

# Modelling Association Football Scores and Inefficiencies in the Football Betting Market

By MARK J. DIXON† and STUART G. COLES

*Lancaster University, UK*

[Received November 1995. Revised September 1996]

## SUMMARY

A parametric model is developed and fitted to English league and cup football data from 1992 to 1995. The model is motivated by an aim to exploit potential inefficiencies in the association football betting market, and this is examined using bookmakers' odds from 1995 to 1996. The technique is based on a Poisson regression model but is complicated by the data structure and the dynamic nature of teams' performances. Maximum likelihood estimates are shown to be computationally obtainable, and the model is shown to have a positive return when used as the basis of a betting strategy.

**Keywords:** Betting strategy; Expected return; Football (soccer); Maximum likelihood; Poisson distribution

## 1. Introduction

Betting on the outcome of football (soccer) matches has a long tradition in the UK, most popularly in the form of football pools, which typically involve the selection of matches that are thought to be those most likely to be a draw. A recently introduced type of betting, *fixed odds betting*, is also rapidly increasing in popularity. Bookmakers offer odds on the various outcomes of a match. The simplest version of this uses just the outcome of the match, in the sense of it being a win by either the team playing at home or the team playing away, or a draw. More complicated bets can also be placed on the score or on the half-time and full-time results. In making bets the challenge then is to find 'good bets', in which the considered probability of occurrence is higher than the corresponding probability determined by the bookmakers' odds, so that there is a positive expected return. Unlike in other types of betting, such as in horse-racing, the odds are fixed around one week before the matches are played. This allows a detailed comparison of the bookmakers' odds with estimated probabilities so that any perceived weaknesses in the bookmakers' specification can be exploited. Consequently, a statistical model that is capable of accurately predicting probabilities of the outcome of football matches has the potential to form the basis of a profitable betting strategy. This paper develops a model that meets this requirement.

†Address for correspondence: Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK.  
E-mail: m.dixon@lancaster.ac.uk

Various proposals have been made for modelling the outcome of football matches; these are reviewed in Section 2. For a betting strategy, however, probabilities must be estimated on a team-specific basis, so that the probabilities of the various match outcomes between two specific teams on a particular date can be calculated. This degree of resolution falls outside the scope of most of the published models. An exception to this is the model due to Maher (1982), that assumes independent Poisson distributions for the number of goals scored by each of the home and away teams, with means that are specific to each team's past performance. This forms the basis of our modelling approach. However, in attempting to derive a model which is not just a reasonable description of the data, but which also has the potential to provide better estimates of probabilities than the subjective estimates ascribed by bookmakers, we have had to modify and enhance this basic model structure. These modifications account for the fluctuating performance of individual teams and also enable the estimation of match outcomes for cup competitions in which teams from different leagues play one another. One consequence of these modifications is that simple equations for the maximum likelihood estimators are no longer available, but despite the high dimensionality of the model we show that maximum likelihood estimators are still available numerically. From the fitted model, the probabilities of the outcomes of each match are calculated and compared with the bookmakers' odds; this underlies the specification of a betting strategy which, using historical data, we show to have a positive return.

Section 2 reviews the literature discussing the use of statistical methodology in summarizing data from football matches. The data available to us are described in Section 3. Section 4 develops a statistical model, building on the basic model structure of Maher (1982). The application of the model to our assimilated data is described together with some sample results in Section 4. The utility of the model as the basis for a betting strategy is outlined in Section 5. Finally Section 6 suggests refinements which, we believe, would lead to further improvements in return.

## 2. Literature Review

Surprisingly few papers have examined the use of statistical techniques for modelling football data. American National Football League (NFL) football has received much more attention, but the differences between the two sports mean that modelling techniques for NFL football do not naturally transfer to association football.

Early references to statistical modelling of football data concentrate mainly on the distribution of the number of goals scored in a game. Moroney (1956) briefly examined this problem and suggested that, although the Poisson distribution provided an adequate fit to scores, improvements could be obtained by working with the negative binomial distribution. Reep *et al.* (1971) similarly examined the fit of the negative binomial distribution to scores from football matches and other goal scoring games. They concluded that 'chance dominates the game', finding no way of predicting outcomes within their class of models given the inherent noise in observed data. In contrast Hill (1974) applied a simple comparisons test for final league placings with expert predictions and demonstrated a significant correlation. A more sophisticated analysis of this type was by Fahrmeir (1994), who applied newly developed techniques for time-dependent, ordered, paired comparisons to German football data.

These points illustrate an apparent dichotomy: in the long run, it is not difficult to predict fairly accurately which teams are likely to be successful, but the development of models that have a sufficiently high resolution to exploit this long run predictive capability for individual matches is substantially more difficult. To our knowledge, the only paper that derives a model for football scores in a match between specific teams, accounting for the different quality of the teams involved, is that by Maher (1982). He obtained maximum likelihood estimates for a model in which the scores of the home and away teams in any game are independent Poisson distributions, with means modelled as functions of the respective teams' previous performances. This approach forms the basis of our model in Section 4.

With somewhat different applications in mind, several papers have looked at the effect of specific circumstances on team performances: Barnett and Hilditch (1993) applied standard nonparametric tests to see whether artificial pitches, subsequently banned in the English league, gave a significant advantage to the home team; Ridder *et al.* (1994) investigated the effect of the sending-off of a player on the outcome of a football match. Other papers have used statistical models to describe aspects of individual matches themselves: Chedzoy (1995) informally investigated times when goals are scored; Reep and Benjamin (1968) modelled the number and type of passing moves within a game; Clarke and Norman (1995) investigated the advantage of playing at home.

