

Original citation:

Cattelan, Manuela , Varin, Cristiano and Firth, David. (2013) Dynamic Bradley–Terry modelling of sports tournaments. Journal of the Royal Statistical Society : Series C, Volume 62 (Number 1). pp. 135-150. ISSN 0035-9254

Permanent WRAP url:

<http://wrap.warwick.ac.uk/54660/>

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

This article is made available under the Creative Commons Attribution 3.0 Unported (CC BY 3.0) license and may be reused according to the conditions of the license. For more details see: <http://creativecommons.org/licenses/by/3.0/>

A note on versions:

The version presented in WRAP is the published version, or, version of record, and may be cited as it appears here.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

warwick**publications**wrap

highlight your research

<http://go.warwick.ac.uk/lib-publications>



Dynamic Bradley–Terry modelling of sports tournaments

Manuela Cattelan,

Università di Padova, Italy

Cristiano Varin

Università Ca' Foscari, Venice, Italy

and David Firth

University of Warwick, Coventry, UK

[Received September 2010. Final revision February 2012]

Summary. In the course of national sports tournaments, usually lasting several months, it is expected that the abilities of teams taking part in the tournament will change over time. A dynamic extension of the Bradley–Terry model for paired comparison data is introduced to model the outcomes of sporting contests, allowing for time varying abilities. It is assumed that teams' home and away abilities depend on past results through exponentially weighted moving average processes. The model proposed is applied to sports data with and without tied contests, namely the 2009–2010 regular season of the National Basketball Association tournament and the 2008–2009 Italian *Serie A* football season.

Keywords: Bradley–Terry model; Cumulative logit model; Exponentially weighted moving average process; Paired comparisons; Sports tournaments

1. Introduction

The analysis of sports data has always aroused great interest among statisticians. Albert *et al.* (2005) collected a number of articles that summarize various statistical aspects of interest in sports data including rating of players or teams, evaluation of sport strategies, enhancement of sport rules, illustration of statistical methods and forecasting of results.

Sports data have been investigated from different perspectives, often with the aim of forecasting the results. A first approach consists in modelling the scores of the two opposing teams. Maher (1982) employed independent Poisson distributions for the score of each team with means that depend on the attack and defence strength of teams. Dixon and Coles (1997) proposed an *ad hoc* adjustment of the Poisson distribution introducing a dependence parameter that modifies the probabilities of the results 0–0, 0–1, 1–0 and 1–1. Dixon and Coles (1997) introduced also a dynamic element in the model updating the parameter estimates including the results up to the last observation and downweighting observations that are distant in time. Karlis and Ntzouras (2003) suggested applying a bivariate Poisson distribution with a dependence parameter between the number of goals scored by the two teams and then extended the model to inflate the probabilities of draws.

Address for correspondence: Manuela Cattelan, Department of Statistical Sciences, University of Padua, Via C. Battisti 241–243, Padova 35121, Italy.
E-mail: manuela.cattelan@stat.unipd.it

McHale and Scarf (2007) modelled the number of shots of the two teams. They proposed two different types of Archimedean copula with either Poisson or negative binomial distributions for the marginals to account for the negative dependence between shots for and shots against.

Extensions allowing dynamic developments of abilities of the teams were proposed by Rue and Salvesen (2000) and Crowder *et al.* (2002). Rue and Salvesen (2000) assumed that the attack and defence strength parameters of each team follow a Brownian motion process. The model is estimated by employing Bayesian inference through Markov chain Monte Carlo methods. Crowder *et al.* (2002) suggested an auto-regressive model for the attack and defence abilities of teams. The original model is then replaced by a derived version that is easier to handle by maximum likelihood.

A second approach to the analysis of sports data consists of modelling the difference in scores. Clarke and Norman (1995) performed a linear regression of the difference in scores on the difference in strength of the two teams. Harville (2003) employed a similar specification but eliminated the incentives for running up the score beyond a predetermined number of points. A dynamic specification of strength in this context was considered in Harville (1980) who proposed an auto-regressive process for the strength of teams in different seasons. Also Glickman and Stern (1998) assumed that the evolution of week-by-week and seasonal strength follows a first-order auto-regressive process; inference is carried out in a Bayesian framework through Markov chain Monte Carlo algorithms.

Finally, sports data can be analysed by considering only the outcomes of the matches (win–draw–loss). Goddard and Asimakopoulous (2004) used an ordered probit model to determine which covariates, e.g. importance of the match, fouls, yellow and red cards, affect the result of the match. An ordered probit model was adopted also by Koning (2000) who specified the probability of the outcome as a function of the difference of abilities of the two teams. Kuk (1995) introduced two strength parameters for each team: one denoting the strength when playing at home and the other when playing away.

Barry and Hartigan (1993) proposed a dynamic extension for the ability parameters of teams; they employed a choice model assuming a prior distribution for strength of teams that changes slowly in time. Fahrmeir and Tutz (1994) considered three possible specifications for the development of abilities: a first- and second-order random walk and a local linear trend model. These models are estimated by using empirical Bayes methods. Glickman (1999) specified a logit model assuming a prior with normal increments for abilities of teams and proposed an approximate Bayesian algorithm for ranking purposes. Knorr-Held (2000) employed a logit model assuming random-walk priors for abilities of teams. The variance of the random walk is estimated through four different predictive criteria whereas the abilities are estimated by means of the extended Kalman filter and smoother.

In this paper, we analyse the results of sport tournaments from the last perspective, i.e. modelling the outcomes of matches. Since we are interested in studying how the strengths of the teams evolve during the season, we develop a dynamic paired comparison model. In particular, we model the evolution in time of the abilities in home and away matches of each team through an exponentially weighted moving average process.

The paper is organized as follows. Section 2 presents two motivating data sets regarding the American National Basketball Association (NBA) league and the Italian major men's football league. Section 3 describes the proposed dynamic version of the Bradley–Terry model, discusses maximum likelihood estimation and considers model validation by Brier and ranking probability scores. The methodology is applied in Section 4 to the data for the two sports. Concluding remarks and future research are summarized in Section 5.

The data and R (R Development Core Team, 2011) code written for implementing the analyses are available from

<http://www.blackwellpublishing.com/rss>

2. Description of the data and analyses with non-dynamic abilities

2.1. National Basketball Association

As our first motivating example we consider the 2009–2010 regular season of the NBA league. There are 30 teams in the league playing 82 games each: 41 at home and 41 away. The total number of matches is 1230. The schedule of the tournament includes a greater number of contests against teams in the same division and in the same conference, whereas competitions between teams in different conferences are less numerous. The regular season started at the end of October 2009 and ended in mid-April 2010. Matches were played on 164 different days. The number of matches per day ranges from 1 to 14; the mean number is 7.5.

The description of the proposed methodology for the analysis of tournaments is simplified by the assumption of an order for the $m = 1230$ matches among the $n = 30$ teams that are involved in the tournament. A convenient choice is to arrange the matches in chronological order, with those played at the same time in alphabetic order of the home team. Let Y_i be the binary random variable which denotes the result of the i th match, $i = 1, \dots, m$, played by the home team h_i against the visiting team v_i , with $h_i, v_i = 1, \dots, n$, $h_i \neq v_i$. We arbitrarily code $Y_i = 1$ if the home team wins and $Y_i = 0$ if the visitors win.

