0.000 1.000
15.		GHA	0.000 1.000	15.		ENG	0.000 1.000
16.		GRE	0.000 1.000	16.		GHA	0.000 1.000
17.		ITA	0.000 1.000	17.		GRE	0.000 1.000
18.		CIV	0.000 1.000	18.		ITA	0.000 1.000
19.		JPN	0.000 1.000	19.		CIV	0.000 1.000
20.		MEX	0.000 1.000	20.		JPN	0.000 1.000
21.		NED	0.000 1.000	21.		MEX	0.000 1.000
22.		NEW	0.000 1.000	22.		NED	0.000 1.000
23.		NGA	0.000 1.000	23.		NEW	0.000 1.000
24.		RUS	0.000 1.000	24.		NGA	0.000 1.000
25.		KOR	0.000 1.000	25.		RUS	0.000 1.000
26.		ESP	0.000 1.000	26.		KOR	0.000 1.000
27.		SWE	0.000 1.000	27.		ESP	0.000 1.000
28.		USA	0.000 1.000	28.		SWE	0.000 1.000
29.		SVN	-0.002 0.998	29.		USA	0.000 1.000
30.		PAR	-0.003 0.997	30.		SVN	-0.009 0.991
31.		POL	-0.005 0.995	31.		PAR	-0.012 0.988
32.		IRN	-0.040 0.960	32.		POL	-0.012 0.988
33.		SRB	-0.047 0.954	33.		IRN	-0.019 0.981
34.		HON	-0.083 0.921	34.		SRB	-0.022 0.978
35.		FRA	-0.198 0.821	35.		HON	-0.085 0.918
36.		SUI	-0.202 0.817	36.		FRA	-0.090 0.914
37.		CMR	-0.204 0.815	37.		SUI	-0.097 0.907
38.		TUN	-0.234 0.791	38.		CMR	-0.153 0.858
39.		ALG	-0.340 0.712	39.		TUN	-0.297 0.743
40.		KSA	-0.495 0.610	40.		ALG	-0.526 0.591
						KSA	-0.788 0.455

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. Table 5 also provides the exponentials of the ability parameters. Due to the used (log-)link, they represent the multiplicative (or divisive) effects of the respective parameters on the response scale. In the current example, this means that the number of goals Switzerland concedes are divided by 1.8 compared to the case where Switzerland would not have an additional defense parameter.

3.5 Probabilities for FIFA World Cup 2014 winner

































We used both estimates from the two models fitted on the WC1994 and the WC2002 data to simulate the tournament progress of the FIFA World Cup 100,000 times. As we have seen above that the model on the WC2002 data performs slightly better than the WC1994 model with respect to all regarded goodness-of-fit and prediction criteria, we present in this section only the prediction results of the model based on the WC2002 data. The results corresponding to the WC1994 data can be found in the Appendix.

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 goals of both competing teams from the predicted Poisson distributions, i.e., $G_{ij} \sim \text{Poisson}(\hat{\lambda}_{ij})$, $G_{ji} \sim \text{Poisson}(\hat{\lambda}_{ji})$, with estimates $\hat{\lambda}_{ij}$ and $\hat{\lambda}_{ji}$ from the WC2002 model. This allows us to precisely follow the official FIFA rules when determining the final group standings.⁸ If a match in the knockout stage ended in a draw, we simulated another 30 min of extra time using scoring rates equal to 1/3 of the 90 min rates. If the match then still ended in a draw, the winner was calculated simply by coin flip, reflecting a penalty shoot out.

Based on these simulations, for each of the 32 participating teams probabilities to reach the next stage and, finally, to win the tournament are obtained. These are summarized in Table 6 together with the winning probabilities based on the ODDSET odds for comparison. In contrast to most other prediction approaches for the FIFA

⁸ 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.

































Table 6: Estimated probabilities (in %) for reaching the different stages in the FIFA World Cup 2014 for all 32 teams based on 100,000 simulation runs of the FIFA World Cup 2014 and based on the estimates of the WC2002 data together with winning probabilities based on the ODDSET odds.

