The Statistician (1998)
47, *Part* 3, *pp.* 523–538

A birth process model for association football matches

Mark J. Dixon†

University of Newcastle, UK

and Michael E. Robinson

University of Surrey, Guildford, UK

[Received May 1997. Revised January 1998]

Summary. Data from over 4000 recent association football (soccer) matches from the main English competitions show clear evidence that the rate of scoring goals changes over the course of a match. This rate tends to increase over the game but is also influenced by the current score. We develop a model for a soccer match that incorporates parameters for both the attacking and the defensive strength of a team, home advantage, the current score and the time left to play. This model treats the number of goals scored by the two teams as interacting birth processes and shows a satisfactory fit to the data. We also investigate football *clichés* and find evidence that contradicts the *cliché* that a team is more vulnerable just after it has scored a goal. Our model has applications in the football spread betting market, where prices are updated during a match, and may be useful to both bookmakers and bettors.

Keywords: Exponential distribution; Football (soccer); Goal times; Likelihood function; Maximum likelihood; Spread betting; Two-dimensional birth process

1. Introduction

There are various reasons for applying statistical techniques to model sporting events. Models are often used to suggest strategic improvements for either individual competitors (e.g. Ladany and Machol (1977)) or to improve the excitement from the spectator's viewpoint (Ridder *et al*., 1994). Sometimes models are developed to test the fairness of either the rules of the game or the structure of competitions, such as leagues and cups (Barnett and Hilditch, 1993; Appleton, 1995); often sporting events provide a novel application of recently developed statistical theory (Robinson and Tawn, 1995; Smith, 1988). Predicting the probability of future outcomes is probably the most popular requirement of analyses, e.g. in newspaper or media predictions (Stefani and Clarke, 1992; Dixon and Coles, 1997). Our motivations for modelling football goal times are to develop an increased understanding of the scoring process and to provide probability estimates of future outcomes that are of use in spread betting.

Betting on the outcome of football matches can take many forms. In the UK, *football pools*, which typically involve the selection of a number of matches that are thought to be those most likely to be a draw, have been popular for many years. *Fixed odds betting* where bets are made on

†*Address for correspondence:* Department of Statistics, University of Newcastle, Newcastle upon Tyne, NE1 7RU, UK.
E-mail: mark.dixon@newcastle.ac.uk

 0039–0526/98/47523

a home win, draw or an away win is another popular form of gambling. *Spread* or *index* betting is a relatively new, more complicated, type of gambling for association football. Although a full description is deferred until Section 5, it is useful to point out here that certain spread bets require accurate estimates of the distribution of the final score conditional on the current score at *any* given time during a football match between specified teams on a specified date.

In Sections 2 and 3 we develop a model that gives information on the score behaviour over 90 minutes for future matches. The model is based on Dixon and Coles (1997) who examined fixed odds betting using a Poisson regression model for full-time scores. Although the Poisson distribution is a reasonable fit to the full-time results, an examination of goal time data shows a clear deviation from a homogeneous model over the 90 minutes of a match. We develop a non-homogeneous model and fit to league and cup goal time data from 1993 to 1996.

In Section 4 we use the model to investigate some interesting footballing *clichés* (often used by commentators):

(a) a team is never more vulnerable than just after they have scored a goal—we test for this effect that is sometimes called the 'immediate strike back';
(b) more goals are scored as the game progresses, perhaps because of tiredness of players—the model is used to assess whether this is evident and the significance of the effect if it exists;
(c) teams tend to score more or fewer goals depending on the current score—for example teams may try to defend a lead or to restore equality if they are ahead or behind respectively.

In Section 2 we present an empirical study of the available data. In Section 3 we develop the model and apply it to the data of Section 2. Applications of the model and the *clichés* above are considered in Section 4, and in Section 5 we describe spread betting and give an example of how the model can be used to set prices or to bet in this market.

2. Data

A wealth of information is available from each football match played. Obviously scores are recorded, but also available are the times of the goals, the goal scorers, the team's league position at the time of playing and so on. An individual team's performance in any particular game could also be affected by many external factors: newly signed players or the sacking of a manager for example. Though this information is also available, it is less easily formalized and its qualitative value is subjective. Consequently, our model exploits only each team's history of match scores, and the goal times within each match.

Data on 10 409 goal times have been collected, from newspapers and weekly football magazines, for 4012 league and cup matches over the period 1993–1996 for 92 English league clubs from four divisions: the Football Association premiership, and Divisions 1–3 of the Football League. Fig. 1 summarizes these data by aggregating goal times over all teams. Matches are played over two periods, each of 45 minutes. Some cup competitions allow extra time of 30 minutes, but these data have been ignored. Goal times are generally recorded to the integer part of the time of the goal although there are often discrepancies between sources. Thus, there is an occasional repeated time, where two goals have been scored in the same minute.

In this section we consider only aggregated features of the data; it is stressed that this is a much simpler method than modelling match-specific effects which is the approach developed later in the paper. Examining aggregated data over teams is the approach taken by many researchers for modelling sporting events; for example Chedzoy (1995) examined aggregated goal time data.

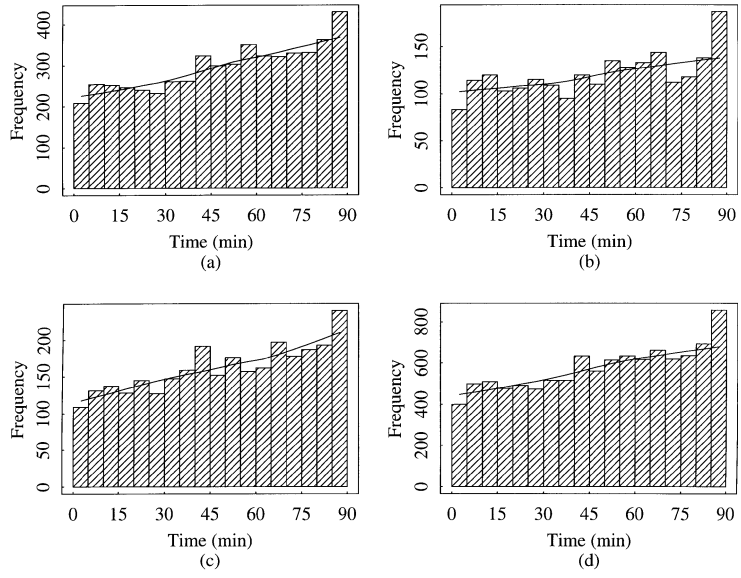

Fig. 1. Histograms of goal times: distributions of all goal times for matches that ended in (a) a home win, (b) a draw and (c) an away win; (d) distribution for all matches (———, kernel density estimate)

However, this can often lead to spurious findings: for instance Clarke and Norman (1995) showed how spurious home advantage effects can appear when averaging over teams.