In relation to betting strategies, papers on the efficiency and exploitation of betting markets are numerous in the economics literature. Many papers address horse-racing and NFL football betting, and a few also consider betting on football matches, though little use of statistical methodology is made in these. Discussions of various betting markets can be found in Golec and Tamarkin (1991), Hausch *et al.* (1981) and, specifically in the context of football betting, Pope and Peel (1989).

### 3. Data

A wealth of information is available from each football match played. Obviously scores are recorded, but also 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 subjective. Consequently, our model exploits only each team's history of match scores, which we have assimilated over a 3-year period, though the possibility of including other forms of data is investigated in Section 6.

The available data, which comprise 6629 full-time league and cup match results from the seasons 1992–93, 1993–94 and 1994–95, each consist of a home score and an away score. Data from 1995–96 are also available but are used as a validation sample to test the utility of the model subsequently when used as a basis for a betting strategy. The data from 1992 to 1995 provide accurate empirical estimates of various aggregated features. Table 1 gives the relative frequency, expressed as a percentage, of the scores from 0–0 to 4–4. Standard errors on the basis of an underlying multinomial model are shown in parentheses. Aggregating, the ratio of frequencies of home wins, draws and away wins is found to be 46:27:27. Thus, an empirical estimate of the probability of a randomly selected match resulting in a home win, for example,

TABLE 1

*Empirical estimates of each score probability for joint and marginal probability functions†*

Home goals	Estimates of score probabilities (%) for the following numbers of away goals:					
	Away	0	1	2	3	4
0	22.1 (0.36)	33.4 (0.74)	36.4 (0.57)	19.5 (0.49)	7.9 (0.42)	2.1 (0.16)
1	33.0 (0.65)	8.2 (0.32)	7.4 (0.28)	4.5 (0.23)	1.4 (0.13)	0.4 (0.06)
2	24.5 (0.51)	10.3 (0.38)	12.7 (0.30)	6.4 (0.24)	2.7 (0.15)	0.6 (0.07)
3	12.6 (0.40)	8.2 (0.31)	9.1 (0.25)	4.8 (0.22)	1.9 (0.14)	0.5 (0.09)
4	5.3 (0.31)	4.2 (0.25)	4.5 (0.25)	2.3 (0.19)	1.2 (0.11)	0.4 (0.06)
5		1.6 (0.14)	1.8 (0.13)	1.1 (0.13)	0.6 (0.07)	0.1 (0.04)

†Standard errors are given in parentheses.

is 0.46. Because of the size of the database, these empirical estimates provide accurate estimates of random match probabilities. Our objective in later sections is to obtain estimates in matches which are not randomly selected but are team specific.

The assumption that the marginal distribution of random match scores is Poisson can be examined at this stage. Fitting a Poisson distribution to the aggregated home and away scores in Table 1 reveals that by any criterion the Poisson model is a near perfect fit to the aggregated score data. This gives some reassurance that the Poisson regression model developed in Section 4 is at least plausible for our data, despite concerns raised by other researchers about the general appropriateness of the Poisson assumption. A further assumption of the basic model in Section 4 is that the home and away scores are independent. To assess the validity of this assumption, Table 2 displays

$$\frac{\tilde{f}(i, j)}{\tilde{f}_H(i)\tilde{f}_A(j)}$$

for each home and away score  $(i, j)$ ,  $i = 0, \dots, 6$  and  $j = 0, \dots, 5$ , where  $\tilde{f}$ ,  $\tilde{f}_H$  and  $\tilde{f}_A$  are the joint and marginal empirical probability functions for home and away scores respectively. Bootstrap standard errors are given in parentheses. Table 2 suggests

TABLE 2

*Estimates of the ratios of the observed joint probability function and the empirical probability function obtained under the assumption of independence between the home and away scores†*

Home goals	Estimates of ratios for the following numbers of away goals:					
	0	1	2	3	4	5
0	111.5 (3.52)	92.0 (2.87)	103.4 (4.18)	82.1 (7.67)	96.4 (15.31)	96.8 (28.12)
1	93.7 (2.43)	105.7 (2.00)	99.3 (3.74)	103.7 (6.31)	86.9 (13.15)	108.3 (19.99)
2	99.6 (2.91)	101.7 (2.11)	99.2 (3.78)	97.4 (7.41)	95.9 (17.4)	106.7 (23.77)
3	100.3 (4.25)	98.5 (3.61)	91.8 (6.51)	116.6 (11.03)	139.8 (23.85)	75.4 (40.5)
4	91.0 (7.07)	93.8 (7.16)	108.6 (10.74)	138.0 (16.31)	111.7 (32.86)	90.4 (55.33)
5	94.1 (13.24)	102.3 (12.28)	114.3 (20.6)	73.3 (31.01)	120.8 (74.71)	130.4 (129.7)
6	139.1 (31.95)	49.1 (23.66)	146.4 (41.33)	45.3 (57.84)	174.1 (122.2)	—

†The numbers are multiplied by 100 for clarity. Standard errors are given in parentheses.

that the assumption of independence between scores is reasonable, except for the scores 0–0, 1–0, 0–1 and 1–1. Based on the estimates and errors alone, the score 0–3 seems to be significantly underestimated by the independence model. However, viewed within the context of all the other results, we regard this as due to sampling error. A modification of the independence assumption in the light of these observations is considered in Section 4.