			Round of 16	Quarter finals	Semi finals	Final	World Champion	Oddset
1.		GER	91.4	77.9	57.0	39.2	27.6	14.2
2.		BRA	91.8	67.9	54.4	30.9	20.0	20.3
3.		SUI	84.2	62.0	35.0	21.6	12.5	0.7
4.		ESP	84.2	52.0	37.8	21.6	12.1	10.9
5.		ARG	90.6	53.2	26.7	15.5	7.3	14.2
6.		POR	60.2	38.6	20.2	10.4	3.6	2.4
7.		BEL	82.5	36.3	19.8	9.3	3.4	5.9
8.		ENG	70.4	41.2	14.7	5.5	1.8	3.5
9.		CRO	58.1	26.1	15.5	5.1	1.6	0.7
10.		FRA	51.2	26.5	9.8	4.6	1.1	3.5
11.		ITA	56.8	31.8	10.8	4.4	1.3	3.5
12.		NED	55.7	21.3	11.8	4.1	1.2	3.5
13.		URU	65.1	37.3	11.6	4.1	1.2	2.8
14.		COL	60.6	31.5	10.7	4.0	1.2	3.9
15.		CIV	58.3	26.3	9.2	2.9	0.6	0.7
16.		CHI	42.9	13.5	7.3	2.5	1.0	2.0
17.		GRE	53.2	22.5	7.3	2.1	0.4	0.7
18.		USA	27.2	13.0	5.4	2.0	0.4	0.7
19.		MEX	42.0	14.7	5.8	1.9	0.3	0.7
20.		GHA	21.2	8.7	4.0	1.8	0.5	0.7
21.		RUS	51.3	11.5	4.3	1.4	0.3	1.2
22.		HON	28.3	10.5	3.6	1.2	0.3	0.1
23.		KOR	41.4	9.7	3.4	1.0	0.0	0.2
24.		BIH	48.2	17.1	3.8	0.8	0.1	0.5
25.		ECU	36.3	14.8	3.4	0.8	0.1	0.7
26.		JPN	27.9	7.7	1.7	0.4	0.0	0.5
27.		ALG	24.8	4.3	1.1	0.4	0.1	0.1
28.		NGA	39.4	12.5	2.1	0.2	0.0	0.4
29.		IRN	21.8	3.4	0.4	0.2	0.0	0.1
30.		AUS	17.2	2.8	0.8	0.1	0.0	0.2
31.		CMR	8.1	1.7	0.5	0.0	0.0	0.2
32.		CRC	7.7	1.7	0.1	0.0	0.0	0.1

World Cup 2014 clearly favoring Brazil, we get a neck-and-neck race between Germany and Brazil, finally with better chances for Germany. The major reason for this is that with a high probability in the simulations 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 thin advantage with a winning probability of 51.7% against 48.3%. The favorites Germany and Brazil are followed by the teams of Switzerland, Spain, Argentina and Portugal. Similarly, for the WC1994 data

Table 7: Estimated (adapted) probabilities (in %) for reaching the next stages in the FIFA World Cup 2014 for all 32 teams based on 100,000 simulation runs of the FIFA World Cup 2014.

