A Study on European Football Championships in the GLMM Framework with an Emphasis on UEFA Champions League Experience

Andreas Groll[1] and Jasmin Abedieh[2]

[1] Department of Mathematics, Workgroup Financial Mathematics,
Ludwig-Maximilians-University, Theresienstr. 39, 80333 Munich, Germany
(E-mail: andreas.groll@math.lmu.de)
[2] Jasmin Abedieh
(E-mail: jasmin.abedieh@hotmail.de)

Abstract. This article has two major objectives. First, the results of a preceding article are revised, where all matches of the European football championship (EURO) 2012 have been predicted on the quite small data basis of the two preceding EUROs, resulting in a possible course of the tournament. There, a pairwise Poisson model for the number of goals scored by national teams competing in EURO matches was established, incorporating two approaches for variable selection, which was then used for prediction. Including the data of the EURO 2012, in the present article this analysis is replicated on a more reliable data basis and the set of selected influence variables is compared to the results of the preceding analysis. Besides, the course of the EURO 2012 suggests a positive correlation between a national team's success at a EURO and the number of its players that have been successful in the preceding Union of European Football Associations (UEFA) Champions League (CL) season. Hence, a second objective of this article is to check, if in fact a significant influence of this covariate can be detected.

Keywords: Football, European football championships, UEFA Champions League, Sports tournaments, Generalized linear mixed model, Lasso, Variable selection.

1 Introduction

Recently, the statistical analysis of major soccer events such as the UEFA CL (see Leitner et al. [15], Eugster et al. [4]), the EURO (see Leitner et al. [12], Leitner et al. [13], Zeileis et al. [17] or Groll and Abedieh [10]) or the Fédération Internationale de Football Association (FIFA) World Cup (see Leitner et al. [14], Stoy et al. [16], Dyte and Clarke [3]) has gained more and more attention. A major and challenging objective in this context is to predict the respective tournament winner. In general, the existing approaches can be divided into two major categories: approaches based on the easily available source of "prospective" information contained in bookmakers' odds (compare Leitner et al. [12], Leitner et al. [14] and Zeileis et al. [17]) and regression based models (compare Stoy et al. [16], Dyte and Clarke [3] or Groll and Abedieh [10]).

Based on the 62 matches of the EUROs 2004 and 2008, in Groll and Abedieh [10] a pairwise Poisson model for the number of goals scored by national teams in the single matches of the tournaments is used for prediction of the EURO

2012. There, 32 potential influence variables[1] are considered and team-specific random effects are included, resulting in a flexible generalized linear mixed model. Each match occurs in the data set in the form of two different rows, one for each team, containing both the variables corresponding to the team whose goals are considered as well as those of its opponent. The matched-pair design is accounted for by including a second match-specific random intercept, following Carlin et al. [2], which is assumed to be independent from the team-specific random intercept. Two different methods for the selection of relevant predictors, an L_1-penalization based technique (see Groll and Tutz [11]; implemented in the `glmmLasso` function of the corresponding R-package from Groll [9]) as well as subset selection, are used to obtain a sparse final model, which is then used for the prediction of the whole tournament outcome of the EURO 2012: it contains only the four variables *market value* (MV; of both teams), the *maximum number of teammates* and *UEFA points*. Note here that in contrast to other team sports such as basketball, ice-hockey or handball, in soccer pure chance plays a larger role, as in soccer fewer points (goals) are scored and thus single game situations can have a tremendous effect on the outcome of the match. Hence, the prediction of soccer tournaments is especially demanding. Nevertheless, the forecast of the EURO 2012 tournament outcome in Groll and Abedieh [10] shows surprisingly many accordances with the true one: seven of the eight teams that qualified for the knockout stage were predicted correctly, three of the four teams that qualified for the half-finals and finally, the tournament winner Spain.

However, their results base on the quite small data basis of only two preceding EUROs. Besides, intuitively it is somewhat surprising that the covariate *maximum number of teammates* has a significant effect at all and that this effect is negative. Therefore, a major objective of the present article is to revise the results of Groll and Abedieh [10], including the data of the EURO 2012. We replicate their analysis on a more reliable data basis and compare the set of selected influence variables with the results of the preceding analysis.