Figs 1(a)–1(c) are histograms of goal times for matches that ended in a home win, a draw and an away win respectively and Fig. 1(d) shows all the goal times. Two features are evident from Fig. 1. Firstly, a noticeably high number of goals are scored in the last part of each half (around 45 and 90 minutes). This increased number of goals scored is due to injury time, usually between 0 and 5 minutes, added on by the referee. Goals scored in injury time are recorded as 45 or 90 minutes for the first and second halves respectively.

The second feature is that an increasing number of goals are scored throughout the 90 minutes. On the basis of Fig. 1, it is tempting to conclude that, for each particular match, scoring rates increase throughout time, perhaps because of tiredness of players. However, this may be a spurious effect due to averaging. For example, if scoring rates remained constant while the score was (0, 0) but increased as soon as a goal had been scored, then, averaged over matches, this could lead to the gradual increases observed in Fig. 1. Thus two possible reasons for the inhomogeneity over time are

(a) a gradual increase in scoring rates (e.g. because of tiredness) and
(b) variation due to dependence on the current score.

Applying standard survival analysis techniques (e.g. Crowder *et al.* (1991)) we can examine these effects at this stage by using aggregated data.

In a match picked at random, let T_{xy} be the time to the next goal while the current score is (x, y) for $x, y = 0, 1, 2, \ldots$, and let δ_{xy} be a censoring indicator that is 0 if the match ends before the next goal is scored and is 1 if a goal is observed. Then, assuming that $T_{xy} \sim \exp(\nu_{xy})$, standard survival analysis gives the maximum likelihood estimate of ν_{xy} as

$$\hat{\nu}_{xy} = \sum_{i=1}^{N} \delta_{xy,i} \Big/ \sum_{i=1}^{N} t_{xy,i}$$

where $N = 4012$ is the number of matches, $t_{xy,i}$ and $\delta_{xy,i}$ are the observed times to the next goal and censoring indicators respectively, at score (x, y) in match i. In match i, if score (x, y) occurs and is the final score, then $\delta_{xy,i} = 0$ and $t_{xy,i} = 90$ minus the time of the final goal scored; if the score (x, y) never occurs, then $\delta_{xy,i}$ and $t_{xy,i}$ are both taken to be 0. Fig. 2 displays the observed times to the first goal, i.e. $t_{00,i}$ for i such that $\delta_{00,i} = 1$, and Table 1 gives the estimates of the rates ν_{xy}, for $x, y \leqslant 2$. Although in general the rates are increasing, which could be due to a gradual increase in scoring rates with time, there is some evidence that the rate depends on the current score. For example, the rates when the score is (1, 1) and (2, 0) are significantly different, and these scores should occur at approximately the same time in a (random) match. We investigate this further in Section 3.

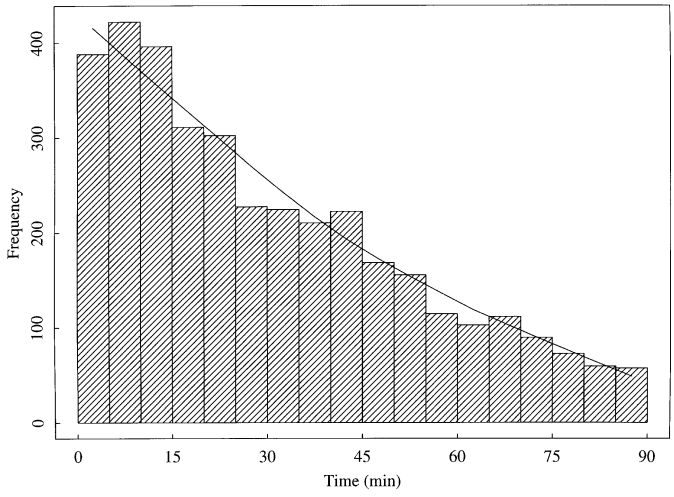

Fig. 2. Histogram of the time to the first goal: censored observations, i.e. matches which ended (0, 0), are not shown (in this data set there are 310 (0, 0) results) (———, kernel density estimate of the times to the first goal)

Table 1. Estimates and standard errors of the rate of the time to the next goal, in a match picked at random, when the score is (0, 0), (1, 0), (0, 1), (1, 1), (2, 0), (0, 2) (2, 1), (1, 2) and (2, 2)†

Rate	*Estimate*	*Standard error*
ν_{00}	0.0250	0.0004
ν_{10}	0.0289	0.0007
ν_{01}	0.0293	0.0009
ν_{11}	0.0302	0.0010
ν_{20}	0.0353	0.0014
ν_{02}	0.0315	0.0018
ν_{21}	0.0327	0.0019
ν_{12}	0.0369	0.0026
ν_{22}	0.0372	0.0029

†For example, $\nu_{00} = 0.025$ corresponds to a scoring rate of one goal every 40 minutes on average while the score is (0, 0).

3. Modelling match-specific goal times

With the aim of developing a model that can provide estimates for a future match between specified teams, several features are required of a statistical model. The model should take into account the different abilities of both teams in a match, with an allowance for home advantage. In addition a measure of a team's ability is likely to be based on their recent performance and the ability of the teams that they have played against. Finally the model should be sufficiently flexible that the scoring rate or intensity within a match can vary with time and with knowledge of the result 'so far'.