			Round of 16	Quarter finals	Semi finals	Final	World Champion
1.		GER	91.4	78.5	71.2	48.7	72.5
2.		ARG	90.6	58.9	49.9	57.6	27.5
3.		BRA	91.8	76.7	54.2	51.3	0.0
4.		NED	55.7	59.2	67.1	42.4	0.0
5.		BEL	82.5	65.6	50.1	0.0	0.0
6.		COL	60.6	81.1	45.8	0.0	0.0
7.		CRC	7.7	42.8	32.9	0.0	0.0
8.		FRA	51.2	71.8	28.8	0.0	0.0
9.		GRE	53.2	57.2	0.0	0.0	0.0
10.		SUI	84.2	41.1	0.0	0.0	0.0
11.		MEX	42.0	40.8	0.0	0.0	0.0
12.		USA	27.2	34.4	0.0	0.0	0.0
13.		NGA	39.4	28.2	0.0	0.0	0.0
14.		CHI	42.9	23.3	0.0	0.0	0.0
15.		ALG	24.8	21.5	0.0	0.0	0.0
16.		URU	65.1	18.9	0.0	0.0	0.0
17.		ESP	84.2	0.0	0.0	0.0	0.0
18.		ENG	70.4	0.0	0.0	0.0	0.0
19.		POR	60.2	0.0	0.0	0.0	0.0
20.		CIV	58.3	0.0	0.0	0.0	0.0
21.		CRO	58.1	0.0	0.0	0.0	0.0
22.		ITA	56.8	0.0	0.0	0.0	0.0
23.		RUS	51.3	0.0	0.0	0.0	0.0
24.		BIH	48.2	0.0	0.0	0.0	0.0
25.		KOR	41.4	0.0	0.0	0.0	0.0
26.		ECU	36.3	0.0	0.0	0.0	0.0
27.		HON	28.3	0.0	0.0	0.0	0.0
28.		JPN	27.9	0.0	0.0	0.0	0.0
29.		IRN	21.8	0.0	0.0	0.0	0.0
30.		GHA	21.2	0.0	0.0	0.0	0.0
31.		AUS	17.2	0.0	0.0	0.0	0.0
32.		CMR	8.1	0.0	0.0	0.0	0.0

After each round, the data set (WC2002) is extended with by the matches already played and the model is refitted. Only actual matches from the World Cup are simulated.

Germany has the highest probability to win the trophy, followed by Spain and Brazil, see Table 9.

In a second step, we investigate how the model (and the respective winning probabilities) change when the

data set is successively extended by the completed matches of the current World Cup in each stage. For example, after the group stage the model is refitted including all 48 matches from the group stage. Then, for the round of 16 the

Table 8: Most probable final group standings together with the corresponding probabilities for the FIFA World Cup 2014 based on 100,000 simulation runs and on the estimates of the WC2002 data.

Group A 39%	Group B 26%	Group C 15%	Group D 19%
1. BRA	1. ESP	1. COL	1. ENG
2. CRO	2. NED	2. GRE	2. ITA
MEX	CHI	JPN	URU
CMR	AUS	CIV	CRC
Group E 19%	Group F 29%	Group G 38%	Group H 23%
1. SUI	1. ARG	1. GER	1. BEL
2. FRA	2. BIH	2. POR	2. RUS
ECU	NGA	GHA	ALG
HON	IRN	USA	KOR

qualified teams from the group stage are known and used for the prediction of the round of 16. For example, according to the initial model Costa Rica appeared to be a clear underdog and only had low chances to reach the round of 16 (7.7%). Based on the initial model, in the upcoming knockout match against Greece, Costa Rica's probability to qualify for the quarter finals was estimated to be 27.8%, whereas the adapted model yields an increased probability of 42.8%. Therefore, the model accounted for the good performance of Costa Rica in the group stage and, indeed, Costa Rica actually defeated Greece in a penalty shootout. A similar effect comes up for the following quarter final

between Costa Rica and the Netherlands where the chances of Costa Rica are increased from 19.3% to 32.9%. Again, the real match was actually quite close with Netherlands winning in another penalty shootout. Table 7 summarises the adapted probabilities for all stages, again based on 100,000 simulation runs. In the appendix, Table 10 shows the respective (adapted) probabilities for the WC1994 data.

3.6 Most probable tournament outcome

Finally, based on the 100,000 simulations, we also provide the most probable tournament outcome, exemplarily for the WC2002 data. 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.