Traditional paired comparison models describe the outcome probability as $\text{pr}(Y_i = 1) = F(a_{h_i} - a_{v_i})$, where F is a cumulative distribution function and a_{h_i} and a_{v_i} are the parameters representing the abilities of the home and the visiting teams in match i . This simple choice model is commonly termed the Bradley–Terry model (Bradley and Terry, 1952) or the Thurstone–Mosteller model (Thurstone, 1927; Mosteller, 1951) depending on whether F is the cumulative distribution function of a logistic or of a standard normal random variable respectively. In the rest of the paper, we shall consider the Bradley–Terry specification.

The advantage deriving from playing at home is commonly taken into account by including a common home effect parameter η for all teams (Fahrmeir and Tutz, 1994; Knorr-Held, 2000; Harville, 2003), thus leading to the model

$$\text{pr}(Y_i = 1) = \frac{\exp(\eta + a_{h_i} - a_{v_i})}{1 + \exp(\eta + a_{h_i} - a_{v_i})}. \quad (1)$$

Parameter identifiability requires one constraint in the set of abilities, such as the sum constraint $\sum_{k=1}^n a_k = 0$ or the reference team constraint $a_k = 0$ for some $k \in \{1, \dots, n\}$.

Table 1 shows the estimates of the abilities \hat{a}_k from model (1) with the sum constraint on team abilities. Teams are ranked on the basis of the number of matches won during the season out of the total 82 matches played (the second column). The third column indicates which percentage of the matches won was played at home. On average *circa* 60% of the matches are won by the home team; therefore the advantage in playing at home seems not negligible. The fourth and sixth columns in Table 1 report the estimated abilities and the ranking according to the estimated abilities. There is a very close agreement between the ranking that is obtained by the estimated abilities and the number of matches won; indeed the Kendall rank correlation τ is 0.97. The estimated home effect (with standard error in parentheses) is $\hat{\eta} = 0.487 (0.067)$. The fifth column of Table 1 reports the quasi-standard errors (Firth and de Menezes, 2004) of the abilities. The quasi-standard errors allow us to reconstruct approximately the uncertainty of pairwise differences $\hat{a}_k - \hat{a}_{k'}$ used for comparing teams k and k' without the need to report

Table 1. 2009–2010 American NBA league†

<i>Team</i>	<i>Won</i>	<i>% home</i>	<i>Results for static model</i>			<i>Results for dynamic model</i>	
			<i>Ability</i>	<i>qse</i>	<i>Rank</i>	<i>Ability</i>	<i>Rank</i>
Cleveland Cavaliers	61	0.57	1.189	0.267	1	0.769	1
Orlando Magic	59	0.58	1.048	0.260	2	0.568	3
Los Angeles Lakers	57	0.60	0.997	0.254	3	0.650	2
Dallas Mavericks	55	0.51	0.835	0.249	4	0.451	5
Phoenix Suns	54	0.59	0.767	0.248	5	0.382	8
Atlanta Hawks	53	0.64	0.692	0.247	8	0.424	7
Denver Nuggets	53	0.64	0.744	0.246	7	0.482	4
Utah Jazz	53	0.60	0.749	0.248	6	0.348	9
Boston Celtics	50	0.48	0.493	0.243	12	0.447	6
Oklahoma City Thunder	50	0.54	0.565	0.242	10	0.268	11
Portland Trail Blazers	50	0.52	0.538	0.242	11	0.253	12
San Antonio Spurs	50	0.58	0.580	0.242	9	0.270	10
Miami Heat	47	0.51	0.307	0.239	13	0.070	14
Milwaukee Bucks	46	0.61	0.241	0.239	14	0.054	15
Charlotte Bobcats	44	0.70	0.155	0.239	15	0.019	16
Houston Rockets	42	0.55	0.140	0.237	16	0.080	13
Chicago Bulls	41	0.59	0.000	0.237	17	−0.065	20
Memphis Grizzlies	40	0.58	−0.011	0.237	18	0.019	17
Toronto Raptors	40	0.63	−0.074	0.240	19	0.017	18
New Orleans Hornets	37	0.65	−0.177	0.240	20	−0.037	19
Indiana Pacers	32	0.72	−0.544	0.246	21	−0.382	22
Los Angeles Clippers	29	0.72	−0.651	0.249	22	−0.300	21
New York Knicks	29	0.62	−0.749	0.250	23	−0.415	23
Detroit Pistons	27	0.63	−0.821	0.252	25	−0.446	25
Philadelphia 76ers	27	0.44	−0.830	0.253	26	−0.457	26
Golden State Warriors	26	0.69	−0.804	0.254	24	−0.586	28
Washington Wizards	26	0.58	−0.904	0.257	28	−0.494	27
Sacramento Kings	25	0.72	−0.888	0.258	27	−0.436	24
Minnesota Timberwolves	15	0.67	−1.616	0.302	29	−0.834	29
New Jersey Nets	12	0.67	−1.972	0.327	30	−1.119	30

†The table displays the number of matches won, the percentage of matches won played at home, % home, estimated abilities, quasi-standard errors qse and ranks based on the static Bradley–Terry model, and estimated mean abilities and ranks based on the dynamic Bradley–Terry model.

also the covariance between \hat{a}_k and $\hat{a}_{k'}$. For example, if it is of interest to test whether the ability of Cleveland is significantly higher than the ability of Orlando, the standard error of the difference between the estimators can be approximated by using the quasi-variances simply as $(0.267^2 + 0.260^2)^{1/2} = 0.373$. In this case, the abilities of the best two teams do not appear statistically different.

2.2. Association football

The second application concerns the 2008–2009 Italian *Serie A* football league. This tournament comprises $n = 20$ teams with matches played between August 2008 and May 2009. The tournament has a double-round-robin structure, so each team competes twice against all the other teams in the league: once at home and once away. The total number of matches is thus $m = n(n - 1) = 380$. These matches were played on 74 different days. As in the NBA example, there are days with just one match and days with up to 10 matches played. The average number of matches per day is 5.14.

Table 2. 2008–2009 Italian *Serie A* football league†

Team	pts	% home	Results for static model			Results for dynamic model	
			Ability	qse	Rank	Ability	Rank
Internazionale	84	0.56	1.380	0.348	1	0.462	1
Juventus	74	0.53	0.928	0.324	2	0.303	2
Milan	74	0.61	0.913	0.326	3	0.294	3
Fiorentina	68	0.65	0.643	0.327	5	0.205	4
Genoa	68	0.60	0.693	0.314	4	0.177	5
Roma	63	0.68	0.422	0.310	6	0.108	6
Udinese	58	0.66	0.231	0.306	7	−0.007	11
Palermo	57	0.75	0.124	0.319	8	0.071	7
Cagliari	53	0.70	−0.010	0.315	9	0.004	10
Lazio	50	0.56	−0.233	0.329	12	0.028	8
Atalanta	47	0.70	−0.224	0.317	11	−0.009	12
Napoli	46	0.76	−0.259	0.307	13	0.009	9
Sampdoria	46	0.70	−0.177	0.295	10	−0.098	14
Siena	44	0.73	−0.425	0.312	15	−0.109	15
Catania	43	0.79	−0.409	0.319	14	−0.070	13
Chievo	38	0.45	−0.527	0.306	16	−0.243	16
Bologna	37	0.57	−0.627	0.315	17	−0.252	17
Torino	34	0.74	−0.760	0.316	18	−0.259	18
Reggina	31	0.58	−0.850	0.311	20	−0.337	20
Lecce	30	0.63	−0.833	0.303	19	−0.276	19

†The table displays final points pts, the percentage of points won at home, % home, estimated abilities ability, quasi-standard errors qse and ranks based on the static Bradley–Terry model, and estimated mean abilities and ranks based on the dynamic Bradley–Terry model.

