Andreas Groll*, Cristophe Ley, Gunther Schauberger and Hans Van Eetvelde

# A hybrid random forest to predict soccer matches in international tournaments

https://doi.org/10.1515/jqas-2018-0060

**Abstract:** In this work, we propose a new hybrid modeling approach for the scores of international soccer matches which combines *random forests* with *Poisson ranking methods*. While the random forest is based on the competing teams' covariate information, the latter method estimates ability parameters on historical match data that adequately reflect the current strength of the teams. We compare the new *hybrid random forest* model to its separate building blocks as well as to conventional Poisson regression models with regard to their predictive performance on all matches from the four FIFA World Cups 2002–2014. It turns out that by combining the random forest with the team ability parameters from the ranking methods as an additional covariate the predictive power can be improved substantially. Finally, the hybrid random forest is used (in advance of the tournament) to predict the FIFA World Cup 2018. To complete our analysis on the previous World Cup data, the corresponding 64 matches serve as an independent validation data set and we are able to confirm the compelling predictive potential of the hybrid random forest which clearly outperforms all other methods including the betting odds.

**Keywords:** FIFA World Cup 2018; random forests; soccer; sports tournaments; team abilities.

## 1 Introduction

The use of statistical models to predict the outcome of international soccer tournaments, such as European championships (EUROs) or FIFA World Cups, has become more and more popular in recent years, as documented by the growing scientific literature in this research field. In this context, a frequently used model class that has proved to be of value is the class of Poisson regression models which directly model the number of goals scored by both competing teams in the single matches of the tournaments. Let $X_{ij}$ and $Y_{ij}$ denote the goals of the first and second team, respectively, in a match between teams $i$ and $j$, where $i, j \in \{1, \dots, n\}$ and $n$ denotes the total number of teams in the regarded tournaments. One assumes $X_{ij} \sim Po(\lambda_{ij})$ and $Y_{ij} \sim Po(\mu_{ij})$ where $\lambda_{ij}$ and $\mu_{ij}$ denote the intensity parameters (i.e. the expected number of goals) of the respective Poisson distributions. For these intensity parameters several modeling strategies exist, which incorporate playing abilities or covariates of the competing teams in different ways.

In the simplest case, the Poisson distributions are treated as (conditionally) independent, conditional on the teams' abilities or covariates. For example, Dyte and Clarke (2000) applied this model to data from FIFA World Cups and let the Poisson intensities of both competing teams depend on their FIFA ranks. Groll and Abedieh (2013) and Groll, Schauberger, and Tutz (2015) considered a large set of potentially influential variables for EURO and World Cup data, respectively, and used $L_1$-penalized approaches to detect a sparse set of relevant covariates. Based on these, predictions for the EURO 2012 and FIFA World Cup 2014 tournaments were provided. These approaches showed that, when many covariates are regarded and/or the predictive power of the single variables is not clear in advance, regularized estimation approaches can be beneficial.

Many researchers have relaxed the strong assumption of conditional independence and have introduced different approaches to allow for dependent scores. Dixon and Coles (1997) were the first to identify a (slightly negative) correlation between the scores. As a consequence, they introduced an additional dependence parameter. However, they ignored the fact that the intensity parameters in models including abilities (or covariates) of both teams are themselves correlated. Even though the Poisson distributions are assumed to be independent conditional on the abilities, they are marginally correlated. Karlis and Ntzoufras (2003) proposed to model the scores of both teams by a bivariate Poisson distribution, which is able to account for (positive) dependencies between the scores. While the bivariate Poisson distribution can only account for positive dependencies, copula-based models

***Corresponding author: Andreas Groll,** TU Dortmund University, Faculty Statistics, Vogelpothsweg 87, 44227 Dortmund, Germany, e-mail: groll@statistik.tu-dortmund.de
**Cristophe Ley and Hans Van Eetvelde:** Ghent University, Department of Applied Mathematics, Computer Science and Statistics, Krijgslaan 281, S9, Campus Sterre, Ghent 9000, Belgium
**Gunther Schauberger:** Technische Universitaet Muenchen, Department of Sport and Health Sciences, Munich, Bavaria, Germany

also allow for negative dependencies (see, for example, McHale and Scarf 2007, 2011 or Boshnakov, Kharrat, and McHale 2017).

However, with regard to the bivariate Poisson case, Groll et al. (2018) provide some evidence that, if highly informative covariates of both competing teams are included into the intensities of both (conditionally) independent Poisson distributions, the dependence structure of the match scores can already be appropriately modeled. They included a large set of covariates for EURO data and used a boosting approach to select a sparse model for the prediction of the EURO 2016. As the dependency parameter of the bivariate Poisson distribution was never updated by the boosting algorithm, two (conditionally) independent Poisson distributions were sufficient.

Closely related to the covariate-based Poisson regression models are Poisson-based ranking methods for soccer teams. The main idea is to find adequate ability parameters that reflect the current strength of the teams. On basis of a set of matches, those parameters are then estimated by means of maximum likelihood. Ley, de Wiele, and Eetvelde (2019) have investigated various models from the literature and compared them in terms of their predictive performance because the latter is directly related to a team's current strength. The resulting best model for this purpose is the simplest bivariate Poisson distribution of Karlis and Ntzoufras (2003). Interestingly, Ley et al. (2019) found that those models outperform their competitors both for domestic league matches and national team matches. These statistical strength-based rankings present an interesting alternative to the FIFA ranking. An alternative approach solely based on bookmakers' odds was proposed by Leitner, Zeileis, and Hornik (2010), who developed a ranking method that can be used for the prediction of international football tournaments.

A fundamentally different modeling approach is based on a random forest – an ensemble learning method for classification, regression and other tasks proposed by Breiman (2001). The method originates from the machine learning and data mining community and operates by first constructing a multitude of decision trees (see, e.g. Breiman et al. 1984; Quinlan 1986) on different training data sets, which are resampled from the original dataset. The predictions from the individual trees are then summarized, either by taking the mode of the predicted classes (in classification) or by averaging the predicted values (in regression). Random forests reduce the tendency of overfitting and the variance compared to regular decision trees, and are a common powerful tool for prediction. In preliminary work from Schauberger and Groll (2018) the predictive performance of different types of random forests has been compared on data containing all matches of the FIFA World Cups 2002–2014 with conventional regression methods for count data, such as the Poisson models mentioned above. It turned out that random forests provided very satisfactory results and generally outperformed the regression approaches. Moreover, their predictive performances were either close to or even outperforming those of the bookmakers, which serve as natural benchmark.

In the present work, we propose a new modeling approach that combines random forests with the Poisson ranking methods, leading to what we call a *hybrid random forest model*. Using FIFA World Cup data, we show that the predictive power of the random forests can be further increased if adequate estimates of team ability parameters, reflecting the current strength of the national teams, are incorporated as additional covariates. Though the main purpose of the proposed method is to create a good prediction of the tournament outcome (comprising winning probabilities for all teams, but also, for example, the results in single groups or the probabilities for all teams to reach certain tournament stages), we evaluate the predictive performance of the method on single matches. The motivation is that an improvement of the prediction of single matches will typically improve the prediction of the tournament outcome. We are most interested in the prediction of the three match outcomes *win*, *draw* and *loss*, but as in the group stage of FIFA World Cups sometimes also the goal difference or the number of goals scored is of importance (namely in case two teams have achieved the same number of points), we also investigate the prediction quality for exact scores. The results motivated us to use this hybrid random forest model to calculate predictions of the FIFA World Cup 2018 in advance of the tournament. Now that the FIFA World Cup 2018 is finished, we use the corresponding 64 matches as an independent validation data set to investigate the predictive performance of the new hybrid method.

The rest of the manuscript is structured as follows. In Section 2 we describe the two underlying data sets. The first covers all matches of the four preceding FIFA World Cups 2002–2014 including covariate information, the second consists of the match results of all international matches played by all national teams during certain time periods. Next, in Section 3 we briefly explain the basic idea of random forests and ranking methods and, finally, how they can be combined to a hybrid random forest model. The performance of this hybrid random forest is then compared to the performance of its single building blocks as well as to conventional (regularized) Poisson regression approaches in Section 4. In Section 5, we fit the hybrid random forest model to the complete World Cup 2002–2014 data and evaluate its predictive performance with respect to the results of the FIFA World Cup 2018 tournament. Finally, we conclude in Section 6.

## 2 Data

In this section, we briefly describe two fundamentally different types of data that can be used to model and predict international soccer tournaments such as the FIFA World Cup. The first type of data covers variables that characterize the participating teams of the single tournaments and connects them to the results of the matches that were played during these tournaments. The second type of data is simply based on the match results of all international matches played by all national teams during certain time periods. These data do not only cover the matches from the specific tournaments but also all qualifiers and friendly matches.

### 2.1 Covariate data

The first type of data we describe covers all matches of the four FIFA World Cups 2002–2014 together with several potential influence variables. Basically, we use a very similar set of covariates as introduced in Groll et al. (2015). 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), and, therefore, vary from one World Cup to another.

