Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 26 (2010) 471–481
international journal of forecasting
www.elsevier.com/locate/ijforecast

Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008

Christoph Leitner*, Achim Zeileis, Kurt Hornik

Department of Statistics and Mathematics, WU Wirtschaftsuniversität Wien, Augasse 2–6, 1090 Wien, Austria

Abstract

Different methods for assessing the abilities of participants in a sports tournament, and their corresponding winning probabilities for the tournament, are embedded in a common framework and their predictive performances compared. First, ratings of abilities (such as the Elo rating) are complemented with a simulation approach which yields winning probabilities for the full tournament. Second, tournament winning probabilities are extracted from bookmakers' odds using a consensus model, and the underlying abilities of the competitors are then derived by an "inverse" application of the tournament simulation. Both techniques are employed for forecasting the results of the *European football championship 2008* (UEFA EURO 2008), for which the consensus model based on bookmakers' odds outperforms methods based on both the Elo rating and the FIFA/Coca Cola World rating. Moreover, the bookmaker consensus model correctly predicts that the final will be played by the teams from Germany and Spain (with a probability of about 20.5%), while showing that both finalists profit from being drawn in groups with relatively weak competitors.

Keywords: Sports forecasting; EURO 2008; Bookmakers odds; Elo rating; Abilities

1. Introduction

Due to the growing popularity of online sports betting, there has been an increasing amount of interest in the analysis and forecasting of competitive sports. Forecasts of sports events are often based on one of two types of information: ratings or rankings of the competitors' ability/strength, and bookmakers' odds for winning a competition between two or more contestants. Here, we show how both types of forecasts – winning probabilities and underlying abilities – can be derived from both types of information – ability ratings and bookmakers odds. Their predictive performances are assessed in an empirical study forecasting the winner of the European football championship 2008 (UEFA EURO 2008).

Sports ratings or rankings are typically derived by suitably aggregating information about the competitors' previous performances, and are often found to have predictive power in forecasting tasks. Boulier and Stekler (1999) show that rankings provide useful

* Corresponding author.
E-mail address: christoph.leitner@wu.ac.at (C. Leitner).

0169-2070/$ - see front matter © 2009 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2009.10.001

information for forecasting basketball tournaments and tennis matches. Lebovic and Sigelman (2001) analyze the predictive accuracy of college football rankings, while Suzuki and Ohmori (2008) use the FIFA/Coca Cola World rating (Fédération Internationale de Football Association, 2008), one of the most popular rating systems in soccer, as a tool for forecasting the last four FIFA World Cups (1994, 1998, 2002, 2006). In addition, Dyte and Clarke (2000) use the FIFA ratings to predict the distribution of scores in international soccer matches. Another popular rating system is the Elo rating system, which was originally developed for calculating the relative skills of chess players (e.g., Elo, 2008), and which has subsequently also been applied to various sports, including soccer. For example, Song, Boulier and Stekler (2009) apply it as one method for forecasting the winner of single American Football games. Edmans, García and Norli (2007) also select important soccer games based on the World Football Elo Ratings.

Bookmakers' odds represent a rather different type of rating to the methods above. Based on the bookmakers' expert judgments (which typically include, but are not limited to, knowledge about past performances) the odds reflect the expected outcomes in a particular competition, where the bookmakers have strong economic incentives to rate the competitors correctly. A bias (in either direction, too good or too bad) will cost them money, or, in other words, will reduce their profits. Hence, bookmakers can be seen as experts in the matter of sports ratings (see Pope & Peel, 1989), and are likely to provide good predictions (Forrest & Simmons, 2000). This has been confirmed by various empirical studies in which fixed odds have been found to be an efficient forecasting instrument for the outcome of single matches (e.g., Boulier & Stekler, 2003; Dixon & Pope, 2004; Forrest, Goddard & Simmons, 2005; Song, Boulier & Stekler, 2007; Spann & Skiera, 2009; Vlastakis, Dotsis & Markellos, 2009).

One advantage of employing bookmakers' odds is that winning probabilities for the corresponding competition can easily be derived, but this is not straightforward for many of the ability ratings. However, if abilities are measured on a ratio scale (or can be transformed to such), winning probabilities for pairwise matches can be derived using the approach of Bradley and Terry (1952). The Elo rating, from which pairwise winning probabilities for single matches can be obtained (e.g., Edmans et al., 2007; Stefani & Pollard, 2007), is notable in this respect. Thus, when the competition of interest is a single match, forecasts based on ability ratings and bookmakers' odds can easily be compared. However, this is not the case if the competition is a more complex tournament for which the bookmakers' odds, by their prospective nature, can include additional effects such as group draws or seedings. In order to link forecasts of abilities (associated with pairwise winning probabilities) with winning probabilities for sports tournaments, we suggest a simulation approach that allows us to (approximately) map abilities to winning probabilities, and vice versa.