The teams, ranked according to the final points order, are listed in Table 2. In the football tournament, the winning team gains 3 points whereas the losing team receives nothing. If the match is drawn, both teams gain 1 point. On average, 65% of the total points are gained in home matches, with percentages ranging from 45% to 79%, and it is then evident that home advantage of teams should be included in the model.

In contrast with basketball, football matches can also end in a draw; hence the random variable Y_i has three categories arbitrarily coded as 2 if the home team wins, 1 in the case of a draw and 0 in the case of victory of the visiting team. Accordingly, model (1) is extended to account for draws with a cumulative link specification (Agresti, 2002):

$$\text{pr}(Y_i \leq y_i) = \frac{\exp(\delta_{y_i} + \eta + a_{h_i} - a_{v_i})}{1 + \exp(\delta_{y_i} + \eta + a_{h_i} - a_{v_i})}, \quad y_i \in \{0, 1, 2\}, \quad (2)$$

where $-\infty < \delta_0 < \delta_1 < \delta_2 = \infty$ are cut point parameters. Parameter identifiability is achieved by imposing the ‘symmetrical’ constraints $\delta_0 = -\delta$ and $\delta_1 = \delta$, with $\delta \geq 0$. These constraints are needed to ensure that two teams with the same ability playing on a neutral field (no home advantage) have the same probability of winning the match. If Y_i assumes only two possible values, then the cumulative logit model (2) reduces to the standard Bradley–Terry model (1), and δ becomes 0.

The fourth column of Table 2 shows the estimates of the abilities \hat{a}_k again with the sum constraint on team abilities. The estimates of the abilities range from −0.850 for Reggina to 1.380 for Internazionale. The ranking that is derived from the estimated abilities is very similar to the final points ranking, as the Kendall τ rank correlation is 0.94. The estimated home effect parameter is

$\hat{\eta} = 0.661$ (0.105) and the estimated cut point parameter is $\hat{\delta}_0 = -0.652$ (0.061). The fifth column in Table 2 reports quasi-standard errors for teams' abilities. In this case, if, for example, it is of interest to test whether the ability of Internazionale is significantly higher than the ability of Juventus, the standard error of the difference can be computed as $(0.348^2 + 0.324^2)^{1/2} = 0.475$, so there is no evidence in this model that they are statistically different even though Internazionale ended the tournament 10 points ahead of Juventus.

In the above static models, parameters a_k measure the overall abilities of the teams over a complete season. However, team abilities are expected to change during the season because of injuries to players, tiredness due to participation also in international competitions, team psychology and other factors. In the next section we develop a dynamic version of the Bradley–Terry model in which abilities are allowed to change and to depend on the past performance of the team.

3. Dynamic Bradley–Terry model

3.1. The model

We model the match results with the following dynamic Bradley–Terry model:

$$\text{pr}(Y_i \leq y_i | Y_{i-1} = y_{i-1}, \dots, Y_1 = y_1) = \frac{\exp\{\delta_{y_i} + a_{h_i}(t_i) - a_{v_i}(t_i)\}}{1 + \exp\{\delta_{y_i} + a_{h_i}(t_i) - a_{v_i}(t_i)\}}, \quad (3)$$

where $a_{h_i}(t_i)$ describes the ability of the home team h_i in match i played against the visiting team v_i at time t_i . We specify an evolution of team ability in home matches which depends only on past matches played at home, whereas the ability when playing away depends only on past matches played as visitors. First, consider the ability in home matches and let $t_i^{(-1)}$ be the time of the match previous to match i in which h_i was the home team. The ability of the home team $a_{h_i}(t_i)$ is assumed to evolve in time following the exponentially weighted moving average (EWMA) process

$$a_{h_i}(t_i) = \lambda_1 \mu_{h_i}(t_i) + (1 - \lambda_1) a_{h_i}(t_i^{(-1)}), \quad (4)$$

for some home-specific smoothing parameter $\lambda_1 \in [0, 1]$. The term $\mu_{h_i}(t_i)$ denotes the mean home ability of team h_i based only on the result of the nearest previous match played at home by h_i

$$\mu_{h_i}(t_i) = \beta_1 r_{h_i}(t_i^{(-1)}), \quad (5)$$

with β_1 being a home-specific parameter and $r_{h_i}(t_i^{(-1)})$ a variable measuring the result of team h_i in the match played at time $t_i^{(-1)}$. In the NBA application, we specify $r_{h_i}(t_i^{(-1)})$ as the binary variable equal to 1 if team h_i won its previous home match and 0 if it was defeated. Thus, if j denotes the nearest match previous to match i , which was played by h_i at home, i.e. $t_i^{(-1)} = t_j$, then $r_{h_i}(t_i^{(-1)}) = y_j$. Instead, in the *Serie A* application, we specify $r_{h_i}(t_i^{(-1)})$ as the number of points earned by team h_i at time $t_i^{(-1)}$ of its previous home match: 3 points for a victory, 1 for a draw and 0 for a loss.

The ability model (4) must be complemented by an initial condition. We assume that all teams start with the same home ability equal to $\beta_1 \bar{r}_h$, where \bar{r}_h is an average of variables $r_{h_i}(t)$ over the *previous* season. In the analysis of the NBA 2009–2010 season, \bar{r}_h is 0.608, the frequency of victories at home during the NBA 2008–2009 regular season. In the analysis of the *Serie A* 2008–2009 data, \bar{r}_h is 1.676 points, which is the average number of points gained by home teams during the *Serie A* 2007–2008 season.

Suppose that the home team h_i has played K matches at home before the match played at time t_i . Then, by iterated back-substitution, the model based on the pair of equations (4)–(5) can be reformulated as

$$a_{h_i}(t_i) = \beta_1 \left\{ \lambda_1 \sum_{k=0}^{K-1} (1 - \lambda_1)^k r_{h_i}(t_i^{(-k-1)}) + (1 - \lambda_1)^K \bar{r}_h \right\} = \beta_1 x_{h_i}(t_i; \lambda_1), \quad (6)$$

with $t_i^{(-r)}$ denoting the time of the r th previous match played at home by team h_i . Thus, the ability $a_{h_i}(t_i)$ is a function of the entire past history of home matches, $r_{h_i}(t_i^{(-1)})$, $r_{h_i}(t_i^{(-2)})$, \dots , $r_{h_i}(t_i^{(-K)})$. The derived covariate $x_{h_i}(t_i; \lambda_1)$ is a weighted mean of these past results with weights λ_1 , $\lambda_1(1 - \lambda_1)$, \dots , $\lambda_1(1 - \lambda_1)^K$, geometrically decreasing to 0. The smoothing parameter λ_1 specifies the persistence of the dependence on previous home matches. In the limiting case $\lambda_1 = 1$, the home team's ability depends only on the previous home match, $a_{h_i}(t_i) = \beta_1 r_{h_i}(t_i^{(-1)})$. In contrast, if $\lambda_1 = 0$ the home ability is constant in time and equal for all teams, $a_{h_i}(t_i) = \beta_1 \bar{r}_h = \eta$. Values of $\lambda_1 \in (0, 1)$ specify different levels of smoothing. In particular, home abilities smoothed in time are obtained when λ_1 approaches 0.

