

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

A. Groll \*      T. Kneib †      G. Schauberger ‡

June 14, 2016

**Keywords** Football, EURO 2016, Bivariate Poisson Model.

## 1 Introduction

Based on all matches from the three previous UEFA European championships, the number of goals a team scores against a specific opponent is modeled by a joint bivariate Poisson model, including covariate information of both competing teams. Based on the estimates, the current tournament is simulated 100 000 times to obtain winning probabilities for all participating national teams.

## 2 A Bivariate Poisson-Model for Soccer Data

### 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  $X = X_1 + X_3$  and  $Y = X_2 + X_3$  follow a joint bivariate Poisson distribution, denoted by  $biPoi(\lambda_1, \lambda_2, \lambda_3)$ , with a joint probability function

$$\begin{aligned} P_{X,Y}(x, y) &= P(X = x, Y = y) \\ &= \exp(-(\lambda_1 + \lambda_2 + \lambda_3)) \frac{\lambda_1^x \lambda_2^y}{x! y!} \sum_{k=0}^{\min(x,y)} \binom{x}{k} \binom{y}{k} k! \left( \frac{\lambda_3}{\lambda_1 \lambda_2} \right)^k. \end{aligned} \quad (1)$$

The bivariate Poisson distribution allows for dependence between the two random variables  $X$  and  $Y$ . Marginally each random variable follows a univariate Poisson distribution with  $E[X] = \lambda_1 + \lambda_3$  and  $E[Y] = \lambda_2 + \lambda_3$ . Moreover, the dependence of  $X$  and  $Y$  is expressed by  $cov(X, Y) = \lambda_3$ . 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 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  $\lambda_k$ .

---

\*Department of Statistics, Ludwig-Maximilians-University Munich, Akademiestr. 1, 80799 Munich, Germany, [andreas.groll@stat.uni-muenchen.de](mailto:andreas.groll@stat.uni-muenchen.de)

†Department of Statistics and Econometrics, Georg-August-University Goettingen, Humboldtallee 3, 37073 Goettingen, Germany, [tkneib@uni-goettingen.de](mailto:tkneib@uni-goettingen.de)

‡Department of Statistics, Ludwig-Maximilians-University Munich, Akademiestr. 1, 80799 Munich, Germany, [gunther@stat.uni-muenchen.de](mailto:gunther@stat.uni-muenchen.de)

### 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  $\lambda_1$  and  $\lambda_2$ , respectively, and some extra information reflecting the match conditions of the corresponding match in  $\lambda_3$ . However, the covariate effects  $\beta_k, k = 1, 2$ , usually should be the same for both competing teams. Then, one obtains the model representation

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

with  $\mathbf{x}_1$  and  $\mathbf{x}_2$  denoting the covariates of team 1 and team 2. In contrast, the covariance parameter  $\lambda_3$  could generally depend on different covariates and effects, i.e.

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

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:

$$\begin{aligned} P_{X,Y}(x, y) &= P(X = x, Y = y) \\ &= \exp(-(\gamma_1(\gamma_2 + \gamma_2^{-1}) + \lambda_3)) \frac{(\gamma_1 \gamma_2)^x}{x!} \frac{(\frac{\gamma_1}{\gamma_2})^y}{y!} \sum_{k=0}^{\min(x,y)} \binom{x}{k} \binom{y}{k} k! \left(\frac{\lambda_3}{\gamma_1^2}\right)^k, \end{aligned}$$

with  $\lambda_1 = \gamma_1 \gamma_2$ ,  $\lambda_2 = \frac{\gamma_1}{\gamma_2}$ . The new parameters  $\gamma_1, \gamma_2$  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 and, as before,  $\lambda_3 = \exp(\alpha_0 + \mathbf{z}^T \boldsymbol{\alpha})$ . In the current analysis, we used the same covariates in the linear predictor of  $\lambda_3$  and set  $\lambda_3 = \exp(\alpha_0 + \tilde{\mathbf{x}}^T \boldsymbol{\alpha})$ .

























### 2.4 Estimation

The model was estimated using the R-package `gamboostLSS` (Hofner et al., 2016; Mayr et al., 2012). With `gamboostLSS` the model family of GAMLSS (Generalized Additive Models for Location, Scale and Shape) is combined with the boosting estimation technique. It allows to use multi-parametric distributions in regression models in combination with implicit variable selection. From a set of potential influence variables (for a detailed description of all possible variables see Groll and Abedieh, 2013) for  $\gamma_2$  only the covariates *bookmakers' odds* (odds for winning the title before the tournament) and *market value* were chosen. For  $\lambda_3$ , no covariates were chosen.

## 3 Simulation Results

























Based on the final model different simulation studies were applied. For each match, the model is used to calculate the two-dimensional distribution of the scores of both matches und the result can be drawn randomly from this distribution. First, the whole tournament was simulated 100 000 times. As the exact match outcomes were known, the official UEFA rules for the final standings in the groups could be applied in case of equal numbers of points.

