Statistical Modelling 2018; **18(5-6)**: 460–482

Predicting matches in international football tournaments with random forests

Gunther Schauberger[1,2] and Andreas Groll[3]
[1]Chair of Epidemiology, Department of Sport and Health Sciences, Technical University of Munich, Germany.
[2]Department of Statistics, Ludwig-Maximilians-Universität München, Germany.
[3]Faculty of Statistics, Technische Universität Dortmund, Germany.

Abstract: Many approaches that analyse and predict results of international matches in football are based on statistical models incorporating several potentially influential covariates with respect to a national team's success, such as the bookmakers' ratings or the FIFA ranking. Based on all matches from the four previous FIFA World Cups 2002–2014, we compare the most common regression models that are based on the teams' covariate information with regard to their predictive performances with an alternative modelling class, the so-called random forests. Random forests can be seen as a mixture between machine learning and statistical modelling and are known for their high predictive power. Here, we consider two different types of random forests depending on the choice of response. One type of random forests predicts the precise numbers of goals, while the other type considers the three match outcomes—win, draw and loss—using special algorithms for ordinal responses. To account for the specific data structure of football matches, in particular at FIFA World Cups, the random forest methods are slightly altered compared to their standard versions and adapted to the specific needs of the application to FIFA World Cup data.

Key words: random forests, football, FIFA World Cups, Poisson regression, regularization

1 Introduction

In the last decade, an increasing interest in the modelling and prediction of major international football events can be observed. As a consequence, many different statistical techniques and approaches have been applied and adapted to deal with different types of football data. Among these, an essential class of models is based on regression methods, which incorporate covariate information of the opposing teams.

In particular, Poisson regression models have gained a lot of attention, where the numbers of goals of both competing teams can be directly linked to a set of influence variables. Early references in this context are, for example, Lee (1997) and Dyte and Clarke (2000). The latter researchers focus on scores in international football

Address for correspondence: Gunther Schauberger, Chair of Epidemiology, Department of Sport and Health Sciences, Technical University of Munich, Georg-Brauchle-Ring 56, 80992 München, Germany.
E-mail: gunther.schauberger@tum.de

 10.1177/1471082X18799934

matches, treating each team's goals as (conditionally) independent Poisson variables depending on two influence variables, namely the team's FIFA ranking and the match venue. More recently, Groll and Abedieh (2013) and Groll et al. (2015) have further extended these Poisson models by incorporating a large set of potential influence variables as well as (either random or fixed) team-specific ability parameters. By using different regularization techniques they discovered a sparse set of relevant covariates, which were then used to predict the European championship (EURO) 2012 and FIFA World Cup 2014 winners, respectively. In both cases, the actual tournament winner was identified as the most likely one by the model.

Note that, implicitly, all of these models treat the two numbers of goals scored in a match as independent (conditioned on covariate information). First approaches to account for possible dependencies between the scores by using adjusted Poisson models are proposed by Dixon and Coles (1997) and Rue and Salvesen (2000). Alternatively, the bivariate Poisson distribution allows to explicitly model (positive) dependence within the Poisson framework. One of the first works dealing with this distribution in the context of football data is Maher (1982). Furthermore, an extensive study for the use of the bivariate Poisson distribution for the modelling of football data can be found in Karlis and Ntzoufras (2003). However, in Groll et al. (2018) it has been shown by the help of gradient boosting techniques that (at least for their setting of EURO data) no additional modelling of the covariance structure is necessary: for suitably designed linear predictors, which are based on highly informative covariates, two (conditionally) independent Poisson distributions are adequate.

We also want to mention a completely different approach, which is solely based on the easily available source of 'prospective' information contained in bookmakers' odds, compare Leitner et al. (2010) and their follow-up papers. They obtain winning probabilities for each team by aggregating winning odds from several online bookmakers and then use inverse tournament simulation to compute team-specific abilities by paired comparison models. Based on these abilities, pairwise probabilities for each possible game at the corresponding tournament can be calculated and, finally, the whole tournament can be simulated.

In this work, we pursue a different approach and investigate an alternative tool for the prediction of the outcomes of football matches, namely random (decision) forests – an ensemble learning method for classification, regression and other tasks proposed by Breiman (2001). The method stems from the machine learning and data mining community and operates by first constructing a multitude of so-called decision trees (see, e.g., Quinlan, 1986; Breiman et al., 1984) on a training dataset. For prediction, the predictions from the individual trees are summarized, either by taking the mode of the predicted classes (in classification) or by averaging the predicted values (in regression). This way, random forests reduce the tendency of overfitting and the variance compared to regular decision trees, and, hence, are a common powerful tool for prediction. Therefore, random forests might also be a promising alternative for the prediction of football matches. In the present work, we use both random forests for metric (i.e., the number of goals) and ordinal response (i.e., win–draw–loss) as well as the combination of both. On a dataset containing all matches of the FIFA World Cups 2002–2014, we compare the predictive performance of these different

types of random forests with conventional regression methods for count data, such as Poisson generalized linear models (GLMs).

The rest of the manuscript is structured as follows: In Section 2, we describe the underlying dataset covering all matches of the four preceding FIFA World Cups 2002–2014. Next, in Section 3 we explain how random forests can be used as prediction tools for the outcomes of football matches. Alternative, regression-based methods are summarized in Section 4. The main differences between the outputs of random forests and regression models for football predictions are highlighted in Section 5. Both modelling alternatives are then compared with regard to their predictive performance in Section 6. Finally, we conclude in Section 7.

2 Data

In this section, we provide a brief description of the underlying dataset covering all matches of the four preceding FIFA World Cups 2002–2014 together with several potential influence variables. In general, we use essentially the same set of covariates that is introduced in Groll et al. (2015). For each participating team, most of these covariates are observed shortly before the start of the respective World Cup (e.g., the FIFA ranking) or for the same year of the World Cup (e.g., the GDP per capita). Therefore, the covariate values of the teams may vary from one World Cup to another. Several of the variables contain information about the recent performance and sportive success of national teams, as it is reasonable to assume that the current form of a national team has an influence on the team's success in the upcoming tournament. Beside these sportive variables, also certain economic factors as well as variables describing the structure of a team's squad are collected. A detailed description of these variables can be found in Groll et al. (2015).

- **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 as a ratio with the respective global population to account for the general growth of the world population (source: `http://data.worldbank.org/indicator/SP.POP.TOTL`).
- **ODDSET probabilities:** Bookmaker odds provided by the German state betting agency ODDSET are converted into winning probabilities. Therefore, the variable reflects probabilities for each team to win the respective World Cup. (The possibility of betting 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).
- **FIFA Rank:** The FIFA ranking provides a ranking system for all national teams measuring the performance of the team over the last four years (source: `http://de.fifa.com/worldranking/index.html`).

