Hybrid Machine Learning Forecasts for the FIFA Women's World Cup 2019

Andreas Groll [*]   Christophe Ley [†]   Gunther Schauberger [‡]

Hans Van Eetvelde [§]   Achim Zeileis [¶]

June 4, 2019

Abstract In this work, we combine two different ranking methods together with several other predictors in a joint random forest approach for the scores of soccer matches. The first ranking method is based on the bookmaker consensus, the second ranking method estimates adequate ability parameters that reflect the current strength of the teams best. The proposed combined approach is then applied to the data from the two previous FIFA Women's World Cups 2011 and 2015. Finally, based on the resulting estimates, the FIFA Women's World Cup 2019 is simulated repeatedly and winning probabilities are obtained for all teams. The model clearly favors the defending champion USA before the host France.

Keywords: FIFA Women's World Cup 2019, Soccer, Random forests, Team abilities, Sports tournaments.

1 Introduction

While the research on adequate statistical models to predict the outcome of (men's) international soccer tournaments, such as European championships (EUROs) or FIFA World Cups, has substantially advanced in recent years, to the best of our knowledge there exists no significant scientific literature on modeling women's soccer. In this work, we propose a combined ranking-based machine learning approach that is then used to forecast the FIFA Women's World Cup 2019.

A model class frequently used to model soccer results is the class of Poisson regression models. These directly model the number of goals scored by both competing teams in the single soccer matches. 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 let n denote the total number of teams in the regarded set of matches. One assumes $X_{ij} \sim Po(\lambda_{ij})$ and

---

[*]Statistics Faculty, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany, *groll@statistik.tu-dortmund.de*

[†]Faculty of Sciences, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281, 9000 Gent, Belgium, *Christophe.Ley@UGent.be*

[‡]Chair of Epidemiology, Department of Sport and Health Sciences, Technical University of Munich, *g.schauberger@tum.de*

[§]Faculty of Sciences, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281, 9000 Gent, Belgium, *hans.vaneetvelde@ugent.be*

[¶]Department of Statistics, Universität Innsbruck, Austria, *Achim.Zeileis@uibk.ac.at*

$Y_{ij} \sim Po(\mu_{ij})$, with intensity parameters λ_{ij} and μ_{ij} (reflecting the expected numbers of goals). 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 FIFA World Cup data, with Poisson intensities that depend on the FIFA ranks of both competing teams. Groll and Abedieh (2013) and Groll et al. (2015) considered a large set of potential predictors for EURO and World Cup data, respectively, and used L_1-penalized approaches to detect sparse sets of relevant covariates. Based on these, they calculated predictions for the EURO 2012 and FIFA World Cup 2014 tournaments. Their findings show that, when many covariates are regarded and/or the predictive power of the single predictors is not clear in advance, regularization can be beneficial.

These approaches can be generalized in different ways to allow for dependent scores. For example, Dixon and Coles (1997) identified a (slightly negative) correlation between the scores and introduced an additional dependence parameter. Karlis and Ntzoufras (2003) and Groll et al. (2018) model the scores of both teams by a bivariate Poisson distribution, which is able to account for (positive) dependencies between the scores. If also negative dependencies should be accounted for, copula-based models can be used (see, e.g., McHale and Scarf, 2007, McHale and Scarf, 2011 or Boshnakov et al., 2017).

Closely related to the covariate-based Poisson regression models are Poisson-based ranking methods for soccer teams. On basis of a (typically large) set of matches, ability parameters reflecting the current strength of the teams can be estimated by means of maximum likelihood. An overview of the most frequently used Poisson-based ranking methods can be found in Ley et al. (2019).

An alternative ranking approach that is solely based on bookmakers' odds was proposed by Leitner et al. (2010). They calculate winning probabilities for each team by aggregating winning odds from several online bookmakers. Based on these winning probabilities, by inverse tournament simulation team-specific *bookmaker consensus abilities* can be computed by paired comparison models, automatically stripping the effects of the tournament draw. Next, pairwise probabilities for each possible game at the corresponding tournament can be predicted and, finally, the whole tournament can be simulated.