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 39% and 38%, 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 6. 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 $1.49 \cdot 10^{-60}$

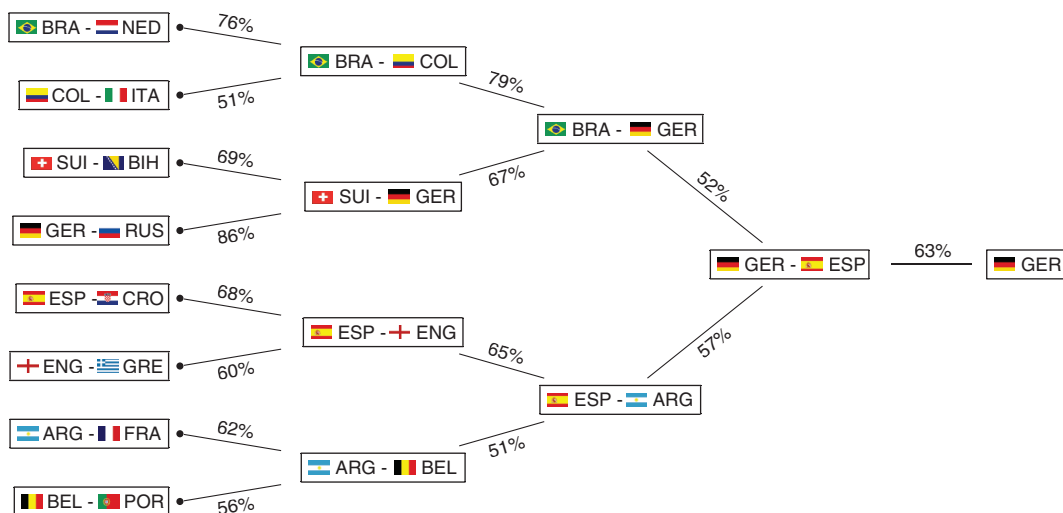


Figure 6: Most probable course of the knockout stage together with corresponding probabilities for the FIFA World Cup 2014 based on 100,000 simulation runs and on the estimates of the WC2002 data.

(given as the product of all single probabilities of Table 8 and Figure 6). Hence, deviations of the true tournament outcome from the model's most probable one are not only possible, but very likely.

In fact, if we compare the most probable tournament outcome of the FIFA World Cup 2014 from Table 8 and Figure 6 with the true one, several differences become obvious. In general, several underdogs, such as e.g., Algeria, Costa Rica, USA or Chile have reached the round of sixteen, while several favorites, such as e.g., Spain, Italy, England or Portugal, dropped out already in the group stage. This could not be adequately represented by the model. Nevertheless, beyond the round of sixteen, the model's predicted tournament course gets closer and closer to the true one, with three out of four semi-finalists predicted correctly and finally, with Germany correctly predicted as the World Champion.

4 Concluding remarks

A team-specific generalized linear Poisson model for the number of goals scored by soccer teams facing each other in international tournament matches is set up. As an application, the FIFA World Cups 1994–2010 and 2002–2010, respectively, 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-package `grplasso`, are used. A detailed goodness-of-fit analysis is presented and suitable “out-of-sample” performance measures for prediction are considered, which are based on the three-way tendencies of the considered matches.

The fitted models were used for simulation of the FIFA World Cup 2014. According to these simulations, Germany and Brazil turned out to be the top favorites for winning the title, with an advantage for Germany. Besides, the most probable tournament outcome is provided.

A major part of the statistical novelty of the presented work lies in the use of penalty terms for covariate effects in combination with team-specific abilities. It allows to include many covariates simultaneously and performs automatic variable selection. In the case of high correlation between certain covariates, the estimation procedure is stabilized by the penalization. If several high correlated variables possibly contain information on the response, the LASSO tends to include the predictor with the highest explanatory power. Furthermore, as the basic model used throughout this article is in general not identified, the penalized likelihood approach nevertheless allows for unique estimates. Theoretically, this would also allow for the estimation of effects of covariates not varying over different tournaments, which are un-separable from team-specific effects in an unpenalized estimation.

