LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK

Andreas Groll & Thomas Kneib & Andreas Mayr & Gunther Schauberger

Who's the Favourite? – A Bivariate Poisson Model for the UEFA European Football Championship 2016

Technical Report Number 195, 2016
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Who's the Favourite? – A Bivariate Poisson Model for the UEFA European Football Championship 2016

A. Groll * T. Kneib † A. Mayr ‡ G. Schauberger §

July 22, 2016

Abstract Many approaches that analyze and predict the results of soccer matches are based on two independent pairwise Poisson distributions. The dependence between the scores of two competing teams is simply displayed by the inclusion of the covariate information of both teams. One objective of this article is to analyze if this type of modeling is appropriate or if an additional explicit modeling of the dependence structure for the joint score of a soccer match needs to be taken into account. Therefore, a specific bivariate Poisson model for the two numbers of goals scored by national teams competing in UEFA European football championship matches is fitted to all matches from the three previous European championships, including covariate information of both competing teams. A boosting approach is then used to select the relevant covariates. Based on the estimates, the current tournament is simulated 1,000,000 times to obtain winning probabilities for all participating national teams.

Keywords Football, EURO 2016, Bivariate Poisson Model, Boosting, Variable selection.

1 Introduction

Many approaches that analyze and predict the results of soccer matches are based on two independent Poisson distributions. Both numbers of goals scored in single soccer matches are modeled seperately, assuming that each score follows its own Poisson distribution, see, e.g., Lee (1997) or Dyte and Clarke (2000). For example, Dyte and Clarke (2000) predict the distribution of scores in international soccer matches, treating each team's goals as conditionally independent Poisson variables depending on two influence variables, the FIFA ranking of each team and the match venue. Poisson regression is used to estimate parameters for the model and based on these parameters the matches played during the 1998 FIFA World Cup were simulated.

However, it is well-known that the scores of two competing teams in a soccer match are correlated. One of the first works investigating the topic of dependency between scores of competing soccer teams is the fundamental article of Dixon and Coles (1997). There it has been shown that the joint distribution of the scores of both teams cannot be well represented by the product of two independent marginal Poisson distributions of the home and away teams. They suggest to use an additional term to adjust for certain under- and overrepresented match results. Along these lines, Rue and Salvesen (2000) propose a similarly adjusted Poisson model with some additional modifications. After all, the findings in Dixon and Coles (1997) are based on the marginal distributions and only hold for models where the predictors of both scores are uncorrelated. However, the model proposed by Dixon and Coles (1997) includes team-specific attack and defense ability parameters and then uses independent Poisson distributions for the numbers of goals scored. Therefore, the linear predictor for the number of goals of a specific team depends both on parameters of the team itself and its competitor. Groll et al. (2015) pointed out that when fitting exactly the same model to FIFA World Cup data the estimates of the attack and defense abilities of the teams are negatively correlated. Therefore, although independent Poisson distributions are used for the scores in one match, the linear predictors and, accordingly, the predicted outcomes are (negatively) correlated.

*Department of Statistics, Ludwig-Maximilians-University Munich, Akademiestr. 1, 80799 Munich, Germany, *andreas.groll@stat.uni-muenchen.de*

†Department of Statistics and Econometrics, Georg-August-University Goettingen, Humboldtallee 3, 37073 Goettingen, Germany, *tkneib@uni-goettingen.de*

‡IMBE, Friedrich-Alexander-University Erlangen-Nuernberg, Waldstrasse 6, 91054 Erlangen, Germany, *andreas.mayr@fau.de*

§Department of Statistics, Ludwig-Maximilians-University Munich, Akademiestr. 1, 80799 Munich, Germany, *gunther@stat.uni-muenchen.de*

These findings already indicate that up to a certain amount the dependence between the scores of two competing teams can simply be displayed by the inclusion of the covariate information of both teams. For example, Groll and Abedieh (2013) use a pairwise (independent) Poisson model for the number of goals scored by national teams in the single matches of the UEFA European football championship (EURO), but incorporate several potential influence variables of *both* competing teams in the linear predictors of the single Poisson distributions together with additional team-specific random effects. Furthermore, in order to additionally account for the matched-pair design, they include a second match-specific random intercept, following Carlin et al. (2005), which is assumed to be independent from the team-specific random intercept. However, it turns out that this additional random intercept is very small ($< 1 \cdot 10^{-5}$) and, hence, can be ignored. This provides further evidence that if highly informative covariates of both competing teams are included into the linear predictor structure of both independent Poisson distributions, this might already appropriately model the dependence structure of the match scores.

These results are further confirmed in Groll et al. (2015). Following Groll and Abedieh (2013), an L_1-regularized independent Poisson model is used on FIFA World Cup data. There, the linear predictors of the single independent Poisson components include, in addition to team-specific attack and defense abilities, the differences of several covariates of both competing teams. In an extensive goodness-of-fit analysis it is investigated if the obtained dependence structure between the linear predictors of the two scores of a match represents the actual correlations in an appropriate manner. For this purpose the correlations between the real outcomes and the model predictions are compared and it turned out that the correlations within the linear predictors for both teams competing in a match fully accounted for the correlation between the scores of those teams and that there was no need for further adjustment. So one major objective of this article is to analyze if this type of modeling is appropriate or if an additional explicit modeling of the dependence structure for the joint score of a soccer match needs to be taken into account.

One possibility to explicitly model (positive) dependence within the Poisson framework is the bivariate Poisson distribution. One of the first works dealing with this distribution in the context of soccer data is Maher (1982). An extensive study for the use of the bivariate Poisson distribution for the modeling of soccer data is found in Karlis and Ntzoufras (2003). There, the three parameters λ_1, λ_2 and λ_3 of the bivariate Poisson distribution are modeled by linear predictors depending on team-specific attack and defense abilities as well as team-specific home effect parameters. In particular, it is illustrated how also the third parameter λ_3, which represents the covariance between the two scores, can be explicitly structured in terms of covariate effects (here: simply team-specific home effects). We adopt this approach in the present work and extend the linear predictors of the three parameters λ_1, λ_2 and λ_3 of the bivariate Poisson distribution to include several covariate effects. We set up a specific bivariate Poisson model for the two numbers of goals scored by national teams competing in EURO tournaments including covariate information of both competing teams.

Note that in addition to the bivariate Poisson, also alternative approaches to handle the correlation of soccer matches have been proposed in the literature. For example, McHale and Scarf (2006, 2011) model the dependence by using bivariate discrete distributions and by specifying a suitable family of dependence copulas.