- **Host:** A dummy variable indicating whether or not 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 confederation of the respective team with (in principle) six possible values: Africa (CAF); Asia (AFC); Europe (UEFA); North, Central America and Caribbean (CONCACAF); Oceania (OFC); South America (CONMEBOL). The confederations OFC and AFC had to be merged because in the dataset only one team (New Zealand, 2006) from OFC participated in a World Cup.
- **(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 a few weeks before the respective World Cup) of the UEFA Champions League (CL) and UEFA Europa League are counted.
- **Number of Players Abroad:** For each squad, the number of players playing in clubs abroad (in the season previous to the respective World Cup) is counted.
- **Factors Describing the Team's Coach:** For the coach of each national team, the 'Age' and the duration of his 'Tenure' are observed. Furthermore, a dummy variable is included, whether the coach has the same 'Nationality' as his team or not.

In total, this adds up to 16 variables which were collected separately for each World Cup and each participating team. For illustration, Table 1 shows exemplarily for the first four matches of the FIFA World Cup 2002 the results (1a) and (parts of) the covariates (1b) of the respective teams. In the remainder of this section, this data excerpt will be used to illustrate how the final datasets are constructed.

Table 1 Exemplary table showing the results of four matches and parts of the covariates of the involved teams

(a) Table of results

FRA	0:1	SEN
URU	1:2	DEN
FRA	0:0	URU
DEN	1:1	SEN
⋮	⋮	⋮

(b) Table of covariates

World Cup	Team	Age	Rank	Oddset	...
2002	France	28.3	1	0.149	...
2002	Uruguay	25.3	24	0.009	...
2002	Denmark	27.4	20	0.012	...
2002	Senegal	24.3	42	0.006	...
⋮	⋮	⋮	⋮	⋮	⋱

For the modelling techniques introduced in the following sections, all of the metric covariates are incorporated in the form of differences. 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 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.

In general, we will consider two different types of response variables which will lead to two fundamentally different datasets. For the first type, the number of goals is directly used as response variable. Therefore, each match corresponds to two different observations, one per team. The second type uses the ordinal variable with categories 1 (win), 2 (draw) and 3 (loss) from the perspective of the first-named team. Therefore, in this case each match represents one observation in the dataset. Also the covariate differences are computed from the perspective of the first-named team. For illustration, the resulting data structures for the exemplary matches from Table 1 are displayed in Table 2, separately for count data response in Table 2a and for ordinal response in Table 2b.

Table 2 Exemplary tables illustrating the data structure for both response types

(a) Data structure for count data response (goals)

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	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

(b) Data structure for ordinal response (1: win; 2: draw; 3: loss).

Result	Team	Opponent	Age	Rank	Oddset	...
3	France	Senegal	4.00	−41	0.14	...
3	Uruguay	Denmark	−2.10	4	−0.00	...
2	France	Uruguay	3.00	−23	0.14	...
2	Denmark	Senegal	3.10	−22	0.01	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

3 Modelling football results using random forests

In this work, we propose to use random forests as prediction tools for the outcomes of football matches. Before introducing possible specific strategies for the application of random forests to football data we start with a general introduction into the basic ideas of random forests.

3.1 Random forests

Random forests were introduced by Breiman (2001) as an extension of the method proposed by Ho (1998). The underlying principle of random forests is the aggregation of a (large) number of classification or regression trees and, therefore, the method can be used both for classification and regression purposes. The single trees are grown independently from each other. To get a final prediction, predictions of single trees are aggregated, either by majority vote (for classification) or by averaging (for regression).

Before going into further details of the principles of random forests we shortly sketch the main essence of classification and regression trees (Breiman et al., 1984). In general, the term classification tree is used for trees with categorical (or binary) response variables, while trees for metric responses are called regression trees. With classification and regression trees the feature space is partitioned, each partition has its own prediction (or its own model, see, e.g., Zeileis et al., 2008). The partitioning of the predictor space is done recursively and can follow different criteria. However, the main goal is always to find the split which provides the strongest difference between the two new partitions with respect to the chosen criterion. Observations within the same partition are supposed to be as similar as possible, observations from different partitions are supposed to be very different (with respect to the response variable). The splits are performed subsequently, each partition can be further partitioned in the following step. The consecutive splitting steps can be visualized using a dendrogram.

For illustration we exemplarily fit a regression tree. We use (a part of) the data introduced in Section 2, which we will use later for an in-depth comparison of the predictive power of different methods. The data contain all matches from the FIFA World Cups 2002–2014. As response, we consider all final scores of the teams, that is, we have two observations per match. For simplicity, we only use three predictor variables, which are the differences between the 'FIFA Rank', the bookmakers' probabilities 'Oddset' and the 'Age' of both teams.

Figure 1 shows the resulting regression tree using the function `ctree` from the `R`-package `party` (Hothorn et al., 2006). The predictor space is partitioned into five partitions, each of the predictors is used at least once for a split. Now, the tree could be used as a prediction method for new observations. In each node, the average value of the response variable of the node members is used as prediction.

Random forests are grown by repeatedly growing different classification/regression trees and applied to new observations by combining the different predictions from the single trees. The main goal is to decrease the variance compared to single trees. Therefore, it is necessary to decrease the correlation between the single trees. For that purpose, two different randomization steps are applied. 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. But, in contrast to regular trees, in random forests the single trees are commonly not pruned. Pruning leads to a lower variance but also increases the bias. Accordingly, an unpruned tree has the advantage of being nearly