In Groll and Abedieh [10] it is already mentioned that for the half-final of the EURO 2012, with the national teams of Spain, Germany and Portugal, exactly those three teams qualified that had the largest proportion of players amongst their squad that reached at least the half-finals of the UEFA CL 2012: Spain with 14, Germany with 10 and Portugal with 4 players. All other national teams, except for France with 3 players, had only 2 or fewer players that reached at least the half-finals of the preceding UEFA CL season. Besides, also Frohwein [5] has already pointed out that there is a connection between the final rounds of the UEFA CL and the FIFA World Cup final. Though this

[1] Several economic (GDP per capita, population size) and sportive factors (unfairness points, home advantage, odds, market value, FIFA points, UEFA points) as well as factors describing the team's structure (maximum number of teammates, second maximum number of teammates, average age, number of CL players, number of Europa League players, age of the national coach, nationality of the national coach, number of legionnaires) have been included in the analysis, both of the team whose goals are considered and of its opponent. For a detailed description of these variables consult Groll and Abedieh [10].

coherence seems too distinct to be just a matter of chance, based on the data of the EUROs 2004 and 2008, no clear significance of the covariate *number of CL players* could be detected in Groll and Abedieh [10]. Hence, a second major objective of this article is to investigate, if the number of a national team's players, which reached at least the half-finals in the preceding UEFA CL season, has now a significant influence on the team's success[2] at the subsequent EURO tournament, if the data of the EURO 2012 are included.

The rest of the article is structured as follows. In Section 2 a pairwise Poisson model for the number of goals is used to determine the covariates of a final model, which serves as control model with regard to the preceding analysis presented in Groll and Abedieh [10]. In Section 3 the explorative power of the covariate *number of CL players* with respect to the success of national teams at EURO tournaments is investigated, before we conclude in Section 4.

2 Poisson Regression on the EUROs 2004-2012

The following regression analysis is based on the mixed Poisson model presented in Section 2 of Groll and Abedieh [10], with 32 covariates and the number of goals scored by national teams in the single matches of the tournaments as response variable. Team-specific random intercepts are included in order to adequately account for different basis levels of the national teams, as well as match-specific random intercepts to model the matched-pair design. We use two different approaches that are both able to perform variable selection, an L_1-penalization technique, which is proposed by Groll and Tutz [11] and implemented in the `glmmLasso` function, and forward subset selection based on the `glmer` function (Bates and Maechler [1]), denoted by `glmer-select`.

For the Lasso approach we obtain different levels of sparseness by changing the determination procedure of the optimal tuning parameter. In the following we consider three techniques: AIC, BIC and leave-one-out cross-validation (LOOCV)[3]. BIC leads to the sparsest models, followed by AIC, whereas the LOOCV yields models that include several covariates. The sparseness of the models obtained by the forward selection procedure `glmer-select` can be controlled directly by the level of significance α in the corresponding model testing, which is based on an analysis of deviance. Table 1 presents the corresponding results for $\alpha \in \{0.01, 0.05, 0.1\}$. We consider the same models as in Groll and Abedieh [10], decreasing step by step the number of given influence variables:

- **Model 1**: A model containing all covariates is fitted.

- **Model 2**: A model containing all covariates except for the variable *ODDSET odds* is fitted.

- **Model 3**: As the variable *fairness* is not available for the prediction of future EUROs, a model containing all covariates except for *ODDSET odds* and *fairness* is fitted.

[2] measured by the number of goals scored in the single matches of the next EURO.

[3] Due to the matched-pair design, not single observations but single matches are excluded from the training data.

	glmmLasso			glmer-select		
	BIC	AIC	LOOCV	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
M1	ODDS	ODDS	ODDS	ODDS	ODDS	ODDS
	-	ODDS opp.	ODDS opp.	MV opp.	MV opp.	MV opp.
	-	fairness	fairness	-	fairness	fairness
	-	fairness opp.	MV opp.	-	-	-
	-	-	MV	-	-	-
	-	-	MV opp.	-	-	-
	-	-	UEFA pts.	-	-	-
	-	-	UEFA pts. opp.	-	-	-
	-	-	# CL players	-	-	-
	-	-	# CL players opp.	-	-	-
M2	MV opp.	fairness	fairness	fairness	fairness	fairness
	-	MV opp.	fairness opp.	MV opp.	MV opp.	MV opp.
	-	-	MV	-	MV	MV
	-	-	MV opp.	-	-	-
	-	-	FIFA pts.	-	-	-
	-	-	UEFA pts.	-	-	-
	-	-	UEFA pts. opp.	-	-	-
	-	-	# CL players	-	-	-
	-	-	# CL players opp.	-	-	-
	-	-	nat. coach	-	-	-
M3	MV opp.	MV	MV	MV	MV	MV
	-	MV opp.	MV opp.	MV opp.	MV opp.	MV opp.
	-	-	FIFA pts.	-	-	-
	-	-	UEFA pts.	-	-	-
	-	-	UEFA pts. opp.	-	-	-
	-	-	# CL players	-	-	-
	-	-	# CL players opp.	-	-	-
	-	-	nat. coach	-	-	-

Table 1: Selected variables for `glmmLasso` and `glmer-select` for Model 1-3 and different levels of sparseness (covariates have been standardized).