In order to compare forecasts based on ability ratings and bookmakers' odds, we apply them to the UEFA EURO 2008, one of the world's biggest sporting events, that took place in June 2008 in Austria and Switzerland. For the odds-based forecasts, quoted bookmakers' odds for the 16 participating teams were obtained from 45 international bookmakers prior to the tournament (on 2008-04-21), and aggregated to form a consensus model. This is compared with both the forecasts from the World Football Elo rating (also considered in a note from UBS Wealth Management Research Switzerland, 2008, for the prediction of the EURO 2008) and the ranking implied by the FIFA/Coca Cola World rating (also employed in a note from Raiffeisen Zentralbank, 2008), both of which were also obtained on 2008-04-21. Forecasts based on these approaches were first obtained by an analysis in May 2008, prior to the tournament (see Leitner, Zeileis & Hornik, 2008, for a technical report with these preliminary findings), and then reassessed after the end of the tournament. In this ex post comparison, the bookmaker consensus model performs best, and predicts the correct final pair of teams (Germany vs. Spain, with a probability of about 20.5%). Furthermore, the results provide many further insights into the effects of the group draw in the tournament, clearly showing that the two finalists come from groups with relatively weak competitors.

The remainder of this paper is organized as follows: Section 2 discusses some basic features of sports ratings, bookmakers' odds, and sports tournaments. Section 3 provides a description of the data and the tournament for the EURO 2008, for which the various forecasts are obtained and assessed in Section 4. Section 5 concludes the paper.

2. Ratings of (prob)abilities in sports tournaments

2.1. Sports ratings

Ratings of "abilities" or "strengths". In competitive sports, players or teams, as well as their supporters, are interested in ratings of the competitors as a measure of their abilities or strengths. A common strategy for deriving suitable ratings employs adaptive schemes which update assessments based on historic performances following the availability of data about current performances. Typical examples of this include the FIFA/Coca Cola World rating in soccer and the ATP (Association of Tennis Professionals) rating in tennis (see Stefani, 1997, for an overview). Some ratings are based on a simple points system, while others employ statistical models, such as the Elo rating (Elo, 2008), which implies pairwise winning expectancies (see Joe, 1991). A natural application of ability ratings is to employ them for forecasting performances in future matches (e.g., Song et al., 2009). In some sports, ratings are also used for deriving seedings, which can be used for forecasting in turn, as was done by Boulier and Stekler (1999).

Bookmakers' odds as ratings of winning probabilities. A rather different source of "ratings" of competitors in sports are bookmakers' odds. Unlike the ratings discussed above, these are not derived directly from past performances, but emerge from "expert" knowledge. Of course, this typically encompasses knowledge about past results, but may also take into account expectations about future events. Due to the increasing popularity of online sports betting, bookmakers' odds are a type of data that is both abundant and easily available, and that has been successfully employed in forecasts of single matches (e.g., Boulier & Stekler, 2003; Dixon & Pope, 2004; Forrest et al., 2005; Song et al., 2007; Spann & Skiera, 2009; Vlastakis et al., 2009). Another important difference between bookmakers' odds and the ability ratings discussed above is that they are an assessment of outcome probabilities (e.g., winning probabilities in the case of sports tournaments), rather than of the underlying abilities. However, the raw quoted bookmakers' odds are not "honest" odds, but are the payout amounts for successful bets. This has two important implications: (1) they still contain the stake, i.e., the payment for placing the bet (the "1" in Eq. (1) below); and (2) more importantly, the bookmakers' odds contain a profit margin, the so-called "overround", which means that the "true" underlying odds are actually larger (see e.g., Forrest et al., 2005; Henery, 1999). Assuming that the overround δ is constant across all possible outcomes (e.g., the same for all competitors winning a tournament), it can be computed by restricting the corresponding probabilities to sum to unity. More precisely, the raw quoted odds $rawodds_i$ for event i can be adjusted to $odds_i$, and then transformed to probabilities p_i via:

$$odds_i = (rawodds_i - 1)\,\delta, \tag{1}$$

$$p_i = 1 - \frac{odds_i}{1 + odds_i}. \tag{2}$$

Then, δ can be chosen such that $\sum_i p_i = 1$. (Note that the complementary probabilities have to be used, as the bookmakers' odds represent the expectations of an outcome not occurring.) In the case of winning odds for a tournament, this means that the implied winning probabilities can easily be derived from the quoted odds for all competitors.

2.2. Sports tournaments

Pairwise comparisons. In many sports disciplines, winners and losers are determined by pairwise comparisons, called matches or games. Clearly, the outcome of a match depends on the current abilities of the two competitors. Given abilities which are measured on a ratio scale, the classical method of computing winning probabilities from abilities is the Bradley and Terry (1952) approach, which derives the probability of competitor i beating competitor j as:

$$\pi_{i,j} = \frac{ability_i}{ability_i + ability_j} \quad (i \neq j), \tag{3}$$

where $ability_i$ is the ability of team i on a ratio scale. However, for many sports rating systems it is not clear what the underlying measurement scale is. A notable exception is the Elo rating (Elo, 2008), which uses a similar approach for obtaining winning expectations. Hence, as is discussed in detail below, Elo ratings can easily be transformed to abilities in the sense of Eq. (3).