## 4. Model and Inference

### 4.1. Model Specification

With the aim of developing a profitable betting strategy, various features are required of a statistical model for football matches. For example:

- (a) the model should take into account the different abilities of both teams in a match;
- (b) there should be allowance for the fact that teams playing at home generally have some advantage—the so-called ‘home effect’;
- (c) the most reasonable measure of a team’s ability is likely to be based on a summary measure of their recent performance;
- (d) the nature of football is such that a team’s ability is likely to be best summarized in separate measures of their ability to attack (to score goals) and their ability to defend (not to concede goals);
- (e) in summarizing a team’s performance by recent results, account should be taken of the ability of the teams that they have played against.

It is not practical to obtain empirical estimates of probabilities of match outcomes that account for all these constraints. Instead, we use a statistical model that structurally incorporates each of these features. Our basis is the model proposed by Maher (1982), with modifications to enable the inclusion of non-complete data sets, and data from different divisions simultaneously, and to allow for fluctuations in team performance.

The basic assumption of Maher’s 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 respective 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

$$\begin{aligned} X_{i,j} &\sim \text{Poisson}(\alpha_i \beta_j \gamma), \\ Y_{i,j} &\sim \text{Poisson}(\alpha_j \beta_i), \end{aligned} \tag{4.1}$$

where  $X_{i,j}$  and  $Y_{i,j}$  are independent,  $\alpha_i, \beta_i > 0, \forall i$ , the  $\alpha_i$  measure the ‘attack’ rate of the teams, the  $\beta_i$  measure the ‘defence’ rates and  $\gamma > 0$  is a parameter which allows for the home effect. In fact, Maher (1982) included a more general model specification than this, allowing for separate home and away, and attack and defence parameters for every team. However, like Maher (1982), we have found model (4.1) to be an adequate simplification, though there are still assumptions in this model that would not be supported by detailed study of match data. The essential point is that, although details of the model may be inaccurate, the global structure should be

sufficiently accurate to enable the development of a betting strategy with a positive expected (and realized) return.

Some aspects of the model are easily improved on, however. Consider first the assumption of independence. Maher (1982) suggested the use of a bivariate Poisson family as an extension of the basic model, but this family is unable to represent the departure from independence for low scoring games that we identified in Section 3. Instead, we propose the following modification of model (4.1):

$$\Pr(X_{i,j} = x, Y_{i,j} = y) = \tau_{\lambda,\mu}(x, y) \frac{\lambda^x \exp(-\lambda)}{x!} \frac{\mu^y \exp(-\mu)}{y!} \quad (4.2)$$

where

$$\begin{aligned} \lambda &= \alpha_i \beta_j \gamma, \\ \mu &= \alpha_j \beta_i \end{aligned}$$

and

$$\tau_{\lambda,\mu}(x, y) = \begin{cases} 1 - \lambda\mu\rho & \text{if } x = y = 0, \\ 1 + \lambda\rho & \text{if } x = 0, y = 1, \\ 1 + \mu\rho & \text{if } x = 1, y = 0, \\ 1 - \rho & \text{if } x = y = 1, \\ 1 & \text{otherwise.} \end{cases}$$

In this model,  $\rho$ , where

$$\max(-1/\lambda, -1/\mu) \leq \rho \leq \min(1/\lambda\mu, 1),$$

enters as a dependence parameter:  $\rho = 0$  corresponds to independence, but otherwise the independence distribution is perturbed for events with  $x \leq 1$  and  $y \leq 1$ . It is easily checked that the corresponding marginal distributions remain Poisson with means  $\lambda$  and  $\mu$  respectively.

Another limitation of the model is that it is static—the attack and defence parameters of each team are regarded as constant through time. This issue will be considered in Section 4.3.

#### 4.2. Model Inference

It follows from model (4.2) that with  $n$  teams there are attack parameters  $\{\alpha_1, \dots, \alpha_n\}$ , defence parameters  $\{\beta_1, \dots, \beta_n\}$ , the dependence parameter  $\rho$  and the home effect parameter  $\gamma$  to be estimated. To prevent the model from being over-parameterized, we impose the constraint

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

For the English league system, which comprises the Premier League and divisions 1–3 of the Football League,  $n = 92$ , so the model has 185 identifiable parameters.

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

$$L(\alpha_i, \beta_i, \rho, \gamma; i = 1, \dots, n) = \prod_{k=1}^N \tau_{\lambda_k, \mu_k}(x_k, y_k) \exp(-\lambda_k) \lambda_k^{x_k} \exp(-\mu_k) \mu_k^{y_k} \quad (4.3)$$

where

$$\begin{aligned} \lambda_k &= \alpha_{i(k)} \beta_{j(k)} \gamma, \\ \mu_k &= \alpha_{j(k)} \beta_{i(k)}, \end{aligned} \quad (4.4)$$

and  $i(k)$  and  $j(k)$  denote respectively the indices of the home and away teams playing in match  $k$ . With complete data, in the sense of each team having played every other team equally often, and in the simpler case of independence between home and away team scores ( $\rho = 0$ ), Maher (1982) obtained a system of linear equations whose roots are the maximum likelihood estimates. To achieve greater generality, we are restricted to direct numerical maximization of equation (4.3). The near orthogonality of many parameter combinations means that this is straightforward, despite the high dimensionality of the model.