Several of the variables contain information about the recent performance and sportive success of national teams, as the current form of a national team should have an influence on the team's success in the upcoming tournament. One additional covariate in this regard, which we will introduce later, is reflecting the national teams' current playing abilities and is related to the second type of data introduced in Section 2.2. The estimates of these ability parameters are based on a separate Poisson ranking model, see Section 3.2 for details. Beside these sportive variables, also certain economic factors as well as variables describing the structure of a team's squad are collected. We shall now describe in more detail these variables.

**Economic factors:**

*GDP per capita.* To account for the general increase of the gross domestic product (GDP) during 2002–2014, a ratio of the GDP per capita of the respective country and the worldwide average GDP per capita is used (source: http://unstats.un.org/unsd/snaama/dnllist.asp).

*Population.* The population size is used in relation to the respective global population to account for the general world population growth (source: http://data.worldbank.org/indicator/SP.POP.TOTL).

**Sportive factors:**

*ODDSET probability.* We convert bookmaker odds provided by the German state betting agency ODDSET into winning probabilities (adjusting for the bookmaker's margin). The variable hence reflects the probability for each team to win the respective World Cup[1].

*FIFA rank.* The FIFA ranking system ranks all national teams based on their performance over the last four years (source: http://de.fifa.com/worldranking/index.html).

*Elo rating.* The World Soccer Elo rating is based on the Elo rating system, originally developed by Dr. Arpad Elo to rate the playing abilities of chess players, and aims at reflecting the current strength of a soccer team relative to its competitors (source: http://www.eloratings.net/). In August 2018 the FIFA changed from its own ranking system to the Elo ratings, which are now the basis for the FIFA World ranking list.

**Home advantage:**

*Host.* A dummy variable indicating if a national team is a hosting country.

*Continent.* A dummy variable indicating if a national team is from the same continent as the host of the World Cup (including the host itself).

*Confederation.* This categorical variable comprises the teams' confederation with six possible values: Africa (CAF); Asia (AFC); Europe (UEFA); North, Central America and Caribbean (CONCACAF); Oceania (OFC); South America (CONMEBOL).

**Factors describing the team's structure:**

The following variables describe the structure of the teams. They were observed with the 23-player-squad nominated for the respective World Cup and were obtained manually both from the website of the

[1] The option to bet on the World Champion before the start of the tournament is rather novel. ODDSET, for example, offered the bet for the first time at the FIFA World Cup 2002.

German soccer magazine *kicker*, http://kicker.de, and from http://transfermarkt.de.

*(Second) maximum number of teammates*. For each squad, both the maximum and second maximum number of teammates playing together in the same national club are counted.

*Average age*. The average age of each squad is collected.

*Number of Champions League (Europa League) players*. As a measurement of the success of the players on club level, the number of players in the semi finals (taking place only few weeks before the respective World Cup) of both the UEFA Champions League (CL) and UEFA Europa League (EL) are counted.

*Number of players abroad/Legionnaires*. For each squad, the number of players playing in clubs abroad (in the season preceding the respective World Cup) is counted.

**Factors describing the team's coach:**

For the coaches of the teams, *Age* and duration of their *Tenure* are observed. Furthermore, a dummy variable indicates if a coach has the same *Nationality* as his team. These variables were also obtained manually from the website of the German soccer magazine *kicker*, http://kicker.de, from http://transfermarkt.de and from https://en.wikipedia.org.

In total, this adds up to 17 variables which were collected separately for each World Cup and each participating team. As an illustration, Table 1 shows the results (1a) and (parts of) the covariates (1b) of the respective teams, exemplarily for the first four matches of the FIFA World Cup 2002. We use this data excerpt to illustrate how the final data set is constructed.

For the modeling techniques that we shall introduce in the following sections, all of the metric covariates are incorporated in the form of differences between the two competing teams. For example, the final variable *Rank* will be the difference between the FIFA ranks of both teams. The categorical variables *Host*, *Continent*, *Confederation* and *Nationality*, however, are included as separate variables for both competing teams. For the variable *Confederation*, for example, this results in two columns of the corresponding design matrix denoted by *Confed* and *Confed.Oppo*, where *Confed* is referring to the confederation of the first-named team and *Confed.Oppo* to the one of its opponent.

As we use the number of goals of each team directly as the response variable, each match corresponds to two different observations, one per team. For the covariates, we consider differences which are computed from the perspective of the first-named team. For illustration, the resulting final data structure for the exemplary matches from Table 1 is displayed in Table 2.

**Table 1:** Exemplary table showing (a) the results of four matches and (b) parts of the covariates of the involved teams.

| Table of results | | |
|---|---|---|
| FRA 🇫🇷 | 0:1 | 🇸🇳 SEN |
| URU 🇺🇾 | 1:2 | 🇩🇰 DEN |
| FRA 🇫🇷 | 0:0 | 🇺🇾 URU |
| DEN 🇩🇰 | 1:1 | 🇸🇳 SEN |
| ⋮ | ⋮ | ⋮ |

| World Cup | Team | Age | Rank | Oddset | … |
|---|---|---|---|---|---|
| Table of covariates | | | | | |
| 2002 | France | 28.3 | 1 | 0.149 | … |
| 2002 | Senegal | 24.3 | 42 | 0.006 | … |
| 2002 | Uruguay | 25.3 | 24 | 0.009 | … |
| 2002 | Denmark | 27.4 | 20 | 0.012 | … |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋱ |

## 2.2 Historic match results

The data used for estimating the abilities of the teams consist of the results of every international game played in the last 8 years preceding the considered World Cup. Besides the number of goals, we also need the information of the venue of the game in order to correct for the home effect, the type of the game (Friendly, Qualification, World Cup, ...), and the moment in time when a match was played. The reason is that, in the ranking method described in Section 3.2, each match is assigned a weight depending on its importance (hence, type of the game) and the time elapsed since the game took place. For example, Table 3

**Table 2:** Exemplary table illustrating the data structure.

| Goals | Team | Opponent | Age | Rank | Oddset | … |
|---|---|---|---|---|---|---|
| 0 | France | Senegal | 4.00 | −41 | 0.14 | … |
| 1 | Senegal | France | −4.00 | 41 | −0.14 | … |
| 1 | Uruguay | Denmark | −2.10 | 4 | −0.00 | … |
| 2 | Denmark | Uruguay | 2.10 | −4 | 0.00 | … |
| 0 | France | Uruguay | 3.00 | −23 | 0.14 | … |
| 0 | Uruguay | France | −3.00 | 23 | −0.14 | … |
| 1 | Denmark | Senegal | 3.10 | −22 | 0.01 | … |
| 1 | Senegal | Denmark | −3.10 | 22 | −0.01 | … |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋱ |

**Table 3:** Historical match result data used for estimating the abilities, exemplarily for the FIFA World Cup 2002.

| Date | Home_team | Away_team | Score | Country | Neutral | Type |
|---|---|---|---|---|---|---|
| 2002-05-30 | Kuwait | Iran | 1:3 | Kuwait | False | Friendly |
| 2002-05-26 | Belgium | Costa Rica | 1:0 | Japan | True | Friendly |
| 2002-05-26 | Cameroon | England | 2:2 | Japan | True | Friendly |
| 2002-05-26 | Denmark | Tunisia | 2:1 | Japan | True | Friendly |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

shows an excerpt of the historic match data used to obtain ability estimates for the teams at the FIFA World Cup 2002.

# 3 A hybrid random forests model

In this section, we propose to use a hybrid random forest approach that combines the information from the two types of data bases introduced above. The proposed method combines a random forest approach for the covariate data with abilities estimated on the historic match results as used by the ranking methods. Before introducing the proposed hybrid method, we first separately present the basic ideas of the two model components.

## 3.1 Random forests

Random forests, originally proposed by Breiman (2001), are an aggregation of a (large) number of classification or regression trees (CARTs). CARTs (Breiman et al. 1984) repeatedly partition the predictor space mostly using binary splits. The goal of the partitioning process is to find partitions such that the respective response values are very homogeneous within a partition but very heterogeneous between partitions. CARTs can be used both for metric response (regression trees) and for nominal/ordinal responses (classification trees). The most frequent visualization tool for CARTs is the so-called dendrogram (see Figure 1). For prediction, all response values within a partition are aggregated either by averaging (in regression trees) or simply by counting and using majority vote (in classification trees).

In this work, we use trees (and, accordingly, random forests) for the prediction of the number of goals a team scores in a match of a FIFA World Cup. As an illustrative example, Figure 1 shows the dendrogram for a regression tree applied to the covariate data introduced in Section 2.1 using the function `ctree` from the `R`-package `party` (Hothorn et al. 2006).

