Forecasting the Winner of the FIFA World Cup 2010

Leitner, Christoph; Zeileis, Achim; Hornik, Kurt

DOI:
10.57938/e23a2b86-03fa-420f-b7ba-b1c2cb131e47

Published: 01/06/2010

Document Version
Publisher's PDF, also known as Version of record

Forecasting the Winner of the FIFA World Cup 2010

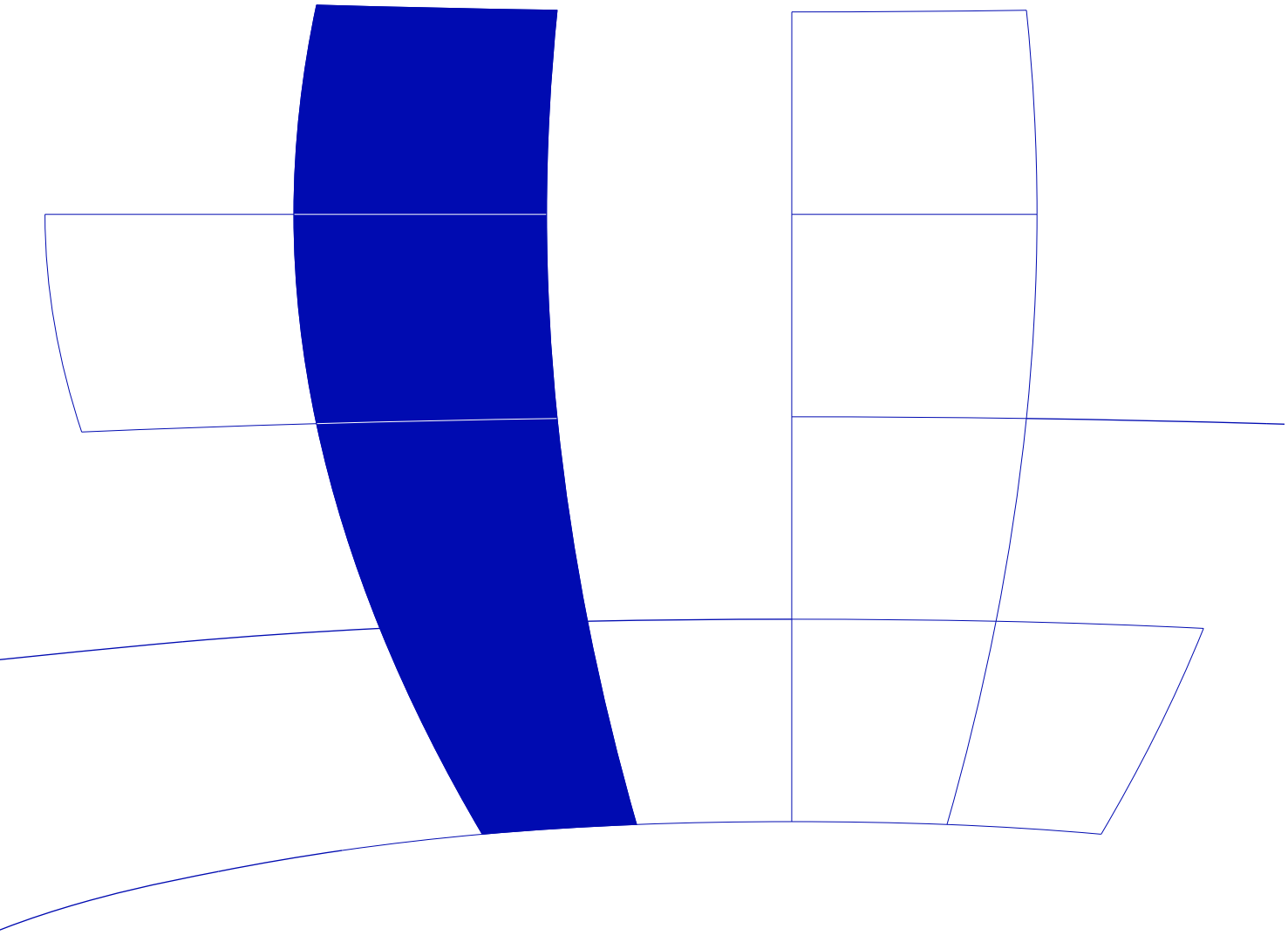
Christoph Leitner, Achim Zeileis, Kurt Hornik