Before describing our model for goal times, we summarize the model of Dixon and Coles (1997) and Maher (1982) for full-time results. The basic assumption of the model is that the number of goals scored by the home and away teams in any particular game are independent Poisson variables, whose means are determined by the attack and defence qualities of each side. More explicitly, in a match between teams indexed i and j, let $X_{i,j}$ and $Y_{i,j}$ be the number of goals scored by the home and away sides respectively. Then the model is

$$\begin{aligned} X_{i,j} &\sim \text{Poisson}(\alpha_i \beta_j \gamma_h), \\ Y_{i,j} &\sim \text{Poisson}(\alpha_j \beta_i), \end{aligned} \tag{3.1}$$

where $X_{i,j}$ and $Y_{i,j}$ are independent and $\alpha_i, \beta_i > 0 \ \forall i$. The α_i measure the 'attack' rate of the teams, the β_i measure the respective 'defence' rates, and $\gamma_h > 0$ is a parameter that allows for the home effect.

It follows from model (3.1) that, with n teams, attack parameters $\{\alpha_1, \ldots, \alpha_n\}$, defence parameters $\{\beta_1, \ldots, \beta_n\}$ and the home effect parameter γ_h are to be estimated. As such, the model is overparameterized, so the constraint

$$n^{-1} \sum_{i=1}^{n} \alpha_i = 1$$

is imposed. For the English league system, that comprises the Premier League and Divisions 1–3 of the Football League, $n = 92$, so the model has 184 identifiable parameters.

The basic framework of inference is the likelihood function. With matches indexed $k = 1, \ldots, N$, and corresponding scores (x_k, y_k), this takes the form, up to proportionality,

$$L(\alpha_i, \beta_i, \gamma_h; i = 1, \ldots, n) = \prod_{k=1}^{N} \exp(-\lambda_k) \lambda_k^{x_k} \exp(-\mu_k) \mu_k^{y_k} \tag{3.2}$$

where

$$\begin{aligned} \lambda_k &= \alpha_{i(k)} \beta_{j(k)} \gamma_h, \\ \mu_k &= \alpha_{j(k)} \beta_{i(k)} \end{aligned}$$

and $i(k)$ and $j(k)$ denote respectively the indices of the home and away teams playing in match k.

A structural limitation of model (3.2) is the fact that the parameters are static, i.e. it is assumed that teams have a constant performance rate, as determined by α_i and β_i, from week to week. Dixon and Coles (1997) extended this model to allow for fluctuations in a team's ability by downweighting the likelihood contributions of past data and defining time-dependent parameters.

For notational simplicity, we restrict attention subsequently to an extension of the static model to within-match goal time modelling and time references will be to time elapsed *during* a match. The extension to the dynamic model is immediate, and all subsequent results are obtained by employing the likelihood weighting technique to allow for ability fluctuations.

3.1. Model for goal times

In this section we develop a model for the home–away scoring process that can be thought of as a two-dimensional birth process with the home and away scores as two different species (for example see Fig. 3). First we consider the goal scoring process for a particular match k between teams $i(k)$ and $j(k)$. There are two scoring processes, H_k and A_k, for home and away goals with intensities $\lambda_k(t)$ and $\mu_k(t)$ that are allowed to vary with time t and with status of the process. Here $t \in [0, 1]$ is the (rescaled) time elapsed during the match. The approach that we take is to specify parametrically the form of the intensities $\lambda_k(t)$ and $\mu_k(t)$ and to use regression techniques to select a suitable family of models. If the processes H_k and A_k are taken to be independent homogeneous Poisson processes with intensities

$$\begin{aligned} \lambda_k(t) &= \lambda_k = \alpha_{i(k)}\beta_{j(k)}\gamma_h, \\ \mu_k(t) &= \mu_k = \alpha_{j(k)}\beta_{i(k)} \end{aligned} \tag{3.3}$$

for all $t \in [0, 1]$ then the model reduces to the full-time scores model of the previous section. Fig. 1 shows that, for within-match modelling, such a homogeneous process is unsuitable; we develop a more structured model in stages.

Within this framework, we model the two types of time variation alluded to in Section 2, these being the continuously increasing scoring rates and the variation due to dependence on the current score. First we consider models that only incorporate the second type, and then we extend the model to incorporate both types of non-stationarity. Firstly we assume that the two processes, H_k and A_k, are independent, conditional on the processes up to time t, and that the scoring rates are piecewise constant in the sense that the home and away intensities are constant until a goal is scored and only change at those times. Then, for a goal-less match, the rates are constant throughout the 90 minutes, i.e. $\lambda_k(t) = \lambda_k$ and $\mu_k(t) = \mu_k$ for all $t \in [0, 1]$, and are taken to be as in equations (3.3). For non-goal-less matches, denote the goal times in match k by

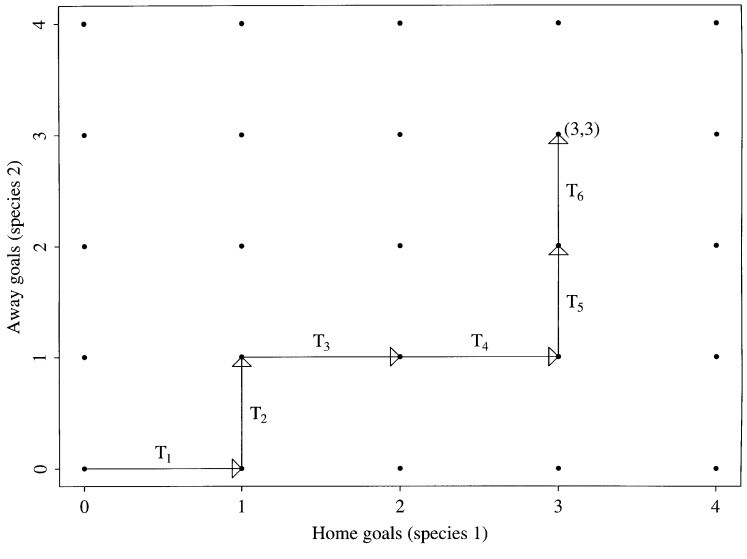

Fig. 3. Graphical representation of the scoring process in a match with final score (3, 3): the home and away goals can be considered as two species in a two-dimensional birth process; the transition (goal) times are denoted by T_i, $i = 1 \ldots, 6$, along the line segments