A fundamentally different modeling approach is based on a random forest – a popular ensemble learning method for classification and regression (Breiman, 2001), which originates from the machine learning and data mining community. Firstly, a multitude of so-called decision trees (Quinlan, 1986; Breiman et al., 1984) is constructed on different training data sets, which are resampled from the original dataset. The predictions from the individual trees are then aggregated, 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. To investigate the predictive potential of random forests, Schauberger and Groll (2018) compared different types of random forests on data containing all matches of the FIFA World Cups 2002–2014 with conventional regression methods for count data, such as the Poisson models from above. The random forests provided very satisfactory results and generally outperformed the regression approaches. Groll et al. (2019) showed on the same FIFA World Cup data that the predictive performance of random forests could be further improved by combining it with the Poisson ranking methods, leading to what they call a *hybrid random forest model*.

In the present work, we carry this strategy forward and combine the random forest with both

the Poisson ranking methods as well as the bookmaker consensus abilities from Leitner et al. (2010). So in a sense, this results in a *doubly-hybrid* or *combined ranking-based* random forest. The model is fitted to all matches of the FIFA Women's World Cups 2011 and 2015 and based on the resulting estimates, the FIFA Women's World Cup 2019 is then simulated 100,000 times to determine winning probabilities for all 24 participating teams.

The rest of the manuscript is structured as follows. In Section 2 we describe the three underlying data sets. The first covers all matches of the two preceding FIFA Women's World Cups 2011 and 2015 including covariate information, the second consists of the match results of all international matches played by all national teams during certain time periods and the third contains the winning odds from several bookmakers for the single World Cups regarded in this analysis. Next, in Section 3 we briefly explain the basic idea of random forests and the two different ranking methods and, finally, how they can be combined to a hybrid random forest model. In Section 4, we fit the hybrid random forest model to the data of the two World Cups 2011 and 2015. Based on the resulting estimates, the FIFA Women's World Cup 2019 is simulated repeatedly and winning probabilities for all teams are presented. Finally, we conclude in Section 5.

2 Data

In this section, we briefly describe three 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. The third type of data contains the winning odds from different bookmakers separately for single World Cups.

2.1 Covariate data

The first type of data we describe covers all matches of the two FIFA Women's World Cups 2011 and 2015 together with several potential influence variables. Basically, we use a similar (but smaller[1]) 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., average age), 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 is supposed to 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

---

[1]It turned out that, compared to men, for women's national teams covariates were generally more difficult to get or were simply not recorded at all, as for women's soccer data archives are less detailed and sometimes incomplete. For example, while for men's national coaches, their age, the duration of their tenure and their nationality could be obtained manually from the website of the German soccer magazine *kicker*, http://kicker.de, from http://transfermarkt.de and from https://en.wikipedia.org, this was not possible for women.

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, and are denoted by *PoisAbil*. Another additional covariate, which is also introduced later, reflects the bookmaker consensus abilities (denoted by *OddsAbil*) from Leitner et al. (2010) and is related to the third type of data introduced in Section 2.3. Details on this ranking method can be found in Section 3.3.

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 2011–2015, a ratio of the GDP per capita of the respective country and the worldwide average GDP per capita is used (source: mostly `http://www.imf.org/external/pubs/ft/weo/2018/01/weodata/index.aspx`, but for England, Scotland, South and North Korea some additional research was necessary).

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:

FIFA rank. The FIFA Coca-Cola Women's World Ranking is based on an Elo-type rating system, which was originally developed by Dr. Arpad Elo to rate the playing abilities of chess players. It aims at reflecting the current strength of a soccer team relative to its competitors (source: `https://de.fifa.com/fifa-world-ranking/ranking-table/women/`).

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 German soccer magazine *kicker*, `http://kicker.de`, and from `http://transfermarkt.de`[2].

(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 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 the UEFA Champions League (CL) is counted.

Number of Major League Soccer players. As the US Major League Soccer (MLS) is supposedly the best national soccer league on the globe in women's soccer, for each national team the number of players in the MLS is 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.

In addition, we include a dummy variable indicating whether a certain match is a group- or a knockout match. The motivation for this is that soccer teams might change their playing style and be more cautious in knockout matches. In total, together with the two ability variables from the two ranking methods this adds up to 15 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 Women's World Cup 2011. We use this data excerpt to illustrate how the final data set is constructed.