A second objective of this work is to provide predictions of the current EURO 2016. Therefore, the proposed specific bivariate Poisson model is fitted to all matches from the three previous EUROs 2004 - 2012, including covariate information of both competing teams. A suitable boosting approach is then used to select a small set of relevant covariates. Based on the obtained estimates, the current EURO 2016 tournament is simulated 1,000,000 times to obtain winning probabilities for all participating national teams.

The rest of the article is structured as follows. In Section 2 we introduce the bivariate Poisson model for soccer data. The boosting methodology for fitting the bivariate Poisson model for the number of goals is introduced in Section 2.4. Next, we present a list of several possible influence variables in Section 3.1 that will be considered in our regression analysis. Based on the boosting approach a selection of these covariates is determined yielding a sparse model, which is then used in Section 4 for the prediction of the EURO 2016.

2 A Bivariate Poisson Model for Soccer Data

In the present section, we set up a specific bivariate Poisson model for the two numbers of goals scored by national teams competing in EURO tournaments including covariate information of both competing teams.

2.1 The Bivariate Poisson Distribution

In the following, we consider random variables $X_k, k = 1, 2, 3$, which follow independent Poisson distributions with parameters $\lambda_k > 0$. Then the random variables $Y_1 = X_1 + X_3$ and $Y_2 = X_2 + X_3$ follow a joint bivariate Poisson distribution, with a joint probability function

$$\begin{aligned} P_{Y_1,Y_2}(y_1, y_2) &= P(Y_1 = y_1, Y_2 = y_2) \\ &= \exp(-(\lambda_1 + \lambda_2 + \lambda_3)) \frac{\lambda_1^{y_1}}{y_1!} \frac{\lambda_2^{y_2}}{y_2!} \sum_{k=0}^{\min(y_1,y_2)} \left(\begin{array}{c} y_1 \\ k \end{array} \right) \left(\begin{array}{c} y_2 \\ k \end{array} \right) k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2} \right)^k . \end{aligned} \tag{1}$$

The bivariate Poisson distribution allows for dependence between the two random variables Y_1 and Y_2. Marginally each random variable follows a univariate Poisson distribution with $E[Y_1] = \lambda_1 + \lambda_3$ and $E[Y_2] = \lambda_2 + \lambda_3$. Moreover, the dependence of Y_1 and Y_2 is expressed by $cov(Y_1, Y_2) = \lambda_3$. If $\lambda_3 = 0$ holds, the two variables are independent and the bivariate Poisson distribution reduces to the product of two independent Poisson distributions. The notation and usage of the bivariate Poisson distribution for modeling soccer data has been described in detail in Karlis and Ntzoufras (2003).

2.2 Incorporation of Covariate Information

In general, each of the three parameters $\lambda_k, k = 1, 2, 3$, in the joint probability function (1) of the bivariate Poisson distribution can be modeled in terms of covariates by specifying a suitable response function, similar to classical generalized linear models (GLMs). Hence, one could use, for example,

$$\lambda_k = \exp(\boldsymbol{\eta}_k) ,$$

with a linear predictor $\boldsymbol{\eta}_k = \beta_{0k} + \mathbf{x}_k^T \boldsymbol{\beta}_k$ and response function $h(\cdot) = \exp(\cdot)$ in order to guarantee positive Poisson parameters λ_k. The vectors $\mathbf{x}_k = (x_{1k}, \ldots, x_{pk})^T$ collect all covariate information of predictor k.

2.3 Re-parametrization of the Bivariate Poisson Distribution

In the context of soccer data a natural way to model the three parameters $\lambda_k, k = 1, 2, 3$, would be to include the covariate information of the competing teams 1 and 2 in λ_1 and λ_2, respectively, and some extra information reflecting the match conditions of the corresponding match in λ_3. However, the covariate effects $\boldsymbol{\beta}_k, k = 1, 2$, usually should be the same for both competing teams. Then, one obtains the model representation

$$\lambda_1 = \exp(\beta_0 + \mathbf{x}_1^T \boldsymbol{\beta}) , \quad \lambda_2 = \exp(\beta_0 + \mathbf{x}_2^T \boldsymbol{\beta}) , \tag{2}$$

with $\mathbf{x}_1$ and $\mathbf{x}_2$ denoting the covariates of team 1 and team 2. In contrast, the covariance parameter λ_3 could generally depend on different covariates and effects, i.e.

$$\lambda_3 = \exp(\alpha_0 + \mathbf{z}^T \boldsymbol{\alpha}) , \tag{3}$$

where $\mathbf{z}$ could contain parts of the covariates $\mathbf{x}_1$ and $\mathbf{x}_2$, or their differences or completely new covariates. If instead in the linear predictors in (2) the differences of the teams' covariates are used, one obtains

$$\lambda_1 = \exp(\beta_0 + (\mathbf{x}_1 - \mathbf{x}_2)^T \boldsymbol{\beta}) , \quad \lambda_2 = \exp(\beta_0 + (\mathbf{x}_2 - \mathbf{x}_1)^T \boldsymbol{\beta}) ,$$

or, with $\tilde{\mathbf{x}} = \mathbf{x}_1 - \mathbf{x}_2$, the simpler model

$$\lambda_k = \exp(\beta_0 \pm \tilde{\mathbf{x}}^T \boldsymbol{\beta}) , \quad k = 1, 2 .$$

This allows to re-parametrize the bivariate Poisson probability function from (1) in the following way:

$$\begin{aligned} P_{Y_1,Y_2}(y_1, y_2) &= P(Y_1 = y_1, Y_2 = y_2) \\ &= \exp(-(\gamma_1(\gamma_2 + \gamma_2^{-1}) + \lambda_3)) \frac{(\gamma_1\gamma_2)^{y_1}}{y_1!} \frac{(\frac{\gamma_1}{\gamma_2})^{y_2}}{y_2!} \sum_{k=0}^{\min(y_1,y_2)} \left(\begin{array}{c} y_1 \\ k \end{array} \right) \left(\begin{array}{c} y_2 \\ k \end{array} \right) k! \left(\frac{\lambda_3}{\gamma_1^2} \right)^k , \end{aligned} \tag{4}$$

with $\lambda_1 = \gamma_1\gamma_2$, $\lambda_2 = \frac{\gamma_1}{\gamma_2}$. The new parameters γ_1, γ_2 are then given as functions of the following linear predictors:

$$\begin{aligned} \gamma_1 &= \exp(\beta_0)\,, \\ \gamma_2 &= \exp(\tilde{\mathbf{x}}^T\boldsymbol{\beta})\,, \end{aligned}$$

with $\tilde{\mathbf{x}} = \mathbf{x}_1 - \mathbf{x}_2$ denoting the difference of both teams' covariates.