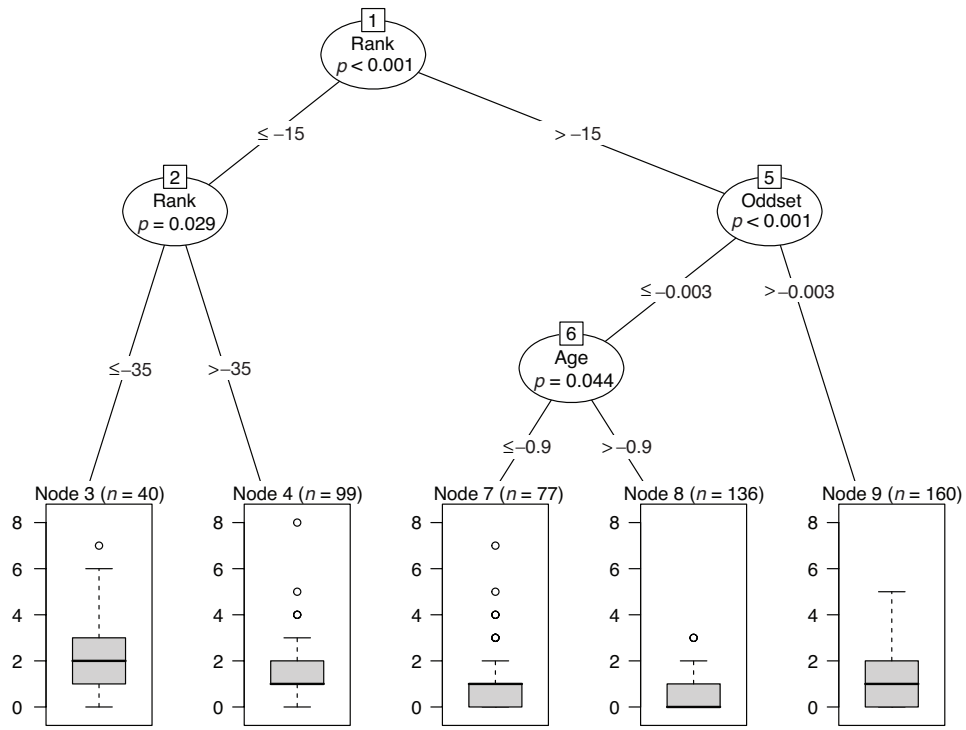

Figure 1 Exemplary regression tree for FIFA World Cup data. Number of goals is used as response variable, FIFA Rank, Oddset and Age are used as predictors

unbiased but the disadvantage of having a high variance. However, the combination of many trees compensates for the high variance of the individual trees. Therefore, by de-correlating and combining many trees, predictions with low bias and reduced variance can be achieved.

3.2 Random forests for football results

In this section, we explain how random forests can be used on football data. In principle, we distinguish between two fundamentally different approaches depending on the type of response (as already described in Section 2). Similar to the methods introduced in the following section, the first approach uses the number of goals as response. Here, in the dataset each match is represented by two rows, one per team. The response is treated as a metric variable and the forest is built using regression trees. However, no explicit distribution assumption is necessary for the application of random forests. The second type of random forests directly tries to classify the ordinal outcomes (win–draw–loss; always from the perspective of the first-named team) of a match. Therefore, each match is represented by a single row in the dataset.

3.2.1 Random forest for the prediction of the number of goals

When the metric variable 'Number of Goals' is considered as the response, we use regression trees for the single trees which are then combined into a random forest. The basic principle is that a predefined number of trees B (e.g., $B = 5\,000$) is fitted based on (bootstrap samples of) the training data. For the prediction of a new observation, the covariate values of the observation are dropped down each of the regression trees, resulting in B predictions. The final prediction is simply the average over all B predictions. This prediction can be seen as a point estimate of the expected value of the response conditioning on the covariate values.

We use two slightly different variants of random forests for this approach. First, we use the variant of the classical random forest algorithm proposed by Breiman (2001) from the `R`-package `ranger` (Wright and Ziegler, 2017). The second variant we use is 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 conditional inference trees is that one can avoid selection bias in cases where the covariates have different scales, for example, numerical versus categorical with many categories. The advantages of these so-called conditional random forests over classical ones are described by Strobl et al. (2007) and Strobl et al. (2008). Conditional forests share the feature of conditional inference trees of avoiding biased variable selection. Furthermore, the single predictions are not aggregated by simple averaging over the single predictions but using observation weights as described in Hothorn et al. (2004).

However, the point estimates for the numbers of goals can not directly be used for the prediction of the outcome of single matches or a whole tournament. Simply plugging in both predictions corresponding to one match does not deliver an integer outcome (i.e., a result) for the match. 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 the next section, we will use the predicted expected value for the number of goals as an estimate for the event rate λ of a Poisson distribution $Po(\lambda)$. 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.

3.2.2 Random forest for the prediction of ordinal match outcomes

If instead of the number of goals the ordinal match outcomes are used as response variable, random forests specifically designed for ordinal responses are applied. They can be seen as an in-between technique of regression and classification forests. In principle, forests for ordinal responses are built as regression forests where the ordinal categories are replaced by metric score values. The determination of the exact score values depends on the type of algorithm, but the score values can also be set by the user. The aggregation of the single trees is then executed analogously to the case of classification forests. This means that the final prediction is determined by majority vote over the single regression tree predictions.

Again, two different variants are used. The first variant is the function `cforest` from the `party` package. By default, for a three-categorical ordinal response it simply uses the values 1, 2 and 3 as score values. Equivalent to conditional forests for metric response variables, `cforest` uses observation weights to aggregate the single predictions. The second variant is an algorithm recently proposed by Hornung (2017b) that is implemented in the `ordinalForest` package (Hornung, 2017a). In contrast to the previous method, it uses a complex pre-processing step to learn sensible values for the scores in a data-driven way. Hence, one can avoid the (rather restrictive) assumption that the differences between the single ordinal categories (i.e., between the scores) are equal.

When the ordinal response is considered, it is essential which team is the first-named team. If the order of the teams was reverted, a value of 1 would be replaced by 3 and the other way around. Of course, such an inversion also needs to be accompanied by a redefinition of the covariates. However, even though 1 and 3 are somehow interchangeable, it turns out that across all four tournaments the relative frequencies of the three results are not balanced. While 40.6% of the matches are wins of the first-named team (i.e., take value 1) only 31.6% of the matches are wins of the second-named team (response value 3). The reason is that due to the specific tournament design of the FIFA there exists a certain structure with respect to first- and second-named teams: especially during the group stage of the World Cups the first-named team is usually the team from the 'stronger' draw bowl. This can be problematic because it implies certain asymmetries in the sense that there is a higher a priori probability of category 1 compared to 3 for each match, even if in a particular match the first-named team is expected to be worse than its opponent. For a good predictive performance it would be preferable if the ordering of first- and second-named team was random. Therefore, we try to escape this issue by additional randomization based on the following steps:

1. Build $T = 50$ versions of the training data with a randomized distribution of first- and second-named team.
2. Learn a separate random forest for each of the T randomized training datasets.
3. Build a second version of the test dataset which contains each match from the original test data with the inverted order of the competing teams.
4. Using each of the T random forests, obtain predictions for each match, separately for both versions of the test data and average the total of $2 \cdot T$ predictions/probabilities for all three response categories, as the final prediction for the respective match.

Obviously, the random forests for ordinal match outcomes cannot directly be used for the simulation of exact match outcomes. Therefore, we propose to combine an ordinal random forest with a random forest predicting the number of goals from Section 3.2.1. In that case, for the simulation of a match result we first randomly draw one of the three match outcomes based on the probabilities obtained from the ordinal forest. Subsequently, we randomly draw exact scores for both teams

using the predictions from a random forest for the number of goals as described in Section 3.2.1. Then the first match result which coincides with the drawn match outcome is accepted. For example, if in the first step we draw a win of the first-named team we accept the first result where the first-named team scores more goals than the second-named team.