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

(a) Table of results

NGA	0:1	FRA
GER	2:1	CAN
CAN	0:4	FRA
GER	1:0	NGA
⋮	⋮	⋮

(b) Table of covariates

World Cup	Team	PoisAbil	OddsAbil	Age	...
2011	France	1.69	0.02	25.86	...
2011	Germany	2.35	1.25	25.95	...
2011	Nigeria	1.39	−0.47	22.24	...
2011	Canada	1.82	−0.17	25.52	...
⋮	⋮	⋮	⋮	⋮	⋱

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 *PoisAbil* will be the difference between the Poisson

---

[2]Note that for the World Cup 2011 the size of the national teams' squads was restricted to 21 players. Hence, all of the following factors that add up players with a certain characteristic (namely all factors except for the *average age*) have been divided by the respective squad size (i.e. 21 or 23) to make them comparable across tournaments.

ranking abilities of both teams. The categorical variables *Host*, *Continent* and *Confederation*, 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. The dummy variable *groupstage* corresponds to a single column in the design matrix and is either zero or one for both rows corresponding to the same match.For illustration, the resulting final data structure for the exemplary matches from Table 1 is displayed in Table 2.

Table 2: Exemplary table illustrating the data structure.

Goals	Team	Opponent	Groupstage	PoisAbil	OddsAbil	Age	...
0	Nigeria	France	1	-0.49	-0.30	-3.62	...
1	France	Nigeria	1	0.49	0.30	3.62	...
2	Germany	Canada	1	1.42	0.53	0.43	...
1	Canada	Germany	1	-1.42	-0.53	-0.43	...
0	Canada	France	1	-0.18	0.13	-0.33	...
4	France	Canada	1	0.18	-0.13	0.33	...
1	Germany	Nigeria	1	1.73	0.96	3.71	...
1	Nigeria	Germany	1	-1.73	-0.96	-3.71	...
⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋱

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 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 the time elapsed since the game took place. For example, Table 3 shows an excerpt of the historic match data used to obtain ability estimates for the teams at the FIFA Women's World Cup 2011.

Table 3: Historical match result data used for estimating the abilities, exemplarily for the FIFA Women's World Cup 2011

Date	Home team	Away team	Score	Country	Neutral
2011-05-19	Iceland	Bulgaria	6:0	Iceland	no
2011-03-09	United States	Iceland	4:2	Portugal	yes
2011-03-09	Portugal	Finland	2:1	Portugal	no
2011-03-09	Wales	China PR	1:2	Portugal	yes
⋮	⋮	⋮	⋮	⋮	⋮

2.3 Bookmaker data

The basis for the bookmaker consensus ranking model from Leitner et al. (2010), which is explained in more detail in Section 3.3, are the winning odds of the bookmakers[3]. These are typically available already a few weeks before the tournament start. The popularity of this specific bet has substantially increased over time: while for the World Cup 2011 we could only obtain the odds from the German state betting agency ODDSET (upon request), for the World Cup 2015 we found already corresponding odds of three different bookmakers publicly available, see Table 4. For the upcoming World Cup 2019 tournament we easily obtained the winning odds from 18 different bookmakers.

Table 4: Winning odds of all 24 participating teams from several bookmakers, exemplarily for the FIFA Women's World Cup 2015

	MyTopSportsbooks	SportsInsights	BovadaSportsbook
United States	4.00	4.00	3.25
Germany	4.50	4.35	4.50
France	5.50	10.00	9.00
Japan	9.00	9.00	8.00
Brazil	9.00	8.50	8.00
Canada	15.00	15.00	11.00
Sweden	17.00	15.00	11.00
England	21.00	21.00	21.00
Norway	34.00	28.00	26.00
Australia	51.00	51.00	41.00
China PR	61.00	66.00	51.00
Spain	67.00	66.00	41.00
Netherlands	81.00	76.00	51.00
South Korea	81.00	76.00	67.00
Switzerland	101.00	86.00	67.00
New Zealand	151.00	151.00	101.00
Nigeria	201.00	201.00	151.00
Colombia	251.00	251.00	151.00
Mexico	251.00	251.00	126.00
Ecuador	501.00	501.00	251.00
Cameroon	501.00	501.00	301.00
Ivory Coast	501.00	501.00	251.00
Costa Rica	1001.00	1001.00	251.00
Thailand	5001.00	5001.00	401.00