Institute for Statistics and Mathematics
Wirtschaftsuniversität Wien

Research Report Series

Report 100
June 2010

http://statmath.wu-wien.ac.at/



Forecasting the Winner of the FIFA World Cup 2010

Christoph Leitner[1], Achim Zeileis[2], Kurt Hornik[1]

[1] Institute for Statistics and Mathematics, WU Wirtschaftsuniversität Wien, Austria
[2] Department of Statistics, Universität Innsbruck, Austria

Abstract

The FIFA World Cup is one of the most prestigious tournament all over the world and hence there is major interest, among fans and experts alike, in forecasting the winner of this tournament. To investigate this issue, a class of linear mixed-effects models for quoted winning odds from various bookmakers is explored. Based on this "prospective" data reflecting the expectations of the bookmakers (as opposed to past performances used in many other forecasting methods) different models for the "true" odds of winning the tournament can be established, capturing both team-specific effects (along with effects for the team's tournament group and continental confederation) and bookmaker-specific variations. A selection among various model specifications yields a model with a fixed team effect plus a random bookmaker-specific deviation. It forecasts team Spain with a probability of 17.86% as the winner of the tournament; the second best team is Brazil with a winning probability of 15.27%. In addition to the forecast of the winning probability, information about the groups of the preliminaries and the different continental confederations can be obtained from the model.

1 Introduction

The *FIFA World Cup* is one of the most prestigious sports tournaments all over the world. Millions of football supporters are interested in the games and the title winner. Various strategies for forecasting the winner of the World Cup 2010 have been proposed (e.g., J.P.Morgan, 2010; UBS Wealth Management Research, 2010; Goldman Sachs, 2010). Here, we employ a general mixed-effects model framework which builds upon the ideas of Leitner et al. (2009, 2010) for forecasting the winner of this tournament. Unlike many other sports prediction methods it is not based on historical data (see e.g., Dyte and Clarke, 2000; Goddard and Asimakopoulos, 2004) but designed for bookmakers odds for winning the World Cup, i.e., reflecting current expectations. The motivation for using bookmakers odds is that (a) they incorporate expectations about a specific tournament, (b) bookmakers have financial incentives to rate teams correctly, (c) other empirical studies have shown that odds provide an efficient forecasting instrument for the outcomes of single games (see e.g., Forrest et al., 2005; Dixon and Pope, 2004). Based on these ideas, Leitner et al. (2009, 2010) use quoted odds to forecast the outcome of whole tournaments, such as the UEFA Champions League 2009/10 and the EURO 2008. Their studies performed successfully, e.g., particular predicting correctly the final for the EURO 2008. Here, we adapt this method to the FIFA World Cup 2010 and establish a model framework for the winning logits capturing different effects associated with the participants, the bookmakers, the groups of the preliminaries and the team's confederations leading to a variety of mixed-effects models. After establishing the general modeling approach a subsequent model selection yields the final model upon which our forecasts for the tournament are based. For this study we use the quoted long-term odds for winning the FIFA World Cup 2010 of 26 international bookmakers which were published online after the group draw but before the tournament started (accessed on 2010-05-29 from the bookmakers' websites).