Tournament schedule. If a winner is determined from a group (rather than just a pair) of competitors, this is typically accomplished by using a sequence

of pairwise comparisons, called a tournament. There are various designs available for constructing suitable schedules for such a tournament (for a discussion, see Scarf & Bilbao, 2006). In a round-robin tournament, where each competitor (player or team) plays each other, it is obvious that the strongest competitor has the highest probability of winning in each pairwise comparison, and therefore the greatest chance of winning the tournament, followed by the second strongest competitor, and so on. However, for other tournament schedules the strongest competitor does not necessarily have the highest probability of winning. For example, if the tournament schedule is based on a draw of a group phase and/or a knockout phase, some competitors might be favored/penalized by being drawn together with relatively weak/strong competitors. However, when the tournament schedule and the abilities of the participants are known, it is straightforward (in principle) to compute the associated probabilities of winning, based on the pairwise probabilities from Eq. (3), by applying conditional probabilities to all possible tournament "paths". As an explicit enumeration of all paths can be burdensome, the winning probabilities can also be approximated easily by simulating a large number of tournament runs (such as 100,000), and then assessing the empirical winning proportions $\widetilde{p}$ for each competitor, as given in Table 1.

The resulting (approximated) winning probabilities $\widetilde{p}(ability)$ then capture all "tournament effects" induced by the schedule as well. Note that this approach currently models the contestants' abilities as constant over the course of the competition, but it could be further enhanced to accommodate hypothesized patterns of changes in abilities. Also, this generic simulation setup might require an adaptation to some details of a specific tournament; e.g., for EURO 2008 potential ties after the group phase need to be resolved (as is described in detail in Section 4).

3. EURO 2008: Data and tournament description

3.1. Data

Elo ratings. The World Football Elo Ratings (Advanced Satellite Consulting Ltd, 2008), Elo ratings for short, have been collected from http://www.eloratings.net/ for all 16 teams participating in the EURO 2008 (accessed 2008-04-21). In contrast to many other sports rating systems (such as the FIFA ratings discussed below), the Elo ratings imply winning expectancies for pairwise comparisons (see Elo, 2008, Equation 46). The probability that team i beats team j can be related to

$$\pi_{i,j} = \frac{1}{10^{-(Elo_i - Elo_j)/400} + 1} \quad (i \neq j), \tag{4}$$

where Elo_i and Elo_j are the Elo ratings for teams i and j, respectively. For the home teams (i.e., Austria and Switzerland in the EURO 2008), 100 rating points are added to the respective Elo ratings (Advanced Satellite Consulting Ltd, 2008). Thus, the Elo ratings are essentially on a $\log_{10}$ scale, which is somewhat different from the standard Bradley and Terry (1952) model. However, using Eqs. (3) and (4), it is easy to provide a transformation to log-abilities in the Bradley–Terry sense, which thus imply the same pairwise winning probabilities $\pi_{i,j}$. As the log-abilities are only defined up to a constant γ, we choose γ such that they are on a logit scale:

$$\log\left(ability_i^{(ELO)}\right) = \frac{\log(10)}{400} Elo_i + \gamma, \tag{5}$$

$$\sum_i \text{logit}^{-1}\left(\log\left(ability_i^{(ELO)}\right)\right) = 1, \tag{6}$$

where Eq. (6) implies $\gamma = -13.496$ for the EURO 2008 data, log is the natural logarithm, and logit^{-1} denotes the inverse of the logit function. The resulting Elo log-abilities are provided in Table 2, where the logit scale facilitates a comparison with the logits of the tournament winning probabilities derived below.

Bookmakers' odds. Long-term odds for winning the EURO 2008 were obtained from the websites of 45 international bookmakers for all 16 participating teams on 2008-04-21. These include all of the 50 European top-selling online sports bookmakers who offered odds for this event. Prior to all further analysis, the odds are adjusted by removing the stake and a bookmaker-specific overround (see Eq. (1)), then transformed to winning probabilities by means of Eq. (2). This yields tournament winning probabilities $p_{i,b}$ for $i = 1, \ldots, 16$ teams and $b = 1, \ldots, 45$ bookmakers, which reflect the bookmakers' beliefs about the outcome of the EURO 2008.

FIFA ratings. The FIFA/Coca Cola World ratings (Fédération Internationale de Football Association,

Table 1
Deriving simulated winning probabilities $\widetilde{p}(ability)$ for a tournament from the abilities of all competitors.

ability abilities of all competitors	→	$\pi_{i,j}$ pairwise winning probabilities for all matches	→	100,000 runs tournament simulations	→	$\widetilde{p}$ simulated winning probabilities for tournament

2008), or FIFA ratings for short, were retrieved from http://www.fifa.com/ for all 16 participating teams on 2008-04-21. These ratings capture the abilities of the teams but use an unknown scale, so that it is not straightforward to compute either pairwise winning probabilities $\pi_{i,j}$ or tournament winning probabilities p_i (see McHale & Davies, 2007, for an approach to building more complex statistical models based on the FIFA rating). Therefore, in the following, the FIFA ratings are only employed for comparisons as a ranking (rather than rating).

3.2. The tournament

The UEFA EURO 2008 is a tournament in which 52 European national football teams (UEFA's members) compete in a multi-stage modus (qualification, group and knockout stages) in order to determine the European champion. First, 16 teams are determined for the group stage, i.e., the main EURO 2008 tournament carried out in June 2008 in Austria and Switzerland, via a qualification stage. Table 2 lists the 16 teams, as drawn into four groups, labeled A through D. 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 quarter-final. The winner of group A plays against the second best team of group B (first quarter-final) and the winner of group B plays against the second best team of group A (second quarter-final). Analogously, the winner of group C plays against the second best team of group D (third quarter-final) and the winner of group D plays against the second best team of group C (fourth quarter-final). The four winners of the quarter-finals reach the semi-finals, where the winner of the first quarter-final plays against the winner of the second one and the winner of the third quarter-final plays against the winner of the fourth. The winners of the semi-finals then play the final, and the winner of the final is the European football champion (Union of European Football Associations, 2009).