$$(\mathbf{t}_k, \mathbf{J}_k) = \{(t_{k,l}, J_{k,l}): l = 1, \ldots, m_k\},$$

where $m_k = x_k + y_k$ and $t_{k,l}$ are the total number of goals and the time of the lth goal in match k respectively and $J_{k,l}$ is an indicator that is 0 for a home goal and 1 for an away goal. Let λ_{xy} and μ_{xy} (for $x, y = 0, 1, \ldots$) be parameters that determine the (homogeneous) scoring rates during which the score is (x, y). Also define the home and away scoring rates in match k at time t and score (x, y) to be

$$\lambda_k(t) = \lambda_{xy}\lambda_k$$

and

$$\mu_k(t) = \mu_{xy}\mu_k$$

where t is the set of times during which the score is (x, y) and λ_k and μ_k are given by equations (3.3). Since the rates depend only on the current score, we can think of the process as a two-dimensional (time homogeneous) birth process with states $E = \{(0, 0), (1, 0), (0, 1), \ldots\}$ and exponential transition times that depend on the current state. For example, the distribution of the time to the first home goal conditionally on the score being (0, 0) is exponential with rate $\lambda_{00}\lambda_k$.

Now consider how this model can be extended to model the injury time effect that is evident in Fig. 1. As no data are available showing how much injury time is added, goal times of 45 and 90 minutes are considered as (possibly) censored observations. We introduce new parameters ρ_i, $i = 1, 2$, that represent a multiplicative adjustment to the scoring intensity over the periods (44, 45] and (89, 90] minutes with respect to that predicted by the particular regression model fitted. Thus for models with this effect added the home scoring rate is taken to be

$$\lambda_k(t) = \begin{cases} \rho_1\lambda_{xy}\lambda_k & \text{for } t \in (44/90, 45/90], \\ \rho_2\lambda_{xy}\lambda_k & \text{for } t \in (89/90, 90/90], \\ \lambda_{xy}\lambda_k & \text{otherwise} \end{cases} \tag{3.4}$$

and similarly for the away rate $\mu_k(t)$.

3.1.1. Likelihood

The likelihood for this process, for a particular match k, is essentially that of a two-dimensional pure birth process (for example see Moller and Sorensen (1994)). It can be derived by considering the process as a sequence of independent times between goals (home or away). In match k, conditional on the score being (x, y), the distribution of the time to the next home or away goal is exponential with rate $\lambda_{xy}\lambda_k$ and $\mu_{xy}\mu_k$ respectively. The likelihood contribution from each interval is then the likelihood of an observed or censored exponential for the home component multiplied by a censored or observed exponential for the away component respectively. The contribution from the final period that does not end in a goal is a censored exponential for both components. Then taking the product over each of the intervals the likelihood for match k is

$$L(\mathbf{t}_k, \mathbf{J}_k) = \exp(-\Lambda[0, 1]) \exp(-\Upsilon[0, 1]) \prod_{l=1}^{m_k} \lambda_k(t_{k,l})^{1-J_{k,l}} \mu_k(t_{k,l})^{J_{k,l}}, \tag{3.5}$$

where

$$\Lambda[t_1, t_2] = \int_{t_1}^{t_2} \lambda_k(t)\, \mathrm{d}t$$

and

$$\Upsilon[t_1,\ t_2] = \int_{t_1}^{t_2} \mu_k(t)\,\mathrm{d}t$$

are the home and away integrated intensities respectively. Scores between matches are assumed to be independent, and so the overall likelihood is given by taking the product over matches. In the remainder of this section, the models are fitted to the 4012 matches described in Section 2.

3.1.2. Regression models

As specified, the model is overparameterized; we reduce the dimension via a hierarchy of regression models. With n teams, in each model there are $2n - 1$ team-specific parameters and a home effect parameter, as in equation (3.2). In addition there are a varying number of within-match parameters, that are defined by our regression models. Setting $\lambda_{xy} = 1$ and $\mu_{xy} = 1$ for all x and y and fixing $\rho_i = 1$, $i = 1, 2$, leads to the same likelihood as the homogeneous Poisson model. Maximum likelihood is used to estimate the $2n = 184$ parameters in this simplest model, termed model I, and standard errors are obtained by using the observed information matrix (Cox and Hinkley, 1974). Table 2 gives the log-likelihood (up to proportionality) and the number of parameters for this and subsequent models. The home parameter estimate is $\hat{\gamma}_h = 1.37$, indicating a home advantage; the relative advantage from playing at home remains approximately constant for all subsequent models. Introducing the injury time parameters, defined in equations (3.4), gives model II; maximum likelihood estimates (with standard errors in parentheses) of the injury time parameters are $\hat{\rho}_1 = 1.48$ (0.14) and $\hat{\rho}_2 = 2.18$ (0.17). Thus, as expected, there is a significant increase in rate for goals recorded as 45 or 90 minutes because of injury time, i.e. $\hat{\rho}_1 > 1$ and $\hat{\rho}_2 > 1$.

This basic model is extended to model III by defining the λ_{xy}, in terms of two parameters λ_{10} and λ_{01}, by

$$\lambda_{xy} = \begin{cases} 1 & \text{for } x - y = 0, \\ \lambda_{10} & \text{for } x - y \geqslant 1, \\ \lambda_{01} & \text{for } x - y \leqslant -1 \end{cases}$$

with μ_{xy} defined similarly for μ_{10} and μ_{01}. The injury time parameters are also included. We have abused the notation slightly here, since, for example, λ_{10} represents the rate for scores other than (1, 0) and indicates that the home team is leading. Likelihood ratio tests in conjunction with parameter estimates show that this model gives a significantly improved fit over model I and model II. Many other models are fitted by defining λ_{xy} and μ_{xy} in this piecewise constant manner, and likelihood ratio tests are used to assess parsimonious fits. The best-fitting model in this time homogeneous case is defined as model IV, which defines λ_{xy} by six parameters as

Table 2. Log-likelihood values for the model fits

Model	*Number of parameters*	*Log-likelihood*
I	184	0.00
II	186	46.33
III	190	92.42
IV	198	114.00
V	188	126.70
VI	196	150.41

$$\lambda_{xy} = \begin{cases} 1 & \text{for } x = 0,\ y = 0, \\ \lambda_{10} & \text{for } x = 1,\ y = 0, \\ \lambda_{01} & \text{for } x = 0,\ y = 1, \\ \lambda_{11} & \text{for } x = 1,\ y = 1, \\ \lambda_{22} & \text{for } x - y = 0,\ x,\ y \geqslant 2, \\ \lambda_{21} & \text{for } x - y \geqslant 1,\ x \geqslant 2, \\ \lambda_{12} & \text{for } x - y \leqslant -1,\ y \geqslant 2. \end{cases}$$

The six away parameters μ_{xy} are defined similarly.