The FIFA World Cup is a tournament where national football teams of all six world's confederations, the Asian Football Confederation (AFC), the Confédération Africaine de Football (CFA), the Confederatión of North, Central America and the Caribbean (CONCACAF), the Conferación Sudamericana de Fútbol

(CONMEBOL), the Oceania Football (OFC) and the Union des Associations Europénnes de Football (UEFA) compete in a multi-stage modus (qualification, group and knockout stage) to determine the World champion (Fédération Internationale de Football Association, 2010b). First, 32 teams are determined via a qualification stage (within the confederations) for the group stage, i.e., the main World Cup 2010 tournament carried out in June and July 2010 in South Africa. Table 2 lists the 32 teams as drawn into eight groups, labeled A through H. Each group of four plays a round-robin—every team plays every other team, for a total of six matches within the group—and the top two teams in each group advance to the next stage, the round of 16. The eight winners of the round of 16 reach the quarter-final. The four winners of the quarter-finals reach the semi-finals. The winners of the semi-finals then play the final and the winner of the final is the World football champion (Fédération Internationale de Football Association, 2010c).

Using the quoted long-term odds for winning the FIFA World Cup 2010 of all 32 participating teams from 26 international bookmakers, our approach predicts Spain as the winner of the tournament with probability 17.86%; the second best team is Brazil with a winning probability of 15.27%.

The paper is organized into four sections: Section 2 gives a description of our method which is applied in Section 3. Section 4 concludes the paper with a brief discussion.

2 Method

2.1 Pre-processing

The quoted odds of the bookmakers do not represent the true chances that a team will win the tournament, because they include the stake and a profit margin, better known as the "overround" on the "book" (for further details see e.g., Henery, 1999; Forrest et al., 2005). Assuming that each bookmaker $b = 1, \ldots, 26$ applies a constant overround δ_b, the implied expected winning probabilities $p_{i,b}$ for team $i = 1, \ldots, 32$ by bookmaker b can be obtained from the raw quoted odds $rawodds_{i,b}$ via

$$p_{i,b} = \frac{1}{rawodds_{i,b}\,(1 + \delta_b)}, \tag{1}$$

where δ_b is chosen such that $\sum_i p_{i,b} = 1$.

2.2 Modeling

In order to model the expected winning probabilities $p_{i,b}$ for each team $i = 1, \ldots, 32$ and bookmaker $b = 1, \ldots, 26$, as derived from the raw quoted odds, straightforward linear models are not appropriate as the $p_{i,b}$ necessarily lie within the unit interval. Therefore, we follow the standard technique of employing a suitable link function to transform probabilities to the real line and then using linear models for the transformed data. Various link functions would be conceivable; standard choices include the logit or probit link function. In the following, we employ the logit link throughout; using the probit link instead would lead to qualitatively similar results.

On the transformed logit scale, an intuitive and straightforward strategy would be to compute team-wise means for the consensus and team-wise standard deviations for the disagreement across bookmakers (as suggested by, e.g., Zarnowitz and Lambros, 1987). In our application, this simple strategy might be appropriate because we could expect the teams to be sufficiently different and the bookmakers to have rather similar information about the teams. However, from a statistical point of view it would be interesting to investigate whether this simple strategy is sufficient or can be improved by including additional effects (e.g., pertaining to the bookmakers), or by using a more parsimonious parametrization which still gives a good approximation of the underlying data-generating process. Therefore, we employ a stochastic model class that captures the underlying probability distribution on a logit scale and contains the simple strategy as a special case. We assume additive and normally distributed "errors" on the logit scale, providing a natural way for assessment of means and variances in the models.

The model relates the expected winning logits $\text{logit}(p_{i,b})$ to the (unobservable) "true" winning logits $\text{logit}(p_i)$ for team i, reflecting the bookmakers consensus, plus an additional (unobservable) normally-distributed error term $\epsilon_{i,b}$ of bookmaker b for team i, reflecting the disagreement across the bookmakers.