3 A combined ranking-based random forest

In this section, we propose to use a hybrid random forest approach that combines the information from all three types of data bases introduced above. The proposed method combines a random forest for the covariate data with both the abilities estimated on the historic

---

[3]The possibility of betting on the overall cup winner before the start of the tournament is quite novel. While for men, the German state betting agency ODDSET offered the bet for the first time at the FIFA World Cup 2002, for women we could not get any odds before the World Cup 2011.

match results as used by the Poisson ranking methods and the abilities obtained from the bookmaker consensus approach. Before introducing the proposed hybrid method, we first separately present the basic ideas of the three 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). 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 already mentioned in the Introduction, random forests are the aggregation of a large number B (e.g., $B = 5000$) of trees, grown on B bootstrap samples from the original data set. Combining 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. For a short introduction to random forests and how they can specifically be used for soccer data, see Groll et al. (2019).

In R (R Core Team, 2018), two slightly different variants of regression forests are available: the classical random forest algorithm proposed by Breiman (2001) from the R-package `ranger` (Wright and Ziegler, 2017), and a modification implemented in the function `cforest` from the `party` package[4]. In Schauberger and Groll (2018) and Groll et al. (2019), the latter package turned out to be superior and will be used in the following.

3.2 Poisson 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 based on time depreciation.

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, for instance, a match played *Half period* days ago only contributes half as much as a match played today. We stress that the *Half period* refers to calendar days in a year, not match days. In the present case we use a Half period of 500 days based on

---

[4]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 if 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.

an optimization procedure to determine which Half period led to the best prediction for women's soccer matches based on the average Rank Probability Score (RPS; Gneiting and Raftery, 2007)

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 ($i, j \in \{1, ..., n\}$) in match m* (where $m \in \{1, ..., M\}$). The joint probability function of the home and away score is then given by the bivariate Poisson probability mass function,

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

where λ_C is a covariance parameter assumed to be constant over all matches and λ_{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 β_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

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

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 $r_1, \ldots, r_n$, which allow ranking the different teams, 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).

3.3 The bookmaker consensus ranking model

Prior to the tournament on 2019-06-03 we obtained long-term winning odds from 18 online bookmakers. However, before these odds can be transformed to winning probabilities, the stake has to be accounted for and the profit margin of the bookmaker (better known as the "overround") has to be removed (for further details see Henery, 1999; Forrest et al., 2005). Here, it is assumed that the quoted odds are derived from the underlying "true" odds as: *quoted odds = odds · δ + 1*, where $+1$ is the stake (which is to be paid back to the bookmakers' customers in case they win) and $\delta < 1$ is the proportion of the bets that is actually paid out by the bookmakers. The overround is the remaining proportion $1 - \delta$ and the main basis of the bookmakers' profits (see also Wikipedia, 2019 and the links therein). Assuming that each bookmaker's δ is constant across the various teams in the tournament (see Leitner et al., 2010, for all details), we obtain overrounds for all bookmakers with a median value of 24.8%.

To aggregate the overround-adjusted odds across the 18 bookmakers, we transform them to the log-odds (or logit) scale for averaging (as in Leitner et al., 2010). The bookmaker consensus is computed as the mean winning log-odds for each team across bookmakers and then transformed back to the winning probability scale.

In a second step the bookmakers' odds are employed to infer the contenders' relative abilities (or strengths). To do so, an "inverse" tournament simulation based on team-specific abilities is used. The idea is the following:

1. If team abilities are available, pairwise winning probabilities can be derived for each possible match using the classical Bradley and Terry (1952) model. This model is similar to the Elo rating (Elo, 2008), popular in sports, and computes the probability that a Team *A* beats a Team *B* by their associated abilities (or strengths):

$$\Pr(A \text{ beats } B) = \frac{ability_A}{ability_A + ability_B}.$$

2. Given these pairwise winning probabilities, the whole tournament can be easily simulated to see which team proceeds to which stage in the tournament and which team finally wins.