4. Forecasting of the EURO 2008

In this section, forecasts of team (log-)abilities and winning probabilities for the EURO 2008 tournament are obtained based on the Elo ratings and the bookmakers' odds, respectively. The resulting four quantities are compared with the actual result of the tournament, and the best-performing method is analyzed in more detail.

4.1. Forecasting based on the Elo ratings

As was argued in Sections 2 and 3, the Elo ratings Elo_i ($i = 1, \ldots, 16$) represent an assessment of the current ability/strength of the teams participating the EURO 2008. By construction, pairwise probabilities $\pi_{i,j}$ for all combinations of participants can be obtained. Furthermore, to approximate winning probabilities that include tournament effects such as the group draw, the empirical winning proportions from 100,000 simulated tournaments are used:

$$ability_i^{(ELO)} = \exp\left(\frac{\log(10)}{400} Elo_i - 13.496\right), \tag{7}$$

$$p_i^{(ELO)} = \widetilde{p}\left(ability^{(ELO)}\right)_i. \tag{8}$$

Thus, $ability^{(ELO)}$ is the vector of abilities (in the Bradley–Terry sense) which the tournament simulations are carried out based on. The results for all teams are reported in Table 2.

By adopting the classical Bradley–Terry model, the simulation of each match yields only a winner and a loser, without the possibility of a tie or any further information about the number of goals or the goal difference. This is sufficient for the knock-out stage of the tournament, as it reflects the fact that the actual

Table 2
Log-abilities, winning probabilities, and corresponding logits of all teams for the EURO 2008 based on the Elo rating (ELO) and the bookmaker consensus model (BCM). The ELO log-abilities are computed directly from the Elo ratings and the winning probabilities are derived via simulation. The BCM logits are estimated by team-wise means of bookmaker log-odds, and the corresponding log-abilities are found by "inverse" simulation. The rows are sorted by the BCM winning probabilities.

	$\log(ability_i)$		p_i (%)		$\text{logit}(p_i)$		
	ELO	BCM	ELO	BCM	ELO	BCM	Group
Germany	−2.34	−2.33	15.99	17.45	−1.66	−1.55	B
Spain	−2.25	−2.41	13.14	12.21	−1.89	−1.97	D
Italy	−1.97	−2.40	18.28	11.34	−1.50	−2.06	C
Portugal	−2.95	−2.54	3.36	9.97	−3.36	−2.20	A
France	−2.09	−2.50	14.08	9.14	−1.81	−2.30	C
Netherlands	−2.33	−2.62	8.29	6.77	−2.40	−2.62	C
Croatia	−2.86	−2.77	5.03	6.72	−2.94	−2.63	B
Czech Republic	−2.67	−2.74	7.17	5.88	−2.56	−2.77	A
Switzerland	−2.79	−2.88	5.18	3.92	−2.91	−3.20	A
Greece	−2.93	−2.91	2.76	3.31	−3.56	−3.37	D
Sweden	−3.32	−2.98	0.77	2.87	−4.86	−3.52	D
Russia	−3.42	−3.00	0.55	2.72	−5.20	−3.58	D
Turkey	−3.27	−3.06	1.30	2.26	−4.33	−3.77	A
Romania	−2.72	−3.04	2.77	2.12	−3.56	−3.83	C
Poland	−3.35	−3.19	1.19	2.05	−4.42	−3.87	B
Austria	−3.93	−3.85	0.14	0.93	−6.55	−4.67	B

matches always have a winner (if necessary through overtime and penalties). However, for the group phase within the simulation this approach might result in tied teams. If necessary, we resolve such ties through additional "fictitious" matches between the tied teams in order to obtain unique winners and runner-ups of the groups.

Our simulation method could be extended by using more elaborate simulation techniques, including ties and numbers of goals, e.g., a model where the team scores follow independent Poisson distributions (e.g., Dixon & Coles, 1997; Dyte & Clarke, 2000; Maher, 1982), or an ordered probit regression model (Goddard & Asimakopoulos, 2004).[1]

According to the Elo rating, Italy is the strongest team ($\log(ability^{(ELO)}) = -1.97$), and also has the highest probability of winning the tournament ($p^{(ELO)} = 18.28\%$). However, the second strongest team, France, has only the third highest winning probability ($\log(ability^{(ELO)}) = -2.09$, $p^{(ELO)} = 14.08\%$), while Germany is only the fifth strongest but has the second highest winning probability ($\log(ability^{(ELO)}) = -2.34$, $p^{(ELO)} = 15.99\%$). Thus, Germany clearly profits from being drawn in a group (B) with weaker competitors, while France is at a certain disadvantage from being placed in a group (C) with strong competitors, such as Italy. This tournament effect can conveniently be assessed by comparing the differences between the teams' log-abilities and their winning logits, respectively (as both measurements have been constructed such that they are on a logit scale). For example, Italy's margin over Germany of 0.37 ($= -1.97 - (-2.34)$) is reduced to 0.16 ($= -1.5 - (-1.66)$) by including tournament effects, while France's margin over Germany of 0.25 is reversed to −0.15. Furthermore, it is worth noting that Spain, the favorite in group D, has the fourth highest winning probability ($p^{(ELO)} = 13.14\%$), while Austria has the lowest chance of winning the EURO 2008 ($p^{(ELO)} = 0.14\%$), notwithstanding its potential home advantage (see e.g., Clarke & Norman, 1995; Forrest, Beaumont, Goddard & Simmons, 2005).