Only two splits are performed in this example, one for the variable *Elo* and one for *Rank*, which leads to a

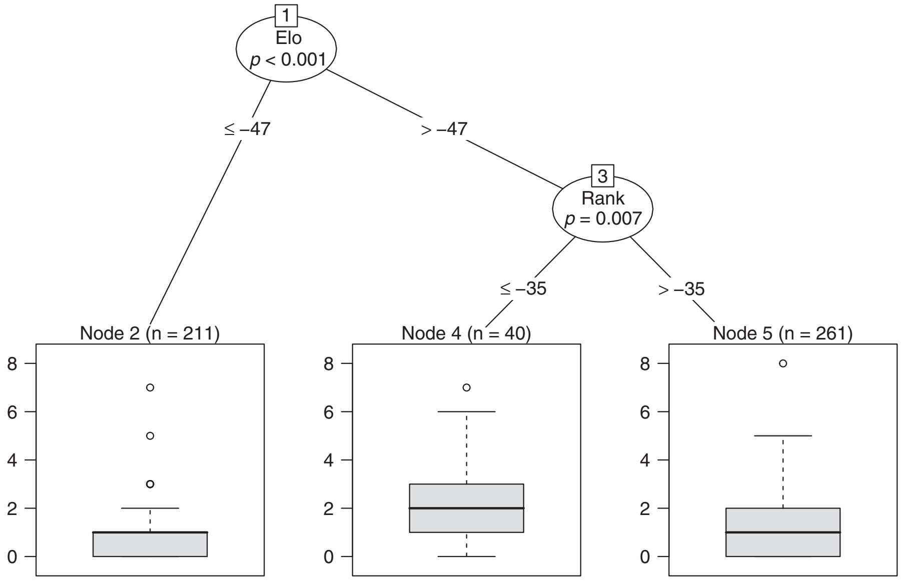

**Figure 1:** Exemplary regression tree for FIFA World Cup 2002–2014 data. Number of goals is used as response variable, variables described in Section 2 are used as predictors.

total of 3 partitions in the predictor space. The boxplots corresponding to each of the 3 final partitions show the distribution of the response (number of goals) for all observations falling into the respective nodes. In principle, one could of course perform many more splits, finally leading to perfectly separated partitions where each partition only contains observations referring to the same value for the response variable (i.e. the same number of goals scored). However, typical regression trees are "pruned" to prevent overfitting to the training data.

As mentioned before, random forests are the aggregation of a large number $B$ (e.g. $B = 5000$) of trees. The combination of many trees has the advantage that the resulting predictions inherit the feature of unbiasedness from the single trees while reducing the variance of the predictions. The single trees are grown independently from each other. To get a final prediction, predictions of single trees are aggregated, in our case of regression trees simply by averaging over all the predictions from the single trees. In order to achieve the goal that the aggregation of trees is less variant than a single tree, it is important to reduce the dependencies between the trees that are aggregated to a forest. Typically, two randomisation steps are applied to achieve this goal. First, the trees are not applied to the original sample but to bootstrap samples or random subsamples of the data. Second, at each node a (random) subset of the predictor variables is drawn which are used to find the best split. These steps de-correlate the single trees and help to lower the variance of a random forest compared to single trees. The size of the random subset of predictors at each node (argument `mtry`) is a tuning parameter; in what follows, we will choose this parameter by cross-validation. Following the suggestions of Probst and Boulesteix (2017) the number of trees $B$ does not have to be tuned as long as it is chosen sufficiently large.

In R (R Core Team 2018), two slightly different variants of regression forests are available. First, the classical random forest algorithm proposed by Breiman (2001) from the R-package `ranger` (Wright and Ziegler 2017). The second variant is implemented in the function `cforest` from the `party` package. Here, the single trees are constructed following the principle of conditional inference trees as proposed in Hothorn et al. (2006). The main advantage of these conditional inference trees is that they avoid selection bias in cases where the covariates have different scales, e.g. numerical vs. categorical with many categories (see, for example, Strobl et al. 2007, and Strobl et al. 2008, for details). Conditional forests share the feature of conditional inference trees of avoiding biased variable selection. Cross-validation of the tuning parameter `mtry` can be done using the machine learning framework provided by the R-package `mlr` (Bischl et al. 2016).

Besides regression forests modeling the exact number of goals, random forests for the categorical (ordinal) match outcome *win*, *draw* and *loss* can be applied. Though these forests cannot directly be used for the simulation of exact match outcomes, Schauberger and Groll (2018) explain how to suitably combine them with a random forest predicting the number of goals. Altogether, in the preliminary work from Schauberger and Groll (2018) the predictive performance of these different random forest approaches has been compared and it turned out that the `cforest` from the `party` package yielded the best results. For this reason, in the remainder of this work we will focus on this specific approach (from now on simply referred to as *Random Forest*).

## 3.2 Ranking methods

In this section we describe how (based on historic match data, see Section 2.2) Poisson models can be used to obtain rankings that reflect a team's current ability. We will restrict our attention to the best-performing model according to the comparison achieved in Ley et al. (2019), namely the bivariate Poisson model. The main idea consists in assigning a strength parameter to every team and in estimating those parameters over a period of $M$ matches via weighted maximum likelihood, where the weights are of two types: time depreciation and match importance.

We start by describing the weights. The time decay function is defined as follows: a match played $x_m$ days back gets a weight of

$$w_{time,m}(x_m) = \left(\frac{1}{2}\right)^{\frac{x_m}{\text{Half period}}},$$

meaning that a match played *Half period* days ago only contributes half as much as a match played today and a match played 3 × *Half period* days ago contributes 12.5% of a match played today. We stress that the *Half period* refers to calendar days in a year, not match days. This ensures that recent matches receive more importance and leads to the desired current-strength ranking. The match importance weights are directly inherited from the official FIFA ranking and can take the values 1 for a friendly game, 2.5 for a confederation or World Cup qualifier, 3 for a confederation tournament (e.g. UEFA EUROs or the Africa Cup of Nations) or the confederations cup, and 4 for World Cup matches. The relative importance of a national match is indicated by $w_{type,\,m}$ for $m = 1, \ldots, M$.

The bivariate Poisson ranking model is based on a proposal from Karlis and Ntzoufras (2003) and can be described as follows. If we have $M$ matches featuring a total of $n$ teams, we write $Y_{ijm}$ the random variable *number of goals scored by team i against team j* ($i, j \in \{1, \dots, n\}$) *in match m* (where $m \in \{1, \dots, M\}$). The joint probability function of the home and away score is then given by the bivariate Poisson probability mass function,

$$P(Y_{ijm} = z, Y_{jim} = y) = \frac{\lambda_{ijm}^{z}\lambda_{jim}^{y}}{z!y!} \exp(-(\lambda_{ijm} + \lambda_{jim} + \lambda_C)) \cdot \sum_{k=0}^{\min(z,y)} \binom{z}{k}\binom{y}{k} k! \left(\frac{\lambda_C}{\lambda_{ijm}\lambda_{jim}}\right)^{k},$$

where $\lambda_C$ is a covariance parameter assumed to be constant over all matches and $\lambda_{ijm}$ is the expected number of goals for team $i$ against team $j$ in match $m$, which we model as

$$\log(\lambda_{ijm}) = \beta_0 + (r_i - r_j) + h \cdot \mathbb{I}(\text{team } i \text{ playing at home}), \tag{1}$$

where $\beta_0$ is a common intercept and $r_i$ and $r_j$ are the strength parameters of team $i$ and team $j$, respectively. Since the ratings are unique up to addition by a constant, we add the constraint that the sum of the ratings has to equal zero. The last term $h$ represents the home effect and is only added if team $i$ plays at home. Note that we have the Independent Poisson model if $\lambda_C = 0$. The overall (weighted) likelihood function then reads

$$L = \prod_{m=1}^{M} \left(P(Y_{ijm} = y_{ijm}, Y_{jim} = y_{jim})\right)^{w_{type,m} \cdot w_{time,m}},$$

where $y_{ijm}$ and $y_{jim}$ stand for the actual number of goals scored by teams $i$ and $j$ in match $m$. The values of the strength parameters $r_1, \dots, r_n$, which determine the resulting ranking, are computed numerically as maximum likelihood estimates on the basis of historic match data as described in Section 2.2. These parameters also allow to predict future match outcomes thanks to the formula (1).

Following the findings of Ley et al. (2019)[2] we use the Bivariate Poisson model with a Half Period of 3 years, since this model was selected as the best ranking model according to the average Rank Probability Score (RPS; Gneiting and Raftery 2007), which is defined in Section 4. From now on this model is simply referred to as *Ranking*.

## 3.3 The hybrid random forest

In order to link the information provided by both the covariate data and historic match data we now combine the random forest approach from Section 3.1 and the ranking method from Section 3.2. We propose to use the ranking approach to generate a new (highly informative) covariate that can be incorporated into the random forest model. For that purpose, for each World Cup we estimate the team abilities $r_i, i = 1, \dots, 32$, of all 32 participating teams shortly before the start of the respective tournament. For example, to obtain ability estimates for the 32 teams that participated in the World Cup 2002, the historic match data for a certain time period preceding the World Cup 2002 (we chose to use 8 years, weighted by the described time depreciation effect) is used. This procedure gives us the estimates $\hat{r}_i$ as an additional covariate covering the current strength of all teams participating in a certain World Cup. Actually, this variable appears to be somewhat similar to the FIFA ranking, but turns out to be much more informative, see Section 5.1. From now on the random forest model augmented by this new covariate is simply referred to as *Hybrid Random Forest*.