4 Alternative approaches

We want to compare the described random forest approaches to more traditional modelling approaches which have already been used for modelling football results, at least in a similar way. In general, the most frequently used modelling approach for football results is to treat the scores of the competing teams as (conditionally) independent variables following a Poisson distribution (possibly conditioning on certain covariates). Especially the works of Dixon and Coles (1997) and Maher (1982) set the basics for this modelling approach. Therefore, all methods described in this section use the number of goals scored by single teams as response variables (compare Table 2a). If, as in our case, one wants to include several covariates of the competing teams into the model it is sensible to use regularization techniques when estimating the models to allow for variable selection and to avoid overfitting. In the following, we will apply three different regularization approaches.

Lasso

The most simple approach will be to use a conventional Lasso (Tibshirani, 1996) penalty for the covariate parameters. In this model, the single scores are used as response variable and (conditionally on the covariates) a Poisson distribution is assumed. Each score is treated as a single observation, so that per match there are two observations. Accordingly, for n teams the respective model has the form

$$\begin{aligned} Y_{ijk} | \boldsymbol{x}_{ik}, \boldsymbol{x}_{jk} &\sim Po(\lambda_{ijk}), \\ log(\lambda_{ijk}) &= \beta_0 + (\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk})^\top \boldsymbol{\beta} + \boldsymbol{z}_{ik}^\top \boldsymbol{\gamma} + \boldsymbol{z}_{jk}^\top \boldsymbol{\delta}. \end{aligned} \tag{4.1}$$

Here, Y_{ijk} denotes the score of team i against team j in tournament k, where $i, j \in \{1, \ldots, n\}$, $i \neq j$. 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. For these variables, it is not sensible to build differences between the respective values. 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 vector $\boldsymbol{\theta}^\top = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\delta}^\top)$.

For estimation, instead of the regular likelihood $l(\beta_0, \boldsymbol{\theta})$ the penalized likelihood

$$l_p(\beta_0, \boldsymbol{\theta}) = l(\beta_0, \boldsymbol{\theta}) + \lambda P(\beta_0, \boldsymbol{\theta}) \tag{4.2}$$

is maximized, where $P(\beta_0, \boldsymbol{\theta}) = \sum_{v=1}^{\tilde{p}} |\theta_v|$ is the ordinary Lasso penalty and λ is a tuning parameter. The optimal value for the tuning parameter λ will be determined by 10-fold cross-validation (CV). The model will be fitted using the function `cv.glmnet` from the `R`-package `glmnet` (Friedman et al., 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 also able to set parameters to exactly zero. Therefore, depending on the chosen value of the tuning parameter, Lasso also enforces variable selection.

In addition to the conventional Lasso solution that minimizes the 10-fold CV error, the `cv.glmnet` function from the `glmnet` package also provides a more sparse solution. Here, a different strategy is applied to choose the optimal value for the tuning parameter λ. Instead of choosing the model with the minimal CV error one chooses the most restrictive model which is within one standard error of the minimum of the CV error. We refer to this method as *Lasso (1se)* in the following.

Gamboost

Compared to the Lasso approach presented earlier, in the Gamboost approach the previous model is extended from linear to smooth covariate effects f_v for all metric covariates, leading to $log(\lambda_{ijk}) = \beta_0 + \sum_{v=1}^{p} f_v(x_{vik} - x_{vjk}) + \boldsymbol{z}_{ik}^{\top}\boldsymbol{\gamma} + \boldsymbol{z}_{jk}^{\top}\boldsymbol{\delta}$. In this case, suitable penalization is not as simple as in the case presented earlier. Therefore, we switch from penalization to boosting and will use the function `gamboost` from the package `mboost`. Boosting is an iterative fitting procedure where in each step the fit is updated only by a small increment. In our case, in every step only one of the covariates is selected (the one associated with the highest improvement of the fit) and only the respective smooth function is updated. Also, each of these respective updates is a rather smooth, only slightly non-linear function to prevent over-fitting. Of course, for all dummy variables the updates are simply linear effects, and to avoid over-fitting these linear updates are also shrunk. For more details on boosting algorithms see, for example, Bühlmann and Hothorn (2007). Similar to Lasso also boosting has to be tuned, in this case the number of boosting steps is the main tuning parameter. Tuning is done by fitting the model to B bootstrap samples (here $B = 25$) and evaluating the out-of-bag prediction error by taking advantage of all left-out observations. Implicitly, the selection of the number of boosting iterations also enforces variable selection, because all covariates for which the respective smooth function (or linear parameter) has never been updated (until reaching the optimal number of boosting steps) are eliminated from the final model.

Group Lasso

Finally, we use a Group Lasso approach which is a different extension of the Lasso approach presented earlier. This approach corresponds to the approach proposed

by Groll et al. (2015) where it was used to predict the FIFA World Cup 2014. Here, the linear predictor from (4.2) is extended by team-specific attack and defense effects for all competing teams and has the form $log(\lambda_{ijk}) = \beta_0 + (\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk})^\top \boldsymbol{\beta} + \boldsymbol{z}_{ik}^\top \boldsymbol{\gamma} + \boldsymbol{z}_{jk}^\top \boldsymbol{\delta} + att_i - def_j$. Note that, as pointed out by Groll et al. (2015), the inclusion of team-specific effects makes the parameters corresponding to the 'Continent' and 'Confed' variables unidentified. Therefore, the respective terms are excluded from $\boldsymbol{z}_{ik}$, $\boldsymbol{\gamma}$, $\boldsymbol{z}_{jk}$ and $\boldsymbol{\delta}$. The attack and defense parameters are considered as fixed effects and are again estimated using a penalized likelihood approach which extends the conventional Lasso penalty term by a Group Lasso (Yuan and Lin, 2006) penalty term. Altogether, the penalty term reads

$$(\beta_0, \boldsymbol{\theta}, \boldsymbol{att}, \boldsymbol{def}) = \sum_{\nu=1}^{\tilde{p}} |\theta_\nu| + \sqrt{2} \sum_{i=1}^{n} \sqrt{att_i^2 + def_i^2},$$

where $\boldsymbol{att}^\top = (att_1, \ldots, att_n)$ and $\boldsymbol{def}^\top = (def_1, \ldots, def_n)$, and the factor $\sqrt{2}$ accounts for the respective group size. The second part of the penalty term represents a group penalty on the team-specific effects, such that both effects corresponding to the same team form a group of parameters. In Group Lasso, groups of parameters can be defined where variable selection is then applied to the group as a whole. Therefore, either all parameters from a certain variable group enter the model or none. If selected, the additional team-specific effects can cover effects which are constant for the respective national team across all World Cups of the training data and which are not yet covered by the covariate effects.