4.2. Forecasting based on bookmakers' odds

When appropriately adjusted and transformed, as described in Sections 2 and 3, the bookmakers' odds yield expected winning probabilities $p_{i,b}$ for each

[1] However, all approaches should give reasonable approximations of the probabilities of being promoted to the next round.

team $i = 1, \ldots, 16$ and bookmaker $b = 1, \ldots, 45$. In the following, a single forecast of the winning probability of each team is obtained by the aggregation of the $p_{i,b}$ values across bookmakers. Subsequently, a vector of underlying team abilities is found by the "inverse" application of the simulation approach adopted above.

The bookmakers' odds are prospective ratings of the performances of the 16 participating teams in the EURO 2008, and vary between the 45 bookmakers. To obtain an aggregated measure for each team, some sort of consensus between the different ratings has to be computed (for discussions of various strategies for the aggregation of forecasts, see e.g., Kolb & Stekler, 1996; Schnader & Stekler, 1991; Song et al., 2007, 2009; and Zarnowitz & Lambros, 1987). Here, we adopt a simple additive model on a logit scale

$$\text{logit}(p_{i,b}) = \text{logit}(p_i) + \epsilon_{i,b}, \tag{9}$$

where p_i is the latent winning probability for team i and $\epsilon_{i,b}$ is the deviation of bookmaker b for team i. In principle, it would be possible to refine this model further by including group effects in the winning logits $\text{logit}(p_i)$ or bookmaker-specific bias and variance in the deviation $\epsilon_{i,b}$. See Leitner, Zeileis and Hornik (2009) for an exploration of several mixed-effects models (e.g., Pinheiro & Bates, 2000) capturing different team- and bookmaker-specific effects. However, as the bookmakers' expectations about the EURO 2008 are fairly homogeneous, a straightforward fixed-effects model with zero-mean deviations $\epsilon_{i,b}$ should be appropriate. Thus, the consensus winning logits are simply means across bookmakers:

$$\widehat{\text{logit}(p_i)} = \frac{1}{45} \sum_{b=1}^{45} \text{logit}\left(p_{i,b}\right). \tag{10}$$

Transforming these logits back to the probability scale yields the bookmakers' consensus winning probabilities $p_i^{(BCM)}$. Both the probabilities and the corresponding logits for this bookmaker consensus model (BCM) are shown in Table 2. The model captures 98.21% of the variance in $p_{i,b}$, and the associated estimated standard error of $\epsilon_{i,b}$ is 0.11396.

Although forecasting the winning probabilities for the EURO 2008 is the main concern in our investigation, we are also interested in the team abilities underlying the bookmakers' expectations. The tournament schedule was already known at the time that the bookmakers' odds were retrieved, and hence should be included in the expectations about the outcome of the tournament. To strip the "tournament effects" (see Section 2.2) from this measure, we employ an "inverse" application of the simulation approach described in the previous sections. More precisely, we want to find the set of team abilities $ability_i$ ($i = 1, \ldots, 16$) that result in the simulated winning probabilities $\widetilde{p}(ability)_i$ that are as similar as possible to the consensus winning probabilities $p_i^{(BCM)}$:

$$p_i^{(BCM)} = \text{logit}^{-1}\left(\widehat{\text{logit}(p_i)}\right), \tag{11}$$

$$ability^{(BCM)} = \underset{ability}{\text{argmin}} \sum_{i=1}^{16} \left| p_i^{(BCM)} - \widetilde{p}\,(ability)_i \right|. \tag{12}$$

The minimum in the second line is determined using a local search strategy for the full vector $ability^{(BCM)}$, where 100,000 tournament runs are employed in each evaluation of $\widetilde{p}(\cdot)$. The results are reported in Table 2.

According to the BCM, Germany has the greatest chance of winning the EURO 2008 ($p^{(BCM)} = 17.45\%$), with a margin over Spain of 12.21% and over Italy of 11.34%. Thus, although there is a considerable overlap among the listings of the top five teams according to the BCM and Elo, the rankings and associated winning probabilities of these teams are rather different. Also, France (which was the second strongest team according to the Elo rating) has only the fifth highest winning probability (9.14%). Finally, the host country Austria is again expected to have the lowest winning probability (0.93%), but it is somewhat larger than the Elo forecast in absolute terms.

To investigate the tournament effect, differences in the teams' winning logits can again be compared with differences in their log-abilities. Again, this shows that Germany profits greatly from the group draw, because its margins over Spain and Italy in terms of winning logits (0.42 and 0.51, respectively) are greatly reduced in terms of log-abilities (0.08 and 0.07). Note also that this reduction is larger for Italy than for Spain, showing that Italy suffers particularly from being drawn in the strong group C (often referred to as the "group of death").

4.3. Ex post comparison of all forecasts

The previous subsections present two different types of forecasts (abilities and winning probabilities) derived from two different types of ratings (Elo rating and bookmakers' odds). As is common in forecasting, our central interest is in which strategy performs best in practice. Although this is difficult to answer because there are no "real" replications of the tournament, we can compare the forecasts with the single real outcome of the EURO 2008.