3. Such a tournament simulation can then be run sufficiently often (here 1,000,000 times) to obtain relative frequencies for each team winning the tournament.

Here, we use the iterative approach of Leitner et al. (2010) to find team abilities so that the resulting simulated winning probabilities (from 1,000,000 runs) closely match the bookmaker consensus probabilities. This allows to strip the effects of the tournament draw (with weaker/easier and stronger/more difficult groups), yielding a log-ability measure (on the log-odds scale) for each team.

3.4 The combined ranking-based random forest

In order to link the information provided by the covariate data, the historic match data and the bookmakers' odds, we now combine the random forest approach from Section 3.1 and the ranking methods from Section 3.2 and Section 3.3. We propose to use both ranking approaches to generate two new (highly informative) covariates that can be incorporated into the random forest model. For that purpose, for each World Cup we estimate the team abilities $r_i, i = 1, \ldots, 24$, of all 24 participating teams shortly before the start of the respective tournament. For example, to obtain ability estimates for the 24 teams that participated in the World Cup 2011, the historic match data for a certain time period preceding the World Cup 2011 (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 for all teams participating in a certain World Cup. Actually, this variable appears to be somewhat similar to the Elo ranking, but turns out to be much more informative, see Section 4.1. Moreover, based on the winning odds provided by the bookmakers, we also calculate the additional abilities $s_i, i = 1, \ldots, 24$, of all 24 participating teams shortly before the start of the respective tournament corresponding to the bookmaker consensus model. The corresponding estimates $\hat{s}_i$ again serve as another additional covariate. Also this variable turns out to be more important than the Elo ranking, see again Section 4.1.

The newly generated variables 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 from 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. In order to be able to use these point estimates for the prediction of the outcome of single matches or a whole tournament, we follow Groll et al. (2019) and treat the predicted expected value for the number of goals as an estimate for the intensity 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.

4 Modeling the FIFA Women's World Cup 2019

We now fit the proposed combined ranking-based random forest model to the data of the World Cups 2011 and 2015. Next, we calculate the Poisson ranking ability parameters based on historic match data over the 8 years preceding the World Cup 2019 as well as the bookmaker consensus abilities based on the winning odds from 18 different bookmakers. Based on these ability estimates, the fitted random forest will be used to simulate the FIFA Women's World Cup 2019 tournament 100,000 times to determine winning probabilities for all 24 participating teams.

4.1 Fitting the combined ranking-based random forest to the data of the World Cups 2011 and 2015

We fit the hybrid random forest approach with $B = 5000$ single trees to the complete data set covering the two World Cups 2011 and 2015. The best way to understand the role of the single predictor variables in a random forest 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 2011 could be assigned to the average age of the US team in 2015. 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. Figure 1 shows bar plots of the variable importance values for all variables in the hybrid random forest applied to the data of the World Cups 2011 and 2015. Interestingly, the Poisson abilities are by far the most important predictor in the random forest and carry clearly more information than all other predictors. But also the abilities from the bookmaker consensus approach seem to be slightly more informative compared to the *FIFA rank*. Even though the *FIFA rank*, the teams' average *age* or the number of *CL players* also contain some information concerning the current strengths of the teams, it is definitely worth the effort to estimate such abilities in separate models. For a more detailed comparison of the team abilities and the *FIFA rank*, see Table 5.
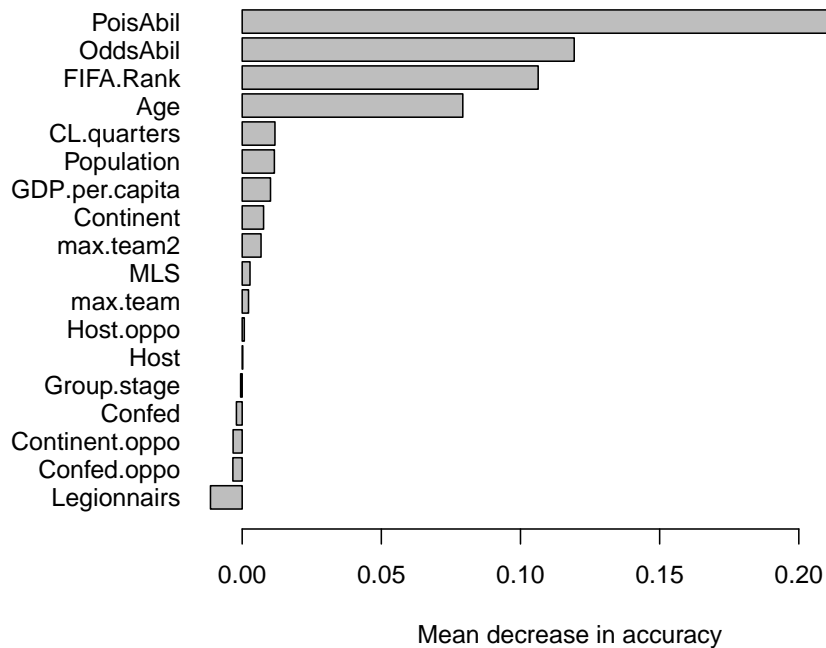
Figure 1: Bar plot displaying the variable importance in the hybrid random forest applied to FIFA World Cup data.