As before, we set $\lambda_3 = \exp(\alpha_0 + \mathbf{z}^T\boldsymbol{\alpha})$. In the current analysis, we base the linear predictor of λ_3 in general on the same covariate differences. However, as we generally don't want to prefer any specific direction of these differences, we use their absolute value and set $\lambda_3 = \exp(\alpha_0 + |\tilde{\mathbf{x}}|^T\boldsymbol{\alpha})$, where $|\tilde{\mathbf{x}}| = (|x_{11} - x_{21}|, \ldots, |x_{1p} - x_{2p}|)^T$.

2.4 Estimation

We apply a statistical boosting algorithm to estimate the linear predictors for γ_1, γ_2 and λ_3. The concept of boosting emerged from the field of machine learning (Freund and Schapire, 1996) and was later adapted to estimate predictors for statistical models (Friedman et al., 2000; Friedman, 2001). Main advantages of statistical boosting algorithms are their flexibility for high-dimensional data and their ability to incorporate variable selection in the fitting process (Mayr et al., 2014a). Furthermore, due to the modular nature of the algorithm, they are relatively easy to extend to new regression settings (Mayr et al., 2014b). The aim of the algorithm is to find estimates for the predictors

$$\exp(\widehat{\eta}_{\gamma_1}) = \exp(\hat{\beta}_0) = \hat{\gamma}_1\,, \tag{5}$$

$$\exp(\widehat{\eta}_{\gamma_1}) = \exp(\tilde{\mathbf{x}}^T\hat{\boldsymbol{\beta}}) = \hat{\gamma}_2\,, \tag{6}$$

$$\exp(\widehat{\eta}_{\lambda_3}) = \exp(\hat{\alpha}_0 + |\tilde{\mathbf{x}}|^T\hat{\boldsymbol{\alpha}}) = \hat{\lambda}_3 \tag{7}$$

that optimize the multivariate likelihood of $L(Y_1, Y_2, \gamma_1, \gamma_2, \lambda_3) := P_{Y_1,Y_2}(y_1, y_2)$ with $P_{Y_1,Y_2}(y_1, y_2)$ from Equation (4), leading to the optimization problem

$$(\widehat{\eta}_{\gamma_1}, \widehat{\eta}_{\gamma_2}, \widehat{\eta}_{\lambda_3}) = \underset{(\widehat{\eta}_{\gamma_1}, \widehat{\eta}_{\gamma_2}, \widehat{\eta}_{\lambda_3})}{\operatorname{argmax}} \mathrm{E}\left[L(Y_1, Y_2, \exp(\widehat{\eta}_{\gamma_1}), \exp(\widehat{\eta}_{\gamma_2}), \exp(\widehat{\eta}_{\lambda_3}))\right]\,.$$

The algorithm cycles through the different predictors and carries out one boosting iteration for each. In every boosting iteration, only one component of the corresponding predictor is selected to be updated, leading to automated variable selection for the covariates. For more on boosting for multiple dimensions see Schmid et al. (2010) and Mayr et al. (2012). Let the data now be given by $(y_{1i}, y_{2i}, \tilde{\mathbf{x}}_i^T), i = 1, \ldots, n$. Then, the following cyclic boosting algorithm is applied:

(1) **Initialize**

Initialize the additive predictors with starting values, e.g. $\widehat{\eta}_{\gamma_1}^{[0]} := \log(\bar{y}_1); \widehat{\eta}_{\gamma_2}^{[0]} := 0; \widehat{\eta}_{\lambda_3}^{[0]} := \log(0.0001)$. Set iteration counter to $m := 1$.

(2) **Boosting for γ_1**

Increase iteration counter: $m := m + 1$

If $m > m_{\text{stop}\gamma_1}$ set $\widehat{\eta}_{\gamma_1}^{[m]} := \widehat{\eta}_{\gamma_1}^{[m-1]}$ and skip step (2).

Compute $\mathbf{u}^{[m]} = \left(\frac{\partial}{\partial\eta_{\gamma_1}} L(y_{1i}, y_{2i}, \exp(\widehat{\eta}_{\gamma_1}^{[m-1]}), \exp(\widehat{\eta}_{\gamma_2}^{[m-1]}), \exp(\widehat{\eta}_{\lambda_3}^{[m-1]}))\right)_{i=1,\ldots,n}$

Estimate ${\hat{\beta}_0}^{[m]}$ for $\mathbf{u}^{[m]}$ by ${\hat{\beta}_0}^{[m]} = \bar{u}^{[m]}$.

Update $\widehat{\eta}_{\gamma_1}^{[m]}$ with $\hat{\beta}_0^{[m]} := \hat{\beta}_0^{[m-1]} + \nu \cdot \hat{\beta}_0^{[m]}$, where ν is a small step length (e.g., $\nu = 0.1$)

(3) **Boosting for γ_2**

If $m > m_{\text{stop}\gamma_2}$ set $\widehat{\eta}_{\gamma_2}^{[m]} := \widehat{\eta}_{\gamma_2}^{[m-1]}$ and skip step (3).

Compute $\mathbf{u}^{[m]} = \left(\frac{\partial}{\partial \eta_{\gamma_2}} L(y_{1i}, y_{2i}, \exp(\widehat{\eta}_{\gamma_1}^{[m]}), \exp(\widehat{\eta}_{\gamma_2}^{[m-1]}), \exp(\widehat{\eta}_{\lambda_3}^{[m-1]}))\right)_{i=1,\ldots,n}$

Fit all components of $\tilde{\mathbf{x}}$ separately to $\mathbf{u}^{[m]}$, leading to $\hat{\beta}_1^{[m]}, \ldots, \hat{\beta}_p^{[m]}$.

Select component j^* that best fits $\mathbf{u}^{[m]}$ with

$$j^* = \underset{1 \leq j \leq p}{\operatorname{argmin}} \sum_{i=1}^{n} (u_i^{[m]} - \hat{\beta}_j^{[m]} x_j)^2$$

Update $\widehat{\eta}_{\gamma_2}^{[m]}$ with $\hat{\beta}_{j^*}^{[m]} = \hat{\beta}_{j^*}^{[m-1]} + \nu \cdot \hat{\beta}_{j^*}^{[m]}$, keeping all other components fixed.

(4) **Boosting for λ_3**

If $m > m_{\text{stop}\lambda_3}$ set $\widehat{\eta}_{\lambda_3}^{[m]} := \widehat{\eta}_{\lambda_3}^{[m-1]}$ and skip step (4).