Table 3 assesses the predictive performance of all four forecasts by comparing them with the actual tournament outcome using Spearman's rank correlation. For the actual results, a total ranking including ties, as is commonly used in rankings of such incomplete tournaments, is employed.[2] First, this shows that the winning probabilities (including the tournament effects) have a higher correlation with the actual outcome (0.525 for BCM and 0.304 for ELO) than the corresponding (log-)abilities (0.441 and 0.203). Second, the forecasts based on the bookmakers' odds clearly outperform those based on the Elo ratings. This shows that the prospective ratings of experts (i.e., the bookmakers) are more useful than the retrospective performance-based Elo ratings.

In addition to the four forecasts derived in this paper, Table 3 also provides correlations with the ranking implied by the FIFA/Coca Cola World rating. Interestingly, this has a higher Spearman correlation (0.373) with the tournament outcome than the Elo forecasts. Furthermore, it is more closely associated with the two (log-)ability measurements (0.841 and 0.815) than with the corresponding winning probabilities (0.809 and 0.809). This confirms that the (retrospective) FIFA rating is a good assessment of the teams' current abilities but also shows that its predictive power could be enhanced if the corresponding winning probabilities could be computed or simulated. However, as no rigorous method for computing pairwise winning probabilities $\pi_{i,j}$ based on the FIFA rating is known to us, we cannot pursue this approach here.

In order to investigate the reasons for the good performance of the BCM for the winning probabilities, it is useful to extract the two best-ranked teams from each group in Table 2. This shows that the consensus winning probabilities correctly predict five of the teams (Germany, Spain, Italy, Portugal, Croatia) which played in the quarter-finals, as well as in the actual final (played by Germany and Spain). The big surprises of the tournament were Russia and Turkey, which both reached the semi-finals rather unexpectedly. Whereas the BCM ranked Russia better than the Elo and the FIFA rating, the opposite is true for Turkey. Furthermore, France, surprisingly, did not reach the quarter-finals, which was expected by neither the bookmakers nor the Elo and FIFA ratings. However, the BCM showed it to be somewhat more likely.

4.4. Tournament analysis based on the BCM forecast

In addition to the team abilities and winning probabilities (Table 2), some further insights can be gained from the best-performing BCM forecast due to the adoption of the simulation approach. So far, we have only considered the empirical winning proportions of each team in the 100,000 tournament runs, but the empirical proportions of reaching the quarter-final, semi-final and final can, of course, be extracted as well. Fig. 1 shows the performance of each team in the simulations based on $ability^{(BCM)}$ as a performance curve (or a "survival" curve over the course of the tournament). The endpoints of the curves are the simulated winning probabilities, which are by construction (Eq. (12)) (roughly) identical to the probabilities derived from the BCM (Table 2).

The performance curves in Fig. 1 show that groups B and D are rather heterogeneous, with weaker teams and clear favorites (Germany and Spain, respectively), while groups A and C are rather homogeneous. This group effect can also be quantified at an aggregated level by considering deviations of the mean group winning logits (computed from Table 2) from the overall mean winning logits across all teams. Despite the fact that group B includes the bookmakers' favorite for winning the European championship (Germany), group B is clearly the weakest group and has the smallest chance of including the winner (with a deviation of -0.187 on the logit scale). This is followed by group D, with a deviation of -0.116. Group C, on the other hand, is clearly the toughest group and has the greatest probability of including the champion (0.293). Group A can be interpreted as the average group, with a deviation of 0.010 from the overall mean.

[2] Various strategies for dissolving the ties have been explored, but they did not lead to qualitatively different results.

Table 3
Spearman's rank correlations between the actual tournament ranking and the rankings based on the estimated BCM winning probabilities and (log-)abilities, the simulated Elo winning probabilities and (log-)abilities (equivalent to the original Elo rating), and the FIFA/Coca Cola World rating.

	$p^{(BCM)}$	$ability^{(BCM)}$	$p^{(ELO)}$	$ability^{(ELO)}$	FIFA rating
Tournament ranking	0.525	0.441	0.304	0.203	0.373
$p^{(BCM)}$		0.988	0.871	0.771	0.809
$ability^{(BCM)}$			0.909	0.826	0.841
$p^{(ELO)}$				0.956	0.809
$ability^{(ELO)}$					0.815