Similarly, the ability of the visiting team is modelled by a second EWMA process

$$a_{v_i}(t_i) = \lambda_2 \mu_{v_i}(t_i) + (1 - \lambda_2) a_{v_i}(t_i^{(-1)}),$$

where $\lambda_2 \in [0, 1]$ is the visitor-specific smoothing parameter and $\mu_{v_i}(t_i) = \beta_2 r_{v_i}(t_i^{(-1)})$ for a visitor-specific coefficient β_2 . The starting values for r are computed similarly to those for the home abilities. In the NBA 2009–2010 data \bar{r}_a is set equal to 0.392, which is the frequency of visitors' victories during season 2008–2009, whereas in the *Serie A* 2008–2009 data $\bar{r}_a = 1.029$ points, which is the average number of points gained by visitors in the *Serie A* 2007–2008 tournament.

Thus, in the proposed dynamic Bradley–Terry model, the EWMA specification is used to account for the serial association between match results of the same team, with suitable differences between home and away matches.

3.2. Likelihood inference

EWMA processes are routinely used in time series forecasting (Holt, 2004) and in statistical quality control charts (Montgomery, 2005). In these contexts, the smoothing parameter is often chosen by trials or by using *ad hoc* methods based on previous experience. However, many have argued that automatic, data-driven choices of the smoothing parameter would be preferable. For example, in classical time series, the smoothing parameter 'is often estimated by minimizing the sum of squared one-step-ahead forecast errors' (Chatfield (2000), page 98).

Here, we follow the recommendation to identify the smoothing parameters by using available observations and consider maximum profile likelihood estimation of the two smoothing parameters λ_1 and λ_2 .

Let $\gamma = (\beta_1, \beta_2, \delta)^T$ be the vector of parameters of interest and let $\lambda = (\lambda_1, \lambda_2)^T$ be the vector of nuisance smoothing parameters. Under the chosen order for the match results, the likelihood function for $\theta = (\gamma^T, \lambda^T)^T$ is written as

$$\mathcal{L}(\theta; \mathbf{y}) = \text{pr}(Y_1 = y_1; \theta) \prod_{i=2}^n \text{pr}(Y_i = y_i | Y_{i-1} = y_{i-1}, \dots, Y_1 = y_1; \theta).$$

Given the home smoothing parameter λ_1 , the home ability can be written as $a_{h_i}(t_i) = \beta_1 x_{h_i}(t_i; \lambda_1)$, where $x_{h_i}(t_i; \lambda_1)$ is the weighted average of past home results $r_{h_i}(t_i^{(-k)})$ with weights $\lambda_1(1 - \lambda_1)^{k-1}$; see formula (6). In parallel, given the visitors' smoothing parameter λ_2 , the visitors' ability is $a_{v_i}(t_i) = \beta_2 x_{v_i}(t_i; \lambda_2)$ where $x_{v_i}(t_i; \lambda_2)$ has a specification that is analogous to the home case. Accordingly, the conditional probability for the result of match i is expressed by the cumulative

logit model

$$\text{pr}(Y_i \leq y_i | Y_{i-1} = y_{i-1}, \dots, Y_1 = y_1; \theta) = \frac{\exp\{\delta_{y_i} + \beta_1 x_{h_i}(t_i; \lambda_1) - \beta_2 x_{v_i}(t_i; \lambda_2)\}}{1 + \exp\{\delta_{y_i} + \beta_1 x_{h_i}(t_i; \lambda_1) - \beta_2 x_{v_i}(t_i; \lambda_2)\}}.$$

Thus, if the smoothing parameters are known, the likelihood function for the interest parameter γ corresponds to that of a standard logistic regression model if draws are not allowed (e.g. basketball) or that of a cumulative logistic regression model with constrained cut points in the case of draws (e.g. football). The simplicity of computation of $\hat{\gamma}_\lambda$, the estimates of γ given λ , suggests a two-step maximization of the likelihood. First, the smoothing parameter vector λ is estimated by maximizing the profile likelihood $\mathcal{L}(\hat{\gamma}_\lambda, \lambda; \mathbf{y})$, then γ is estimated as $\hat{\gamma}_{\hat{\lambda}}$. This approach is employed in the applications that are illustrated in Section 4.

3.3. Model validation

Model validation can be based on comparison of the fitted probabilities from the proposed model with fitted probabilities from the unstructured model (2). The model proposed aims to capture the evolution in time of all teams' abilities with only four parameters (five in the case of draws), whereas the unstructured model has n free parameters ($n+1$ when draws are allowed), with n being the number of teams. Clearly, the unstructured model is expected to fit the data better, and thus it may be viewed as a benchmark. The closer the fitted probabilities of the model proposed are to those of the unstructured model, the better is the fit. To summarize the fitted probabilities we consider the Brier score (Brier, 1950) which is defined for match i as

$$\text{BS}_i = \sum_{q=0}^{Q-1} \{\text{pr}(Y_i = q | Y_{i-1} = y_{i-1}, \dots, Y_1 = y_1; \hat{\theta}) - \mathbf{1}(y_i = q)\}^2, \quad i = 1, \dots, m,$$

where $\mathbf{1}(y_i = q)$ is the indicator function of the event $\{y_i = q\}$, $Q = 2$ for sports without draws and $Q = 3$ when draws are allowed, and $\hat{\theta}$ is the maximum likelihood estimate of θ based on the results of all the matches played, i.e. y_1, \dots, y_m . When the fit is perfect, giving probability 1 to the observed outcome, the Brier score is equal to 0, whereas a completely erroneous fit produces a Brier score equal to 2.

Some researchers have suggested that in the case of more than two categories it is better to employ an index which accounts for the whole distribution of probabilities, such as the rank probability score (Czado *et al.*, 2009)

$$\text{RPS}_i = \sum_{q=0}^{Q-1} \{\text{pr}(Y_i \leq q | Y_{i-1} = y_{i-1}, \dots, Y_1 = y_1; \hat{\theta}) - \mathbf{1}(y_i \leq q)\}^2.$$

In the analysis of sport tournaments the real interest usually lies in forecasting future results. Hence, it may be more relevant to evaluate the fitted model from a predictive point of view. In this case, we quantify the BS_i and RPS_i using the maximum likelihood estimate $\hat{\theta}_{(i-1)}$ computed only with matches played before the forecasted match i , i.e. only with results y_1, \dots, y_{i-1} .