The newly generated variable can be added to the covariate data based on previous World Cups and a random forest can be fitted to these data. Based on this random forest, new matches (e.g. matches from an upcoming World Cup) can be predicted. To predict a new observation, its covariate values are dropped down each of the $B$ regression trees, resulting in $B$ distinct predictions. The average of those is then used as a point estimate of the expected numbers of goals conditioning on the covariate values. However, these point estimates cannot directly be used for the prediction of the outcome of single matches or a whole tournament. First of all, plugging in both predictions corresponding to one match does not necessarily deliver an integer outcome (i.e. a result). For example, one might get predictions of 2.3 goals for the first and 1.1 goals for the second team. Furthermore, as no explicit distribution is assumed for these predictions it is not possible to randomly draw results for the respective match. Hence, similar to the regression methods described in Appendix B, we will treat the predicted expected value for the number of goals as an estimate for the intensity $\lambda$ of a Poisson distribution $Po(\lambda)$. This procedure could be motivated by assuming that within each terminal node of

[2] Ley et al. (2019) have compared the bivariate Poisson model to the Independent Poisson model and other models from the literature such as Thurstone-Mosteller and Bradley-Terry models, and found that the Bivariate Poisson model is the best to use for the purpose of building current-strength rankings at both national team and domestic league levels.

a tree we fit a simple intercept-only Poisson model where the average of all scores equals the maximum likelihood estimate of the intercept parameter. This way we can randomly draw results for single matches and compute probabilities for the match outcomes *win*, *draw* and *loss* by using two independent Poisson distributions (conditional on the covariates) for both scores.

## 4 Model performance

In the following, we investigate the predictive performance of the proposed hybrid random forest model by comparing it to a series of other models. On the one hand, we compare the hybrid model to its separate components, namely the pure random forest (without additional team abilities) and to the ranking method. On the other hand, we use a Lasso regression approach as an additional reference method. This method is also based on the covariate data from Section 2.1 and links the covariate information to the number of goals in a log-linear Poisson model. The model is estimated using a penalized likelihood approach. The details for this method can be found in Appendix B. Analogous to the proposed random forest method, the Lasso regression method can also be extended by incorporating the team abilities from the ranking method as a further covariate. Accordingly, these two methods are from now on referred to as *Lasso* and *Hybrid Lasso*, respectively.

These five different approaches are now compared with regard to their predictive performance. For this purpose, we apply the following general procedure on the World Cup 2002–2014 data for all methods except for the ranking approach:

1. *Form a training data set containing three out of four World Cups.*
2. *Fit each of the methods to the training data.*
3. *Predict the left-out World Cup using each of the prediction methods.*
4. *Iterate steps 1–3 such that each World Cup is once the left-out one.*
5. *Compare predicted and real outcomes for all prediction methods.*

This procedure ensures that each match from the total data set is once part of the test data and we obtain out-of-sample predictions for all matches. For the ranking model, we do not have to apply the described iterative procedure and can directly jump to step 5. Instead, the model was fit to a large data set covering all historic matches from the past 8 years up to the start of the respective World Cup that shall be predicted. The corresponding ability parameter estimates can then be used directly for the prediction. In step 5, several different performance measures for the quality of the predictions are investigated.

Let $\tilde{y}_i \in \{1, 2, 3\}$ be the true ordinal match outcomes for all $i = 1, \dots, N$ matches from the four considered World Cups. Additionally, let $\hat{\pi}_{1i}, \hat{\pi}_{2i}, \hat{\pi}_{3i}$, $i = 1, \dots, N$, be the predicted probabilities for the match outcomes obtained by one of the different methods mentioned above. These can be computed by assuming that the numbers of goals follow (conditionally) independent Poisson distributions, where the event rates $\lambda_{1i}$ and $\lambda_{2i}$ for the scores of match $i$ are estimated by the respective predicted expected values. Let $G_{1i}$ and $G_{2i}$ denote the random variables representing the number of goals scored by two competing teams in match $i$. Then, the probabilities $\hat{\pi}_{1i} = P(G_{1i} > G_{2i})$, $\hat{\pi}_{2i} = P(G_{1i} = G_{2i})$ and $\hat{\pi}_{3i} = P(G_{1i} < G_{2i})$, which are based on the corresponding Poisson distributions $G_{1i} \sim Po(\hat{\lambda}_{1i})$ and $G_{2i} \sim Po(\hat{\lambda}_{2i})$ with estimates $\hat{\lambda}_{1i}$ and $\hat{\lambda}_{2i}$, can be easily calculated via the Skellam distribution. For a short description of the Skellam distribution, see Appendix A. Based on these predicted probabilities, we use three different performance measures to compare the predictive power of the methods:

– the multinomial *likelihood*, which for a single match outcome is defined as $\hat{\pi}_{1i}^{\delta_{1\tilde{y}_i}} \hat{\pi}_{2i}^{\delta_{2\tilde{y}_i}} \hat{\pi}_{3i}^{\delta_{3\tilde{y}_i}}$, with $\delta_{r\tilde{y}_i}$ denoting Kronecker's delta, which is defined in Appendix A. The multinomial likelihood reflects the probability of a correct prediction. Hence, a large value reflects a good fit.

– the *classification rate*, based on the indicator functions $\mathbb{I}(\tilde{y}_i = \underset{r \in \{1,2,3\}}{\arg\max}\,(\hat{\pi}_{ri}))$, indicating whether match $i$ was correctly classified. Again, a large value of the classification rate reflects a good fit. However, note that along the lines of Gneiting and Raftery (2007) the classification rate does not constitute a proper scoring rule.

– the *rank probability score* (RPS), which, in contrast to both measures introduced above, explicitly accounts for the ordinal structure of the responses. For our purpose, it can be defined as $\frac{1}{3-1} \sum_{r=1}^{3-1} \left( \sum_{l=1}^{r} (\hat{\pi}_{li} - \delta_{l\tilde{y}_i}) \right)^2$. As the RPS is an error measure, here a low value represents a good fit.

Odds provided by bookmakers serve as a natural benchmark for these predictive performance measures. For this purpose, we collected the so-called "three-way" odds for (almost) all matches of the FIFA World

**Table 4:** Comparison of the prediction methods for ordinal match outcomes.

| | Likelihood | Class. rate | RPS |
|---|---|---|---|
| Hybrid random forest | 0.422 | 0.548 | 0.187 |
| Random forest | 0.408 | 0.536 | 0.191 |
| Ranking | 0.413 | 0.532 | 0.191 |
| Lasso | 0.422 | 0.524 | 0.199 |
| Hybrid Lasso | 0.429 | 0.552 | 0.194 |
| Bookmakers | 0.425 | 0.524 | 0.188 |

**Table 5:** Comparison of the prediction methods for the exact number of goals and the goal difference based on mean quadratic error.

| | Goal difference | Goals |
|---|---|---|
| Hybrid random forest | 2.487 | 1.290 |
| Random forest | 2.553 | 1.323 |
| Ranking | 2.582 | 1.353 |
| Lasso | 2.959 | 1.479 |
| Hybrid Lasso | 2.833 | 1.449 |

Cups 2002–2014[3]. By taking the three quantities $\tilde{\pi}_{ri} = 1/\text{odds}_{ri}, r \in \{1, 2, 3\}$, of a match $i$ and by normalizing with $c_i := \sum_{r=1}^{3} \tilde{\pi}_{ri}$ in order to adjust for the bookmaker's margins, the odds can be directly transformed into probabilities using $\hat{\pi}_{ri} = \tilde{\pi}_{ri}/c_i$ [4].

Table 4 displays the results for these (ordinal) performance measures for the five prediction methods as well as for the bookmakers, averaged over 250 matches from the four FIFA World Cups 2002–2014. It turns out that the hybrid random forest outperforms both of its separate components with respect to all performance measures. Hence, the basic idea of combining these two methods appears to be beneficial. The hybrid random forest is even able to match up to the bookmakers as the natural benchmark. While the average likelihood is only slightly lower than for the bookmakers, the RPS of the hybrid random forest is the best among all competitors. In general, the random forest approaches also stand out for their high classification rates, only being slightly outperformed by the hybrid Lasso. The two Lasso methods yield satisfactory results with respect to most criteria, only in terms of RPS they are outperformed by the other methods. Furthermore, the hybrid Lasso approach also shows that the incorporation of the ability estimates from the ranking method can provide valuable additional information compared to the remaining covariates.

As the proposed method can also be used to simulate the tournament course of an upcoming tournament (see also Appendix D for a prediction of the FIFA World Cup 2018), we are also interested in the performance of the regarded methods with respect to the prediction of the exact number of goals. In order to identify the teams that qualify for the knockout stage, the precise final group standings need to be determined. To be able to do so, the precise results of the matches in the group stage play a crucial role[5].