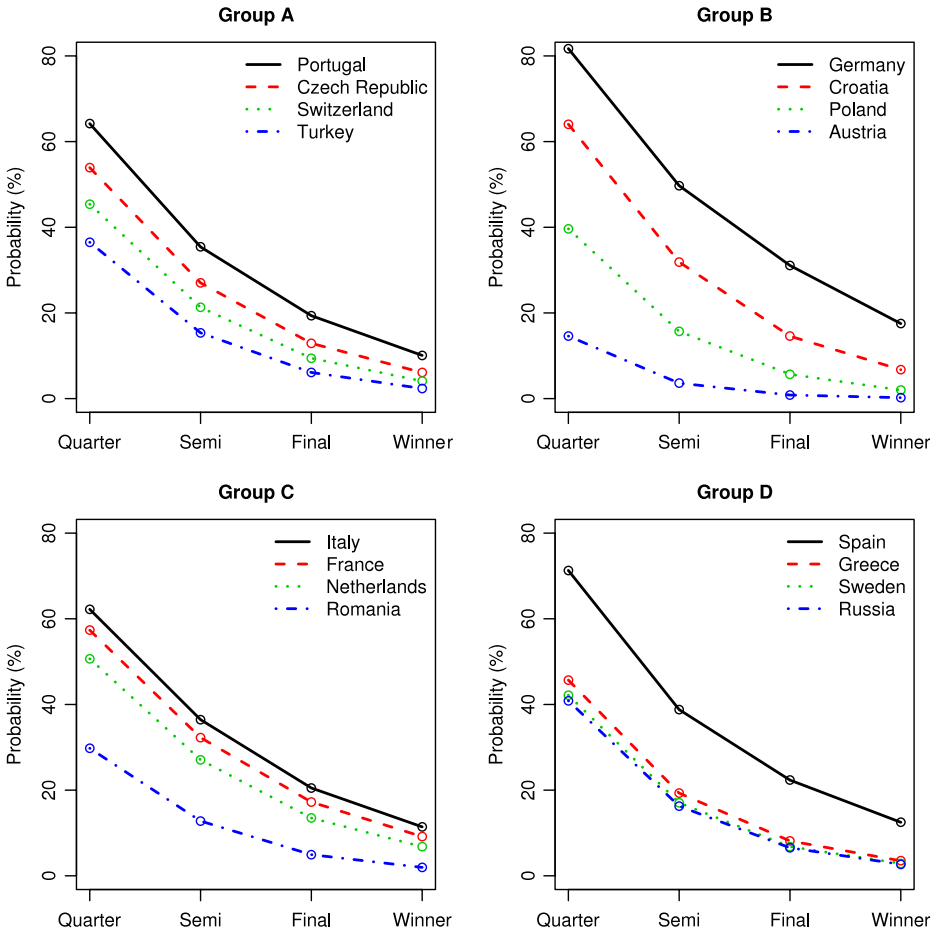

Fig. 1. Simulated probabilities (from 100,000 tournament runs based on the BCM consensus abilities) of each team reaching the quarter-final, the semi-final and the final, and winning the EURO 2008.

The simulation also provides information about the most likely coupling for the final: a match between Germany and Spain, the actual finalists, has the highest probability (20.45%). Given this coupling in the final, the winning probabilities of the two teams are given by the Bradley–Terry model (Eq. (3)), based on

the teams' estimated abilities $ability_i^{(BCM)}$. Although Germany has a slight advantage, with a winning probability of 52.08%, this essentially shows that no clear favorite exists in this final. This is confirmed by the actual EURO 2008 final, which ended with a very close result: Germany 0, Spain 1.

5. Summary and outlook

We have embedded various methods for rating players/teams in competitive sports into a common framework that allows us to forecast probabilities of winning in sports tournaments (rather than single matches), and obtain the competitors' underlying strengths/abilities. The link between abilities and winning probabilities is established by means of a simulation approach that takes into account potential tournament effects such as group draws or seedings. Specifically, these methods are applied to the World Football Elo rating and the odds from a set of international bookmakers, and are assessed using forecasts of the European football championship 2008. A consensus model for the bookmakers' odds performs best in this comparison, correctly predicting the actual final of the tournament and revealing clear tournament effects due to the group draw.

Although the model forecasts provide promising results for the EURO 2008, various improvements are conceivable and deserve further study. For example, the tournament simulation could be enhanced to provide not only winners and losers but more realistic results (such as numbers of goals or goal differences in a soccer tournament). The bookmaker consensus model adopted here only includes a fixed team effect, but could be extended to encompass further fixed or random effects, capturing, for example, group strengths, bookmaker biases, or differences in variances.

Furthermore, our results show that the prospective rating based on aggregated expert judgments in the bookmaker consensus model provides more accurate forecasts of sports tournament outcomes than retrospective ratings that derive current team/player abilities from past performances. The application of both approaches to future tournaments will continue to explore the potential of these methods and help to establish a more complete picture.

6. Computational details

All computations were carried out in the R system (version 2.8.1) for statistical computing (R Development Core Team, 2009).

References

Advanced Satellite Consulting Ltd (2008) The world football Elo rating system. Online; accessed 2008-04-21. URL: http://www.eloratings.net/.

Boulier, B. L., & Stekler, H. O. (1999). Are sports seedings good predictors?: An evaluation. *International Journal of Forecasting*, *15*, 83–91.

Boulier, B. L., & Stekler, H. O. (2003). Predicting the outcomes of national football league games. *International Journal of Forecasting*, *19*, 257–270.

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

Clarke, S. R., & Norman, J. M. (1995). Home ground advantage of individual clubs in English soccer. *Journal of the Royal Statistical Society, Series D*, *44*(4), 509–521.

Dixon, M., & Coles, S. (1997). Modelling association football scores and inefficiencies in the UK football betting market. *Journal of the Royal Statistical Society, Series C*, *46*(2), 265–280.

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

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

Edmans, A., García, D., & Norli, O. (2007). Sports sentiment and stock returns. *Journal of Finance*, *62*(4), 1967–1998.

Elo, A. E. (2008). *The rating of chess players, past and present*. San Rafael, United States: Ishi Press.