4. Applications

4.1. Application to the National Basketball Association tournament

We fit the proposed model to the NBA 2009–2010 regular season. Fig. 1 shows the profile log-likelihood for the smoothing parameters λ_1 and λ_2 . This is maximized at $\hat{\lambda}_1 = 0.043$ and $\hat{\lambda}_2 = 0.025$. These two values are close to 0, thus supporting the effect also of remote match results

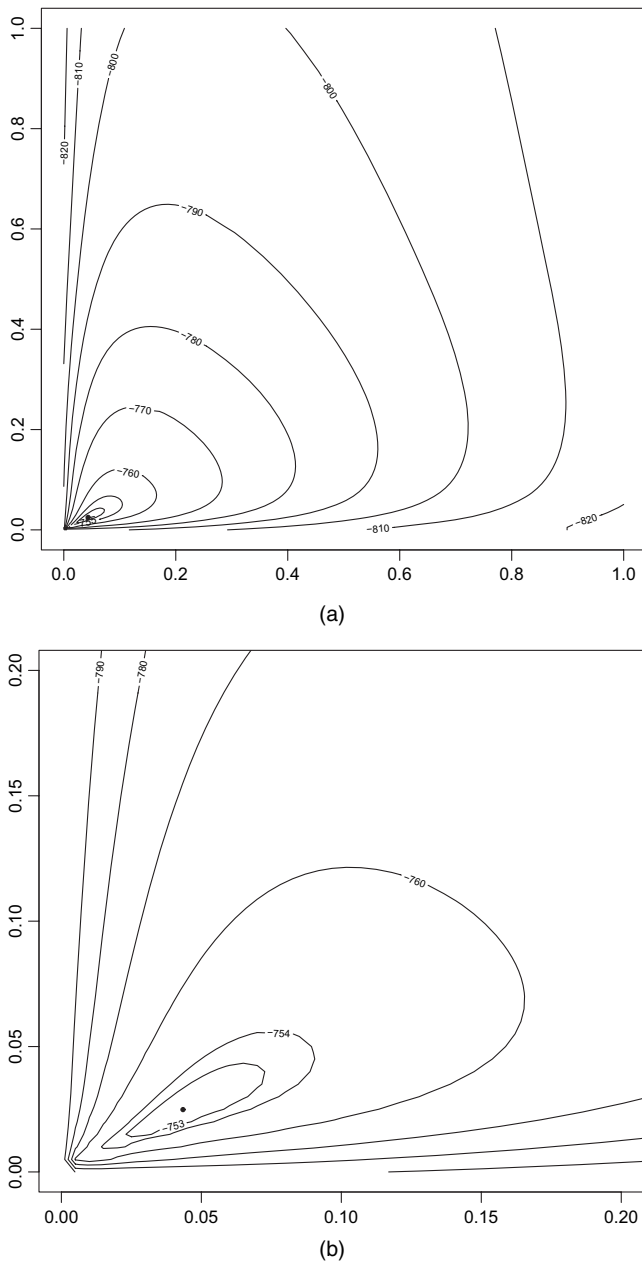


Fig. 1. Contour plots of the profile log-likelihood for the smoothing parameters λ_1 and λ_2 for the NBA 2009–2010 season: (a) plot for all the range of values of λ_1 and λ_2 ; (b) zoom of the area of highest likelihood (•, maximum likelihood estimate)

on the estimation of the present ability. The limiting model with one common home advantage parameter and no evolution in time of abilities corresponds to $\lambda_1 = \lambda_2 = 0$. This pair of values for the smoothing parameters is strongly *not* supported by the data. In fact, the maximized profile log-likelihood is -752.22 whereas the profile log-likelihood for $\lambda_1 = \lambda_2 = 0$ is much smaller, at -830.56 . The maximum likelihood estimates for the home and away coefficients computed at

the estimated smoothing parameters are $\hat{\beta}_1 = 5.503$ and $\hat{\beta}_2 = 7.379$, with estimated standard errors 0.465 and 0.699 respectively. Both estimates appear strongly significant.

To validate the fitted model, we compute the fitted Brier scores for each of the 1230 matches both with the model proposed, which includes only four parameters, and with the unstructured model involving 31 parameters. There is a high positive association between the Brier scores of the two models; in fact their correlation is 0.784. The mean of the Brier scores for the fitted model is 0.424, whereas for the unstructured model it is 0.379; thus the latter is 10.6% smaller in mean. This result was expected given the larger number of parameters of the unstructured model.

However, it is more interesting to consider the appropriateness of the model proposed by evaluation of its predictive performance. For this, we fit the model to the data coming from half of the competition days, i.e. 82 days, then predict the results of the matches taking place in the following day of competition, the 83rd, and compare them with the observed results by the Brier score. Then, the model is refitted including also the matches in day 83 and used to forecast results in day 84, and so on until the last day of matches (day 164). Following this scheme we compute predictions for a total of 638 matches. Predictions from the unstructured model are similarly computed. These predictions are also compared with those obtained by simply using the empirical proportions of wins and losses of the home team computed on the first half of the competition days. These fixed empirical proportions are employed to forecast all the results of the second half of the tournament. Fig. 2(a) shows the boxplots of the Brier scores computed for each forecasted match by using the unstructured model, the EWMA model and the empirical proportions of wins and losses. Visual inspection of the boxplots shows Brier scores of our model are very similar to those of the unstructured model whereas the Brier scores of the empirical proportions are noticeably worse. Specifically, the mean predictive Brier score for the proposed and the unstructured model are very close: 0.421 and 0.409 respectively. The correlation between the two sets of Brier scores is 0.879. The overall conclusion is that the model proposed is very competitive with the unstructured model from a predictive point of view. The mean Brier score of empirical proportions is 0.497, so it is 15% larger than the score of the model proposed. Furthermore, both the unstructured and the proposed models present Brier scores which are lower than the Brier score of empirical proportions in 66% of the forecasts.

Fig. 3 shows the smoothed abilities for nine teams during the complete regular season. The smoothed abilities are computed with parameters estimated from the complete tournament. For each team there is a home and a visiting ability that are plotted in the same graph. The timescale is the sequence of matches; at the end of the regular season each team has played 41 matches at home and 41 away. Cleveland Cavaliers ended the season with the highest number of victories, 61, and, as expected, their ability in both home and away matches increases noticeably from the starting mean ability that is common to all teams. It seems that Cleveland benefits from an important home advantage effect; indeed it won 85% of the matches that it played at home and 63% of the matches played away. Also New Orleans Hornets show an important home effect especially in the first half of the season. This team won 16 matches at home out of the first 20 matches whereas it succeeded only in five matches away out of the first 20. However, in the second part of the season it showed no particular advantage in playing at home since it won eight matches out of the last 21 both at home and away. New Jersey Nets was the team that performed worst during the season. They won a total of 12 matches during the whole tournament and performed poorly both at home and away. The increase in the visiting ability after match 30 is due to their winning three away matches in a row.

Finally, the seventh column in Table 1 reports the ranking that is derived from the proposed model based on the average of the team abilities in each of the 82 days in which the team played. The reported abilities are computed so as to sum to 0, by analogy with the unstructured