Table 1: Mixed-effects models for $\text{logit}(p_{i,b})$ of team i by bookmaker j with different fixed and random effects, where ν is the intercept, μ_j is the effect of bookmaker j, α_i is the effect of team i, $\beta_{g(i)}$ is the effect of group g of team i, $\gamma_{c(i)}$ is the effect of confederation c of team i, and Z_{ij} is a standardized error. Each model is evaluated by the log-likelihood value (logLik), the number of estimated parameters (df), and the BIC.

	Team	Bookmaker	Group	Confederation	logLik	df	BIC
1	none	none	none	none	−1544.63	2	3102.71
2	none	fixed	none	none	−1542.96	27	3267.46
3	none	random	none	none	−1544.67	3	3109.51
4	none	random	fixed	none	−1527.84	10	3122.92
5	none	random	none	fixed	−1329.27	8	2712.32
6	none	random	fixed	fixed	−1283.69	15	2668.24
7	fixed	none	none	none	41.42	33	139.05
8	fixed	fixed	none	none	124.83	58	140.31
9	fixed	random	none	none	86.55	34	55.51
10	random	none	none	none	−87.04	3	194.26
11	random	fixed	none	none	−6.84	28	201.94
12	random	fixed	fixed	none	−6.17	35	247.68
13	random	fixed	none	fixed	1.70	33	218.49
14	random	fixed	fixed	fixed	3.52	40	261.91
15	random	random	none	none	−1544.64	4	3116.18

In order to capture these latent quantities by a linear mixed-effects model, we allow the true winning logits to depend on a team effect α_i, a group effect $\beta_{g(i)}$, a confederation effect $\gamma_{c(i)}$ for confederation c of team i, as well as an overall intercept ν. The error can additionally depend on μ_b, the mean effect of bookmaker b. In summary, this can be written as

$$\begin{aligned} \text{logit}(p_{i,b}) &= \text{logit}(p_i) + \epsilon_{i,b} && (2) \\ &= \nu + \alpha_i + \beta_{g(i)} + \gamma_{c(i)} + \mu_b + \sigma Z_{i,b}, && (3) \end{aligned}$$

where $Z_{i,b}$ is a standardized error and σ is the standard deviation. Even if contrasts are employed, this model is overspecified when all four effects α_i, $\beta_{g(i)}$, $\gamma_{c(i)}$ and μ_b are included as fixed effects due to the the dependence of group $g(i)$ and confederation $c(i)$ on the team i.

In order to overcome this methodological issue, there are various conceivable solutions which can also be motivated by subject-matter considerations: (a) The confederation effects could be omitted signaling that all teams are sufficiently different. Note that the full team effect then still captures confederation differences. (b) Alternatively, the team effect could be specified as a random effect (with zero mean) conveying that the winning logits for each team deviate randomly from the mean as captured by the remaining effects (e.g., by fixed confederation differences). (c) A random effect for the bookmakers would be conceivable implying that the bookmakers' odds deviate randomly from the mean as captured by the remaining effects. Combinations of the ideas (a)–(c) lead to 15 different mixed-effects models. Table 1 specifies the different effects for each model. In order to find a parsimonious model which still gives a good approximation of the underlying data-generating process, standard model selection methods can be employed. We use the Bayesian information criterion (BIC; Pinheiro and Bates, 2000).

3 Results

Based on the modeling approach discussed above, we first choose the final mixed-effects model (Section 3.1) from which the associated probabilities $\widehat{p}_i$ for winning the FIFA World Cup 2010 for all teams are derived

(Section 3.2). In addition, we investigate the models' implications on the strengths of the groups of the preliminaries (Section 3.3) and the confederations of the participating teams (Section 3.4), respectively.

3.1 Model selection

Fitting all 15 conceivable mixed-effects models discussed in the previous sections yields the results in Table 1 which provides the log-likelihood, number of parameters, and associated BIC.