Groll and Abedieh [10] suggest a possible goodness-of-fit criterion to assess the performance of these models, based on the "three-way" odds from the German state betting agency ODDSET for all 93 games of the EUROs 2004-2012, which can be directly transformed into (approximate) probabilities $\hat{p}_i$. On the other hand, let G_k denote the random variables representing the number of goals scored by Team k in a certain match and G_l the goals of its opponent, respectively. Then we can compute the same probabilities by approximating $\hat{p}_1 = P(G_k > G_l), \hat{p}_2 = P(G_k = G_l)$ and $\hat{p}_3 = P(G_k < G_l)$ for each of the 93 matches using the corresponding Poisson distributions, whereas the estimates $\hat{\lambda}_k$ and $\hat{\lambda}_l$ are obtained by our regression models. Hence, we can provide a goodness-of-fit criterion by comparing the values of the log-likelihood of the 93 matches for the ODDSET odds with those obtained for our regression models. For $\omega_j \in I, j = 1, \ldots, 93$, the likelihood is given by the product $\prod_{j=1}^{93} \hat{p}_{1j}^{\delta_{1\omega_j}} \hat{p}_{2j}^{\delta_{2\omega_j}} \hat{p}_{3j}^{\delta_{3\omega_j}}$, with δ_{ij} denoting Kronecker's delta. The log-likelihood scores for `glmmLasso` and `glmer-select` corresponding to Model 1-3 and different levels of sparseness can be found in Table 2. The results show that for all settings the fit obtained by our regression models outperforms the log-likelihood score corresponding to the ODDSET odds (which yields -94.25) and hence, the models seem reasonable.

Table 1 shows that the results for `glmer-select`, which serve as a control for our L_1-penalization approach, generally agree with those obtained by the `glmmLasso` function, but are somewhat sparser. For each setting, the `glmmLasso` approach based on LOOCV chooses several more variables than the

	`glmmLasso`			`glmer-select`		
	BIC	AIC	LOOCV	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
M1	-90.23	-85.85	-83.92	-90.29	-88.12	-88.12
M2	-88.55	-86.52	-85.35	-89.36	-88.14	-88.14
M3	-88.55	-87.14	-86.35	-90.17	-90.17	-90.17

Table 2: Log-likelihood scores for `glmmLasso` and `glmer-select` for Model 1-3 and different levels of sparseness.

other approaches. For Model 1 all methods select the *ODDSET odds*, while for Model 2 almost all methods select *fairness* and the *MV* of the opponent. The *MV* of both teams is selected by almost all methods for Model 3. Note that the covariate *maximum number of teammates*, which was surprisingly selected in the preceding analysis, has not been selected in any of the regarded models and thus will not be incorporated in our final model. The variables *UEFA points* and *number of CL players* (each for both teams) are selected for each setting by `glmmLasso` based on LOOCV, but for no other method. Similar to Groll and Abedieh [10], we focus on the contribution of those covariates that seem to be able to adequately replace the bookmakers' odds and consider all covariates from Model 2 and 3 that have been selected at least three times (and for at least two of the six approaches), except for the variable *fairness* (as it cannot be observed before the start of the tournament and thus cannot be used for prediction). This yields the following sparse predictor

$$\log(\lambda_{it,j\tilde{t}}) = \beta_0 + (\text{MV})_{it,j\tilde{t}}\beta_1 + (\text{MV opp.})_{it,j\tilde{t}}\beta_2 + b_i + c_j, \tag{1}$$

where $\lambda_{it,j\tilde{t}}$ denotes the expected number of goals scored by team i at its t-th game with match number j, $b_i \sim N(0, \sigma_b^2)$ represent team-specific random intercepts and $c_j \sim N(0, \sigma_c^2)$ represent match-specific random intercepts in order to account for the matched pair design with $\tilde{t} \in \{1, 2\}$. This model coincides with the models selected by `glmer` and `glmmLasso` with AIC for Model 3. The corresponding fit is easily obtained, e.g. by using the `glmer` function. The results are presented in Table 3. As expected, the variable *MV* has a clear positive effect on the number of goals a national team scores, while the effect of the opponent's *MV* is negative. Both effects are clearly significant and the final model from equation (1) yields a rather respectable fit with an "in sample" log-likelihood score of -90.17. Both variables have already been selected in Groll and Abedieh [10], but on the basis of the larger data set of the EUROs 2004-2012, now the *maximum number of teammates* and the *UEFA points* of the opponent are not incorporated anymore.