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 probabilities are displayed in the following table:

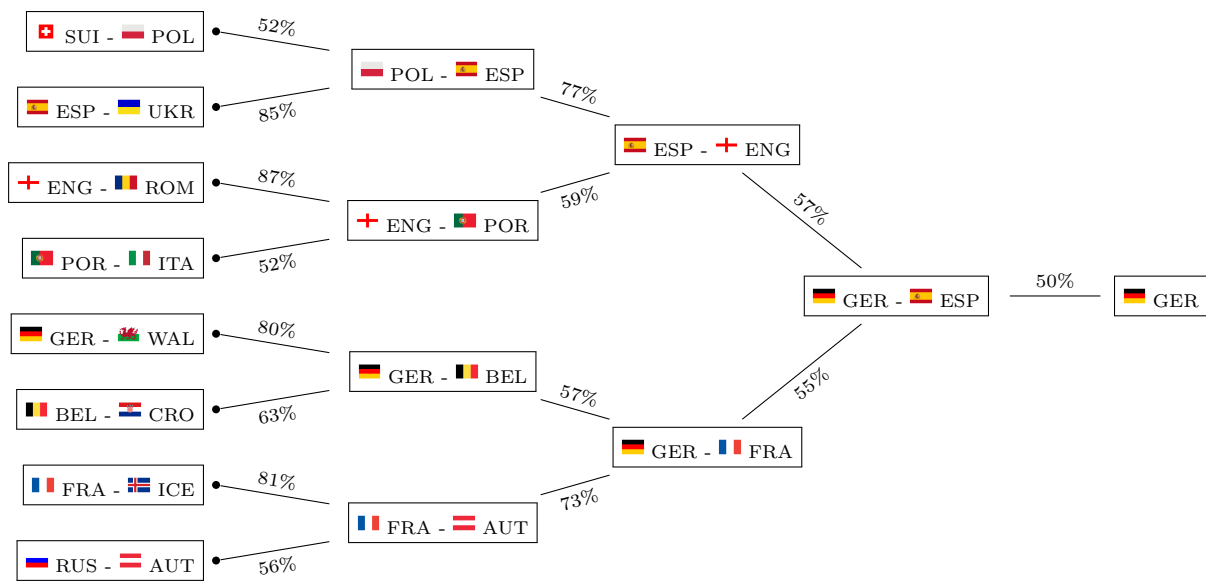
	Round of 16	Quarter Finals	Semi finals	Final	European Champion
Germany 	99.3	79.5	51.3	34.2	21.1
Spain 	95.0	71.2	50.4	33.2	19.8
France 	97.6	72.7	49.4	26.9	14.9
England 	95.2	68.8	42.6	23.5	12.5
Belgium 	94.6	60.8	34.8	20.5	11.0
Portugal 	92.4	52.2	27.4	12.7	5.4
Italy 	85.9	45.2	22.1	10.3	4.2
Croatia 	75.2	36.9	17.8	8.0	3.0
Poland 	86.1	42.8	16.0	5.7	1.8
Austria 	78.5	33.8	13.3	4.3	1.3
Switzerland 	77.8	35.6	13.1	4.3	1.2
Wales 	68.2	29.5	10.6	3.2	0.9
Turkey 	56.2	21.2	8.3	2.9	0.8
Russia 	59.7	23.1	7.5	2.0	0.5
Iceland 	62.2	20.6	6.4	1.7	0.4
Ukraine 	69.0	24.2	7.2	1.7	0.4
Czech Rep. 	41.5	13.0	4.3	1.2	0.3
Slovakia 	45.3	14.2	3.7	0.8	0.2
Ireland 	42.9	11.4	3.5	0.8	0.1
Sweden 	42.4	11.0	3.3	0.8	0.1
Romania 	44.4	11.8	2.6	0.5	0.1
Albania 	42.4	11.0	2.3	0.4	0.1
Hungary 	37.5	8.5	1.8	0.3	0.0
Nor. Ireland 	10.9	1.2	0.1	0.0	0.0

According to our proposed model, Germany is the favourite for the title with a winning probability of 21.1% followed by Spain, France, England and Belgium.

Finally, based on the 100,000 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 the following table.

	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	 Ireland	 Iceland
4	 Albania	 Slovakia	 Nor. Ireland	 Czech Rep.	 Sweden	 Hungary
	21.0%	15.4%	37.7%	18.0%	18.2%	16.5%

Based on the most probable group standings, in the following figure we also provide the most probable course of the knockout stage. According to the most probable tournament course the German team will win the European championship. After all, obviously even this 'most probable' outcome is still extremely unlikely to happen because of the myriad of possible constellations.



## References

- 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.
- Hofner, B., A. Mayr, N. Fenske, and M. Schmid (2016). *gamboostLSS: Boosting Methods for GAMLSS Models*. R package version 1.2-1.
- Karlis, D. and I. Ntzoufras (2003). Analysis of sports data by using bivariate poisson models. *The Statistician* 52, 381–393.
- 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 approach based on boosting. *Journal of the Royal Statistical Society, Series C - Applied Statistics* 61(3), 403–427.