When considering distributions for count data, an alternative to the Poisson distribution is the negative binomial distribution. In general, it is less restrictive than the Poisson distribution, as it overcomes the rather strict assumption of the expectation equating the variance. For this project, we also investigated two different modelling alternatives based on the assumption of negative binomially distributed responses. First, we again used the function `gamboost` from the `mboost` package, allowing for smooth covariate effects for all metric covariates. In contrast to the Poisson case, overdispersion is estimated in the form of an additional scale parameter. However, in our data analyses it turned out that the resulting models correspond to Poisson models because no overdispersion compared to the Poisson assumption was detected: the scale parameter was set to such a large value by the boosting method that the negative binomial model was (approximately) equivalent to the Poisson model. Second, we exploited the very flexible framework of the generalized additive model for location, scale and shape (GAMLSS; Rigby and Stasinopoulos, 2005), where also the second distribution parameter, that is, the scale, can be related to covariates. We applied a boosting approach proposed by Mayr et al. (2012), which is implemented in the `R`-package `gamboostLSS` (Hofner et al., 2016) and which allows to perform variable selection on the predictors of both mean and variance. Hence, in contrast to `gamboost` it would be possible to have also covariates alter the size of the scale parameters. However, the respective parameters were never updated.

Therefore, analogously to the simpler `gamboost` approach, it turned out that the scale parameter was unnecessary and that we again ended up with the Poisson model. Hence, the negative binomial approach is not further pursued for the remainder of this article.

5 Heuristic comparison of model outputs

There are fundamental differences between random forests and regression models with respect to their respective model outputs and the interpretations that these models allow for. Therefore, we want to shortly elaborate on these differences in a rather general manner. Exemplarily, we only pick one random forest for the number of goals and one for the ordinal match result (both based on the `party` package) and compare them to the conventional Lasso and Gamboost approaches. Each of the other methods has strong similarities to one of these approaches and is not treated separately here. The major goal is to highlight the main differences between the outputs of random forests and regression models for football prediction. For that purpose, the four approaches are fitted to the complete dataset introduced in Section 2.

The main goal of random forests is prediction. In contrast to regression models, they are harder to interpret because no explicit relationship between dependent and independent variables can be extracted. In particular, in contrast to the regularized regression methods they do not perform variable selection. Nevertheless, the importance of the single variables can be measured. Typically, this is done by permuting each of the variables separately in the out-of-bag observations of each tree and measuring the prediction accuracy. In this context, permuting a variable means that in 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 loose 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.

Figure 2 shows the respective values for the variable importance of each variable, separately for random forests predicting the number of goals and ordinal match outcomes. In the case of ordinal match outcomes we average over the values from the different (permuted) datasets, see Section 3.2.2. It can be seen that the domains of the importance values differ strongly between 'RF Goals' and 'RF Result', which is simply due to the fact that both models use different response types with different scalings. Besides that, the main outcomes are rather similar. The most important variables are 'FIFA Rank', 'Oddset' and '# CL Players'. Also, GDP and the 'Confed' variables seem to have some explanatory power. The remaining variables show very small or even negative values with respect to their variable importance and, therefore, do not provide additional explanatory power. However, the distinction between influential and non-influential variables is rather heuristic.

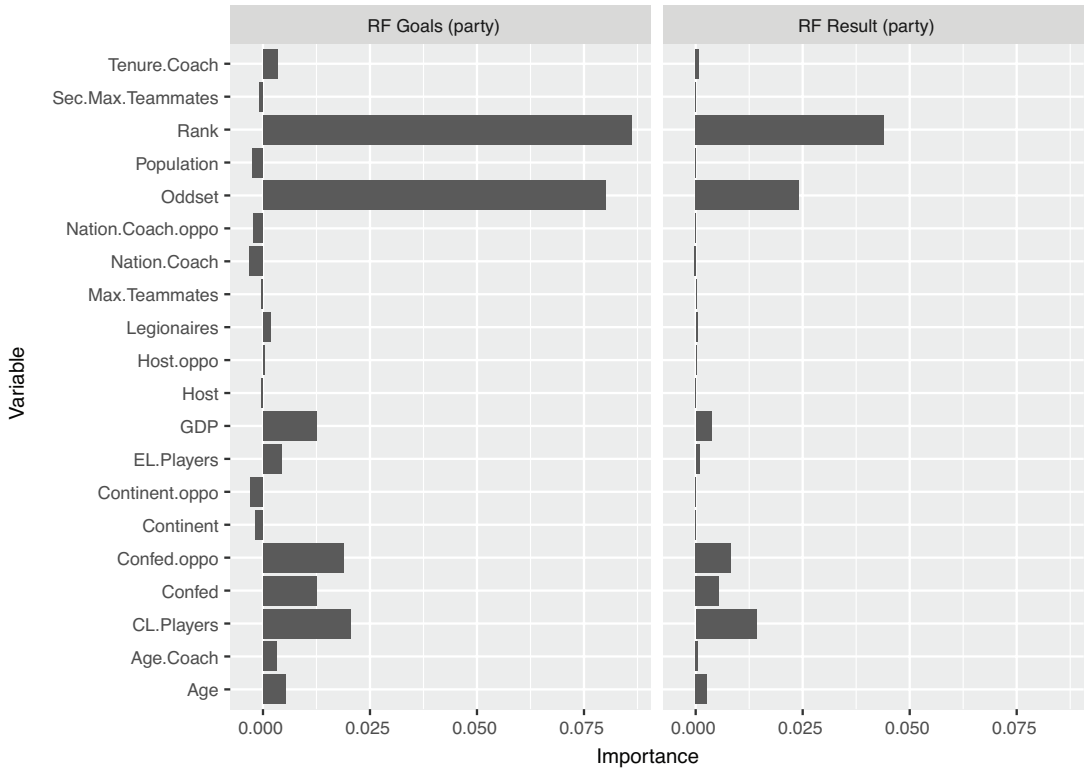

Figure 2 Variable importance for random forests with goals (left) and match results (right) as response variables for World Cup data from 2002–2014

In contrast to random forests, regression models can estimate explicit and interpretable relationships between the covariates and the response and, in our case of regularized regression, can explicitly discriminate between influential and non-influential variables. While the estimated relationships between the covariates and the response are strictly linear in the case of Lasso, Gamboost allows for smooth functions. Table 3 shows all parameter estimates for the Lasso approach which are different from zero.

Table 3 Standardized parameter estimates for Lasso on World Cup 2002–2014 data

# CL Players	Rank	Oddset
0.0210	−0.1773	0.0566

It can be seen that the final model is rather sparse with only three selected covariates. Interestingly, these variables coincide with the three most influential variables from both forest approaches. 'FIFA Rank' exhibits the strongest effect, followed by 'Oddset' and '# CL Players'. The 'FIFA Rank' has a negative effect

because, obviously, high values of this variable are supposed to indicate rather weak teams.