	Coefficients	Standard errors
Intercept	0.162	0.069
MV	0.205	0.060
MV opp.	-0.268	0.076
$\hat{\sigma}_b$	$3.24 \cdot 10^{-6}$	-
$\hat{\sigma}_c$	0	-

Table 3: Estimates for the final model from equation (1) with `glmer`.

3 Explorative Power of the Number CL Players

As already mentioned in the introduction, a second major objective of this article is to investigate, if the number of a national team's players, which reached at least the half-finals in the preceding UEFA CL season, has a significant influence on the team's success at the subsequent EURO tournament, if the data of the EURO 2012 are included. On the one hand, we found in Section 2 that the *number of CL players* for both teams is selected for each setting by `glmmLasso` based on LOOCV, but for no other method. On the other hand, there is a substantial positive correlation (0.827) between the *MV* and the *number of CL players* for the data based on the EUROs 2004-2012.

The results from Section 2 suggest that the only relevant variable for predicting games of a EURO tournament is the *MV* of both teams. Though the variable has gained increasing importance and newly approaches for the prediction of the most renowned soccer events have been based on it (see for example Gerhards and Wagner [6, 7], Gerhards et al. [8]), there are some drawbacks, as its realizations are usually based on estimates. Any registered user of the web-site `http://www.transfermarkt.de`, e.g., is allowed to rate the MVs of single international players, and a player's MV then essentially results as an average of these ratings. Beside the transfer value of a player, the user ratings also cover aspects such as experience, future perspective or prestige of a player. Consequently, there is a certain amount of subjective valuation contained in these estimated MVs and it may be preferable to consider an alternative variable, which is fixed and easy to obtain, such as the *number of CL players*. Of course, such an alternative variable would only help, if it provides at least almost the same explorative power as the *MV*. In the following we show that the *number of CL players* serves a highly precious candidate.

In Table 4 we present the results for the Models 1-3 from Section 2, after excluding the *MV* from the set of potential covariates. We find that, now, in almost all settings for Model 2 and 3 the *number of CL players* is selected. In general, the resulting log-likelihood scores presented in Table 5 are almost indistinguishable compared to those in Table 5. Hence, the variable *number of CL players* seems to be a promising competitor for the *MV*. Following the selection criteria from the preceding section, we end up with the model

$$\log(\lambda_{it,j\tilde{t}}) = \beta_0 + (\text{UEFA pts. opp.})_{it,j\tilde{t}}\beta_1 + (\#\text{ CL players})_{it,j\tilde{t}}\beta_2 + b_i + c_j, \quad (2)$$

which achieves almost the same goodness-of-fit (-90.46) as model (1). The corresponding estimates are presented in Table 6. Again, both effects are clearly significant and, as expected, the opponent's *UEFA points* have a negative effect on the number of goals a national team scores, while the *number of CL players* has a distinct positive effect.

4 Conclusion

In the article two major objectives are treated. First, the results of Groll and Abedieh [10] are revised by replicating their analysis on a more reliable data

		glmmLasso			glmer-select	
	BIC	AIC	LOOCV	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
M1	ODDS	ODDS	ODDS	ODDS	ODDS	ODDS
	-	ODDS opp.	ODDS opp.	UEFA pts. opp.	UEFA pts. opp..	UEFA pts. opp.
	-	UEFA pts. opp.	UEFA pts. opp.	-	fairness	fairness
	-	fairness	fairness	-	-	-
	-	-	# CL players	-	-	-
M2	UEFA pts. opp.	fairness	fairness	fairness	fairness	fairness
	-	UEFA pts. opp.	fairness opp.	UEFA pts. opp.	UEFA pts. opp.	UEFA pts. opp.
	-	-	# CL players	-	# CL players	# CL players
	-	-	# CL players opp.	-	-	-
	-	-	FIFA pts.	-	-	-
	-	-	UEFA pts.	-	-	-
	-	-	UEFA pts. opp.	-	-	-
M3	UEFA pts. opp.	UEFA pts. opp.	UEFA pts. opp.	UEFA pts. opp.	UEFA pts. opp.	UEFA pts. opp.
	-	# CL players	# CL players	# CL players	# CL players	# CL players