Compute $\mathbf{u}^{[m]} = \left(\frac{\partial}{\partial \eta_{\gamma_2}} L(y_{1i}, y_{2i}, \exp(\widehat{\eta}_{\gamma_1}^{[m]}), \exp(\widehat{\eta}_{\gamma_2}^{[m]}), \exp(\widehat{\eta}_{\lambda_3}^{[m-1]}))\right)_{i=1,\ldots,n}$

Fit all components of $|\tilde{\mathbf{x}}|$ separately to $\mathbf{u}^{[m]}$, leading to $\hat{\alpha}_0^{[m]}, \ldots, \hat{\alpha}_p^{[m]}$.

Select component j^* that best fits $\mathbf{u}^{[m]}$ with

$$j^* = \underset{0 \leq j \leq p}{\operatorname{argmin}} \sum_{i=1}^{n} (u_i^{[m]} - \hat{\alpha}_j^{[m]} z_j)^2.$$

Update $\widehat{\eta}_{\lambda_3}^{[m]}$ with $\hat{\alpha}_{j^*}^{[m]} = \hat{\alpha}_{j^*}^{[m-1]} + \nu \cdot \hat{\alpha}_{j^*}^{[m]}$, keeping all other components fixed.

Iterate steps (2) to (4) until $m \geq \max(m_{\text{stop}\gamma_1}, m_{\text{stop}\gamma_2}, m_{\text{stop}\lambda_3})$

Note that the presented algorithm reflects the structure for our re-parametrization of the bivariate Poisson distribution, but could also be easily adapted to estimate $\widehat{\eta}_{\lambda_k}$ corresponding to the original parameters $\lambda_k, k = 1, 2, 3$. Furthermore, we focused on linear predictors in our approach, however, the algorithm's structure stays the same if non-linear base-learners are applied to estimate additive predictors.

The main tuning parameters of the algorithm are the stopping iterations for the different predictors. They display the typical trade-off between small models with small variance and larger models with higher risk of overfitting. The best combination of stopping iterations $(m_{\text{stop}\gamma_1}, m_{\text{stop}\gamma_2}, m_{\text{stop}\lambda_3})$ is typically chosen via cross-validation or resampling procedures or by optimizing the underlying likelihood on separate test data. The specification of the step length ν is of minor importance as long as it is chosen small enough, it mainly affects the convergence speed (Schmid and Hothorn, 2008). The algorithm is implemented with the R add-on package `gamboostLSS` (Mayr et al., 2012; Hofner et al., 2016).

3 Application

In the following, the proposed model is applied to data from the previous EUROs 2004-2012 and is then used to predict the UEFA European championship 2016 in France.

3.1 Data

In this section a description of the covariates is given that are used (in the form of differences) in the bivariate Poisson regression model introduced in the previous sections. As most of these variables have already been used in Groll and Abedieh (2013) a more detailed description is found there. 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 at the start of an European championship has an influence on the team's success in the tournament, and thus on the goals the team will score. Besides these sportive variables, also economic factors, such as a country's GDP and population size, are taken into account. Furthermore, variables are incorporated that describe the structure of a team's squad. Note that several of these variables exhibit a substantial amount of correlation. The corresponding correlation matrix for all considered (differences of) covariates is presented in Table 5 in Appendix A.

Economic Factors:

- *GDP*[1] *per capita.* The GDP per capita represents the economic power and welfare of a nation. Hence, countries with great prosperity might tend to focus more on sports training and promotion programs than poorer countries. The GDP per capita (in US Dollar) is publicly available on the website of The World Bank (see `http://data.worldbank.org/indicator/NY.GDP.PCAP.CD`).

- *Population*[2]. In general, larger countries have a deeper pool of talented soccer players from which a national coach can recruit the national team squad. Hence, the population size might have an influence on the playing ability of the corresponding national team. However, as this potential effect might not hold in a linear relationship for arbitrarily large numbers of populations and instead might diminish (compare Bernard and Busse, 2004), the logarithm of the quantity is used.

Sportive Factors:

- *Home advantage.* There exist several studies that have analyzed the existence of a home advantage in soccer (see for example Pollard and Pollard, 2005; Pollard, 2008; Brown et al., 2002, for FIFA World Cups or Clarke and Norman, 1995, for the English Premier league). Hence, there might also exist a home effect in European championships. For this reason a dummy variable is used, indicating if a national team belongs to the organizing countries.

- *ODDSET odds.* The analyses in Groll and Abedieh (2013) and Groll and Abedieh (2014) indicate that bookmakers' odds play an important role in the modeling of international soccer tournaments such as the EURO as they contain a lot of information with respect to the success of soccer teams. They include the bookmakers' expertise and cover big parts of the team specific information and market appreciation with respect to which teams are amongst the tournament's favorites. For the EUROs from 2004 to 2012 the 16 odds of all possible tournament winners before the start of the corresponding tournament have been obtained from the German state betting agency ODDSET.

- *Market value.* The market value recently has gained increasing attention and importance in the context of predicting the success of soccer teams (see, for example, Gerhards and Wagner, 2008, 2010; Gerhards et al., 2012, 2014). Estimates of the teams' average market values can be found on the webpage `http://www.transfermarkt.de`[3]. For each national team participating in a EURO these market value estimates (in Euro) have been collected right before the start of the tournament.

- *FIFA ranking.* The FIFA ranking provides a ranking system for all national teams measuring the performance of the teams over the last four years. The exact formula for the calculation of the underlying FIFA points and all rankings since implementation of the FIFA ranking system can be found at the official FIFA website (`http://de.fifa.com/worldranking/index.html`). Since the calculation formula of the FIFA points changed after the World Cup 2006, the rankings according to FIFA points are used instead of the points[4].

- *UEFA points.* The associations' club coefficients rankings are based on the results of each association's clubs in the five previous UEFA CL and Europa League (previously UEFA Cup) seasons. The exact formula for the calculation of the underlying UEFA points and all rankings since implementation of the UEFA ranking system can be found at the official UEFA website (`http://www.uefa.com/memberassociations/uefarankings/country/index.html`). The rankings determine the number of places allocated to an association (country) in the forthcoming UEFA club competitions. Thus, the UEFA points represent the strength and success of a national league in comparison to other European national leagues. Besides, the more teams of a national league participate in the UEFA CL and the UEFA Europa League, the more experience the players from that

[1]The GDP per capita is the gross domestic product divided by midyear population. The GDP is the sum of gross values added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products.