Another important aspect is that the team-specific ability parameters need not necessarily be constant, but instead could evolve over time since composition and performance of the teams might change over time. In this context we want to mention a very recent publication of Koopman and Lit (2015). They assume a bivariate Poisson distribution for the goals in English Premier League matches, with intensity coefficients that change stochastically over time by modeling the teams' ability parameters as first order auto-regressive processes. However, due to certain general differences in the structure of national league and FIFA World Cup data it is not straightforward, how this approach can be adopted to the present data situation. Nevertheless, the idea of time-varying ability parameters in modeling international soccer data sounds promising to us and could be the starting point for future research.

Acknowledgments: 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 the pleasant collaboration. 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 expert Tim Frohwein also helped a lot to improve the article.

Appendix

Prediction results and most probable tournament outcome for the WC1994 data

Table 9: Estimated probabilities (in %) for reaching the different stages in the FIFA World Cup 2014 for all 32 teams based on 100,000 simulation runs of the FIFA World Cup 2014 and based on the estimates of the WC1994 data together with winning probabilities based on the ODDSET odds.

































































			Round of 16	Quarter finals	Semi finals	Final	World Champion	Oddset
1.		GER	86.1	68.1	52.3	32.8	20.5	14.2
2.		ESP	91.3	64.1	47.5	31.7	19.5	10.9
3.		BRA	93.0	64.9	48.2	30.8	19.1	20.3
4.		POR	73.3	51.1	35.1	18.7	9.3	2.4
5.		URU	71.3	50.7	22.5	11.5	5.1	2.8
6.		BEL	82.8	36.9	22.4	10.2	4.3	5.9
7.		ITA	67.2	46.3	19.5	9.4	4.0	3.5
8.		SUI	72.3	45.6	19.7	8.5	3.5	0.7
9.		ARG	77.6	44.5	18.9	7.8	3.1	14.2
10.		CRO	64.9	26.2	13.8	6.0	2.1	0.7
11.		FRA	62.2	35.4	13.7	5.3	1.9	3.5
12.		COL	76.3	33.4	10.9	4.1	1.3	3.9
13.		ENG	47.3	28.1	9.5	3.7	1.3	3.5
14.		CHI	50.1	18.0	8.6	3.3	1.0	2.0
15.		NED	44.9	15.1	6.9	2.5	0.7	3.5
16.		BIH	56.6	25.2	7.9	2.4	0.7	0.5
17.		ALG	49.3	13.2	5.6	1.6	0.4	0.1
18.		CIV	61.3	21.4	5.5	1.7	0.4	0.7
19.		USA	23.2	10.7	4.8	1.5	0.4	0.7
20.		ECU	38.8	17.3	4.8	1.3	0.3	0.7
21.		NGA	39.3	14.2	3.4	0.8	0.2	0.4
22.		RUS	42.7	9.0	3.5	0.8	0.2	1.2
23.		GHA	17.4	7.2	2.9	0.7	0.2	0.7
24.		MEX	28.0	6.9	2.4	0.7	0.2	0.7
25.		JPN	43.0	11.5	2.2	0.5	0.1	0.5
26.		HON	26.6	9.8	2.2	0.5	0.1	0.1
27.		IRN	26.4	7.9	1.6	0.3	0.1	0.1
28.		KOR	25.2	3.8	1.1	0.2	0.0	0.2
29.		CRC	14.2	5.6	1.0	0.2	0.0	0.1
30.		CMR	14.0	2.3	0.6	0.1	0.0	0.2
31.		AUS	13.7	2.4	0.6	0.1	0.0	0.2
32.		GRE	19.4	3.1	0.3	0.1	0.0	0.7