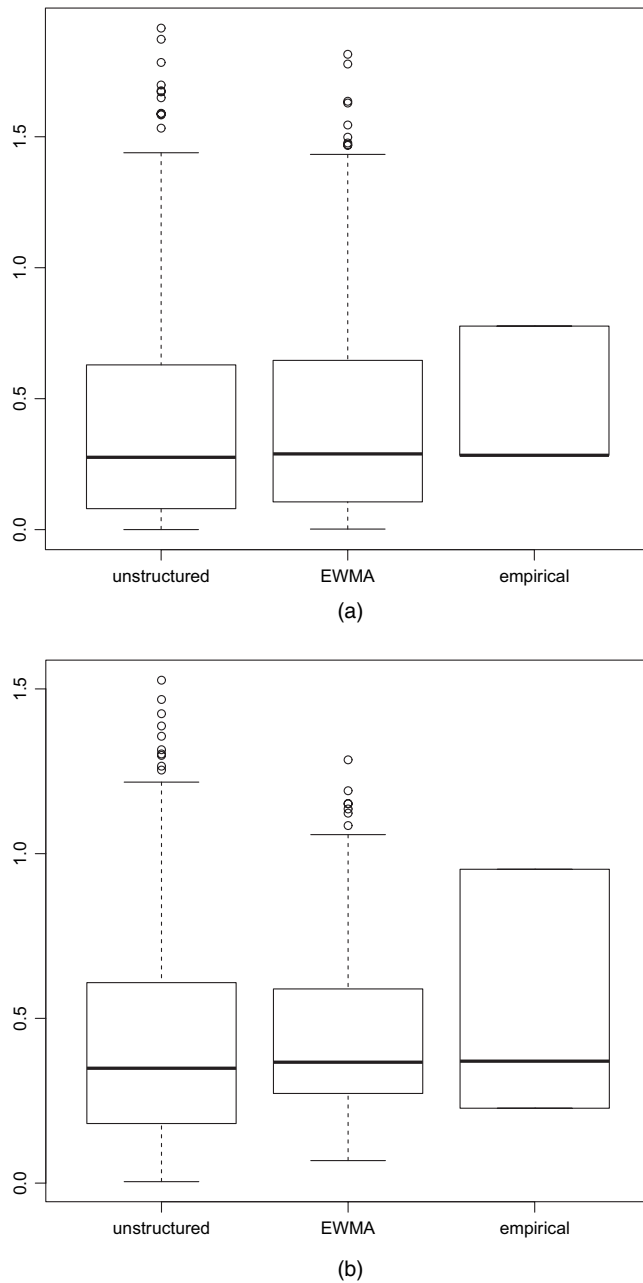


Fig. 2. (a) Boxplots of the predictive Brier scores for the NBA data and (b) boxplots of rank probability scores for the football data computed for the unstructured model, the EWMA model and the forecasts based on empirical proportions

model. The Kendall rank correlation between the rankings of the unstructured and the proposed model is 0.89. In this case we cannot report quasi-standard errors since the abilities are not individual parameters as in the static Bradley–Terry model. Suppose that we are interested in testing whether the ability of Cleveland Cavaliers is significantly higher than the ability of Orlando Magic. The ability of the former team is 0.769 whereas the ability of the latter is 0.568,

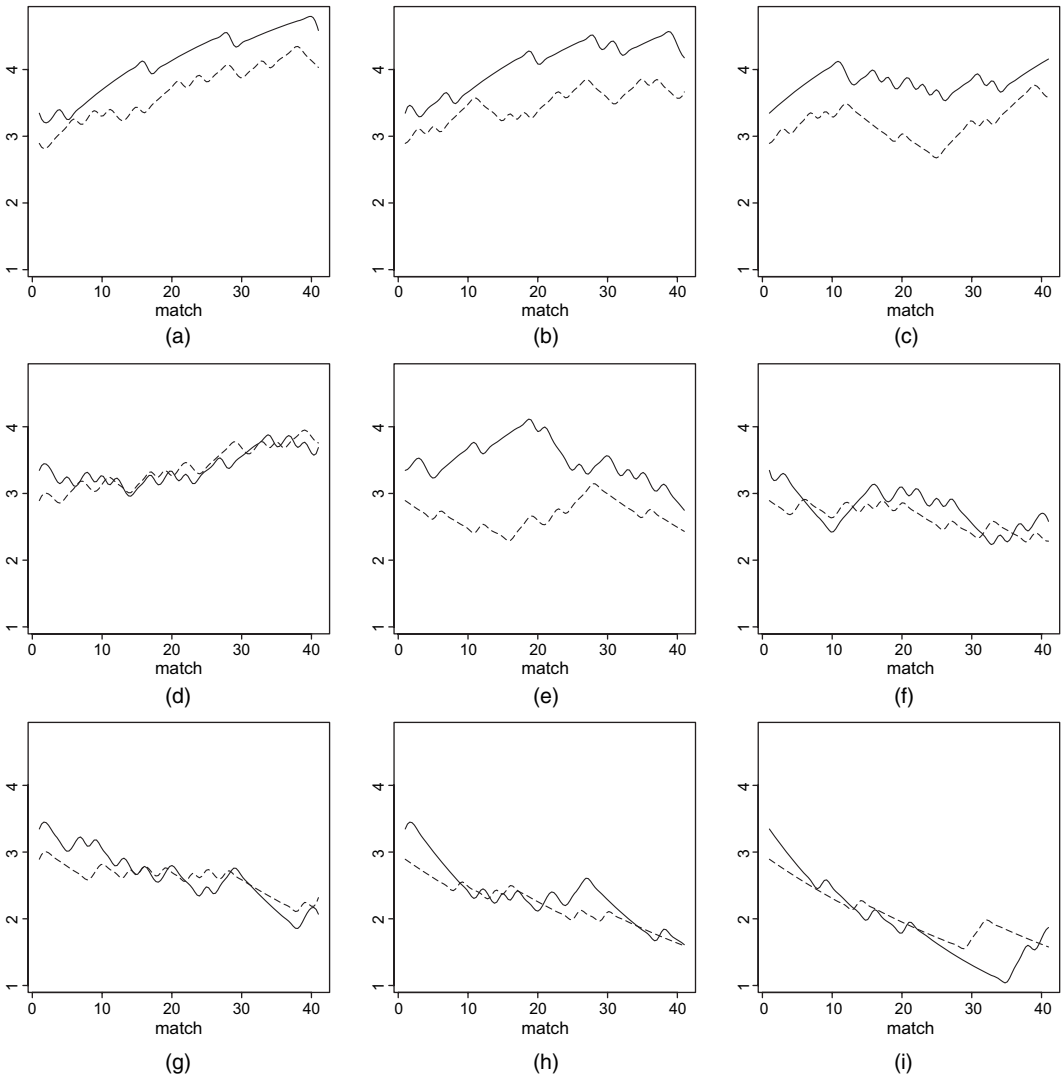


Fig. 3. Smoothed home (—) and visiting (-----) abilities for nine teams for the NBA 2009–2010 season: (a) Cleveland Cavaliers; (b) Los Angeles Lakers; (c) Phoenix Suns; (d) Oklahoma City Thunder; (e) New Orleans Hornets; (f) New York Knicks; (g) Washington Wizards; (h) Minnesota Timberwolves; (i) New Jersey Nets

so their difference amounts to 0.201. This difference appears statistically significant since its standard error is 0.018. This result is different from the result that is given by the unstructured model because the model proposed has fewer parameters and thus the ability estimates are more precise.

4.2. Application to the *Serie A* tournament

The second data set concerns the 2008–2009 *Serie A* football tournament which allows also for ties. The profile log-likelihood is maximized at $\hat{\lambda}_1 = 0.108$ and $\hat{\lambda}_2 = 0.078$: again a pair of values that are quite close to 0. The support for this model, in contrast with the static model

resulting when $\lambda_1 = \lambda_2 = 0$, is given by the maximized profile log-likelihood which is equal to -383.27 , whereas it takes the value -393.68 in the model with smoothing parameters equal to 0. The maximum likelihood estimates for the coefficients computed at the estimated values of the smoothing parameters are $\hat{\beta}_1 = 0.925$ and $\hat{\beta}_2 = 1.031$, with standard error 0.173 and 0.294 respectively, whereas the estimated cut point parameter is $\hat{\delta}_0 = -0.579$ with standard error 0.054.

Since in football data there are three possible results, it is more appropriate to employ the rank probability score to validate the model. The mean rank probability score for the whole tournament, which comprises $m = 380$ matches, is 0.416 for the proposed model and 0.369 for the unstructured model; hence the proposed model has a mean rank probability score which is 11.2% higher than that of the unstructured model. The correlation between rank probability scores for the two models is 0.715.