For this reason, we also evaluate the methods' performances with regard to the quadratic error between the observed and predicted number of goals for each match and each team, as well as between the observed and predicted goal difference. Now let $y_{ijk}$, for $i, j = 1, \dots, n$ and $k \in \{2002, 2006, 2010, 2014\}$, denote the observed numbers of goals scored by team $i$ against team $j$ in tournament $k$ and $\hat{y}_{ijk}$ a corresponding predicted value, obtained by one of the compared methods. Then we calculate the two quadratic errors $(y_{ijk} - \hat{y}_{ijk})^2$ and $\left((y_{ijk} - y_{jik}) - (\hat{y}_{ijk} - \hat{y}_{jik})\right)^2$ for all $N$ matches of the four FIFA World Cups 2002–2014. Finally, per method we calculate (mean) quadratic errors. Note that in this case the odds provided by the bookmakers cannot be used for comparison. So, in contrast to Table 4 where six matches had to be left out due to missing bookmaker information, now all $N = 256$ matches are used. Table 5 shows that the hybrid random forest outperforms all competitors with respect to both criteria and, hence, combining random forests with ranking methods again appears to be profitable.

[3] Three-way odds consider only the match tendency with possible results *victory team 1*, *draw* or *defeat team 1* and are usually fixed some days before the corresponding match takes place. This allows the bookmakers to incorporate current information (e.g. injuries of important players) into the odds during the run of a tournament. The three-way odds were obtained from the website http://www.betexplorer.com/. Unfortunately, for 6 matches from the FIFA World Cup 2006 no odds were available, hence, the results from Table 4 are based on 250 matches only.

[4] The transformed probabilities implicitely assume that the bookmaker's margins are equally distributed on the three possible match tendencies.

[5] 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 induced by lot.

# 5 Modeling the FIFA World Cup 2018

We now fit the proposed hybrid random forest model to the World Cup 2002–2014 data, and calculate the ability parameters based on historic match data over the 8 years preceding the World Cup 2018. The fitted random forest will then be used to evaluate its predictive performance with respect to the results of the FIFA World Cup 2018 tournament.

## 5.1 Fitting the hybrid random forest to World Cups 2002–2014

We fit the hybrid random forest approach with $B = 5000$ single trees to the complete data set covering all World Cups from 2002 to 2014. The optimal number of input variables randomly sampled as candidates at each node is determined by cross-validation and is set to `mtry=5`. Of course, it would be appealing to visualize and interpret the obtained results in order to learn about the relationship between the sporting success of a soccer team and the set of possible influence variables. However, in contrast to regression trees, random forests are much harder to visualize and to interpret. While for individual trees the effect of a single predictor can (almost) be seen at one glance when looking at the respective dendrogram (see Figure 1), this is almost impossible for random forests. Each predictor may have different effects (or no effect at all) in different trees. The best way to nevertheless understand the role of the single predictor variables is the so-called variable importance, see Breiman (2001). Typically, the variable importance of a predictor is measured by permuting each of the predictors separately in the out-of-bag observations of each tree. Out-of-bag observations are observations which are not part of the respective subsample or bootstrap sample that is used to fit a tree. Permuting a variable means that within the variable each value is randomly assigned to a location within the vector. If, for example, *Age* is permuted, the average age of the German team in 2002 could be assigned to the average age of the Brazilian team in 2010. When permuting variables randomly, they lose their information with respect to the response variable (if they have any). Then, one measures the loss of prediction accuracy compared to the case where the variable is not permuted. Permuting variables with a high importance will lead to a higher loss of prediction accuracy than permuting values with low importance. To illustrate the concept of variable importance, Figure 2 shows bar plots of the variable importance values for all variables in the hybrid random forest applied to the World Cup 2002–2014 data. Interestingly, the abilities are by far the most important predictor in the random forest and carry clearly more information than all other predictors. In particular, the information contained in the bookmakers' odds to win the World Cup (covariate *Oddset*), as well as the *Elo ratings* or the *FIFA rank* are much less important. Even though the *Elo ratings*, the *FIFA rank* and the bookmakers' odds certainly contain some information concerning the current playing abilities of the teams it seems to be worth the effort to estimate such abilities in a separate model. For a more detailed comparison of the team abilities, the *Elo ratings*, and the *FIFA rank*, see Appendix C.

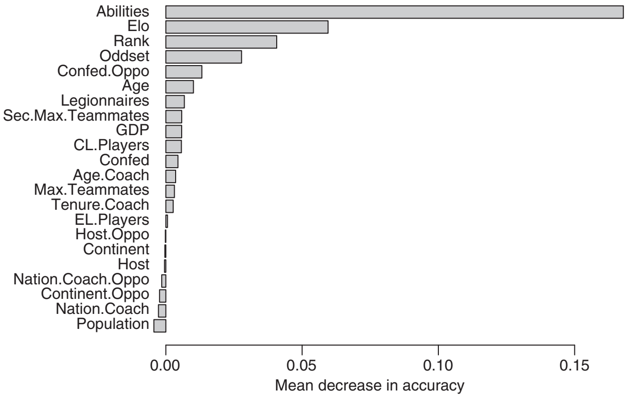

**Figure 2:** Bar plot displaying the variable importance in the hybrid random forest applied to FIFA World Cup data.

## 5.2 Prediction performance

Based on the hybrid random forest fitted in the previous section, we have simulated the FIFA World Cup 2018 100,000 times. These simulations allowed to compute tournament-winning probabilities for all participating teams as well as the most probable tournament course. The corresponding results can be found in Appendix D. It is not possible to evaluate the quality of the estimated winning probabilities for the single teams based on only one replication of the World Cup. However, in retrospective we can compare the results of the 64 matches to the model's predictions. Hence, the matches of the FIFA World Cup 2018 serve as an independent validation data set for our previous analyses from Section 4. We use the same goodness-of-fit metrics as introduced there and use again the Skellam distribution to compute the probabilities for *win, draw* or *loss* for each match. We show the corresponding results in Tables 6 and 7. It can be seen, that the performance of our new hybrid random forest is very satisfactory,

**Table 6:** Comparison of the prediction methods for ordinal match outcomes for the validation data from the World Cup 2018.

| | Likelihood | Class. rate | RPS |
|---|---|---|---|
| Hybrid random forest | 0.442 | 0.609 | 0.190 |
| Random forest | 0.430 | 0.609 | 0.193 |
| Lasso | 0.424 | 0.562 | 0.207 |
| Hybrid Lasso | 0.438 | 0.594 | 0.197 |
| Ranking | 0.422 | 0.578 | 0.194 |
| Bookmakers | 0.438 | 0.562 | 0.194 |

**Table 7:** Comparison of the prediction methods for the exact number of goals and the goal difference based on mean quadratic error for the validation data from the World Cup 2018.

| | Goal difference | Goals |
|---|---|---|
| Hybrid random forest | 2.159 | 1.194 |
| Random forest | 2.198 | 1.197 |
| Lasso | 2.355 | 1.225 |
| Hybrid Lasso | 2.192 | 1.172 |
| Ranking | 2.134 | 1.250 |

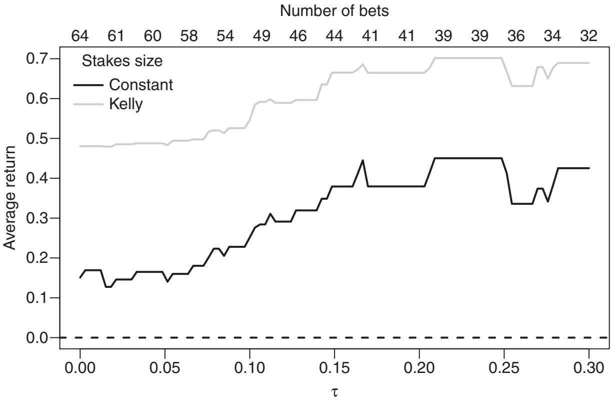

**Figure 3:** Average returns of the hybrid random forest model for constant (black line) and Kelly-based (gray line) stake sizes for varying threshold value $\tau$.

as it exhibits the best results with respect to all criteria except for the precise number of goals (see last column of Table 7).

In addition, using the (average) betting odds of the 64 matches as well as the corresponding predicted probabilities of our hybrid random forest model, certain betting strategies can be applied. For every match $i$ and each of the possible three outcomes $r \in \{1, 2, 3\}$ one can calculate the expected return as follows: $E[return_{ri}] = \hat{\pi}_{ri} * odds_{ri} - 1$. In general, one would choose the outcome with the highest expected return and only place the bet if the expected return is positive, i.e. if $\max\limits_{r\in\{1,2,3\}} E[return_{ri}] > \tau$, with $\tau = 0$. Koopman and Lit (2015) use different values of the threshold $\tau > 0$ and showed that this way the overall mean return could be increased. However, they use constant stake sizes (one unit) for each bet. In contrast, Boshnakov et al. (2017) applied a betting strategy with varying stake sizes based on the Kelly criterion (Kelly 1956). This criterion is a strategy to determine the optimal stake for single bets in order to maximize the return considering the size of the odds and the winning probability.

Figure 3 depicts the average return percentage (i.e. the ratio between profit and investment) of both strategies for varying threshold sizes $\tau \geq 0$. Note that with increasing value of $\tau$ the number of matches on which bets are placed is decreasing (see top axis). For the FIFA World Cup 2018 both betting strategies lead to positive returns for all threshold values $\tau$, where the Kelly strategy turned out to be more profitable. An investment of $100 would yield a profit up to $70 using the Kelly strategy in combination with a large value of the threshold $\tau$. However, these results have to be treated with caution. Due to the rather small sample size the results very much depend on single match results and are probably highly variable. Despite this limitation the results of the betting strategies turn out to be very favorable for our new model.