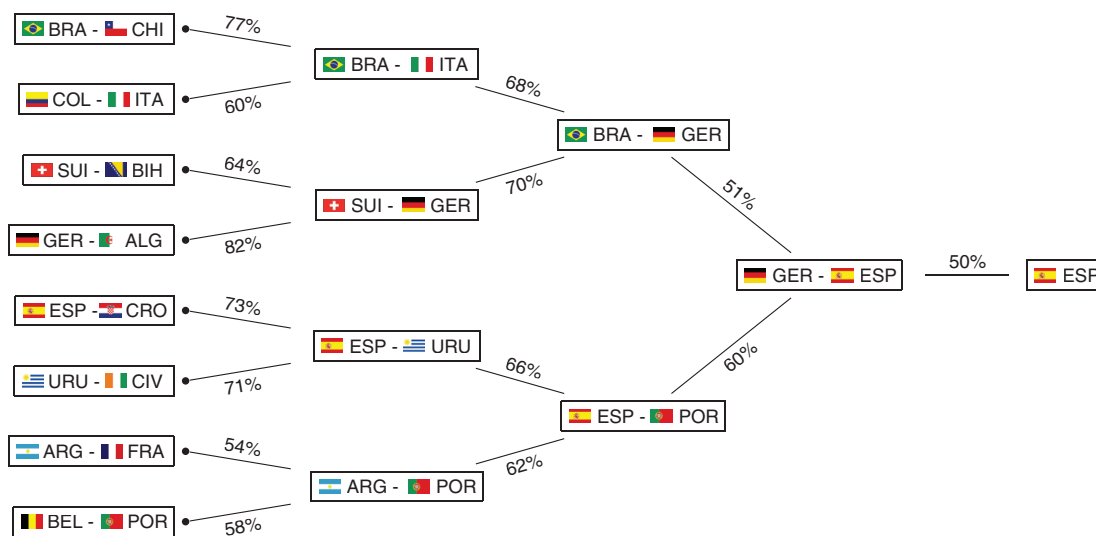
Table 10: Estimated (adapted) probabilities (in %) for reaching the next stages in the FIFA World Cup 2014 for all 32 teams based on 100,000 simulation runs of the FIFA World Cup 2014.

			Round of 16	Quarter finals	Semi finals	Final	World Champion
1.		GER	86.1	81.4	68.4	53.2	73.9
2.		ARG	77.6	48.4	47.5	54.8	26.1
3.		BRA	93.0	76.6	73.3	46.8	0.0
4.		NED	44.9	66.0	67.0	45.2	0.0
5.		BEL	82.8	65.7	52.5	0.0	0.0
6.		CRC	14.2	68.0	33.0	0.0	0.0
7.		FRA	62.2	68.8	31.6	0.0	0.0
8.		COL	76.3	41.6	26.7	0.0	0.0
9.		URU	71.3	58.4	0.0	0.0	0.0
10.		SUI	72.3	51.6	0.0	0.0	0.0
11.		USA	23.2	34.3	0.0	0.0	0.0
12.		MEX	28.0	34.0	0.0	0.0	0.0
13.		GRE	19.4	32.0	0.0	0.0	0.0
14.		NGA	39.3	31.2	0.0	0.0	0.0
15.		CHI	50.1	23.4	0.0	0.0	0.0
16.		ALG	49.3	18.6	0.0	0.0	0.0
17.		ESP	91.3	0.0	0.0	0.0	0.0
18.		POR	73.3	0.0	0.0	0.0	0.0
19.		ITA	67.2	0.0	0.0	0.0	0.0
20.		CRO	64.9	0.0	0.0	0.0	0.0
21.		CIV	61.3	0.0	0.0	0.0	0.0
22.		BIH	56.6	0.0	0.0	0.0	0.0
23.		ENG	47.3	0.0	0.0	0.0	0.0
24.		JPN	43.0	0.0	0.0	0.0	0.0
25.		RUS	42.7	0.0	0.0	0.0	0.0
26.		ECU	38.8	0.0	0.0	0.0	0.0
27.		HON	26.6	0.0	0.0	0.0	0.0
28.		IRN	26.4	0.0	0.0	0.0	0.0
29.		KOR	25.2	0.0	0.0	0.0	0.0
30.		GHA	17.4	0.0	0.0	0.0	0.0
31.		CMR	14.0	0.0	0.0	0.0	0.0
32.		AUS	13.7	0.0	0.0	0.0	0.0

After each round, the data set (WC1994) is extended with by the matches already played and the model is refitted. Only actual matches from the World Cup are simulated.