The best model emerging from the BIC selection is Model 9 (BIC = 55.51), containing a fixed team effect (and hence no additional group or confederation effect) and a random bookmaker effect. Moreover, the three best models (7–9) all have a fixed team effect, followed by Models 10–14 which have a random team effect and perform slightly worse. Finally, all models which have no team effect at all (or just try to capture it by group and/or confederation effects) perform clearly worse. In summary, this conveys that the bookmakers employ knowledge about each individual team when fixing their odds (rather than being mainly determined by group or confederation considerations). Furthermore, the fact that the bookmaker effect can be captured well as a random effect suggests that there are no large systematic deviations between the bookmakers. In retrospect, this model probably comes at no surprise because its interpretation is so intuitive: All teams are expected to perform differently and the bookmakers' expectations just vary randomly around some common overall $\text{logit}(\widehat{p}_i) = \widehat{\nu} + \widehat{\alpha}_i$. It is reassuring that this intuitive model has been selected from a more general class of models, where some of the alternatives would have also had appealing interpretations.

3.2 Probability of winning the FIFA World Cup 2010

The bookmaker consensus for the FIFA World Cup 2010 can be derived from the best model, Model 9 by using the estimated winning logits $\text{logit}(\widehat{p}_i) = \widehat{\nu} + \widehat{\alpha}_i$. This consensus information on the logit scale can be easily transformed to the associated winning probabilities $\widehat{p}_i$ of winning the tournament for all 32 participating teams which are shown in Table 2. This bookmaker consensus information is compared with the FIFA/Coca Cola World rating (Fédération Internationale de Football Association, 2010a) as well as the World Football Elo Rating (Advanced Satellite Consulting Ltd, 2008) of the 32 participating teams. Additionally, the eight origin groups of the preliminaries, and the football confederation of the teams are shown. Spain is the best team of the 32 teams and has the highest probability (17.86%) of winning the tournament. The expected runner-up is Brazil (15.27%), the favorite according to the FIFA and ELO rating. The top two are followed by England (winning probability 11.95%), and Argentina (10.80%), who are expected to play an exciting match for the third place. The team with the lowest winning probability (0.06%) is New Zealand. The first nine teams are assigned to the UEFA or CONMEBOL. Three teams out of the first ten are members of group G which implies that this group is the strongest, i.e., most competitive group, only two teams can advance to the next round. Using the information about the given tournament schedule in combination with the winning probabilities of the participating teams (Table 2) the following 16 teams (eight group-winners and eight runners-up) are expected to play the first knock-out round: France, Mexico (group A), Argentina, Nigeria (B), England, United States (C), Germany, Serbia (D), Netherlands, Cameroon (E), Italy, Paraguay (F), Brazil, Portugal (G), Spain, and Chile (H). These 16 teams are not the 16 participants with the highest winning probabilities implying that the group drawn has an effect to the tournament outcome. In this paper we focus on predicting the winner of the tournament. Nevertheless, if someone is interested in the dynamic of the tournament we suggest to use a simulation approach (see Leitner et al., 2010) to determine, e.g., the probability that a specific team reach the semi-finals.

3.3 Which is the strongest group of the preliminaries?

The forecast of the expected 16 teams qualifying for the first knock-out round implies a group effect. Although our model contains a fixed team effect and hence no group effect (redundant information), we can answer the question: "Which is the strongest group of the preliminaries?". The estimated team effects $\widehat{\alpha}_i$ imply a "group effect". To derive this group effect we calculate the difference on the logit scale between the average team effect in group g and the overall mean ν of all 32 participating teams of the tournament. Table 3 shows these group effects for all eight groups (A–H). The group with the best chance

Table 2: Estimated winning probabilities $\widehat{p}_i$, associated winning logits $\text{logit}(\widehat{p}_i)$ (reflecting the bookmakers consensus), the FIFA/Coca Cola World rating (Fédération Internationale de Football Association, 2010a) and the World Football Elo Rating (Advanced Satellite Consulting Ltd, 2008) for all 32 participating teams of the FIFA World Cup 2010. Additionally, the eight origin groups of the preliminaries, and the football association of the teams are shown.