## 6 Concluding remarks

In this work, we proposed a new hybrid modeling approach for the scores of international soccer matches which combines random forests with Poisson ranking methods. While the former component is based on the competing teams' covariate information, the latter provides ability parameters, which serve as adequate estimates of the current team strengths. In order to combine both methods, the ranking method needs to be repeatedly applied to historical match data preceding each World Cup from the training data. This way, for each World Cup in the training data and each participating team current ability estimates are obtained. These estimates can be added as an additional covariate to the set of covariates used in the random forest procedure.

We compared the new hybrid random forest model to its separate building blocks as well as to conventional Poisson regression models with regard to their predictive performance on all matches from the four FIFA World Cups 2002–2014. The comparison revealed that by combining the random forest with the team ability parameters from the ranking methods the predictive power could be substantially improved. In fact, the hybrid random forest was even capable of competing with the bookmakers.

Next, we fitted the hybrid random forest to a training data set containing all matches of the four FIFA World Cups 2002–2014 and calculated the ability parameters over all historic matches in the last 8 years preceding the FIFA World Cup 2018, and used our new model (in advance of the tournament) to predict that World Cup. To complete our analysis, the corresponding 64 matches served as an independent validation data set and the compelling predictive potential of the hybrid random forest could be confirmed: it clearly outperformed all other methods including the betting odds.

Additionally, based on the estimates of the hybrid random forest on the training data, we repeatedly simulated the FIFA World Cup 2018 100,000 times. According to these simulations, Spain and Germany were supposed to be the top favorites for winning the title, with a slight advantage for Spain. Furthermore, survival probabilities for all teams and at all tournament stages as well as the most probable tournament course are provided.

# Appendix

## A Some notations and definitions

Kronecker's delta, which is used in Section 4 in the formula of the multinomial likelihood and the RPS, is defined as follows:

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j\,, \\ 0, & \text{otherwise}\,. \end{cases}$$

The Skellam distribution, which is also used in Section 4, is the discrete probability distribution of the integer random variable that is defined as the difference $K := Y_1 - Y_2$ of two independent Poisson distributed random variables $Y_1, Y_2$ with respective event rates $\lambda_1, \lambda_2$. The corresponding probability mass function is given by

$$P(K = k) = e^{-(\lambda_1+\lambda_2)} \left(\frac{\lambda_1}{\lambda_2}\right)^{k/2} I_k(2\sqrt{\lambda_1\lambda_2}), \quad k \in \mathbb{Z},$$

where $I_k(\cdot)$ is the modified Bessel function of the first kind (for more details, see Skellam 1946). Now let $Y_1$ and $Y_2$ denote the (conditionally independent) Poisson-distributed numbers of goals of two soccer teams competing in a match. Then, the three probabilities $P(Y_1 > Y_2)$, $P(Y_1 = Y_2)$ and $P(Y_1 < Y_2)$ can be easily obtained by computing $P(K > 0)$, $P(K = 0)$ and $P(K < 0)$ via the Skellam distribution.

## B Lasso regression for soccer data

An alternative, more traditional approach which is often applied for modeling soccer results is based on regression. In the most popular case the scores of the competing teams are treated as (conditionally) independent variables following a Poisson distribution (conditioned on certain covariates), as introduced in the seminal works of Maher (1982) and Dixon and Coles (1997). Similar to the random forests, the methods described here can also be directly applied to data in the format of Table 2 from Section 2.1. Hence, each score is treated as a single observation and one obtains two observations per match. Accordingly, for $n$ teams the respective model has the form

$$\begin{aligned} Y_{ijk}|\boldsymbol{x}_{ik}, \boldsymbol{x}_{jk} &\sim Po(\lambda_{ijk})\,, \\ \log(\lambda_{ijk}) &= \beta_0 + (\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk})^\top \boldsymbol{\beta} + \boldsymbol{z}_{ik}^\top \boldsymbol{\gamma} + \boldsymbol{z}_{jk}^\top \boldsymbol{\delta}\,, \end{aligned} \tag{2}$$

where $Y_{ijk}$ denotes the score of team $i$ against team $j$ in tournament $k$ with $i, j \in \{1, \ldots, n\},\ i \neq j$. The metric characteristics of both competing teams are captured in the $p$-dimensional vectors $\boldsymbol{x}_{ik}, \boldsymbol{x}_{jk}$, while $\boldsymbol{z}_{ik}$ and $\boldsymbol{z}_{jk}$ capture dummy variables for the categorical covariates *Host*, *Continent*, *Confed* and *Nation.Coach* (built, for example, by reference encoding), separately for the considered teams and their respective opponents. Furthermore, $\boldsymbol{\beta}$ is a parameter vector which captures the linear effects of all metric covariate differences and $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ collect the effects of the dummy variables corresponding to the teams and their opponents, respectively. For notational convenience, we collect all covariate effects in the $\tilde{p}$-dimensional real-valued vector $\boldsymbol{\theta}^\top = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\delta}^\top)$.

Due to a rather large number of potential covariates in our data, we use regularization techniques when estimating the models to allow for variable selection and to avoid overfitting. In the following, we will introduce such a basic regularization approach, namely the conventional Lasso (Tibshirani 1996). For estimation, instead of the regular likelihood $l(\beta_0, \boldsymbol{\theta})$ the penalized likelihood

$$l_p(\beta_0, \boldsymbol{\theta}) = l(\beta_0, \boldsymbol{\theta}) - \lambda P(\boldsymbol{\theta}) \tag{3}$$

is maximized, where $P(\boldsymbol{\theta}) = \sum_{v=1}^{\tilde{p}} |\theta_v|$ denotes the ordinary Lasso penalty with tuning parameter $\lambda$. The optimal value for the tuning parameter $\lambda$ will be determined by (standard) 10-fold cross-validation (CV) simply as the parameter that minimizes the CV error. The model will be fitted using the function `cv.glmnet` from the `R`-package `glmnet` (Friedman, Hastie, and Tibshirani 2010). In contrast to the similar ridge penalty (Hoerl and Kennard 1970), which penalizes squared parameters instead of absolute values, Lasso does not only shrink parameters towards zero, but is able to set them to exactly zero. Therefore, depending on the chosen value of the tuning parameter, Lasso also enforces variable selection.

### Possible extensions

While the Lasso method described above was chosen as the reference method to compare the predictive power of the hybrid model, in the literature also several alternatives and extensions are discussed. In the following, we shortly sketch some possible modifications. As a first possible extension of the model (2), the linear predictor can be augmented by team-specific attack and defense effects for all competing teams. This extension was used in Groll et al. (2015) to predict the FIFA World Cup 2014. There, each couple of attack and defense parameters corresponding to a team has been treated as a group and, hence, the Group Lasso penalty proposed by Yuan and Lin (2006) has been applied on those parameter groups.

Alternatively, if the model (2) shall be extended from linear to smooth covariate effects $f(\cdot)$ for metric covariates, boosting techniques designed for generalized additive models could be used, such as the `gamboost` algorithm from the `mboost` package (Hothorn et al. 2017). Instead of the Poisson distribution the negative binomial distribution could be used as the response distribution when considering distributions for count data, which is less restrictive as it overcomes the rather strict assumption of the expectation equating the variance. Schauberger and Groll (2018) investigated two different boosting approaches for this model class. However, no overdispersion compared to the Poisson assumption was detected and the models reduced back to the Poisson case.

Altogether, in Schauberger and Groll (2018) the simple Lasso from (3) with predictor structure (2) turned out to be the best-performing regression approach, though slightly outperformed by the random forests from Section 3.1.

## C Comparison of FIFA ranking, Elo rating and estimated abilities

Table 8 compares the ranking of the 32 participating teams in the FIFA World Cup 2018 according to estimated abilities (left column), Elo rating (center column) and FIFA ranking (right column). The ranking according to the estimated abilities and the Elo ratings are very similar (Spearman correlation of 0.94), while both have a smaller correlation with the FIFA ranking (Spearman correlation of 0.86 and 0.90, respectively).