Fédération Internationale de Football Association (2008). FIFA/Coca-Cola World Ranking. Online; accessed 2008-04-21. URL: http://www.fifa.com/.

Forrest, D., Beaumont, J., Goddard, J., & Simmons, R. (2005). Home advantage and the debate about competitive balance in professional sports leagues. *Journal of Sports Sciences*, *23*(4), 439–445.

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

Forrest, D., & Simmons, R. (2000). Forecasting sport: The behaviour and performance of football tipsters. *International Journal of Forecasting*, *16*, 317–331.

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

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

Joe, H. (1991). Rating systems based on paired comparison models. *Statistics and Probability Letters*, *11*, 343–347.

Kolb, R., & Stekler, H. O. (1996). Is there a consensus among financial forecasters? *International Journal of Forecasting*, *12*, 455–464.

Lebovic, J. H., & Sigelman, L. (2001). The forecasting accuracy and determinants of football rankings. *International Journal of Forecasting*, *17*, 105–120.

Leitner, C., Zeileis, A., & Hornik, K. (2008). *Who is going to win the EURO 2008? (A statistical investigation of bookmakers odds)*. Report 65, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series. URL: http://epub.wu.ac.at/.

Leitner, C., Zeileis, A., & Hornik, K. (2009). Forecasting the winner of the UEFA Champions League 2008/09. In R. Koning, & P. Scarf (eds), *Proceedings of the 2nd International Conference on Mathematics in Sport—IMA Sport 2009* (pp. 94–99).

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

McHale, I., & Davies, S. (2007). Statistical analysis of the effectiveness of the FIFA world rankings. In J. Albert, & R. H. Koning (Eds.), *Statistical thinking in sports* (pp. 77–90). Boca Raton, Florida: Chapman & Hall/CRC.

Pinheiro, J., & Bates, D. (2000). Mixed-effects models in S and S-PLUS. In *Statistics and comupting*. New York, USA: Springer-Verlag, ISBN: 0-387-98957-9.

Pope, P. F., & Peel, D. A. (1989). Information, prices and efficiency in fixed-odds betting market. *Economica*, *56*, 323–341.

R Development Core Team, (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Raiffeisen Zentralbank (2008). RZB-Analyse: Wer wird Fußball-Europameister? Online; accessed 2008-05-09. URL: http://www.rzb.at/.

Scarf, P., & Bilbao, M. (2006). *The optimal design of sporting contests*. Report 320, Salford Business School, Working Paper Series. URL: http://www.mams.salford.ac.uk/.

Schnader, M., & Stekler, H. O. (1991). Do consensus forecasts exist? *International Journal of Forecasting*, *7*, 165–170.

Song, C., Boulier, B. L., & Stekler, H. O. (2007). The comparative accuracy of judgmental and model forecasts of American football games. *International Journal of Forecasting*, *23*, 405–413.

Song, C., Boulier, B. L., & Stekler, H. O. (2009). Measuring consensus in binary forecasts: NFL game predictions. *International Journal of Forecasting*, *25*, 182–191.

Spann, M., & Skiera, B. (2009). Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, *28*, 55–72.

Stefani, R. T. (1997). Survey of the major world sports rating systems. *Journal of Applied Statistics*, *24*(6), 635–646.

Stefani, R. T., & Pollard, R. (2007). Football rating systems for top-level competition: A critical survey. *Journal of Quantitative Analysis in Sports*, *3*(3), 1–20.

Suzuki, K., & Ohmori, K. (2008). Effectiveness of FIFA/Coca-Cola World Ranking in predicting the results of FIFA World Cup finals. *Football Science*, *5*, 18–25.

UBS Wealth Management Research Switzerland (2008). European champions for 2008 will be.... Online; accessed 2008-04-21. URL: http://www.ubs.com/.

Union of European Football Associations (2009). UEFA Champions League. Online; accessed 2009-04-23. URL: http://en.euro2008.uefa.com/tournament/index.html.

Vlastakis, N., Dotsis, G., & Markellos, R. N. (2009). How efficient is the European football betting market? Evidence from arbitrage and trading strategies. *Journal of Forecasting*, *28*, 426–444.

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

Christoph Leitner is Research Assistant at the Department of Statistics and Mathematics, WU Wirtschaftsuniversität Wien. His research interests focus on the analysis of ratings, in both finance and sports. He has recently contributed to several conferences/workshops on sports forecasting, including the proceedings of the 2nd International Conference on Mathematics in Sport (IMA Sport 2009 in Groningen, Netherlands).

Achim Zeileis is Assistant Professor at the Department of Statistics and Mathematics, WU Wirtschaftsuniversität Wien. His research interests include statistical computing, applied econometrics, and statistical learning. He is co-editor of the *Journal of Statistical Software* and associate editor of *Computational Statistics & Data Analysis* and *Advances in Statistical Analysis*.

Kurt Hornik is Professor of Statistics and Mathematics at the Department of Statistics and Mathematics, WU Wirtschaftsuniversität Wien. His research interests include statistical computing, statistical learning, and management science. He is a member of the core development team of the R system for statistical computing and graphics, and a maintainer of the Comprehensive R Archive Network (CRAN). He is also associate editor of the *Journal of Computational & Graphical Statistics*, *Journal of Statistical Software*, *Computational Statistics & Data Analysis*, and *Computational Statistics*.