In equation (4.3), teams from all four divisions are included in the likelihood. This has two consequences: firstly, the parameters for each team should reflect the relative quality of the different divisions and, secondly, the parameters will be estimable only if there is information from matches between teams of different divisions. Fortunately, because there is some mobility between the teams of different divisions at the start of a new season due to promotion and relegation, the issue of parameter identifiability is resolved. The situation is also helped by the inclusion of results from cup games which involve teams of different divisions playing each other. Then, because the parameters are calibrated across the divisions, the model can validly be used to estimate the probabilities of match outcomes involving teams of different divisions, as in cup games for example. These points are illustrated by Table 3, which shows the mean attack and defence parameters for teams in each division. As expected, the average attack and defence rating of teams increases with higher league status, as measured by increasing and decreasing mean values of  $\alpha$  and  $\beta$  respectively.

#### 4.3. Model Enhancement

A structural limitation of model (4.3) is that the parameters are static, i.e. teams are assumed to have a constant performance rate, as determined by  $\alpha_i$  and  $\beta_i$ , over

TABLE 3  
*Mean attack and defence parameters for teams within each division*

<i>League</i>	<i>Mean attack parameter <math>\bar{\alpha}</math></i>	<i>Mean defence parameter <math>\bar{\beta}</math></i>
Premier	1.38	0.68
Division 1	1.07	0.86
Division 2	0.83	1.14
Division 3	0.73	1.32

time. In reality, a team's performance tends to be dynamic, varying from one time period to another, and this behaviour should be incorporated in the model. In particular a team's performance is likely to be more closely related to their performance in recent matches than in earlier matches. In principle, this behaviour could be modelled by formalizing a stochastic development of the model parameters; this is considered in Section 6. In view of the dimensionality of the model, however, and since we shall always need to estimate the parameters at the fixed time point of making a bet rather than, say, forecasting ahead, we take a more simplistic approach here. Thus we assume that the parameters are, in a loose sense, locally constant through time and that historical information is of less value than recent information, and we determine parameter estimates for each time point  $t$  that are based on the history of match scores up to time  $t$ . Modifying equation (4.3) we construct a 'pseudolikelihood' for each time point  $t$ ,

$$L_t(\alpha_i, \beta_i, \rho, \gamma; i = 1, \dots, n) = \prod_{k \in A_t} \{\tau_{\lambda_k, \mu_k}(x_k, y_k) \exp(-\lambda_k) \lambda_k^{x_k} \exp(-\mu_k) \mu_k^{y_k}\}^{\phi(t-t_k)} \quad (4.5)$$

where  $t_k$  is the time that match  $k$  was played,  $A_t = \{k: t_k < t\}$ ,  $\lambda_k$  and  $\mu_k$  are as in equations (4.4) and  $\phi$  is a non-increasing function of time. This represents a slight abuse of notation since the parameters  $\alpha_i, \beta_i, \rho$  and  $\gamma$  are themselves time dependent.

Maximizing equation (4.5) at time  $t$  leads to parameter estimates which are based on games up to time  $t$  only. In this way, the model has the capacity to reflect changes in team performance. Moreover, varying the choice of  $\phi$  allows historical data to be downweighted in the likelihood to a greater or lesser degree.

#### 4.4. Choice of Weighting Function $\phi$

There are various possible choices for the weighting function  $\phi$  in equation (4.5). One possibility would be

$$\phi(t) = \begin{cases} 1 & t \leq t_0, \\ 0 & t > t_0, \end{cases}$$

in which case, at time  $t$ , all results within the last  $t_0$  time units would be given equal weight in the inference. Instead, we work with the model

$$\phi(t) = \exp(-\xi t),$$

in which all previous results, downweighted exponentially according to a parameter  $\xi > 0$ , are included in the inference at time  $t$ . The static model (4.3) arises as the special case  $\xi = 0$ , whereas taking increasingly large values of  $\xi$  gives relatively more weight to the most recent results.

Optimizing the choice of  $\xi$  is problematic, since equation (4.5) defines a sequence of non-independent 'likelihoods', whereas we require  $\xi$  such that the overall predictive capability of the model is maximized. In fact, in subsequent sections, we restrict attention to the prediction of match outcomes rather than match scores. Therefore it is pragmatic to choose  $\xi$  to optimize the prediction of outcomes. First note that the probability of a home win in match  $k$  is estimated as



$$p_k^H = \sum_{l,m \in B_H} \Pr(X_k = l, Y_k = m) \quad (4.6)$$

where  $B_H = \{(l, m): l > m\}$ , and the score probabilities are determined from the maximization of model (4.5) at  $t(k)$ , the time of match  $k$ . Similar expressions hold for  $p_k^A$  and  $p_k^D$ , the probabilities of an away win and a draw respectively. Now define

$$S(\xi) = \sum_{k=1}^N (\delta_k^H \log p_k^H + \delta_k^A \log p_k^A + \delta_k^D \log p_k^D) \quad (4.7)$$

where, for example,  $\delta_k^H = 1$  if match  $k$  is a home win and  $\delta_k^H = 0$  otherwise, and  $p_k^H$ ,  $p_k^A$  and  $p_k^D$  are the maximum likelihood estimates from model (4.5), with weighting parameter set at  $\xi$ . Considering only the outcomes, and not the scores, equation (4.7) is the analogue of a predictive profile log-likelihood. A plot of  $S(\xi)$  against  $\xi$ , with time units taken to be half-weeks, is given in Fig. 1. The function is maximized at  $\xi = 0.0065$ , and all subsequent results are given with respect to this choice of  $\xi$ , though in fact the results are robust across a range of  $\xi$ -values.