In contrast to the simple linear model assumed for Lasso estimation, in Gamboost the effects can also be non-linear (smooth) functions. The estimated (partial) effects for all selected variables are depicted in Figure 3. It can be seen that, with 15

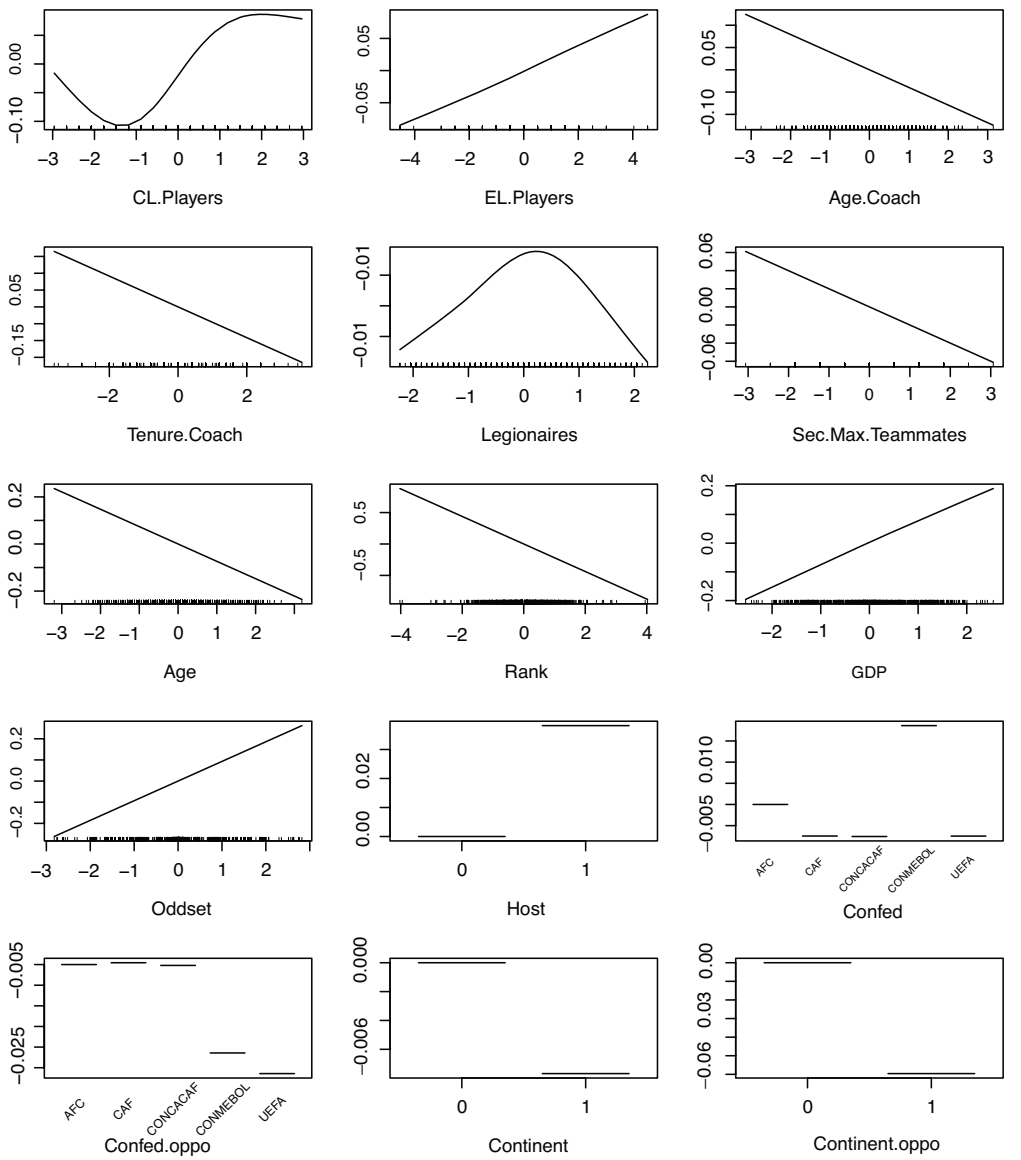

Figure 3 Partial effects for all (selected) covariates in the Gamboost approach for World Cup data from 2002–2014

variables included in the final model, the Gamboost model is clearly less sparse compared to the Lasso solution. However, only the variables '# CL Players' and 'Legionaires' explicitly show non-linear effects. Here, the effect sizes are somewhat harder to determine. Nevertheless, comparing the domains of the effects again shows that 'Oddset', 'FIFA Rank' and '# CL Players' are (among) the most important variables.

Overall, it turns out that, although the results obtained by either random forests or regression models need to be interpreted fundamentally differently, a coinciding set of major influence variables can be identified for both approaches. In particular, the sportive success of national teams in matches of the FIFA World Cups 2002–2014 is mainly determined by the 'Oddset', the 'FIFA Rank' and the '# CL Players'.

6 Comparison of predictive performance

In the following, we want to perform an in-depth comparison of the predictive power of all methods introduced in Sections 3.2 and 4. In particular, we are interested in the question whether the random forest approaches or the more traditional regression approaches perform better. This will be done by using the FIFA World Cup 2002–2014 dataset introduced in Section 2. We apply the following procedure:

1. Form a training dataset 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 guarantees that each match from the total dataset is once part of the test data. Therefore, we get out-of-sample predictions for all matches. In step 5, different performance measures for the quality of the predictions are investigated, separately for the prediction of the (ordinal) match outcomes and the number of goals.

6.1 Prediction of match outcomes

In the following, let $\tilde{y}_1, \ldots, \tilde{y}_N$ be the true ordinal match outcomes, that is, $\tilde{y}_i \in \{1, 2, 3\}$, for all matches N from the four considered World Cups. Additionally, let $\hat{\pi}_{1i}, \hat{\pi}_{2i}, \hat{\pi}_{3i}$, $i = 1, \ldots, N$, be the predicted probabilities for the match outcomes obtained by one of the different methods presented in Sections 3.2 and 4. While these probabilities are directly available from the ordinal random forests from Section 3.2.2, they need to be computed in an additional step for all methods predicting the number of goals. Both for the regression methods and the random forests from Section 3.2.1 we assume that the numbers of goals follow independent Poisson distributions, where the event rates λ_{1i} and λ_{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, we can compute the probabilities via $\hat{\pi}_{1i} = P(G_{1i} > G_{2i})$, $\hat{\pi}_{2i} = P(G_{1i} = G_{2i})$ and $\hat{\pi}_{3i} = P(G_{1i} < G_{2i})$ 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 $\hat{\lambda}_{2i}$. Based on these predicted probabilities, three different performance measures were used to compare the predictive power of the methods.