Now consider models that have continuously varying rates with time, but which do not have rates that depend on the current score. In this case the varying intensities are incorporated by considering the process as a time inhomogeneous birth process. The likelihood of this process is taken to be equation (3.5), only now the intensities are allowed to vary as a function of time. In particular, we model a linear change by defining

$$\lambda_k^*(t) = \lambda_k(t) + \xi_1 t,$$

$$\mu_k^*(t) = \mu_k(t) + \xi_2 t$$

and using λ_k^* (and μ_k^*) in place of λ_k (and μ_k) in equation (3.5). This model, termed model V, is similar to model II with two time variation parameters ξ_1 and ξ_2 added. Other models, such as quadratic variation, match-specific rates or an exponential form to guarantee positivity, can easily be fitted. However, we found the above linear model to be adequate in practice and this gives estimates $\hat{\xi}_1 = 0.70\,(0.08)$ and $\hat{\xi}_2 = 0.65\,(0.07)$.

Finally, fitting a model that incorporates both types of intensity variation and fitting a variety of regression models suggests that both score dependence and time variation effects are evident, and the best-fitting model is model VI. This model has parameters and estimates as defined in Table 3. Examples of team-specific parameter estimates are given in Table 4. The conclusions drawn from this final model are as follows.

(a) The scoring rate generally increases for both teams throughout the match ($\hat{\xi}_1$ and $\hat{\xi}_2$ are significantly greater than 0). This is most likely due to tiredness of players that leads to mistakes in defending.
(b) The attack and defence parameters generally decrease and increase respectively from the premiership down to the third division.
(c) The scoring rates of home and away teams depend on the current score. If the scores are level, the scoring rates are similar to those at (0, 0). If the home team is leading, the home and away rates generally decrease and increase respectively. This may be due to defending a lead, or trying to restore equality. If the away team is leading, the rates of both home and away teams tend to increase. A possible explanation for this is that a draw for the away

Table 3. Maximum likelihood estimates, with standard errors in parentheses, obtained by using model VI

Match state	*Home team parameters*	*Away team parameters*
(1, 0)	$\hat{\lambda}_{10} = 0.86\ (0.05)$	$\hat{\mu}_{10} = 1.33\ (0.09)$
(0, 1)	$\hat{\lambda}_{01} = 1.10\ (0.08)$	$\hat{\mu}_{01} = 1.07\ (0.08)$
(x, y), $x + y > 1$ and $x - y \geqslant 1$	$\hat{\lambda}_{21} = 1.01\ (0.06)$	$\hat{\mu}_{21} = 1.53\ (0.11)$
(x, y), $x + y > 1$ and $x - y \leqslant -1$	$\hat{\lambda}_{12} = 1.13\ (0.10)$	$\hat{\mu}_{12} = 1.16\ (0.11)$
Time variation	$\hat{\xi}_1 = 0.67\ (0.08)$	$\hat{\xi}_2 = 0.47\ (0.07)$

Table 4. Maximum likelihood estimates of a sample of the team-specific parameters, with standard errors in parentheses, obtained by using model VI†

Team	*Division*	$\hat{\alpha}$	$\hat{\beta}$
Manchester United	Premier	2.55 (0.31)	0.28 (0.06)
Newcastle United	Premier	2.27 (0.29)	0.36 (0.07)
Birmingham City	1	1.10 (0.14)	0.60 (0.10)
Sheffield United	1	1.17 (0.17)	0.68 (0.11)
Carlisle United	2	0.86 (0.12)	0.91 (0.15)
Stockport County	2	0.89 (0.13)	0.84 (0.13)
York City	2	0.76 (0.12)	0.90 (0.14)
Leyton Orient	3	0.35 (0.08)	1.34 (0.20)
Colchester United	3	0.68 (0.10)	1.25 (0.20)

†A full set of parameter estimates for 1997 is available from M. Robinson.

team is considered a good result, whereas for the home team a draw is a bad result. This would lead to the home team defending a narrow lead more strongly than the away team in a similar situation.

(d) The scoring rates generally increase once a goal has been scored. This supports the commonly held opinion that teams play more openly once 'the deadlock has been broken'.

It is possible that the scoring rates are independent of the current score for most of the game and change only in the final part of the match. Although we found no evidence for this, this may be due to the limited information that was available.

Table 5 summarizes the results from model VI for a particular match (Newcastle United (home) *versus* Manchester United); for example, the home and away scoring rates at the kick-off

Table 5. Estimated scoring rates conditional on the scores for Newcastle United *versus* Manchester United on March 2nd, 1996†

Score	*Estimated scoring rates for the following times and scores:*					
	Time = 0 min			*Time = 30 min*		
	0	*1*	*2*	*0*	*1*	*2*
0	1.00 (0.26)	1.10 (0.29)	1.14 (0.31)	1.23 (0.26)	1.32 (0.29)	1.36 (0.31)
	0.92 (0.20)	0.98 (0.23)	1.06 (0.25)	1.07 (0.20)	1.14 (0.23)	1.22 (0.25)
1	0.87 (0.23)	1.00 (0.26)	1.14 (0.31)	1.09 (0.23)	1.23 (0.26)	1.36 (0.31)
	1.22 (0.28)	0.92 (0.20)	1.06 (0.25)	1.38 (0.28)	1.07 (0.20)	1.22 (0.25)
2	1.01 (0.27)	1.01 (0.27)	1.00 (0.26)	1.24 (0.27)	1.24 (0.27)	1.23 (0.26)
	1.40 (0.32)	1.40 (0.32)	0.92 (0.20)	1.56 (0.32)	1.56 (0.32)	1.07 (0.20)
	Time = 60 min			*Time = 89 min*		
0	1.45 (0.26)	1.55 (0.30)	1.58 (0.31)	1.67 (0.27)	1.77 (0.30)	1.80 (0.31)
	1.23 (0.20)	1.30 (0.23)	1.38 (0.25)	1.38 (0.21)	1.45 (0.24)	1.53 (0.25)
1	1.32 (0.23)	1.45 (0.26)	1.58 (0.31)	1.53 (0.24)	1.67 (0.27)	1.80 (0.31)
	1.54 (0.27)	1.23 (0.20)	1.38 (0.25)	1.69 (0.28)	1.38 (0.21)	1.53 (0.25)
2	1.46 (0.27)	1.46 (0.27)	1.45 (0.26)	1.68 (0.27)	1.68 (0.27)	1.67 (0.27)
	1.72 (0.32)	1.72 (0.32)	1.23 (0.20)	1.87 (0.32)	1.87 (0.32)	1.38 (0.21)