[2]In order to collect data for all participating countries at the EURO 2004, 2008 and 2012, different sources had to be used. Amongst the most useful ones are `http://www.wko.at`, `http://www.statista.com/` and `http://epp.eurostat.ec.europa.eu`. For some years the populations of Russia and Ukraine had to be searched individually.

[3]Unfortunately, the archive of the webpage was established not until 4th October 2004, so the average market values of the national teams that we used for the EURO 2004 can only be seen as a rough approximation, as market values certainly changed after the EURO 2004.

[4]The FIFA ranking was introduced in August 1993.

national league are able to earn on an international level. As usually a relationship between the level of a national league and the level of the national team of that country is supposed, the UEFA points could also affect the performance of the corresponding national team.

Factors describing the team's structure:

- *(Second) maximum number of teammates*[5]. If many players from one club play together in a national team, this could lead to an improved performance of the team as the teammates know each other better. Therefore, both the maximum and the second maximum number of teammates from the same club are counted and included as covariates.
- *Average age.* The average age of all 23 players is collected from the website `http://www.transfermarkt.de` to include possible differences between rather old and rather young teams.
- *Number of Champions League (Europa League) players*[5]. The European club leagues are assessed to be the best leagues in the world. Therefore, the competitions between the best European teams, namely the UEFA CL and Europa League, can be seen as the most prestigious and valuable competitions on club level. As a measurement of the success of the players on club level, the number of players in the semi finals (taking place only weeks before the respective EURO) of these competitions are counted.
- *Number of players abroad*[5]. The national teams strongly differ in the numbers of players playing in the league of the respective country and players from leagues of other countries. For each team, the number of players playing in clubs abroad (in the season previous to the respective EURO) is counted.

Factors describing the team's coach.

Also covariates of the coach of the national team may have an influence on the performance of the team. Therefore, the *age* of the coach is observed together with a dummy variable[6] , indicating if the coach has the same *nationality* as his team or not.

4 Bivariate Poisson Regression on the EUROs 2004 - 2012:

We now applied the boosting approach introduced in Section 2.4 with linear predictors as specified in Equations (5)-(7). The vectors of covariate differences $\tilde{\mathbf{x}}$ and of absolute covariate differences $|\tilde{\mathbf{x}}|$ in (6) and (7), respectively, incorporate all 15 potential influence variables from Section 3.1. An extract of the design matrix, which corresponds to the covariate differences, is presented in Table 1. The optimal numbers of boosting steps have been determined by a three-dimensional 10-fold cross validation.

	Team 1	Team 2	Goals 1	Goals 2	Year	odds	market value	$\cdots$
1	Portugal	Greece	1	2	2004	-39.0	7.85	$\cdots$
2	Spain	Russia	1	0	2004	-33.5	7.67	$\cdots$
3	Greece	Spain	1	1	2004	38.5	-7.58	$\cdots$
4	Russia	Portugal	0	2	2004	34.0	-7.94	$\cdots$
5	Spain	Portugal	0	1	2004	0.5	-0.27	$\cdots$
6	Russia	Greece	2	1	2004	-5.0	-0.09	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

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

The intercept $\hat{\beta}_0$, corresponding to the linear predictor of $\hat{\gamma}_1$, was updated several times by the boosting algorithm. Also the linear predictor of $\hat{\gamma}_2$ was updated several times and from the set of potential influence variables only the covariates *ODDSET odds*, *market value* and *UEFA points* were chosen.

In contrast, the linear predictor of $\hat{\lambda}_3$ was not updated. No covariates were chosen and the intercept $\hat{\alpha}_0$ was set to a large negative value, leading to $\hat{\lambda}_3 \approx 0$. This reflects an important result from the modeling perspective. It shows that no additional covariance needs to be considered, if the linear predictors of the two Poisson parameters λ_1 and

[5]Note that this variable is not available by any soccer data provider and thus had to be collected "by hand".
[6]These two variables are available on several soccer data providers, see, for example, `http://www.kicker.de/`.

λ_2, in our re-parametrization reflected by γ_2, already contain informative covariate information from both teams and, hence, already induce a certain amount of (negative) correlation. Instead, on the EURO 2004-2012 data two independent Poisson distributions can be used for the two numbers of goals of the matches, if the linear predictor of both Poisson parameters each contains covariate information (here in the form of differences) of both competing teams.

Altogether, we obtained a quite simple model with the following estimates (corresponding to scaled covariate information):

- $\gamma_1 = \exp(\beta_0)$ with $\hat{\beta}_0 = 0.176$

  $\Longrightarrow \hat{\gamma}_1 = \exp(\hat{\beta}_0) = 1.192$; the parameter reflects the average number of goals, if two teams with equal covariates play against each other

- $\gamma_2 = \exp((\mathbf{x}_1 - \mathbf{x}_2)^T\boldsymbol{\beta}) = \exp(\tilde{\mathbf{x}}^T\boldsymbol{\beta})$ with $(\hat{\beta}_{odds}, \hat{\beta}_{marketvalue}, \hat{\beta}_{UEFApoints}) = (-0.120, 0.143, 0.029)$

- $\lambda_3 = \exp(\alpha_0 + |\mathbf{x}_1 - \mathbf{x}_2|^T\boldsymbol{\alpha}) = \exp(\alpha_0 + |\tilde{\mathbf{x}}|^T\boldsymbol{\alpha})$ with $\alpha_0 = -9.21$, $\boldsymbol{\alpha} = \mathbf{0}$

  $\Longrightarrow \lambda_3 \approx 0$; no (additional) covariance between scores of both teams

Hence, our final model for match scores is based on these estimates and, regarding the findings with respect to λ_3, on two independent Poisson distributions with parameters $\lambda_1 = \gamma_1\gamma_2$ and $\lambda_2 = \gamma_1/\gamma_2$. Based on this final model, in the following different simulation studies were applied.

4.1 Probabilities for the UEFA European Championship 2016 Winner

For each match of the EURO 2016, the final model from the previous section is used to calculate the two distributions of the scores of the two competing teams. For this purpose, for the two competing teams in a match the covariate differences of the three selected covariates *ODDSET odds*, *market value* and *UEFA points* have to be calculated in order to be able to compute an estimate of the linear predictor of the parameter $\hat{\gamma}_2$. Then, the match result can be drawn randomly from these predicted distributions, i.e. $G_1 \sim Poisson(\hat{\lambda}_1)$, $G_2 \sim Poisson(\hat{\lambda}_2)$, with estimates $\hat{\lambda}_1 = \hat{\gamma}_1\hat{\gamma}_2$ and $\hat{\lambda}_2 = \hat{\gamma}_1/\hat{\gamma}_2$. Note here that being able to draw exact match outcomes for each match constitutes an advantage in comparison to several alternative prediction approaches, as this allows to precisely follow the official UEFA rules when determining the final group standings[7]. If a match in the knockout stage ended in a draw, we simulated another 30 minutes of extra time using scoring rates equal to 1/3 of the 90 minutes rates, i.e. using Poisson parameters $\hat{\lambda}_1/3$ and $\hat{\lambda}_2/3$. If the match then still ended in a draw, the winner was calculated simply by coin flip, reflecting a penalty shoot out.