A classical performance measure for categorical responses is 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. It reflects the probability of a correct prediction. Hence, a large value of the multinomial likelihood reflects a good fit.

Furthermore, to later calculate the classification rate of each method we consider whether match i was correctly classified using the indicator function $\mathbb{I}(\tilde{y}_i = \underset{r \in \{1,2,3\}}{\arg\max}\,(\hat{\pi}_{ri}))$. Again, a large value of the classification rate reflects a good fit.

Gneiting and Raftery (2007) proposed to use the so-called 'rank probability score' (RPS) as a performance measure which, in contrast to both measures introduced earlier, 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.

As a natural benchmark for these predictive performance measures the predictions based on bookmakers' odds can be considered. For this purpose, we collected the so-called 'three-way' odds for (almost) all matches of the FIFA World Cups 2002–2014. Three-way odds consider only the match tendency with possible results victory of team 1, draw or defeat of team 1 and are usually fixed some days before the corresponding match takes place. 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. 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 $\hat{\pi}_{ri} = \tilde{\pi}_{ri}/c_i$. The transformed probabilities only serve as an approximation, based on the assumption that the bookmaker's margins follow a discrete uniform distribution on the three possible match tendencies. Using these predicted probabilities $\hat{\pi}_{ri}$, we can evaluate the three performance measures for (ordinal) match outcomes introduced earlier also for the information contained in bookmakers' odds.

Table 4 displays the results for these (ordinal) performance measures for all methods introduced in Sections 3.2 and 4 as well as for the bookmakers' odds, averaged over 250 matches from the four FIFA World Cups 2002–2014. It turns out that in terms of the mean multinomial likelihood score, all forest-based methods and also the conventional Lasso achieve a fit that is close to the one obtained

by the bookmakers, which here fulfill their role as a benchmark. With respect to the classification rate, the four forest-based methods clearly outperform all other approaches, remarkably even the bookmakers in this case. Again, the conventional Lasso is best performing among the regression methods with a performance equal to the bookmakers. Finally, in terms of the RPS again the four forest-based methods perform best, yielding clearly lower values than all regression approaches. In particular, the two random forests that directly model the number of goals achieve error rates very close to those of the bookmakers, which here again serve as the benchmark. The regression methods (except for Lasso (1se)) on the contrary yield all very similar results for RPS.

To sum up, all methods based on random forests provide very satisfactory results, which are either close to or even outperforming those obtained by the bookmakers, which can be seen as a natural benchmark. Altogether, the random forests that directly model the number of goals slightly outperform those for ordinal responses. Among the regression approaches, conventional Lasso clearly performs best and overall seems to be a convincing competitor to the forest-based methods.

Table 4 Comparison of the different prediction methods for the ordinal match outcome based on multinomial likelihood, classification rate and ranked probability score (RPS)

	Likelihood	Class. Rate	RPS
RF Goals (party)	0.409	0.548	0.192
RF Goals (ranger)	0.413	0.536	0.192
RF Result (party)	0.404	0.532	0.194
RF Result (ordinalForest)	0.419	0.532	0.195
Lasso	0.414	0.524	0.202
Lasso (1se)	0.364	0.516	0.214
Group Lasso	0.399	0.504	0.201
Gamboost	0.401	0.520	0.200
Bookmakers	0.425	0.524	0.188

6.2 Prediction of exact numbers of goals

Beside the ordinal match outcome (win, draw, loss), we are also interested in the performance of the regarded methods with respect to the prediction of the exact number of goals. This is important, for example, if one wants to predict the whole tournament course or the winning probabilities for a FIFA World Cup before the start of the tournament. The reason is that in order to identify the teams that qualify for the knockout stage, one has to determine the precise final group standings. To be able to do so, the precise results of the matches in the group stage play a crucial role. The final group standings are determined by (a) the number of points, (b) the goal difference and (c) 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 (a)–(c). If still no distinct decision can be taken, the decision is induced by lot.

For this reason, we also evaluate the performance of all introduced methods 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. Let now y_{ijk}, for $i, j = 1, \ldots, n$ and $k \in \{2002, 2006, 2010, 2014\}$, denote the observed number 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 methods from Sections 3.2 and 4. For those methods that combine ordinal and metric random forests we do not directly obtain a fixed predicted number of goals, because the predicted number of goals depends on the drawn match outcome. Therefore, for every method we randomly simulate 100 results (0:1,1:2,2:2,...) for each match we want to predict. Then we calculate the two quadratic errors $(y_{ijk} - \hat{y}_{ijk})^2$ (here we get two errors per simulated match, 200 errors per predicted match) and $\big((y_{ijk} - y_{jik}) - (\hat{y}_{ijk} - \hat{y}_{jik})\big)^2$ (here we get one error per simulated match, 100 errors per predicted match) for all N matches of the four FIFA World Cups 2002–2014. Finally, per method we average over these errors. Note that in this case the odds provided by the bookmakers can not be used for comparison. So in contrast to Table 4, where six matches had to be left out due to missing bookmaker information, we now can calculate both (mean) quadratic errors based on all $N = 256$ matches.

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

	Goal Difference	Goals
RF Goals (party)	4.924	2.519
RF Goals (ranger)	5.092	2.594
RF Result (party)	4.884	2.534
RF Result (ordinalForest)	5.097	2.537
Lasso	5.370	2.681
Lasso (1se)	5.325	2.586
Group Lasso	4.968	2.521
Gamboost	5.201	2.579

Table 5 summarizes the corresponding results. In general, the overall trend that the random forest methods outperform the regression-based approaches is confirmed. However, in contrast to the results from the previous subsection, here the disparities between the forest-based methods are less evident. Now, the best-performing method with respect to the quadratic error of the goal difference is based on a random forest for ordinal responses, namely the one implemented in the `party` package, while for the quadratic error of the goals a random forest that directly models the number of goals performs best, namely the `cforest` from the `party` package. However, note that in both ordinal approaches we use the `party` package to draw final results as described in Section 3.2.2. Therefore, it seems that the prediction of the number of goals works rather well with the random forests which are specifically designed for that purpose, but in some cases they can still be slightly improved by combining them

with a random forest prediction of the ordinal match outcome. Altogether, it is not possible to clearly identify a best-performing class of random forests with respect to the two different types of responses (ordinal or metric). The only regression approach that can compete with the forest-based methods with regard to the two quadratic errors is the Group Lasso, while the conventional Lasso here performs rather bad.