Table 4: Selected variables for glmmLasso and glmer-select for Model 1-3 and different levels of sparseness (excluding the *MV* from the set of potential covariates).

	glmmLasso			glmer-select		
	BIC	AIC	LOOCV	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
M1	-90.23	-85.32	-84.85	-89.98	-87.60	-87.60
M2	-88.52	-86.09	-85.33	-89.21	-87.82	-87.82
M3	-88.52	-87.18	-87.18	-90.46	-90.46	-90.46

Table 5: Log-likelihood scores for glmmLasso and glmer-select for Model 1-3 and different levels of sparseness (excluding the *MV* from the set of potential covariates).

	Coefficients	Standard errors
Intercept	0.171	0.069
UEFA pts. opp.	-0.236	0.071
# CL players	0.179	0.056
$\hat{\sigma}_b$	$6.94 \cdot 10^{-6}$	-
$\hat{\sigma}_c$	$4.73 \cdot 10^{-7}$	-

Table 6: Estimates for the final model from equation (2) with glmer.

basis, including the data of the EURO 2012. Still, the *MV* of both competing teams plays a major role for the teams' success, but on basis of the larger data set of the EUROs 2004-2012, the variables *maximum number of teammates* and *UEFA points opponent* are not incorporated anymore. Secondly, it is shown that the variable *number of CL players* provides almost the same explorative power with respect to the number of goals scored by national teams at EUROs as the *MV*, and hence serves as a reliable and competitive alternative.

References

[1] D. Bates and M. Maechler. *lme4: Linear mixed-effects models using S4 classes*, 2010. URL http://CRAN.R-project.org/package=lme4. R package version 0.999999-0.

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

[3] D. Dyte and S. R. Clarke. A ratings based Poisson model for World Cup soccer simulation. *Journal of the Operational Research Society*, 51 (8), 2000.

[4] M. J. A. Eugster, J. Gertheiss, and S. Kaiser. Having the second leg at home - advantage in the UEFA Champions League knockout phase? *Journal of Quantitative Analysis in Sports*, 7(1), 2011.

[5] T. Frohwein. Die falschen Pferde. In: e-politik.de (08.06.2010), available at: `http://www.e-politik.de/lesen/artikel/2010/die-falschen-pferde` (12.06.2012), 2010.

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

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

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

[9] A. Groll. *glmmLasso: Variable Selection for generalized linear mixed models by L_1-penalized estimation*, 2011. URL `http://CRAN.R-project.org/package=glmmLasso`. R package version 1.1.1.

[10] A. Groll and J. Abedieh. Spain retains its title and sets a new record - generalized linear mixed models on European football championships. *Journal of Quantitative Analysis in Sports*, 2013. To appear.

[11] A. Groll and G. Tutz. Variable selection for generalized linear mixed models by L_1-penalized estimation. *Statistics and Computing*, 2012. To appear.

[12] C. Leitner, A. Zeileis, and K. Hornik. Who is going to win the EURO 2008? (A statistical investigation of bookmakers odds). Research report series, Department of Statistics and Mathematics, University of Vienna, 2008.

[13] C. Leitner, A. Zeileis, and K. Hornik. Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008. *International Journal of Forecasting*, 26 (3):471–481, 2010.

[14] C. Leitner, A. Zeileis, and K. Hornik. Forecasting the winner of the FIFA World Cup 2010. Research report series, Department of Statistics and Mathematics, University of Vienna, 2010.

[15] C. Leitner, A. Zeileis, and K. Hornik. Bookmaker concensus and agreement for the UEFA Champions League 2008/09. *IMA Journal of Management Mathematics*, 22 (2):183–194, 2011.

[16] V. Stoy, R. Frankenberger, D. Buhr, L. Haug, B. Springer, and J. Schmid. Das Ganze ist mehr als die Summe seiner Lichtgestalten. Eine ganzheitliche Analyse der Erfolgschancen bei der Fußballweltmeisterschaft 2010. Working Paper 46, Eberhard Karls University, Tübingen, Germany, 2010.

[17] A. Zeileis, C. Leitner, and K. Hornik. History repeating: Spain beats Germany in the EURO 2012 final. Working paper, Faculty of Economics and Statistics, University of Innsbruck, 2012.