4.2 Probabilities for FIFA World Cup 2019 Winner

In this section, the hybrid random forest is applied to (new) data for the World Cup 2019 in France (in advance of the tournament) to predict winning probabilities for all teams and to predict the tournament course.

The Poisson abilities were estimated by a bivariate Poisson model with a half period of 500 days. All matches of the 167 national teams played since 2011-06-01 up to 2019-06-01 are used for the estimation, what results in a total of 3418 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). The bookmaker consensus abilities are based on the average odds of 18 bookmakers.

For each match in the World Cup 2019, 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 36 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

---

[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 table 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.

Table 5: Ranking of the participants of the FIFA World Cup 2019 according to estimated bookmaker consensus abilities (left; in logs), Poisson abilities (center; in logs) and FIFA ranking (right).

	bookmaker consensus abilities	Poisson abilities	FIFA ranking
1	France	United States	United States
2	United States	Germany	Germany
3	Germany	France	England
4	England	England	France
5	Netherlands	Netherlands	Canada
6	Japan	Spain	Australia
7	Australia	Sweden	Japan
8	Spain	Canada	Netherlands
9	Brazil	Australia	Sweden
10	Sweden	Norway	Brazil
11	Canada	Japan	Norway
12	Norway	Brazil	Spain
13	Italy	Italy	South Korea
14	China PR	China PR	Italy
15	South Korea	South Korea	China PR
16	New Zealand	Scotland	New Zealand
17	Scotland	New Zealand	Scotland
18	Argentina	Chile	Thailand
19	Chile	Nigeria	Argentina
20	Nigeria	South Africa	Nigeria
21	South Africa	Jamaica	Chile
22	Cameroon	Thailand	Cameroon
23	Thailand	Cameroon	South Africa
24	Jamaica	Argentina	Jamaica

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 24 participating teams probabilities to reach the single knockout stages and, finally, to win the tournament are obtained. These are summarized in Table 6 together with the (average) winning probabilities based on 18 different bookmakers for comparison.

We can see that, according to our hybrid random forest model, USA is the favored team with a predicted winning probability of 28.1% followed by France, England and Germany. Overall, this result seems in line with the probabilities from the bookmakers, as we can see in the last column. However, while the bookmakers slightly favor France, the random forest model predicts a clear advantage for USA. Beside the probabilities of becoming world champion, Table 6 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 USA and France have similar chances to at least reach the round-of-sixteen (98.4% and 95.9%, respectively), while the probabilities to at least reach the quarter finals differ significantly. While USA has a probability of 75.5% to reach at least the quarter finals and of 53.4 to reach at least the semi finals, France only achieves respective probabilities of 66.8% and 40.7. Obviously, in contrast to USA, France has a rather high chance to meet a strong opponent in the

Table 6: Estimated probabilities (in %) for reaching the different stages in the FIFA World Cup 2019 for all 24 teams based on 100,000 simulation runs of the FIFA World Cup together with (average) winning probabilities based on the odds of 18 bookmakers.