†The rates are given at four times of the match, and row labels represent the home team score. For example the scoring rate of Newcastle in the 89th minute if the score is (1, 0) is 1.53. Note that the scoring intensities are based on the goals per match and not goals per minute. Approximate standard errors, in parentheses, are obtained by using the delta method.

are 1.00 and 0.92, with approximate standard errors, obtained by using the delta method, 0.26 and 0.20. At time 30 minutes, if the score is (0, 1), the scoring rate of the home team goes up to 1.32 goals per 90 minutes.

3.1.3. Assessing the model fit

The last 600 games of the 1995–1996 season have been retained to be used as a cross-validation assessment of the model fit. These games have not been included in our model development of Section 3 and are treated as a hold-out sample. The data are in addition to the 4012 games of Section 2. The column headed 'Observed' in Table 6 gives the proportion of these remaining 600 games that ended in the scores given in the first column. The other columns give the mean proportions of these games that ended in each score obtained by simulating from model VI. The general agreement of the simulated probabilities with the observed proportions, each being contained within the 95% confidence intervals, goes some way to showing that model VI is a good fit to future data.

4. Applications

4.1. Match outcome probabilities

The motivation for the development of the full-time scores model of Dixon and Coles (1997) was to estimate the probability of full-time outcomes (home win, draw or away win) for future matches. We now consider how to use our score model to estimate such probabilities. Recalling the birth process formulation of Fig. 3, we require the probability of being in each state $\{(x, y): x, y = 0, 1, \ldots\}$ at 90 minutes. This is given by integrating over all possible times and for each possible route to arrive at the point (x, y). Since the heavy computation makes direct calculation infeasible, we use Monte Carlo techniques and for each particular match we simulate the goal process from our fitted model. This leads to estimates of the distribution of the final score, and hence probabilities of match outcomes. Table 7 compares a selection of match outcome probabilities for model VI and model I. Accounting for the standard errors, the home and away win probability estimates are similar; the draw probabilities generally differ. Dixon and Coles (1997) examined the full-time score results and found a complex dependence structure between home and away scores that they modelled approximately in an *ad hoc* way. Model VI successfully captures the observed behaviour without having explicitly modelled such dependence. For example the

Table 6. Observed and simulated score probabilities†

Score	*Model VI probability*	*Approximate confidence interval*	*Observed*
(0, 0)	0.09	(0.08, 0.11)	0.09
(1, 0)	0.12	(0.11, 0.14)	0.13
(0, 1)	0.08	(0.07, 0.09)	0.07
(1, 1)	0.13	(0.12, 0.14)	0.13
(x, 0)	0.13	(0.11, 0.16)	0.16
(x, 1)	0.16	(0.13, 0.17)	0.15
(0, y)	0.06	(0.05, 0.07)	0.07
(1, y)	0.09	(0.08, 0.10)	0.09
(x, y)	0.14	(0.11, 0.16)	0.11

†The x and y represent scores 2, 3, ..., over which proportions have been aggregated. Approximate 95% confidence intervals were obtained by simulation from the asymptotic distribution of the maximum likelihood estimates.

Table 7. Maximum likelihood estimates of match outcome probabilities for five example matches obtained by using model I and model VI†

Match	*Estimates from model I*			*Estimates from model VI*		
	Home win	*Draw*	*Away win*	*Home win*	*Draw*	*Away win*
Newcastle United *versus* Manchester United	0.39 (0.07)	0.26 (0.02)	0.35 (0.07)	0.38 (0.08)	0.30 (0.02)	0.32 (0.07)
Birmingham City *versus* Sheffield United	0.48 (0.07)	0.26 (0.02)	0.26 (0.07)	0.47 (0.08)	0.29 (0.03)	0.24 (0.07)
Carlisle United *versus* Stockport County	0.43 (0.07)	0.26 (0.02)	0.31 (0.06)	0.44 (0.08)	0.28 (0.02)	0.28 (0.07)
Leyton Orient *versus* Colchester United	0.29 (0.06)	0.28 (0.02)	0.43 (0.06)	0.29 (0.07)	0.32 (0.02)	0.39 (0.08)
Manchester United *versus* York City	0.92 (0.03)	0.06 (0.02)	0.02 (0.01)	0.93 (0.04)	0.05 (0.03)	0.02 (0.01)

†Approximate standard errors, obtained by simulation from the asymptotic distribution of the maximum likelihood estimates, are given in parentheses.

estimated probability that a random match will end in a home win, a draw or an away win is given in Table 8. The row labelled 'Observed' is the proportion of all matches that ended in that outcome and estimates from the models are obtained by simulation from the fitted model. It is seen that model I generally underestimates the probability of draws and overestimates away win probabilities whereas model VI more accurately reflects the observed proportions. Similar findings are obtained for the probabilities of exact scores. The close agreement with the observed probabilities provides further evidence that the model captures many features of the observed data.

4.2. Footballing clichés

The results of Section 3.1 indicate both a general increase in scoring rates throughout the match, possibly due to fatigue of players, and a change in scoring rate which depends on the current score and which may be due to defending a lead and/or trying to restore equality.