**Table 8:** Ranking of the participants of the FIFA World Cup 2018 according to estimated abilities (left), Elo rating (center) and FIFA ranking (right).

| | Abilities | Elo rating | FIFA ranking |
|---|---|---|---|
| 1 | Brazil | Brazil | Germany |
| 2 | Germany | Germany | Brazil |
| 3 | Spain | Spain | Belgium |
| 4 | Argentina | France | Portugal |
| 5 | Colombia | Argentina | Argentina |
| 6 | Belgium | Portugal | Switzerland |
| 7 | France | England | France |
| 8 | Portugal | Belgium | Poland |
| 9 | Uruguay | Colombia | Spain |
| 10 | England | Peru | Peru |
| 11 | Croatia | Uruguay | Denmark |
| 12 | Poland | Switzerland | England |
| 13 | Denmark | Denmark | Uruguay |
| 14 | Peru | Croatia | Mexico |
| 15 | Sweden | Mexico | Colombia |
| 16 | Switzerland | Poland | Croatia |
| 17 | Mexico | Sweden | Tunisia |
| 18 | Serbia | Iran | Iceland |
| 19 | Russia | Serbia | Costa Rica |
| 20 | Iceland | Iceland | Sweden |
| 21 | Senegal | Senegal | Senegal |
| 22 | Morocco | Costa Rica | Serbia |
| 23 | Costa Rica | Australia | Australia |
| 24 | Iran | Morocco | Iran |
| 25 | Japan | South Korea | Morocco |
| 26 | Egypt | Japan | Egypt |
| 27 | Australia | Nigeria | Nigeria |
| 28 | Nigeria | Russia | Panama |
| 29 | Tunisia | Panama | South Korea |
| 30 | South Korea | Tunisia | Japan |
| 31 | Panama | Egypt | Saudi Arabia |
| 32 | Saudi Arabia | Saudi Arabia | Russia |

All three methods rank Germany and Brazil as the two top teams. Notable differences between the rankings can be seen, for example, for Spain and Belgium. Both the estimated abilities and the Elo rating rank Spain third while it is ranked ninth by FIFA. Belgium is ranked rather inhomogenously in positions 6, 8 and 3 by the different methods. More details on the comparison of estimated team abilities and the FIFA rank can be found in Ley et al. (2019).

## D Probabilities for FIFA World Cup 2018 Winner

In this section, the hybrid random forest is applied to (new) data for the World Cup 2018 in Russia (in advance of the tournament) to predict winning probabilities for all teams and to predict the tournament course.

The abilities were estimated by the bivariate Poisson model with a half period of 3 years. All matches of the 228 national teams played since 2010-06-13 up to 2018-06-06 are used for the estimation, what results in a total of more than 7000 matches. All further predictor variables are taken as the latest values shortly before the World Cup (and using the finally announced squads of 23 players for all nations).

For each match in the World Cup 2018, the hybrid random forest can be used to predict an expected number of goals for both teams. Given the expected number of goals, a real result is drawn by assuming two (conditionally) independent Poisson distributions for both scores. Based on these results, all 48 matches from the group stage can be simulated and final group standings can be calculated. Due to the fact that real results are simulated, we can precisely follow the official FIFA rules when determining the final group standings[5]. This enables us to determine the matches in the round-of-sixteen and we can continue by simulating the knockout stage. In the case of draws in the knockout stage, we simulate extra-time by a second simulated result. However, here we multiply the expected number of goals by the factor 0.33 to account for the shorter time to score (30 min instead of 90 min). In the case of a further draw in extra-time we simulate the penalty shootout by a (virtual) coin flip.

Following this strategy, a whole tournament run can be simulated, which we repeat 100,000 times. Based on these simulations, for each of the 32 participating teams probabilities to reach the single knockout stages and, finally, to win the tournament are obtained. These are summarized in Table 9 together with the winning probabilities based on the ODDSET odds for comparison.

**Table 9:** Estimated probabilities (in %) for reaching the different stages in the FIFA World Cup 2018 for all 32 teams based on 100,000 simulation runs of the FIFA World Cup together with winning probabilities based on the ODDSET odds.

| | | | Round of 16 | Quarter finals | Semi finals | Final | World Champion | Oddset |
|---|---|---|---|---|---|---|---|---|
| 1. | 🇪🇸 | ESP | 80.5 | 61.2 | 38.0 | 22.7 | 13.7 | 11.8 |
| 2. | 🇩🇪 | GER | 78.0 | 49.0 | 30.4 | 19.1 | 11.5 | 15.0 |
| 3. | 🇫🇷 | FRA | 77.8 | 49.9 | 32.1 | 18.5 | 10.8 | 11.8 |
| 4. | 🇧🇷 | BRA | 75.0 | 44.1 | 28.0 | 17.6 | 10.3 | 15.0 |
| 5. | 🇧🇪 | BEL | 75.9 | 52.5 | 30.1 | 17.7 | 9.9 | 8.3 |
| 6. | 🏴󠁧󠁢󠁥󠁮󠁧󠁿 | ENG | 73.1 | 49.8 | 26.6 | 14.7 | 7.5 | 4.6 |
| 7. | 🇦🇷 | ARG | 71.8 | 39.9 | 22.3 | 11.1 | 5.4 | 8.3 |
| 8. | 🇭🇷 | CRO | 66.4 | 33.5 | 18.3 | 8.5 | 3.8 | 3.0 |
| 9. | 🇵🇹 | POR | 61.1 | 39.8 | 18.7 | 8.0 | 3.2 | 3.8 |
| 10. | 🇨🇴 | COL | 71.4 | 32.2 | 15.5 | 7.4 | 3.2 | 1.8 |
| 11. | 🇨🇭 | SUI | 55.4 | 29.5 | 14.1 | 6.6 | 2.9 | 1.0 |
| 12. | 🇺🇾 | URU | 82.7 | 38.5 | 16.9 | 7.1 | 2.8 | 2.8 |
| 13. | 🇩🇰 | DEN | 56.0 | 27.2 | 14.4 | 6.3 | 2.6 | 1.1 |
| 14. | 🇸🇪 | SWE | 54.5 | 24.3 | 10.7 | 4.8 | 1.9 | 0.8 |
| 15. | 🇷🇸 | SRB | 43.6 | 20.5 | 9.1 | 3.8 | 1.6 | 0.6 |
| 16. | 🇵🇱 | POL | 56.6 | 21.5 | 8.9 | 3.7 | 1.3 | 1.5 |
| 17. | 🇵🇪 | PER | 40.9 | 18.3 | 8.7 | 3.3 | 1.3 | 0.4 |
| 18. | 🇮🇸 | ICE | 37.2 | 14.4 | 6.4 | 2.4 | 1.0 | 0.6 |
| 19. | 🇸🇳 | SEN | 45.3 | 17.0 | 7.0 | 2.8 | 1.0 | 0.6 |
| 20. | 🇲🇦 | MOR | 38.0 | 19.6 | 6.9 | 2.4 | 0.8 | 0.3 |
| 21. | 🇲🇽 | MEX | 42.5 | 16.2 | 5.8 | 2.2 | 0.7 | 1.0 |
| 22. | 🇹🇳 | TUN | 31.4 | 13.6 | 5.0 | 1.9 | 0.7 | 0.2 |
| 23. | 🇦🇺 | AUS | 25.3 | 8.4 | 3.2 | 1.0 | 0.3 | 0.3 |
| 24. | 🇳🇬 | NGA | 24.7 | 8.3 | 3.2 | 1.0 | 0.3 | 0.6 |
| 25. | 🇨🇷 | CRC | 26.0 | 9.2 | 3.2 | 1.1 | 0.3 | 0.3 |
| 26. | 🇪🇬 | EGY | 49.1 | 15.3 | 4.2 | 1.2 | 0.3 | 0.6 |
| 27. | 🇷🇺 | RUS | 49.6 | 13.5 | 3.9 | 1.1 | 0.3 | 2.2 |
| 28. | 🇯🇵 | JPN | 26.7 | 7.8 | 2.3 | 0.7 | 0.2 | 0.6 |
| 29. | 🇰🇷 | KOR | 24.9 | 7.1 | 2.1 | 0.6 | 0.2 | 0.6 |
| 30. | 🇮🇷 | IRN | 20.4 | 8.3 | 2.1 | 0.5 | 0.1 | 0.3 |
| 31. | 🇵🇦 | PAN | 19.5 | 5.6 | 1.4 | 0.4 | 0.1 | 0.1 |
| 32. | 🇸🇦 | KSA | 18.6 | 3.7 | 0.7 | 0.1 | 0.0 | 0.1 |

We can see that, according to our hybrid random forest model, Spain was the favored team with a predicted winning probability of 13.7% followed by Germany, France, Brazil and Belgium. Overall, this result seems in line with the probabilities from the bookmakers, as we can see in the last column. While Oddset favors Germany and Brazil, the hybrid random forest model predicts a slight advantage for Spain. However, we can see no clear favorite, as several teams seem to have good chances. In retrospect, the early drop-outs of Germany and Spain seem rather surprising. While Spain at least played a successful group stage finishing in first place, Germany performed unexpectedly bad with two defeats during the group stage. The probability for such an early drop-out of Germany was predicted to be only around 22% and, therefore, could be seen as the biggest surprise of the tournament. Spain failed in the round-of-16 against host Russia in a penalty shoot-out and, hence, did not reach the quarter finals (the probability for this event had been predicted to be about 39%). Beside the probabilities of becoming world champion, Table 9 provides some further interesting insights also for the single stages within the tournament. For example, it is interesting to see that the two favored teams Spain and Germany had almost equal chances to at least reach the round-of-sixteen (80.5% and 78.0%, respectively), while the probabilities to at least reach the quarter finals differ significantly. While Spain had a probability of 61.2% to reach at least the quarter finals, Germany only achieved a probability of 49.0%. Obviously, in contrast to Spain, Germany had a rather high chance to meet a strong opponent in the round-of-sixteen. In case they would have reached the round-of-sixteen, Germany would have faced either Brazil, Switzerland, Serbia or Costa Rica, while Spain would have faced Uruguay, Russia, Saudi Arabia or Egypt. In the following rounds, Germany catches up to Spain finally ending up with almost equal winning probabilities.