			Round of 16	Quarter finals	Semi finals	Final	World Champion	Bookmakers
1.		USA	98.4	75.5	53.4	39.6	28.1	17.7
2.		FRA	95.9	66.8	40.7	25.4	14.3	18.2
3.		ENG	96.1	69.8	45.3	23.8	13.3	11.0
4.		GER	95.4	66.3	36.9	22.9	12.9	12.4
5.		NED	92.7	47.1	25.9	12.0	5.1	6.0
6.		SWE	91.2	50.7	24.8	12.1	4.4	3.3
7.		BRA	88.7	51.2	25.5	10.5	3.9	3.8
8.		AUS	89.0	50.0	24.2	10.1	3.8	4.7
9.		ESP	81.5	43.8	20.1	9.4	3.6	3.6
10.		JPN	82.5	43.3	21.1	8.0	2.7	5.3
11.		CAN	85.4	33.2	14.7	5.7	2.0	3.1
12.		ITA	81.7	38.8	16.7	5.8	1.9	1.6
13.		NOR	75.0	33.7	13.1	4.6	1.5	2.2
14.		CHN	72.5	29.0	9.5	3.1	0.8	1.5
15.		SCO	66.6	24.5	8.3	2.4	0.7	0.9
16.		KOR	64.8	23.6	7.3	2.0	0.5	1.2
17.		NZL	65.4	16.1	4.9	1.2	0.3	1.1
18.		THA	36.9	7.9	1.8	0.3	0.1	0.2
19.		NGA	30.1	6.5	1.3	0.2	0.0	0.4
20.		ARG	22.6	5.2	1.0	0.2	0.0	0.7
21.		CHI	26.2	5.4	1.1	0.2	0.0	0.7
22.		CMR	26.6	5.1	1.1	0.2	0.0	0.2
23.		RSA	19.6	3.9	0.8	0.1	0.0	0.3
24.		JAM	15.1	2.7	0.4	0.1	0.0	0.1

round-of-sixteen and the quarter finals.

5 Concluding remarks

In this work, we proposed a hybrid modeling approach for the scores of international soccer matches which combines random forests with two different ranking methods, a Poisson ranking method and abilities based on bookmakers' odds. While the random forest is principally based on the competing teams' covariate information, the latter two components provide ability parameters, which serve as adequate estimates of the current team strengths a well as of the information contained in the bookmakers' odds. In order to combine the methods, the Poisson 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. Similarly, the bookmaker consensus abilities are obtained by inverse tournament simulation based on the aggregated winning odds from several online bookmakers. These two ability estimates can be added as additional covariates to the set of covariates used in the random forest procedure.

Additionally, based on the estimates of the combined ranking-based random forest on the training data, we repeatedly simulated the FIFA Women's World Cup 2019 100,000 times. According to these simulations, the defending champion USA is the top favorite for winning the title, followed by France, England and Germany. Furthermore, survival probabilities for all teams and at all tournament stages are provided.

Acknowledgment

References

Boshnakov, G., T. Kharrat, and I. G. McHale (2017). A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting 33*(2), 458–466.

Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika 39*(3/4), 324–345.

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*(2), 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.

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

Forrest, D., J. Goddard, and R. Simmons (2005). Odds-setters as forecasters: The case of English football. *International Journal of Forecasting 21*(3), 551–564.

Gneiting, T. and A. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–376.

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

Groll, A., 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., C. Ley, G. Schauberger, and H. Van Eetvelde (2019). A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*. to appear.

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*(2), 97–115.

Henery, R. J. (1999). Measures of over-round in performance index betting. *Journal of the Royal Statistical Society D 48*(3), 435–439.

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

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

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*(1), 55–77.

McHale, I. and P. Scarf (2007). Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica 61*(4), 432–445.

McHale, I. G. and P. A. Scarf (2011). Modelling the dependence of goals scored by opposing teams in international soccer matches. *Statistical Modelling 41*(3), 219–236.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning 1*(1), 81–106.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schauberger, G. and A. Groll (2018). Predicting matches in international football tournaments with random forests. *Statistical Modelling 18*(5-6), 460–482.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. *BMC Bioinformatics 9*(1), 307.

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*(1), 25.

Wikipedia (2019). Odds — wikipedia, the free encyclopedia. Online, accessed 2019-05-24.

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), 1–17.