The model can also be used to examine whether there is any evidence for the immediate strike back by introducing a new parameter δ that measures the scoring rate immediately after a team has conceded a goal, relative to that predicted by our model. For example if the home team concedes a goal at time t_1 then $\lambda_k(t)$ in equation (3.5) is replaced by $\delta\lambda_k(t)$ for $t \in (t_1, t_1 + \epsilon)$ where ϵ is a small time interval. Fitting model VI with this effect included gives $\hat{\delta} = 0.70\,(0.05)$ and $0.97\,(0.04)$ for $\epsilon = 2$ and $\epsilon = 5$ minutes, suggesting that there is no evidence for the immediate strike back, and in fact the opposite appears to be true, i.e. teams are less likely to concede a goal immediately after scoring than in open play. This may be due to the time that it takes for the match to restart following a goal. The immediate strike-back effect has been suggested probably because people have a tendency to overestimate the frequency of surprising events.

Table 8. Observed and simulated average probabilities of a home win, draw and away win

	Probabilities for the following results:		
	Home win	*Draw*	*Away win*
Observed	0.452	0.287	0.260
Model I	0.463	0.256	0.281
Model VI	0.458	0.287	0.255

5. Application to spread betting

5.1. Introduction

This section gives a brief overview of spread betting. For a more detailed examination, see Jackson (1994). Spread bets can be made on a variety of sports, and the easiest way to describe spread betting is via an example in cricket. Imagine that England have just been put into bat in a test-match, and a spread betting company (a bookmaker) is taking bets on how many runs will be scored in England's innings. The way in which the bets are made is as follows. If the bookmakers think that England will score, say, between 200 and 220 runs, then they offer a *spread* of 200–220 runs. As a bettor, you then have two options. Firstly, if you think that England will score *more* than 220 runs then you would bet, say, £1 per run on 'higher than 220' which is termed *buying runs at 220 for £1 per run*. If you buy runs and on completion of their innings England make 350 runs then you make a profit of $350 - 220 = 130 \times £1 = £130$. If, in contrast, you buy at 220 and England make 150, then you make a loss of $220 - 150 = £70$. Alternatively, if you think that England will score fewer than 200, then you would bet 'lower than 200' or *sell runs at 200 for £1 per run*. This time, if England make 350 runs, you make a loss of $350 - 200 = £150$, or if they score 150 runs, you make a profit of $200 - 150 = £50$.

The terminology arises because spread betting has many features in common with trading on the stock-market. For instance each type of spread offered is termed a (betting) market. Prices are offered on many *commodities* such as the number of points that a team will score in a league competition or the number of seats which will be won by a party in a parliamentary election.

One further feature, which occurs in many markets, is something termed 'betting in the running'. Here, the spreads are continuously updated, and trading can continue throughout the course of an event. Thus in the cricket example runs may be bought and sold at any time until the end of the innings at the spread currently being offered, and the spread fluctuates as England's innings unfolds.

Two common markets for spread betting on football matches are the total number of goals to be scored in a match and the difference between the final home and away scores. For these markets, the spreads are usually quoted as fractions of a goal, to represent some form of an average outcome. For example, in a match between two teams A and B, the bookmaker might display a spread for the total number of goals scored as 2.2–2.4. In this case the bettor then has the option either to buy at 2.4 or to sell at 2.2. If four goals are scored, then buying leads to a gain of $4.0 - 2.4 = 1.6$, and selling to a loss of $4.0 - 2.2 = 1.8$. The spread is updated, and trading may continue, throughout the 90 minutes of the match.

A final consideration is that in practice different prices may be quoted by competing spread betting companies for the same event. This affects both bettors and bookmakers, as rational bettors will choose the best price on offer at a given time, whereas bookmakers may want to attract bets by adjusting prices. This aspect is illustrated in the next section.

5.2. Application example

For illustration, we now consider how our model can be used to bet on (or to set) prices in the case where the commodity is the total number of goals scored in a match. The extension to betting on other football markets, such as the difference in the number of home and away goals scored, is immediate. Consider the game between Watford and Norwich City, played on November 27th, 1995. For this match, we have betting prices from two, competing, spread betting firms. We examine the problem from a bettor's viewpoint and exploit the competition in bookmakers by choosing to bet on the bookmaker with the best price. For example, we would buy at the lower of the buying prices (the higher quote in the spread) and sell at the higher of the selling prices. Thus,

in practice, the spread is usually less than the spreads quoted by individual bookmakers. Fig. 4 shows the best buying and selling prices that are available at each time point from 0 to 85 minutes. The upper and lower bounds of the shaded region represent the best buy and sell price respectively at each time point. Thus before the kick-off, at $t = 0$, both firms quoted a spread for the total number of goals as 2.4–2.7. At $t = 18$, the spreads were 2.2–2.5 and 2.0–2.3, giving the best spread width of 0.1 at $t = 18$. Two away goals were scored, at 32 and 46 minutes, and this is reflected in the two jumps at these times.

Model VI is used to compute the expected gain that arises from either selling or buying goals at any given time. Let $W(t)$ be a random variable that denotes the total goals scored at full time conditionally on the number scored at time t, and define $\Pr\{W(t) = i\} = a_i(t)$. Also let $c_s(t)$ and $c_b(t)$ be the best offered selling and buying prices at time t respectively. Then the estimated expected gain from buying at time t is given by

$$\sum_{i=0}^{\infty} \hat{a}_i(t)\{i - c_b(t)\} = E[W(t)] - c_b(t),$$

and for selling is $c_s(t) - E[W(t)]$. Simulation from model VI for this match is used to obtain $E[W(t)]$ for all t, and the resulting expected return for selling is shown in Fig. 5 for every time point $t \in [0, 85]$. If the model estimates are without error, we receive a positive expected return if we sell at times where the line lies above 0. In this example, it is always beneficial to sell and always detrimental to buy. Fig. 5 suggests that the expected gain is generally positive, even accounting for uncertainty in model parameter estimates, although it should be borne in mind that confidence intervals have been calculated assuming that the fitted model is correct, and that the only uncertainty is in the parameter estimates.