The whole tournament was simulated 1,000,000 times. Based on these simulations, for each of the 24 participating teams probabilities to reach the next stage and, finally, to win the tournament are obtained. These are summarized in Table 2 together with the winning probabilities based on the ODDSET odds for comparison. In contrast to most other prediction approaches for the current UEFA European championship favoring France (see, for example, Zeileis et al., 2016; Goldman-Sachs Economics Research, 2016), we get a neck-and-neck race between Spain and Germany, finally with better chances for Spain. The major reason for this is that in the simulations with a high probability both Spain and Germany finish their groups on the first place and then face each other in the final. In a direct duel, the model concedes Spain a thin advantage with a winning probability of 51.1% against 48.9%. The favorites Spain and Germany are followed by the teams of France, England, Belgium and Portugal. This also shows how unlikely

[7]The final group standings are determined by the number of points. If two or more teams are equal on points after completion of the group matches, specific tie-breaking criteria are applied: if two or more teams are equal on points, the first tie-breaking criteria are matches between teams in question (1. obtained points; 2. goal difference; 3. higher number of goals scored). 4. if teams still have an equal ranking, criteria 1 to 3 are applied once again, exclusively to the matches between the teams in question to determine their final rankings. If teams still have an equal ranking, all matches in the group are considered (5. goal difference; 6. higher number of goals scored). 7. if only two teams have the same number of points, and they were tied according to criteria 1-6 after having met in the last round of the group stage, their ranking is determined by a direct penalty shoot-out (this criterion would not be used if three or more teams had the same number of points.). 7. fair play conduct (yellow card: 1 point, red card: 3 points); 8. Position in the UEFA national team coefficient ranking system.

Note that due to the augmentation from 16 to 24 teams, also the four best third-placed teams qualified for the newly introduced round-of-sixteen. For the determination of the four best third-placed teams also specific criteria are used: 1. obtained points; 2. goal difference; 3. higher number of goals scored; 4. fair play conduct; 5. Position in the UEFA national team coefficient ranking system. Note that depending on which third-placed teams qualify from groups, several different options for the round-of-sixteen have to be taken into account.

in advance of the tournament the triumph of the Portuguese team was assessed. While according to the bookmaker ODDSET only a probability of 4.5% was expected, after all our model assigned an increased probability of 5.5% to this event.

	Round of 16	Quarter Finals	Semi Finals	Final	European Champion	Oddset
Spain	95.4	72.9	52.3	35.1	21.8	13.9
Germany	99.3	79.5	51.3	34.4	21.0	16.9
France	97.5	71.9	48.2	25.8	13.8	18.9
England	95.2	69.4	43.4	23.9	12.9	9.2
Belgium	93.9	58.7	32.8	18.7	9.5	7.3
Portugal	92.5	52.3	27.4	12.6	5.5	4.5
Italy	87.7	47.6	23.8	11.4	4.8	5.3
Croatia	73.2	35.3	16.8	7.3	2.7	3.2
Poland	86.0	42.2	15.6	5.5	1.6	2
Austria	79.1	34.0	13.4	4.4	1.3	2.7
Switzerland	77.9	35.8	13.3	4.3	1.2	1.6
Turkey	56.1	21.2	8.3	2.8	0.8	1.6
Wales	65.6	27.4	9.6	2.8	0.8	1.6
Russia	62.3	25.1	8.6	2.5	0.6	1.3
Ukraine	71.0	25.8	7.7	2.0	0.4	1
Iceland	61.7	20.0	6.2	1.5	0.3	1.6
Czech Rep.	42.5	13.6	4.6	1.3	0.3	1.6
Slovakia	44.5	13.6	3.6	0.8	0.2	1
Sweden	42.9	11.2	3.3	0.8	0.1	1
Ireland	41.7	10.6	3.1	0.7	0.1	1
Romania	45.6	12.3	2.8	0.5	0.1	0.8
Albania	41.3	10.4	2.2	0.4	0.1	0.8
Hungary	37.1	8.1	1.8	0.3	0.0	0.8
Nor. Ireland	10.0	1.0	0.1	0.0	0.0	0.4

Table 2: Estimated probabilities (in %) for reaching the different stages in the UEFA European championship 2016 for all 24 teams based on 1,000,000 simulation runs of the UEFA European championship 2016 together with winning probabilities based on the ODDSET odds.

4.2 Most probable tournament outcome

Finally, based on the 1,000,000 simulations, we also provide the most probable tournament outcome. Here, for each of the six groups we selected the most probable final group standing regarding the complete order of the places one to four. The results together with the corresponding probabilities are presented in Table 3.

	A	B	C	D	E	F
1	France	England	Germany	Spain	Belgium	Portugal
2	Switzerland	Wales	Poland	Croatia	Italy	Austria
3	Romania	Russia	Ukraine	Turkey	Sweden	Iceland
4	Albania	Slovakia	Nor. Ireland	Czech Rep.	Ireland	Hungary
	21.2%	15.1%	37.6%	17.7%	17.5%	16.9%

Table 3: Most probable final group standings together with the corresponding probabilities for the UEFA European championship 2016 based on 1,000,000 simulation runs

It is obvious that there are large differences with respect to the groups' balances. While in Group A the model forecasts the most likely final group standing of the four teams with a somewhat higher probabilities of 37.6%, the other groups seem to be closer.

Based on the most probable group standings, we also provide the most probable course of the knockout stage, compare Figure 1. However, note that the most probable round-of-sixteen cannot directly be concluded from the

most probable final group standings shown in Table 3, as it is depending on which teams turn out to be the four best third-placed teams. For this reason, the knockout stage in Figure 1 starts with the most frequent constellation of the round-of-sixteen, which in fact still is extremely unlikely, as it occurred only 25 times out of the 1,000,000 simulation runs. Below each tournament stage the probability for the displayed combination of matches is displayed. Finally, according to the most probable tournament course the Spanish team would have won the European championship 2016. After all, obviously even this 'most probable' outcome is still extremely unlikely to happen because of the myriad of possible constellations. Hence, deviations of the true tournament outcome from the model's most probable one are not only possible, but very likely.