6.3 Comparison of betting returns

Further insight into the predictive performance of the different fitting procedures can be obtained by analysing the success of certain betting strategies. The betting strategy presented in the following is again based on the three-way odds from the website `http://www.betexplorer.com/`. For this reason, again only 250 matches can be considered, as for 6 matches no odds were available. For every match i and each of the possible three outcomes $r \in \{1, 2, 3\}$ one can calculate the expected return of a bet with a betting volume of one (arbitrary) monetary unit as follows: $E[return_{ri}] = \hat{\pi}_{ri} * odds_{ri} - 1$. This automatically follows from the fact that with probability π_{ri} one gets a payout of $odds_{ri}$ when betting on match outcome r. After subtracting the stake of one (arbitrary) unit we end up with the expected return. In general, one would choose the outcome with the highest expected return and only place the bet if the expected return is positive, that is, if $\max_{r\in\{1,2,3\}} E[return_{ri}] > 0$. The corresponding returns (in %) for all methods are shown in Table 6.

Table 6 Betting returns (in %) for different prediction methods

	Return
RF Goals (party)	1.41
RF Goals (ranger)	4.77
RF Result (party)	−1.01
RF Result (ordinalForest)	5.51
Lasso	3.96
Lasso (1se)	−7.53
Group Lasso	0.23
Gamboost	3.21

It turns out that, altogether, the differences between the returns are relatively high across the different methods. Among the random forests, those from the `party` package show rather bad results compared to the random forests from `ordinalForest` and `ranger`, which gain the highest returns among all methods. From the regression methods, the regular Lasso gains a return of 4%, while Lasso (1se) leads to a loss of 7.5%. However, the betting results should probably not be over-interpreted. First, only one betting company is considered. In fact, the odds provided by the website `http://www.betexplorer.com/` do not even correspond to a real betting company, but are average odds from several bookmakers. Also, the margins in the betting odds are rather high, they range between 4% and 13% per match (on average about 7%). In a real betting scenario, it is more realistic that

the player can choose from a variety of companies and place his bet at the company with the most favourable odds. Third, although the results are based on four World Cups the returns are still strongly depending on single matches and can not be seen as predictions for average betting gains for future World Cups.

7 Concluding remarks

In the present work we compared two fundamentally different, covariate-based approaches for the modelling and prediction of matches in international football tournaments, namely random forests and regression methods. We describe the methods and, on a dataset containing all matches of the FIFA World Cups 2002–2014, compare the predictive performance of random forests for both ordinal and metric response to conventional regression methods for count data, such as Poisson GLMs.

In order to evaluate the performance of the methods, several different performance measures for both ordinal match outcomes and the precise number of goals were investigated. For ordinal match outcomes, all methods based on random forests provided very satisfactory results, which were either close to or even outperforming those obtained by the bookmakers (serving as natural benchmark). Moreover, the forest-based methods outperformed the regression approaches. Only conventional Lasso turned out to be a convincing competitor to the forest-based methods. Within the forest-based methods, random forests that directly model the number of goals slightly outperformed those based on ordinal responses.

In terms of the quadratic errors for the precise number of goals, the overall trend that the random forest methods outperform the regression-based approaches was confirmed. However, disparities between the forest-based methods were less clear: while the best-performing method with respect to the goal difference was based on a random forest for ordinal responses, the best approach for the goals was a random forest directly modelling the number of goals. So here it was not possible to clearly identify a best-performing class of random forests, that is, forests for either ordinal or metric responses. The only regression approach able to compete with the forest-based methods was the Group Lasso, while the conventional Lasso performed rather bad.

Finally, we also analysed the performance of the methods in terms of the success of a simple betting strategy. In general, we found relatively high differences between the returns across the different methods. The highest returns among the tree-based methods were obtained by the `ordinalForest`, while again conventional Lasso performed best among the regression approaches.

Overall, our analyses showed that generally random forests slightly outperform regression-based approaches with respect to a variety of prediction performance measures. Only conventional Lasso turned out to be a promising competitor. Based on these findings, we plan to establish a random forest-based prediction model for simulating the FIFA World Cup 2018 tournament, which takes place in Russia. However, as several of the underlying covariates are based on the final squads nominated for the FIFA World Cup 2018, we need to wait until the final official

squad announcements, which the national coaches need to provide by 4th of June 2018.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors received no financial support for the research, authorship and/or publication of this article.

References

Breiman L (2001) Random forests. *Machine Learning*, **45**, 5–32.

Breiman L, Friedman JH, Olshen RA and Stone JC (1984) *Classification and Regression Trees*. Monterey, CA: Wadsworth.

Bühlmann P and Hothorn T (2007) Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, **22**, 477–505.

Dixon MJ and Coles SG (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 Clarke SR (2000) A ratings based Poisson model for World Cup soccer simulation. *Journal of the Operational Research Society*, **51**, 993–98.

Friedman JH, Hastie T and Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.

Gneiting T and Raftery A (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–76.

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

Ho TK (1998) The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 832–44.

Hoerl AE and Kennard RW (1970) Ridge regression: Bias estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hofner B, Mayr A and Schmid M (2016) gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. *Journal of Statistical Software*, **74**, 1–31.

Hornung R (2017a) Ordinal forests: Prediction and variable ranking with ordinal target variables. R package version 2.1. URL https://CRAN.R-project.org/package=ordinalForest(last accessed 6 September 2018).

———. (2017b) *Ordinal forests* (Technical Report 212). Germany: Department of Statistics LMU Munich.

Hothorn T, Hornik K and Zeileis A (2006) Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, **15**, 651–74.

Hothorn T, Lausen B, Benner A and Radespiel-Tröger M (2004) Bagging survival trees. *Statistics in Medicine*, **23**, 77–91.

Hothorn T, Buehlmann P, Kneib T, Schmid M and Hofner B (2017) mboost: Model-Based Boosting. R package version 2.8-1. URL https://CRAN.R-project.org/package=mboost(last accessed 6 September 2018).

Karlis D and Ntzoufras I (2003) Analysis of sports data by using bivariate Poisson models. *The Statistician*, **52**, 381–93.

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

Leitner C, Zeileis A and Hornik K (2010) Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008. *International Journal of Forecasting*, **26**, 471–81.

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

Mayr A, Fenske N, Hofner B, Kneib T and Schmid M (2012) Generalized additive models for location, scale and shape for high-dimensional data: A flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**, 403–27.

Quinlan JR (1986) Induction of decision trees. *Machine Learning*, **1**, 81–106.

Rigby RA and Stasinopoulos DM (2005) Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 507–54.

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

Strobl C, Boulesteix A-L, Kneib T, Augustin T and Zeileis A (2008) Conditional variable importance for random forests. *BMC Bioinformatics*, **9**, 307.

Strobl C, Boulesteix A-L, Zeileis A and Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, **8**, 25.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, **58**, 267–88.

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

Zeileis A, Hothorn T and Hornik K (2008) Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, **17**, 492–514.