	$\widehat{p}_i(\%)$	$\text{logit}(\widehat{p}_i)$	FIFA	ELO	Group	Confederation
Spain	17.86	−1.53	1565	2078	H	UEFA
Brazil	15.27	−1.71	1611	2085	G	CONMEBOL
England	11.95	−2.00	1068	1972	C	UEFA
Argentina	10.80	−2.11	1076	1899	B	CONMEBOL
Netherlands	7.39	−2.53	1231	2011	E	UEFA
Italy	5.84	−2.78	1184	1922	F	UEFA
Germany	5.76	−2.80	1082	1919	D	UEFA
France	4.48	−3.06	1044	1855	A	UEFA
Portugal	3.22	−3.40	1249	1838	G	UEFA
Cote d'Ivoire	2.57	−3.63	856	1725	G	CAF
Serbia	1.50	−4.18	947	1833	D	UEFA
Chile	1.46	−4.21	888	1851	H	CONMEBOL
Paraguay	1.16	−4.45	820	1730	F	CONMEBOL
USA	1.13	−4.47	957	1741	C	CONCACAF
Ghana	1.06	−4.54	800	1682	D	CAF
Mexico	1.03	−4.57	895	1870	A	CONCACAF
Uruguay	0.85	−4.76	899	1819	A	CONMEBOL
Cameroon	0.81	−4.80	887	1698	E	CAF
Nigeria	0.75	−4.89	883	1696	B	CAF
Denmark	0.71	−4.95	767	1761	E	UEFA
South Africa	0.64	−5.05	392	1648	A	CAF
Australia	0.61	−5.09	886	1766	D	AFC
Greece	0.51	−5.28	964	1726	B	UEFA
Switzerland	0.45	−5.40	866	1746	H	UEFA
South Korea	0.38	−5.57	632	1766	B	AFC
Slovakia	0.33	−5.72	777	1626	F	UEFA
Slovenia	0.30	−5.81	860	1648	C	UEFA
Japan	0.27	−5.89	682	1690	E	AFC
Algeria	0.19	−6.26	821	1536	C	CAF
Honduras	0.12	−6.69	734	1725	H	CONCACAF
Korea DPR	0.07	−7.21	285	1533	G	AFC
New Zealand	0.06	−7.41	410	1531	F	OFC

Table 3: Group effects for the eight groups of the preliminaries A–H.

A	B	C	D	E	F	G	H
0.103	−0.002	−0.172	0.309	−0.082	−0.630	0.471	0.003

to include the winner is group G (0.471), followed by group D (0.309). Despite the fact that group H includes the bookmakers' favorite of winning the World Cup (Spain), group H follows just on the forth

Table 4: Confederation effects and number of qualified teams for the 6 team's confederations.

OFC	AFC	CONCACAF	CAF	UEFA	CONMEBOL
−2.95	−1.48	−0.78	−0.40	0.66	1.01
1	4	3	6	13	5

position (0.003). Group B can be interpreted as the average group (−0.002). The smallest chance to include the winner has group F (−0.630).

3.4 Which is the strongest confederation?

In addition to the group effect, the estimated team effects $\widehat{\alpha}_i$ also imply a "confederation effect". We derive this confederation effect by computing the difference on the logit scale between confederation c and the overall mean ν of all 32 participating teams of the tournament. This result can be used to rank the 6 confederations of the participating teams and give an answer to the interesting question: "Which is the strongest European confederation?". Table 4 shows this confederation effects and the number of qualified teams for the 6 confederations. CONMEBOL is with five participating teams the strongest confederation (1.01), followed by the UEFA (0.66) which has the most participating teams (13).