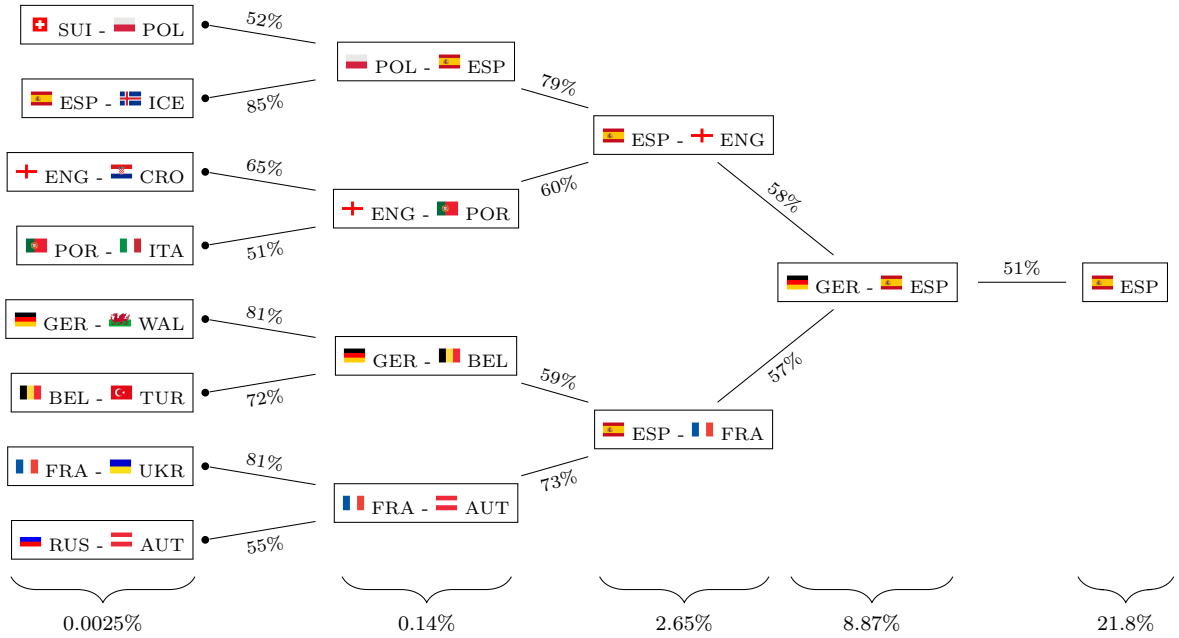

Figure 1: Most probable course of the knockout stage together with corresponding probabilities for the UEFA European championship 2016 based on 1,000,000 simulation runs.

4.3 Prediction power

In the following, we try to asses the performance of our model with respect to prediction. We collected the "three-way" odds[8] for all 51 matches of the EURO 2016 from the online bookmaker *Tipico* (`https://www.tipico.de/de/online-sportwetten/`). By taking the three quantities $\tilde{p}_r = 1/\text{odds}_r, r \in \{1,2,3\}$ and by normalizing with $c := \sum_{r=1}^{3} \tilde{p}_r$ in order to adjust for the bookmaker's margins, the odds can be directly transformed into probabilities using $\hat{p}_r = \tilde{p}_r/c$[9]. On the other hand, let G_1 and G_2 denote the random variables representing the number of goals scored by two competing teams in a match. Then, we can compute the same probabilities by approximating $\hat{p}_1 = P(G_1 > G_2), \hat{p}_2 = P(G_1 = G_2)$ and $\hat{p}_3 = P(G_1 < G_2)$ for each of the 51 matches of the EURO 2016 using the corresponding Poisson distributions $G_1 \sim Poisson(\hat{\lambda}_1)$, $G_2 \sim Poisson(\hat{\lambda}_2)$, where the estimates $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are obtained by our regression model. Based on these predicted probabilities, the average probability of a correct prediction of a EURO 2016 match can be obtained. For the true match outcomes $\omega_m \in \{1,2,3\}, m = 1,\dots,51$, it is given by $\bar{p}_{\text{three-way}} := \frac{1}{51}\sum_{m=1}^{51} \hat{p}_{1m}^{\delta_{1\omega_m}} \hat{p}_{2m}^{\delta_{2\omega_m}} \hat{p}_{3m}^{\delta_{3\omega_m}}$, with δ_{rm} denoting Kronecker's delta. The quantity $\bar{p}_{\text{three-way}}$ serves as a useful performance measure for a comparison of the predictive power of the model and the bookmaker's odds and is shown for both data sets in Table 4. It is striking that the predictive power of our model outperforms the bookmaker's odds, especially if one has in mind that the bookmaker's odds are usually released just some days

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

[9] 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.

before the corresponding match takes place and, hence, are able to include the latest performance trends of both competing teams. This reflects a quite favorable result.

If one puts one's trust into the model and its predicted probabilities, the following betting strategy can be applied: for every match one would bet on the three-way match outcome with the highest expected return, which can be calculated as the product of the model's predicted probability and the corresponding three-way odd offered by the bookmakers. We applied this strategy to our model's results, yielding a return of 30.28%, when for all 51 matches equal-sized bets are placed. This is also a very satisfying result.

boosted bivariate Poisson model	*Tipico* odds
42.22%	39.23%

Table 4: Average probability $\bar{p}_{\text{three-way}}$ of a correct prediction of a UEFA European championship 2016 match for our model and the *Tipico* odds.

5 Concluding remarks

A bivariate Poisson model for the number of goals scored by soccer teams facing each other in international tournament matches is set up. As an application, the UEFA European championships 2004-2012 serve as the data basis for an analysis of the influence of several covariates on the success of national teams in terms of the number of goals they score in single matches. Procedures for variable selection based on boosting methods, implemented in the R-package `gamboostLSS`, are used.

The boosting method selected only three covariates for the two Poisson parameters λ_1 and λ_2, namely the *ODDSET odds*, the *market value* and the *UEFA points*, while for the covariance parameter λ_3 no covariates were selected and the parameter was in fact estimated to be zero. This reflects an important general result for the modeling of soccer data. It shows that on the EURO 2004-2012 data no additional (positive) covariance needs to be considered. Hence, instead of the bivariate Poisson distribution two independent Poisson distributions can be used, if the two corresponding Poisson parameters λ_1 and λ_2 already contain covariate information from both teams, and, in this way already induce a certain amount of (negative) correlation.