#### 4.5. Parameter Estimates and Results

The complete set of parameter estimates, obtained by maximizing equation (4.5) with  $\xi = 0.0065$ , at each time point  $t$ , gives a profile of each team's changing performance in terms of defence and attack abilities. Data from at least 60 half-weeks are required to estimate parameters accurately, so estimates are obtained for  $t$  ranging from 60 to 174. For brevity we show only a subset of the results (for the full set of results for the 1996 season, contact M. Dixon). Tables 4 and 5 give the

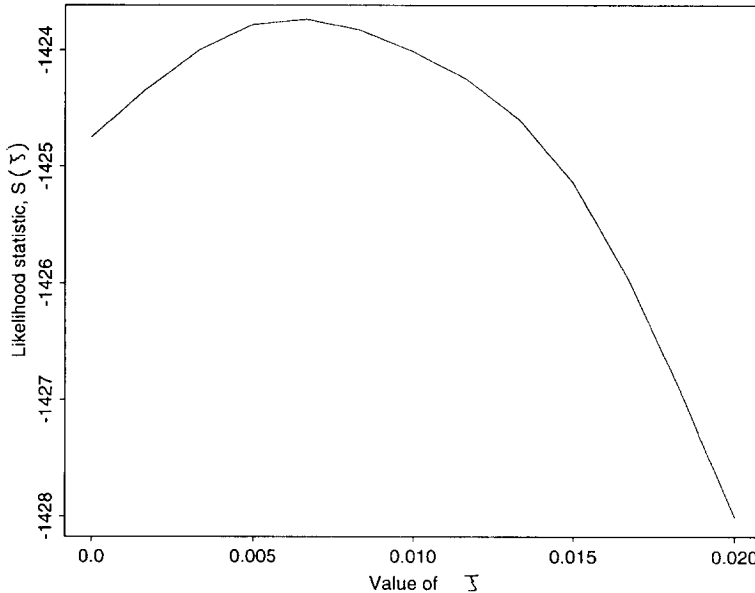


Fig. 1.  $S(\xi)$  versus  $\xi$ : the maximum occurs at  $\xi = 0.0065$

TABLE 4

*Maximum likelihood estimates and standard errors for the attack and defence rate parameters, on August 5th, 1995, for Premiership teams (in the 1995–96 season)*

<i>Team</i>	$\hat{\alpha}$	$se(\hat{\alpha})$	$\hat{\beta}$	$se(\hat{\beta})$
Arsenal	1.235	0.151	0.527	0.078
Aston Villa	1.278	0.178	0.649	0.086
Blackburn	1.730	0.209	0.534	0.082
Bolton	1.183	0.141	0.760	0.100
Chelsea	1.238	0.169	0.658	0.089
Coventry	1.115	0.164	0.699	0.094
Everton	1.177	0.169	0.667	0.091
Leeds	1.510	0.186	0.583	0.088
Liverpool	1.448	0.180	0.561	0.082
Manchester City	1.232	0.170	0.728	0.091
Manchester United	1.869	0.208	0.402	0.067
Middlesbrough	1.244	0.152	0.750	0.109
Newcastle	1.659	0.195	0.578	0.081
Nottingham Forest	1.460	0.170	0.658	0.095
Queen's Park Rangers	1.497	0.195	0.717	0.095
Sheffield Wednesday	1.387	0.179	0.698	0.091
Southampton	1.446	0.183	0.772	0.098
Tottenham	1.622	0.201	0.775	0.100
West Ham	1.192	0.169	0.649	0.087
Wimbledon	1.281	0.174	0.732	0.094

TABLE 5

*Maximum likelihood estimates and standard errors for the attack and defence rate parameters on August 5th, 1995, for teams in division 2 (in the 1995–96 season)*

<i>Team</i>	$\hat{\alpha}$	$se(\hat{\alpha})$	$\hat{\beta}$	$se(\hat{\beta})$
Blackpool	0.858	0.110	1.357	0.163
Bournemouth	0.681	0.096	1.095	0.139
Bradford	0.832	0.107	1.175	0.149
Brentford	0.967	0.115	0.900	0.129
Brighton	0.820	0.107	1.032	0.137
Bristol City	0.825	0.120	0.873	0.114
Bristol Rovers	0.917	0.113	0.965	0.138
Burnley	0.942	0.116	1.067	0.127
Carlisle	0.781	0.100	0.964	0.157
Chesterfield	0.764	0.099	1.024	0.160
Crewe	1.065	0.125	1.265	0.159
Hull	0.822	0.104	1.069	0.146
Notts County	0.985	0.132	0.979	0.120
Oxford	0.956	0.119	0.951	0.119
Peterborough	0.829	0.111	1.161	0.133
Rotherham	0.852	0.106	1.136	0.143
Shrewsbury	0.764	0.104	1.060	0.145
Stockport	0.945	0.115	1.040	0.136
Swansea	0.798	0.106	0.899	0.122
Swindon	1.160	0.154	1.091	0.120
Walsall	0.911	0.114	1.116	0.162
Wrexham	0.957	0.116	1.257	0.150
Wycombe	0.813	0.105	0.984	0.137
York	0.916	0.113	0.926	0.131

maximum likelihood estimates of the attack and defence parameters on August 5th, 1995, for teams which were in the Premier division and division 2 respectively in 1995–96. In addition, Fig. 2 shows the corresponding sequence of estimates of  $\alpha(t)$  and  $\beta(t)$  over time for three teams—the non-uniformity in these estimates suggests that teams' performances are genuinely dynamic. Also shown is the sequence of estimates of the home effect parameter  $\gamma(t)$  which, as might be expected, remains nearly constant over time. The flat portion from  $t = 82$  to  $t = 90$  corresponds to the summer break in the football season. Furthermore, Table 6 gives a sample of matches and the maximum likelihood estimates of the outcome probabilities. The standard errors of the outcome estimates, particularly those of the draw probability estimates, are small relative to the standard errors of the attack and defence parameter estimates.