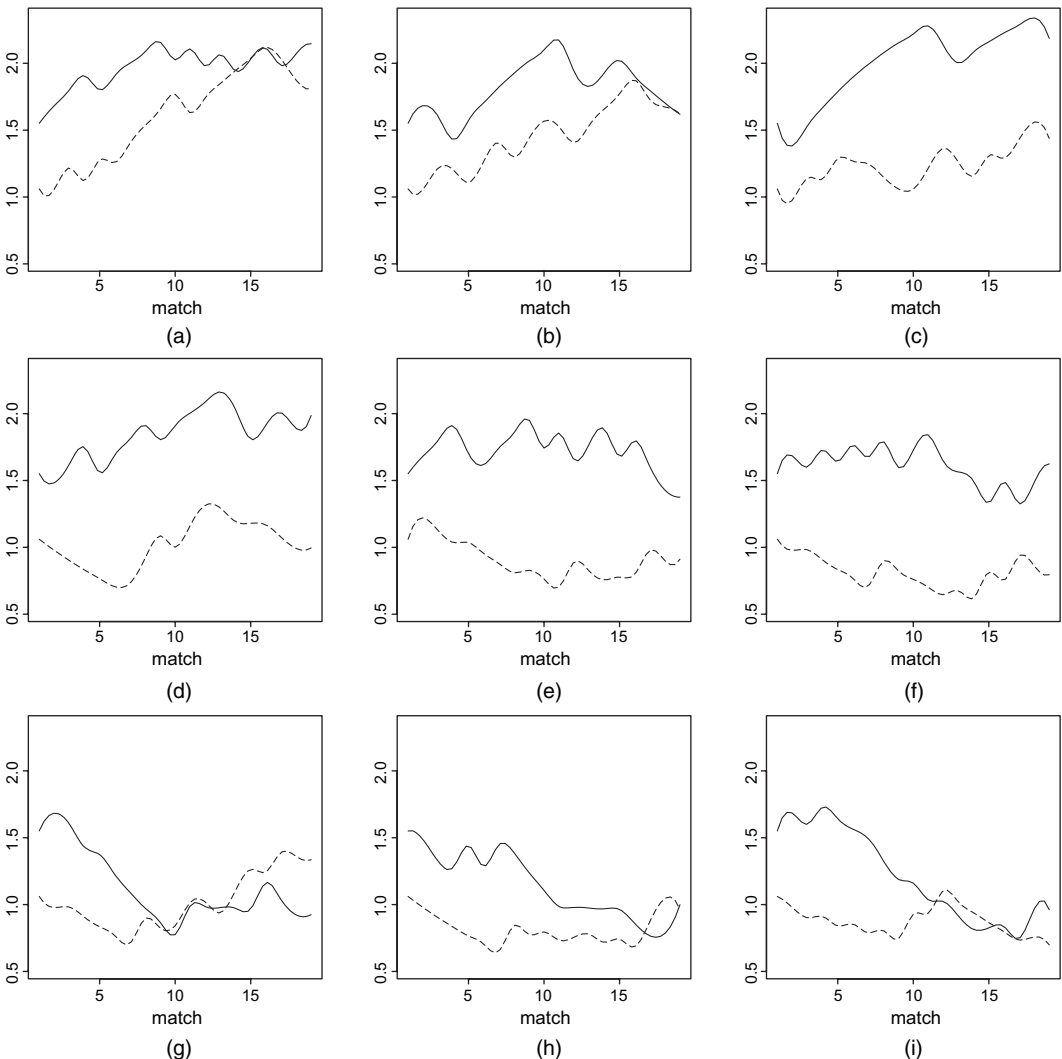


Fig. 4. Smoothed home (—) and visiting (-----) abilities for nine teams for the 2008–2009 football season data: (a) Internazionale; (b) Juventus; (c) Milan; (d) Roma; (e) Atalanta; (f) Siena; (g) Chievo; (h) Reggina; (i) Lecce

As for the NBA example, it is, however, more interesting to consider the predictive performance of the model. To forecast the results we employ the same scheme as used for the basketball data. The model is estimated by using the matches in the first half of the competition days (37 days) and the results of matches occurring in the 38th day of competitions are predicted. Then we updated the model with the matches performed in day 38 and predict the following matches. The procedure is repeated until the end of the tournament with a total of 198 matches predicted. The mean rank probability scores for the unstructured and the EWMA model are essentially equal: 0.450 and 0.451 respectively. The correlation of the RPS_i in the two models is 0.762. Fig. 2(b) shows the boxplots of the rank probability scores of the two models and the scores computed by using the empirical proportions of home wins, draws and losses computed in the first 37 days. It is evident that the medians of the RPS_i for the unstructured and the EWMA models are equal; however, the EWMA model presents a shorter box; thus with respect to the unstructured model it assigns higher probability to results actually observed in fewer cases, but also it less often assigns higher probabilities to results that do not occur. The mean RPS_i of empirical proportions is 0.468, which is 4% larger than the value for the EWMA model. Furthermore, the rank probability scores of the empirical proportions are higher than the scores computed for the proposed model in 58% of the forecast matches.

Fig. 4 shows the estimated abilities for nine of the teams which competed in the 2008–2009 *Serie A* league. Internazionale won the tournament, ending the season with 84 points, whereas Juventus and Milan achieved second position, gaining 74 points each. The plots show that the home performance of Internazionale is more stable during the season; in particular they never lost a home match, whereas Juventus and Milan lost two home matches each. In general, teams tend to perform better at home than away. This is particularly evident for Siena which won nine matches and drew five matches out of the 19 home matches, whereas in the other 19 away matches they won only three times and drew three matches.

The seventh column in Table 2 reports the mean abilities of football teams resulting from the EWMA model. The Kendall rank correlation between the rankings derived from the unstructured and the proposed models is 0.86. In this case, the standard error needed to test whether the teams Internazionale and Juventus have the same ability is 0.037. Since the estimated ability of Internazionale is 0.462 and the estimated ability of Juventus is 0.303, yielding a difference of 0.159, the two teams appear to have statistically different abilities whereas the unstructured model led to the opposite conclusion.

5. Conclusions

We have described a dynamic paired comparison model for the results of matches in sport tournaments. The model specification describes the temporal evolution of teams' abilities by separate EWMA processes for the home and away results. The two applications to basketball and football tournaments show that the model proposed seems to capture the relevant aspects of the evolution in time of team abilities, thus providing sensible forecasts if compared with the unstructured model that fits one ability parameter for each team.

The model proposed uses only information about the final result of previous matches. It seems reasonable that using more detailed information about previous matches may result in more accurate data fitting and improved forecasts. The inclusion of additional information about previous matches is easily handled under the model framework that is described in this paper; for example one could substitute $r_{h_i}(t_i^{(-1)})$ and $r_{v_i}(t_i^{(-1)})$ with vectors $\mathbf{r}_{h_i}(t_i^{(-1)})$ and $\mathbf{r}_{v_i}(t_i^{(-1)})$ and thus consider vector home and visitor parameters β_1 and β_2 . With a possible large number

of covariates from previous matches, it could be sensible to consider some form of shrinkage to avoid overfitting, e.g. by a lasso penalty (Tibshirani, 1996) on parameters β_1 and β_2 .