The obtained sparse model was then used for simulation of the UEFA European championship 2016. According to these simulations, Spain, Germany and France turned out to be the top favorites for winning the title, with an advantage for Spain. Besides, the most probable tournament outcome is provided. An analysis of the predictive power of the model yielded very satisfactory results.

A major part of the statistical novelty of the presented work lies in the combination of boosting methods with a bivariate Poisson model. While the bivariate Poisson model enables explicit modeling of the covariance structure between match scores, the boosting method allows to include many covariates simultaneously and performs automatic variable selection.

Acknowledgement

We are grateful to Falk Barth and Johann Summerer from the ODDSET-Team for providing us all necessary odds data and to Sven Grothues from the Transfermarkt.de-Team for the pleasant collaboration.

References

Bernard, A. B. and M. R. Busse (2004). Who wins the olympic games: Economic developement and medall totals. *The Review of Economics and Statistics 86*(1), 413–417.

Brown, T. D., J. L. V. Raalte, B. W. Brewer, C. R. Winter, A. E. Cornelius, and M. B. Andersen (2002). World cup soccer home advantage. *Journal of Sport Behavior 25*, 134–144.

Carlin, J. B., L. C. Gurrin, J. A. C. Sterne, R. Morley, and T. Dwyer (2005). Regression models for twin studies: a critical review. *International Journal of Epidemiology B57*, 1089–1099.

Clarke, S. R. and J. M. Norman (1995). Home ground advantage of individual clubs in English soccer. *The Statistician 44*, 509–521.

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.

Freund, Y. and R. Schapire (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning Theory*, San Francisco, CA, pp. 148–156. San Francisco: Morgan Kaufmann Publishers Inc.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29*, 1189–1232.

Friedman, J. H., T. Hastie, and R. Tibshirani (2000). Additive logistic regression: A statistical view of boosting (with discussion). *The Annals of Statistics 28*, 337–407.

Gerhards, J., M. Mutz, and G. G. Wagner (2012). Keiner kommt an Spanien vorbei - außer dem Zufall. *DIW-Wochenbericht 24*, 14–20.

Gerhards, J., M. Mutz, and G. G. Wagner (2014). Predictable winners. market value, inequality, diversity, and routine as predictors of success in european soccer leagues. *Zeitschrift f́ur Soziologie 43*(3).

Gerhards, J. and G. G. Wagner (2008). Market value versus accident - who becomes European soccer champion? *DIW-Wochenbericht 24*, 236–328.

Gerhards, J. and G. G. Wagner (2010). Money and a little bit of chance: Spain was odds-on favourite of the football worldcup. *DIW-Wochenbericht 29*, 12–15.

Goldman-Sachs Economics Research (2016). The econometrician's take on euro 2016. `http://www.goldmansachs.com/our-thinking/macroeconomic-insights/euro-cup-2016/`.

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

Groll, A. and J. Abedieh (2014). A study on european football championships in the glmm framework with an emphasis on uefa champions league experience. In J. R. Bozeman, V. Girardin, and C. H. Skiadas (Eds.), *New perspectives on stochastic modeling and data analysis,*, pp. 313–321. ISAST.

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.

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

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

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

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

Mayr, A., H. Binder, O. Gefeller, and M. Schmid (2014a). The evolution of boosting algorithms - from machine learning to statistical modelling. *Methods of Information in Medicine 53*(6), 419–427.

Mayr, A., H. Binder, O. Gefeller, and M. Schmid (2014b). Extending statistical boosting - an overview of recent methodological developments. *Methods of Information in Medicine 53*(6), 428–435.

Mayr, A., N. Fenske, B. Hofner, T. Kneib, and M. Schmid (2012). Generalized additive models for location, scale and shape for high-dimensional data – a flexible aproach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*(3), 403–427.

McHale, I. G. and P. A. Scarf (2006). Forecasting international soccer match re- sults using bivariate discrete distributions. Technical Report 322, Working paper, Salford Business School.

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

Pollard, R. (2008). Home advantage in football: A current review of an unsolved puzzle. *The Open Sports Sciences Journal 1*, 12–14.

Pollard, R. and G. Pollard (2005). Home advantage in soccer: A review of its existence and causes. *International Journal of Soccer and Science Journal 3*(1), 25–33.

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

Schmid, M. and T. Hothorn (2008). Boosting additive models using component-wise P-splines. *Computational Statistics & Data Analysis 53*, 298–311.

Schmid, M., S. Potapov, A. Pfahlberg, and T. Hothorn (2010). Estimation and regularization techniques for regression models with multidimensional prediction functions. *Statistics and Computing 20*, 139–150.

Zeileis, A., C. Leitner, and K. Hornik (2016). Predictive Bookmaker Consensus Model for the UEFA Euro 2016. Working Papers 2016-15, Faculty of Economics and Statistics, University of Innsbruck.

Appendix

A Correlation structure of the EURO 2004 - 2012 data

	home	**GDP**	**max1**	**max2**	**odds**	**population**	**ave. age**	**market value**	**FIFA rank**	**UEFA points**	**CL players**	**UEFA players**	**age coach**	**nation coach**
GDP	-0.01													
max1	0.07	-0.31												
max2	0.03	-0.35	0.64											
odds	0.06	-0.12	-0.03	-0.19										
population	-0.14	-0.07	0.41	0.53	-0.52									
ave age	-0.20	0.11	-0.09	-0.18	0.26	-0.36								
market value	-0.08	0.14	0.27	0.41	-0.75	0.49	-0.27							
FIFA rank	0.58	-0.09	0.03	-0.11	0.70	-0.27	-0.06	-0.52						
UEFA points	0.06	-0.11	-0.33	-0.37	0.72	-0.54	0.12	-0.76	0.44					
CL players	0.07	0.01	0.29	0.33	-0.46	0.24	-0.42	0.82	-0.31	-0.48				
UEFA players	-0.17	-0.08	0.09	0.23	-0.28	0.46	-0.15	0.27	-0.24	-0.20	0.15			
age coach	0.08	0.00	0.00	0.08	0.15	-0.01	-0.06	-0.02	0.13	0.07	0.09	-0.19		
nation coach	0.03	0.23	-0.03	-0.16	-0.17	0.12	-0.15	0.19	0.01	-0.16	0.11	0.10	-0.32	
legionnaires	-0.04	0.15	-0.59	-0.72	0.26	-0.77	0.36	-0.45	0.01	0.55	-0.23	-0.35	0.00	0.01

Table 5: Correlation matrix of the considered variable differences for the EURO 2004 - 2012.