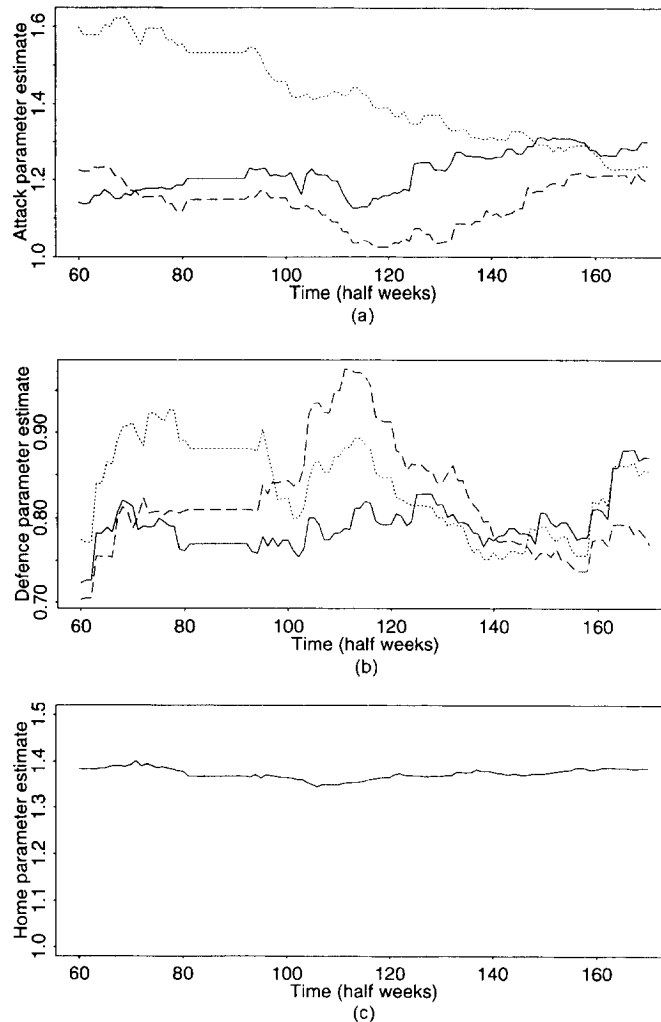


Fig. 2. (a), (b) Time series of the maximum likelihood estimates of attack and defence rate parameters for Sheffield United (—), Norwich (.....) and Everton (---); (c) variation of the common home effect parameter with time

TABLE 6  
Maximum likelihood estimates for match outcome probabilities†

Match	Maximum likelihood estimates for the following outcomes:		
	Home win	Draw	Away win
Arsenal <i>versus</i> Middlesbrough	0.535 (0.069)	0.280 (0.030)	0.184 (0.046)
Aston Villa <i>versus</i> Manchester United	0.214 (0.054)	0.291 (0.029)	0.495 (0.072)
Blackburn <i>versus</i> Queen's Park Rangers	0.615 (0.078)	0.221 (0.033)	0.164 (0.049)
Chelsea <i>versus</i> Everton	0.457 (0.075)	0.298 (0.030)	0.245 (0.057)
Liverpool <i>versus</i> Sheffield Wednesday	0.535 (0.076)	0.262 (0.031)	0.203 (0.052)
Blackpool <i>versus</i> Wrexham	0.428 (0.077)	0.240 (0.018)	0.332 (0.070)
Stockport <i>versus</i> Burnley	0.480 (0.077)	0.259 (0.024)	0.261 (0.062)
Newcastle <i>versus</i> Stoke	0.705 (0.073)	0.198 (0.042)	0.097 (0.034)

†The matches are a subset of the fixtures from August 19th, 1995, plus one other across-division match. Approximate standard errors are calculated using the delta method. Standard errors are given in parentheses.

## 5. Betting Strategy

How useful is the model derived in Section 4, when used as the basis for a betting strategy against odds provided by bookmakers? A detailed investigation into betting strategies for fixed odds football betting is given in Pope and Peel (1989) and Dixon and Pope (1996). Here we address the question by reference to a new set of results, corresponding to the 1995–96 season, for which we have both results and bookmakers' odds. We first use model (4.5), with  $\xi = 0.0065$ , at each new time point  $t$  to obtain current parameter estimates. Then, by comparing estimated result probabilities with bookmakers' odds for the following week, we determine which games are most advantageous to bet on. We then calculate the net return from such a strategy.