The model proposed requires starting values for the covariates. We considered equal values for all teams, \bar{r}_h and \bar{r}_a , based on the results in the previous season. It is possible to consider team-specific starting values, as for example the proportions of wins and losses at home for each team in the previous season. The use of team-specific starting values for the NBA data leads to somewhat different estimates of the model parameters, but the fitting and predictive qualities of the resulting model, as evaluated by the Brier score, are almost identical to those of the model with all equal starting values. This exercise seems not possible for the football data since the teams in the *Serie A* league change season by season: at the end of the regular season the last three teams in the league are demoted to the lower league and the three best teams of the lower league are promoted to *Serie A*. Hence, the teams in the league are not the same in different seasons and it seems inappropriate to use team-specific values computed for teams in different leagues.

The issue of whether team-specific home advantages should be included in paired comparison models has been considered by many researchers with contrasting conclusions. Knorr-Held (1997) did not find much evidence of home advantage heterogeneity among teams in the *Bundesliga*. Nor do the results in Harville and Smith (1994) show much difference in home field advantages among college basketball teams. Analyses for some other contexts do, however, support heterogeneity in home advantages; see Clarke and Norman (1995), Kuk (1995) and Glickman and Stern (1998) for different analyses of the English Premier Football League. The model that is proposed in this paper can be seen as a convenient way to induce home abilities which vary between teams and in time depending on past performances of teams.

We considered sports with matches that may end with two or three different results. There are other competitions where more than three levels of results are possible. An example is volleyball where points are assigned as follows: if a match ends 3–0 or 3–1 the winning team gains 3 points and the losing team remains empty handed, whereas, if the match ends 3–2, the winning team gains 2 points and the losing team is awarded 1 point. As suggested by a referee, the analysis of volleyball matches would require a categorical variable Y_i with four ordered levels. There is no special difficulty in extending our modelling framework to this case. More generally, assume that Y_i is a categorical variable that may assume Q different categories where $Q - 1$ denotes the best result achievable for the home team and 0 denotes the best result for the visiting team. The cumulative logit model (2) is easily extended to handle Q levels

$$\text{pr}(Y_i \leq y_i) = \frac{\exp\{\delta_{y_i} + a_{h_i}(t_i) - a_{v_i}(t_i)\}}{1 + \exp\{\delta_{y_i} + a_{h_i}(t_i) - a_{v_i}(t_i)\}}, \quad y_i \in \{0, 1, \dots, Q - 1\}, \quad (7)$$

where $-\infty < \delta_0 < \dots < \delta_{Q-1} = \infty$ are cut point parameters. To preserve model identifiability, the symmetrical constraint now becomes $\delta_q = -\delta_{Q-2-q}$, $q = 0, \dots, Q - 1$, and $\delta_{Q/2-1} = 0$ when Q is even.

Acknowledgements

The authors thank the Associate Editor and the referee for detailed comments about the first submission. The first two authors have been partially supported by the Progetti di Rilevante Interesse Nazionale 2008 grant of the Italian Ministry of Instruction, University and Research. The third author's work was supported by the UK Engineering and Physical Sciences Research Council.

References

- Agresti, A. (2002) *Categorical Data Analysis*. New York: Wiley.
- Albert, J., Bennett, J. and Cochran, J. J. (eds) (2005) *Anthology of Statistics in Sports*. Philadelphia: Society for Industrial and Applied Mathematics.
- Barry, D. and Hartigan, J. A. (1993) Choice models for predicting divisional winners in major league baseball. *J. Am. Statist. Ass.*, **88**, 766–774.
- Bradley, R. A. and Terry, M. E. (1952) Rank analysis of incomplete block designs: I, The method of paired comparisons. *Biometrika*, **39**, 324–345.
- Brier, G. W. (1950) Verification of forecasts expressed in terms of probabilities. *Monthly Weath. Rev.*, **78**, 1–3.
- Chatfield, C. (2000) *Time Series Forecasting*. London: Chapman and Hall.
- Clarke, S. R. and Norman, J. M. (1995) Home ground advantage of individual clubs in English soccer. *Statistician*, **44**, 509–521.
- Crowder, M., Dixon, M., Ledford, A. and Robinson, M. (2002) Dynamic modelling and prediction of English Football League matches for betting. *Statistician*, **51**, 157–168.
- Czado, C., Gneiting, T. and Held, L. (2009) Predictive model assessment for count data. *Biometrics*, **65**, 1254–1261.
- Dixon, M. J. and Coles, S. G. (1997) Modelling association football scores and inefficiencies in the football betting market. *Appl. Statist.*, **46**, 265–280.
- Fahrmeir, L. and Tutz, G. (1994) Dynamic stochastic models for time-dependent ordered paired comparison systems. *J. Am. Statist. Ass.*, **89**, 1438–1449.
- Firth, D. and de Menezes, R. X. (2004) Quasi-variances. *Biometrika*, **91**, 65–80.
- Glickman, M. E. (1999) Parameter estimation in large dynamic paired comparison experiments. *Appl. Statist.*, **48**, 377–394.
- Glickman, M. E. and Stern, H. S. (1998) A state-space model for national football league scores. *J. Am. Statist. Ass.*, **93**, 25–35.
- Goddard, J. and Asimakopoulous, I. (2004) Forecasting football results and the efficiency of fixed-odds betting. *J. Forecast.*, **23**, 51–66.
- Harville, D. A. (1980) Predictions for national football league games via linear-model methodology. *J. Am. Statist. Ass.*, **75**, 516–524.
- Harville, D. A. (2003) The selection or seeding of college basketball or football teams for postseason competition. *J. Am. Statist. Ass.*, **98**, 17–27.
- Harville, D. A. and Smith, M. H. (1994) The home-court advantage: how large is it, and does it vary from team to team? *Am. Statist.*, **48**, 22–28.
- Holt, C. C. (2004) Forecasting seasonals and trends by exponentially weighted moving averages. *Int. J. Forecast.*, **20**, 5–10.
- Karlis, D. and Ntzoufras, I. (2003) Analysis of sports data by using bivariate Poisson models. *Statistician*, **52**, 381–393.
- Knorr-Held, L. (1997) *Hierarchical Modelling of Longitudinal Data; Applications of Markov Chain Monte Carlo*. Munich: Utz.
- Knorr-Held, L. (2000) Dynamic rating of sports teams. *Statistician*, **49**, 261–276.
- Koning, R. H. (2000) Balance in competitions in Dutch soccer. *Statistician*, **49**, 419–431.
- Kuk, A. Y. C. (1995) Modelling paired comparison data with large numbers of draws and large variability of draw percentages among players. *Statistician*, **44**, 523–528.
- Maher, M. J. (1982) Modelling association football scores. *Statist. Neerland.*, **36**, 109–118.
- McHale, I. and Scarf, P. (2007) Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statist. Neerland.*, **61**, 432–445.
- Montgomery, D. (2005) *Introduction to Statistical Quality Control*. Hoboken: Wiley.
- Mosteller, F. (1951) Remarks on the method of paired comparisons: I, The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, **16**, 3–9.
- R Development Core Team (2011) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Rue, H. and Salvesen, Ø. (2000) Prediction and retrospective analysis of soccer matches in a league. *Statistician*, **49**, 399–418.
- Thurstone, L. L. (1927) A law of comparative judgement. *Psychol. Rev.*, **34**, 273–286.
- Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, **58**, 267–288.