## Most probable tournament course

Finally, based on the 100,000 simulations, we also provide the most probable tournament course. For each of the eight groups we selected the most probable final group standing, while also considering the order of the first two places, but not the irrelevant order of the teams on places three and four. The results together with the corresponding probabilities are presented in Table 10.

Obviously, there are large differences with respect to the groups' balances. While in Group B and Group G the model forecasts Spain followed by Portugal as well as Belgium followed by England with rather high probabilities of 27.7% and 26.7%, respectively, other groups such as Group D, Group E, Group F and Group H seem to be more volatile. Now that we know the true tournament outcome, it is worth a note that indeed in Group B and G the first two places were exactly taken by the two forecasted teams, while in Group F and H there were some surprises.

**Table 10:** Most probable final group standings together with the corresponding probabilities for the FIFA World Cup 2018 based on 100,000 simulation runs.

| Group A<br>25.0% | Group B<br>27.7% | Group C<br>23.8% | Group D<br>22.8% |
|---|---|---|---|
| 1. URU | 1. ESP | 1. FRA | 1. ARG |
| 2. RUS | 2. POR | 2. DEN | 2. CRO |
| KSA | MOR | AUS | ICE |
| EGY | IRN | PER | NGA |
| **Group E**<br>**21.5%** | **Group F**<br>**23.3%** | **Group G**<br>**26.7%** | **Group H**<br>**19.7%** |
| 1. BRA | 1. GER | 1. BEL | 1. COL |
| 2. SUI | 2. SWE | 2. ENG | 2. POL |
| CRC | MEX | PAN | SEN |
| SRB | KOR | TUN | JPN |

Moreover, we provide the most probable course of the knockout stage in Figure 4. The most likely round-of-sixteen directly results from those teams qualifying for the knockout stage in Table 10. For all following matches we compute the probabilities for the respective two teams (say team A and team B) to go to the next stage. This is done by applying the Skellam distribution to first get the probabilities for *A wins*, *draw* and *B wins* after 90 minutes. Second, the probability for *draw* is distributed between teams A and B again following the principles of extra-time and penalty shootouts we already applied for draws in the knockout stage in the previous section. This way the probabilities for *A wins* and *B wins* add up to 1, as is necessary for the knockout stage. In Figure 4, the probabilities accompanying the edges of the tournament tree represent the probability of the favored team to proceed to the next stage.

In the most probable tournament course Germany wins the World Cup. However, again it becomes obvious that with (in that case) Switzerland the German team would have had to face a much stronger opponent than Spain in the round-of-sixteen. Even though they still were the favorite in this match, they would have succeeded to move on to the quarter finals only with a probability of 58%. While in the most probable course of the knock-out stage, though having tough times in all single

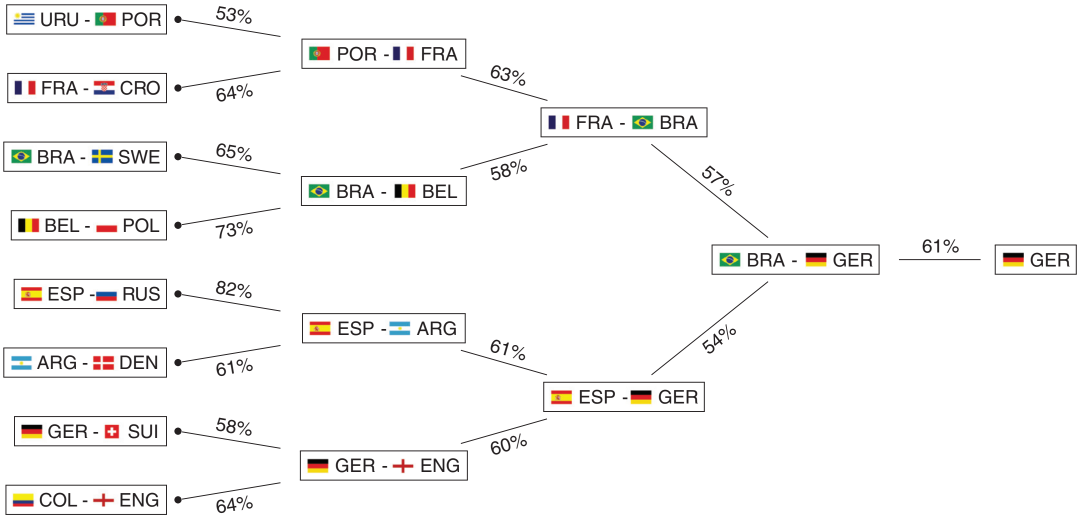

**Figure 4:** Most probable course of the knockout stage together with corresponding probabilities for the FIFA World Cup 2018 based on 100,000 simulation runs.

stages, Germany would have made its way into the final and defended the title, the previous section showed that generally still Spain was the most likely winner.

We wish to attract the reader's attention to the fact that, despite being the most probable tournament course, due to the myriad of possible constellations this exact tournament course still was extremely unlikely: if we take the product of all single probabilities of Table 10 and Figure 4, its overall probability yields $7.63 \cdot 10^{-9}$%. Hence, deviations of the true tournament course from the model's most probable one were not only possible, but very likely.

## References

Bischl, B., M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, and Z. M. Jones. 2016. "mlr: Machine Learning in R." *Journal of Machine Learning Research* 17:1–5. http://jmlr.org/papers/v17/15-066.html.

Boshnakov, G., T. Kharrat, and I. G. McHale. 2017. "A Bivariate Weibull Count Model for Forecasting Association Football Scores." *International Journal of Forecasting* 33:458–466. http://www.sciencedirect.com/science/article/pii/S0169207017300018.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45:5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and J. C. Stone. 1984. *Classification and Regression Trees*. Monterey, CA: Wadsworth.

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.

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.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33:1.

Gneiting, T. and A. E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102:359–378.

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., T. Kneib, A. Mayr, and G. Schauberger. 2018. "On the Dependency of Soccer Scores – A Sparse Bivariate Poisson Model for the UEFA European Football Championship 2016." *Journal of Quantitative Analysis in Sports* 14:65–79.

Groll, A., G. Schauberger, and G. Tutz. 2015. "Prediction of Major International Soccer Tournaments Based on Team-Specific Regularized Poisson Regression: An Application to the FIFA World Cup 2014." *Journal of Quantitative Analysis in Sports* 11:97–115.

Hoerl, A. E. and R. W. Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12:55–67.

Hothorn, T., P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. van der Laan. 2006. "Survival Ensembles." *Biostatistics* 7:355–373.

Hothorn, T., P. Buehlmann, T. Kneib, M. Schmid, and B. Hofner. 2017. *mboost: Model-Based Boosting*. https://CRAN.R-project.org/package=mboost, R package version 2.8-1.

Karlis, D. and I. Ntzoufras. 2003. "Analysis of Sports Data by Using Bivariate Poisson Models." *The Statistician* 52:381–393.

Kelly, J. L. 1956. "A New Interpretation of Information Rate." *Bell System Technical Journal* 35:917–926. http://dx.doi.org/10.1002/j.1538-7305.1956.tb03809.x.

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: Series A (Statistics in Society)* 178:167–186.

Leitner, C., A. Zeileis, and K. Hornik. 2010. "Forecasting Sports Tournaments by Ratings of (Prob)Abilities: A Comparison for the EURO 2008." *International Journal of Forecasting* 26(3):471–481.

Ley, C., T. Van de Wiele, and H. Van Eetvelde. 2019. "Ranking Soccer Teams on the Basis of their Current Strength: A Comparison of Maximum Likelihood Approaches." *Statistical Modelling* 19:55–77. https://doi.org/10.1177/1471082X18817650.

Maher, M. J. 1982. "Modelling Association Football Scores." *Statistica Neerlandica* 36:109–118.

McHale, I. and P. Scarf. 2007. "Modelling Soccer Matches Using Bivariate Discrete Distributions with General Dependence Structure." *Statistica Neerlandica* 61:432–445. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.2007.00368.x.

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.

Probst, P. and A.-L. Boulesteix. 2017. "To Tune or not to Tune the Number of Trees in Random Forest?" *Journal of Machine Learning Research* 18:181:1–181:18.

Quinlan, J. R. 1986. "Induction of Decision Trees." *Machine Learning* 1:81–106.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schauberger, G. and A. Groll. 2018. "Predicting Matches in International Football Tournaments with Random Forests." *Statistical Modelling* 18:460–482. https://doi.org/10.1177/1471082X18799934.

Skellam, J. G. 1946. "The Frequency Distribution of the Difference between Two Poisson Variates Belonging to Different Populations." *Journal of the Royal Statistical Society. Series A (General)* 109:296–296.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics* 8:25.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. 2008. "Conditional Variable Importance for Random Forests." *BMC Bioinformatics* 9:307.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society* B58: 267–288.

Wright, M. N. and A. Ziegler. 2017. "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77:1–17.

Yuan, M. and Y. Lin. 2006. "Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society* B68:49–67.