A typical set of bookmakers' odds for a particular match might be (8:13, 12:5, 4:1) for a home win, draw and away win respectively. Thus, in this example, a stake of 13 units on a home win would yield a profit of 8 units if that outcome occurred. Odds  $o_1:o_2$  transform to a probability  $p$  by using the formula

$$p = o_2 / (o_1 + o_2).$$

The above set of odds then corresponds to the set of probabilities (0.62, 0.29, 0.20), which has a sum of 1.11. This phenomenon is standard in betting markets: if the bookmakers are accurate in their probability specifications, they have an in-built 'take', corresponding to their expected profit, which in the above example is 11%. To win money from bookmakers, in the sense of having a positive expected return, requires a determination of probabilities which is sufficiently more accurate than those obtained from the odds in order to overcome the bookmakers' take. We first rescale multiplicatively the bookmakers' odds so that they sum to 1. Denote these probabilities for match  $k$  by  $b_k^H$ ,  $b_k^D$  and  $b_k^A$  for a home win, draw and away win respectively, and similarly let  $\hat{p}_k^H$ ,  $\hat{p}_k^D$  and  $\hat{p}_k^A$  be the corresponding maximum likelihood estimates for this match under model (4.5). Comparisons of the two sets of probability estimates for each of the result outcomes are given in Fig. 3 for each match in our database. Overall there is reasonable agreement between the probability assessments, but the variability in these plots indicates the potential for positive gain if our model probabilities are accurate.

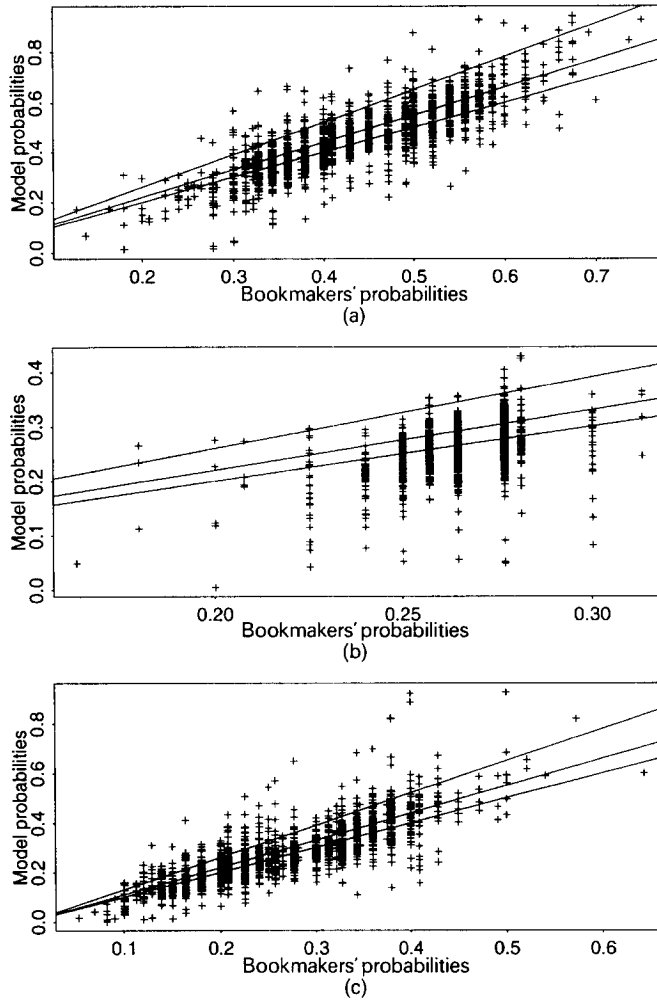


Fig. 3. Model probability estimates plotted against odds for all matches where odds were available: (a) home wins; (b) draws; (c) away wins

If the model probabilities were without error, then the expected gain from a unit stake bet on a home win, for example, is

$$E(G) = p_k^H / b_k^H - 1. \quad (5.1)$$

If  $b_k^H$  is the true probability then the expected gain will be 0.00, or  $-0.11$  if the bookmakers' take is included. In reality neither  $p_k^H$  nor  $b_k^H$  is the true probability, but we obtain a positive return if our estimates are sufficiently more accurate than the bookmakers'.

From equation (5.1), a natural betting strategy for any particular game  $k$  is to bet on a home win, for example, if

$$\hat{p}_k^H / \tilde{b}_k^H > r,$$

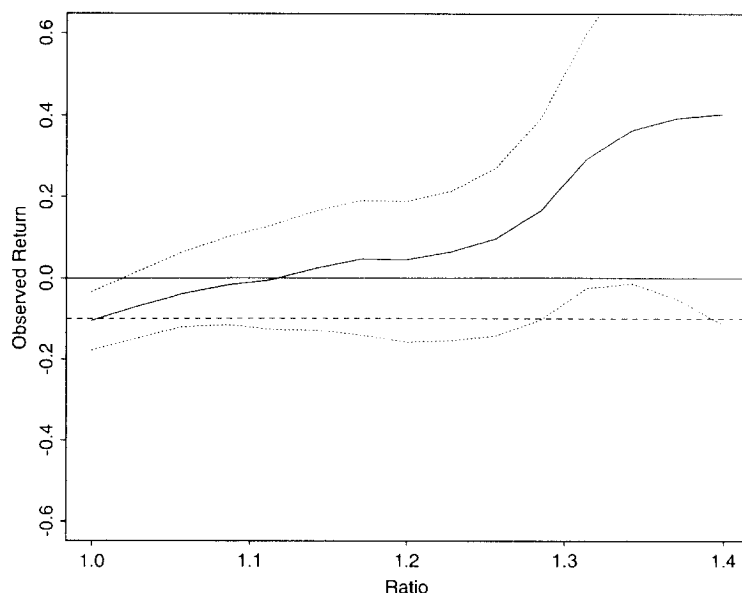


Fig 4. Observed mean return plotted against the ratio of model probabilities to bookmakers' probabilities (---, return of  $-0.11$ , the expected return under random betting, which is due to the bookmakers' take of 11% for each match; ..... , 90% bootstrap intervals): the mean return is calculated only when there are more than 10 sample values

where  $\hat{b}_k^H$  denotes the unscaled bookmakers' probability of a home win in match  $k$ , for some predetermined value of  $r > 1$ , with a corresponding strategy for bets on away wins or draws. Increasing  $r$  leads to a stricter betting regime, but consequently fewer bets. The effect of a range of different choices for  $r$  can be seen in Fig. 3, in which the line  $\hat{p} = rb$  is plotted for  $r = 1.0, 1.1, 1.3$ . For a particular choice of  $r$ , points falling above that line correspond to matches on which this betting strategy would have led to a bet being placed on that particular outcome, with  $r$  as specified.