Table 11: Most probable final group standings together with the corresponding probabilities for the FIFA World Cup 2014 based on 100,000 simulation runs and on the estimates of the WC1994 data.

Group A 43%	Group B 33%	Group C 24%	Group D 22%
1. BRA	1. ESP	1. COL	1. URU
2. CRO	2. CHI	2. CIV	2. ITA
MEX	NED	JPN	ENG
CMR	AUS	GRE	CRC
Group E 22%	Group F 24%	Group G 36%	Group H 24%
1. SUI	1. ARG	1. GER	1. BEL
2. FRA	2. BIH	2. POR	2. ALG
ECU	NGA	GHA	RUS
HON	IRN	USA	KOR

**Figure 7:** Most probable course of the knockout stage together with corresponding probabilities for the FIFA World Cup 2014 based on 100,000 simulation runs and on the estimates of the WC1994 data.

References

- Akaike, H. 1973. "Information Theory and the Extension of the Maximum Likelihood Principle." *Second International Symposium on Information Theory* 267–281.
- Dixon, M. J. and S. G. Coles. 1997. "Modelling Association Football Scores and Inefficiencies in the Football Betting Market." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 46:265–280.
- Dobson, S. and J. Goddard. 2011. *The Economics of Football*. Cambridge: Cambridge University Press.
- 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 Knock-out Phase?" *Journal of Quantitative Analysis in Sports* 7(1).
- Forrest, D. and R. Simmons. 2000. "Forecasting Sport: The Behaviour and Performance of Football Tipsters." *International Journal of Forecasting* 16:317–331.
- Goldman-Sachs Global Investment Research. 2014. "The World Cup and Economics 2014." Accessed February 23, 2015. <http://www.goldmansachs.com/our-thinking/outlook/world-cup-and-economics-2014-folder/world-cup-economics-report.pdf>.
- 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* 9:51–66.
- Groll, A. and G. Tutz. 2014. "Variable Selection for Generalized Linear Mixed Models by L_1 -Penalized Estimation." *Statistics and Computing* 24:137–154.

- 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.
- Karlis, D. and I. Ntzoufras. 2011. "Robust Fitting of Football Prediction Models." *IMA Journal of Management Mathematics* 22:171–182.
- Koopman, S. J. and R. Lit. 2015. "A Dynamic Bivariate Poisson Model for Analysing and Forecasting Match Results in the English Premier League." *Journal of the Royal Statistical Society, A* 178:167–186.
- 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. 2010a. "Forecasting Sports Tournaments by Ratings of (Prob)abilities: A Comparison for the EURO 2008." *International Journal of Forecasting* 26:471–481.
- Leitner, C., A. Zeileis, and K. Hornik. 2010b. "Forecasting the Winner of the FIFA World Cup 2010." Report Series / Department of Statistics and Mathematics, 100. Institute for Statistics and Mathematics, WU Vienna.
- Lloyd's. 2014. "FIFA World Cup: How Much Are Those Legs Worth?" Accessed February 16, 2015. <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.
- McHale, I. G. and P. A. Scarf. 2006. "Forecasting International Soccer Match Results Using Bivariate Discrete Distributions." Technical Report 322, Working paper, Salford Business School.
- McHale, I. G. and P. A. Scarf. 2011. "Modelling the Dependence of Goals Scored by Opposing Teams in International Soccer Matches." *Statistical Modelling* 41:219–236.
- 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.
- R Core Team. 2014. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL <http://www.R-project.org/>.
- 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)* 49:399–418.
- Schwarz, G. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6:461–464.
- Silver, N. 2014. "It's Brazil's World Cup to Lose." Accessed February 18, 2015. <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.
- 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.