4 Discussion

This paper investigates a general model class for the "unknown" true logits for winning a sports tournament based on quoted bookmakers odds. The flexible model framework allows for capturing different effects (e.g., team, group, and confederation). The variety of possible linear mixed-effects models can be derived to a parsimonious model which still gives a good approximation of the underlying data-generating process by a standard model selection approach (BIC). This model is then used to forecast the outcome of a sports tournament. Here we apply this method to forecast the winner of the FIFA World Cup 2010. The model selection approach yields a model with a fixed team effect plus a random bookmaker-specific deviation forecasting team Spain as the winner (winning probability: 17.86%). Furthermore, we give answers to the questions: "Which is the strongest group of the preliminaries?" (Answer: Group G) and "Which is the strongest football confederation?" (Answer: CONMEBOL). Luckily for all football supporters, football is a game and cannot be truly predicted using rational strategies and statistical methods. In fact, Ben-Naim et al. (2006) argue that football (along with baseball) is the most random and competitive popular sport. Thus, one prediction appears to be certain: an exciting FIFA World Cup 2010.

Computational details

All computations were carried out in the R system (version 2.9.2) for statistical computing (R Development Core Team, 2010). In particular, the R package nlme version 3.1-92 (Pinheiro et al., 2009) was used for the estimation of the mixed-effects models (see Pinheiro and Bates, 2000).

References

Advanced Satellite Consulting Ltd. The World Football Elo Rating System, 2008. URL `http://www.eloratings.net/`. [Online; accessed 2010-05-29].

E. Ben-Naim, F. Vazquez, and S. Redner. Parity and predictability of competitions. *Journal of Quantitative Analysis in Sports*, 2(4):1–12, 2006.

M. Dixon and P. Pope. The value of statistical forecasts in the UK association football betting market. *International Journal of Forecasting*, 20:697–711, 2004.

D. Dyte and S. Clarke. A rating based Poisson model for World Cup soccer. *The Journal of the Operational Research Society*, 51(8):993–998, 2000.

Fédération Internationale de Football Association. FIFA/Coca-Cola World Ranking, 2010a. URL http://www.fifa.com. [Online; accessed 2010-05-29].

Fédération Internationale de Football Association. Football confederations, 2010b. URL http://www.fifa.com/aboutfifa/federation/confederations/index.html. [Online; accessed 2010-06-05].

Fédération Internationale de Football Association. FIFA World Cup 2010, 2010c. URL http://www.fifa.com/worldcup/matches/index.html. [Online; accessed 2010-06-05].

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

J. Goddard and I. Asimakopoulos. Modelling football match results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23:51–66, 2004.

Goldman Sachs. The World Cup and economics 2010, 2010. URL http://www2.goldmansachs.com/ideas/global-economic-outlook/the-world-cup-2010.html. [Online; accessed 2010-06-05].

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

J.P.Morgan. England to win the World Cup. Report, Europe Equity Research, May 2010.

C. Leitner, A. Zeileis, and K. Hornik. Forecasting the winner of the UEFA Champions League 2008/09. In R. Koning and P. Scarf, editors, *Proceedings of the 2nd International Conference on Mathematics in Sport – IMA Sport 2009*, pages 94–99, 2009. ISBN: 979-0-905091-21-1.

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. doi: 10.1016/j.ijforecast.2009.10.001.

J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer-Verlag New York, New York, USA, 2000. ISBN 0-387-98957-9.

J. Pinheiro, D. Bates, S. DebRoy, and D. Sarkar. *nlme: Linear and Nonlinear Mixed Effects Models*, 2009. URL http://CRAN.R-project.org/package=nlme. R package version 3.1-92.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL http://www.R-project.org/. ISBN 3-900051-07-0.

UBS Wealth Management Research. And the world champion is. . . , 2010. URL http://www.ubs.com/1/e/bank_for_banks/news/topical_stories/edition_10.html. [Online; accessed 2010-06-05].

V. Zarnowitz and L. A. Lambros. Consensus and uncertainty in economic prediction. *Journal of Political Economy*, 95:561–621, 1987.