Data for continuously updated spread prices are not readily available: we have collected data for a few matches and most seem to exhibit inaccuracies such as that demonstrated above,

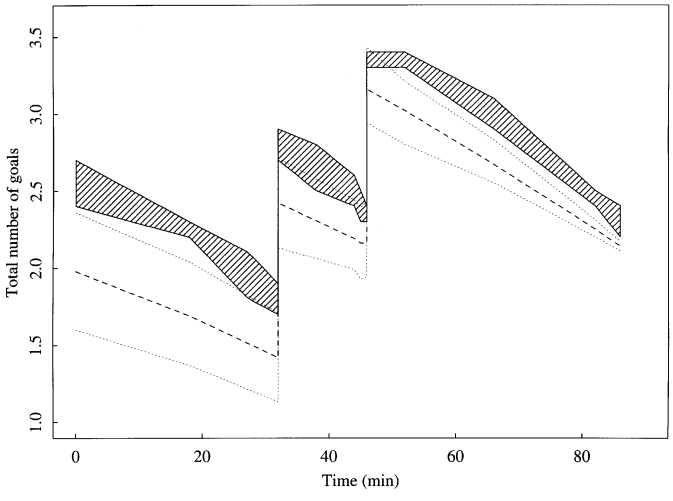

Fig. 4. Best offered selling and buying prices from two competing companies for Watford *versus* Norwich City, played on November 27th, 1995: for a given time point t, the upper and lower bounds of the shaded region represent the best buying and selling prices respectively from the two bookmakers (- - - - - -, estimated expected number of goals (obtained from model VI) that will be scored conditionally on the score at time t; ·····, approximate 90% confidence intervals, obtained from the asymptotic maximum likelihood estimates distribution); the jumps at $t = 32$ and $t = 46$ are due to goals being scored at these times; data were available only up to $t = 85$ minutes

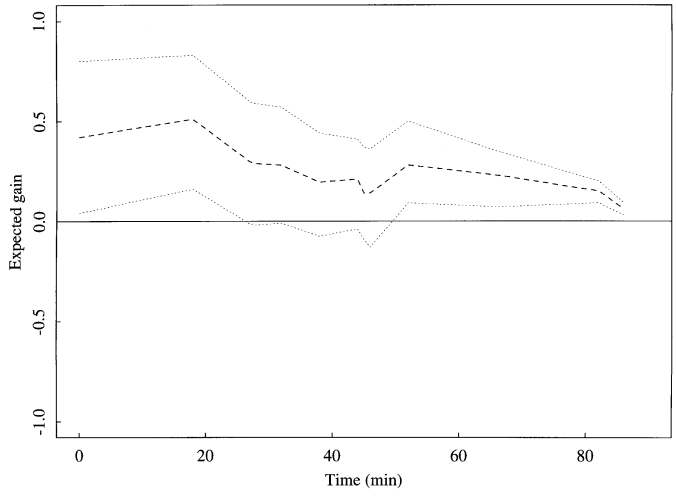

Fig. 5. Expected gain which arises from selling 1 unit at each time point throughout the match (·····, approximate 90% confidence intervals)

although with such small samples it is difficult to draw definite conclusions. We regard this as an important application area for further study, which will be of particular interest to spread betting companies.

6. Conclusions

Our model for the goal scoring process gives an improvement of match outcome estimates over the models of Maher (1982) and Dixon and Coles (1997). It is shown that there are two time inhomogeneous effects in the scoring process, these being a continuously increasing rate for both the home and the away teams, perhaps due to increased defensive mistakes as players become tired, and a variation due to the rate dependence on the current score. The dependence is most noticeable when the home team has a narrow lead when the home and away scoring rates decrease and increase significantly. We have found no evidence for the immediate strike back.

The main use of the model is to investigate setting prices in the spread betting market. Preliminary evidence suggests that there are inaccuracies in current prices.

Acknowledgements

We thank a referee for a more informative wording of the summary. We are grateful to Paul Blackwell at the University of Sheffield and Stuart Coles, Rob Henderson and Jonathan Tawn at Lancaster University for helpful discussions and advice on a previous version of the paper. We thank Sara Morris, formerly at Lancaster University, for assistance with data entry. The data were obtained from Williams (1992, 1993, 1994), Hugman (1991) and *90-minutes* weekly magazine.

References

Appleton, D. R. (1995) May the best man win? *Statistician*, **44**, 529–538.

Barnett, V. and Hilditch, S. (1993) The effect of an artificial pitch surface on home team performance in football (soccer). *J. R. Statist. Soc.* A, **156**, 39–50.

Chedzoy, O. (1995) Influences on the distribution of goals in soccer. *Private communication*.
Clarke, S. R. and Norman, J. M. (1995) Home ground advantage of individual clubs in English soccer. *Statistician*, **44**, 509–521.
Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman and Hall.
Crowder, M. J., Kimber, A. C., Smith, R. L. and Sweeting, T. J. (1991) *Statistical Analysis of Reliability Data*. London: Chapman and Hall.
Dixon, M. J. and Coles, S. G. (1997) Modelling association football scores and inefficiencies in the football betting market. *Appl. Statist.*, **46**, 265–280.
Hugman, B. J. (1991) *The Official Football League Yearbook*. Chichester: Facer.
Jackson, D. A. (1994) Index betting on sports. *Statistician*, **43**, 309–315.
Ladany, S. P. and Machol, R. E. (eds) (1977) *Optimal Strategies in Sport*. Amsterdam: North-Holland.
Maher, M. J. (1982) Modelling association football scores. *Statist. Neerland.*, **36**, 109–118.
Moller, J. and Sorenson, M. (1994) Statistical analysis of a spatial birth-and-death process model with a view to modelling linear dune fields. *Scand. J. Statist.*, **21**, 1–19.
Ridder, G., Cramer, J. S. and Hopstaken, P. (1994) Estimating the effect of a red card in soccer. *J. Am. Statist. Ass.*, **89**, 1124–1127.
Robinson, M. E. and Tawn, J. A. (1995) Statistics for exceptional athletics records. *Appl. Statist.*, **44**, 499–511.
Smith, R. L. (1988) Forecasting records by maximum likelihood. *J. Am. Statist. Ass.*, **83**, 331–338.
Stefani, R. and Clarke, S. (1992) Predictions and home advantage for Australian rules football. *J. Appl. Statist.*, **19**, 251–261.
Williams, T. (1992) *Football Club Directory.* Chichester: Hamsworth Active.
———(1993) *Football Club Directory.* Chichester: Hamsworth Active.
———(1994) *Football Club Directory.* Chichester: Hamsworth Active.