The success of this betting strategy can be assessed by calculating the observed return, if such a strategy had been adopted, given the match results which actually occurred. This is plotted as a function of  $r$  in Fig. 4, together with 90% confidence intervals obtained using the bootstrap. There is considerable variability in the plot which makes it difficult to draw definitive conclusions. However, with  $r = 1.2$  our betting strategy leads to a return which is borderline significantly different from  $-0.11$ , the expected return under a random betting strategy due to the bookmakers' take, and has a positive expected absolute return for any  $r > 1.1$ . It is in this sense that we claim that the model and inference of Section 4 meet our stated objective of deriving a model for estimating football match outcomes which is the basis of a betting strategy with positive returns.

## 6. Conclusions

Our aim has been to derive a method for estimating the probabilities of football results with the potential to achieve a positive expected return when used as the basis

of a betting strategy against bookmakers' odds. Our basic model is simple—a bivariate Poisson distribution for the numbers of goals scored by each team, with parameters related to past performance—but refinements necessary to improve the realism and precision of the model make the associated inference a heavy computational burden. None-the-less, the computations are manageable and the resulting model is accurate in many respects.

Our betting strategy is equally simple: we bet on all outcomes for which the ratio of model to bookmakers' probabilities exceeds a specified level. For sufficiently high levels, we have shown that this strategy yields a positive expected return, even allowing for the in-built bias in the bookmakers' odds.

The simplicity of our model and the associated betting strategy is appealing. However, to improve further on the utility of our approach, we perceive that further modifications may be desirable. One possibility is to refine further the Poisson regression model. Stochastically updated parameters are a natural idea in this context, but the detailed implementation may be difficult. Smith (1981) considered a dynamic regression framework for simple Poisson models, but the generalization of these ideas to the scale of model (4.5) is not immediate. Broadening the scope of our model to incorporate additional covariate information is a second area for development. The quantitative value of such data is not always obvious, so such a development might need a Bayesian structure to exploit their subjective value. A third possibility is the betting regime. We have restricted attention so far to fixed odds bets of match *outcomes*. This leads to a betting strategy in which relatively few bets are actually placed. As bookmakers offer odds on specific match scores, the probabilities of which are also obtained from our model, a betting strategy based on match scores could be developed. More radically, there are several 'market-style' index betting options for football matches, where goal margins are bought and sold like commodities (e.g. Jackson (1994) and Dixon and Robinson (1996)); the implementation of our model for market strategies in such an option is a further possibility.

The lure of scientific improvement of our model and betting strategy, with the purely incidental by-product of winning money from the bookmakers, encourages us to build on the apparent success of the present model in the various ways discussed above.

### Acknowledgements

We thank a referee and the Editor for comments on an earlier version of the paper and Jonathan Tawn for comments that helped to improve the handling of dependence between scores. We are grateful to many colleagues at Lancaster who helped to digitize football league data and bookmakers' odds. For entry of data and odds, we thank Simon Garside, Sara Morris, Mike Robinson and Pete Sewart. The data were obtained from Williams (1992, 1993, 1994), Hugman (1991) and *90-Minutes* weekly magazine. Odds were obtained from local branches of major UK bookmakers. We also thank Mike Robinson and Barry Rowlingson for writing some of the software used in the handling of data.

### References

- 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.
- Dixon, M. J. and Pope, P. (1996) Inefficiency and bias in the UK football betting market. Submitted to *Managmt Sci.*
- Dixon, M. J. and Robinson, M. E. (1996) A point process model for goal times in association football matches. Submitted to *J. Am. Statist. Ass.*
- Fahrmeir, L. (1994) Dynamic-stochastic models for time-dependent ordered paired comparison systems. *J. Am. Statist. Ass.*, **89**, 1438–1449.
- Golec, J. and Tamarkin, M. (1991) The degree of inefficiency in the football betting market. *J. Finan. Econ.*, **30**, 311–323.
- Hausch, D. B., Ziemba, W. T. and Rubinstein, M. (1981) Efficiency of the market for racetrack betting. *Managmt Sci.*, **27**, 1435–1452.
- Hill, I. D. (1974) Association football and statistical inference. *Appl. Statist.*, **23**, 203–208.
- Hugman, B. J. (1991) *The Official Football League Yearbook*. Chichester: Facer.
- Jackson, D. A. (1994) Index betting on sports. *Statistician*, **43**, 309–315.
- Maher, M. J. (1982) Modelling association football scores. *Statist. Neerland.*, **36**, 109–118.
- Moroney, M. J. (1956) *Facts from Figures*, 3rd edn. London: Penguin.
- Pope, P. F. and Peel, D. A. (1989) Information, prices and efficiency in a fixed-odds betting market. *Economica*, **56**, 323–341.
- Reep, C. and Benjamin, B. (1968) Skill and chance in association football. *J. R. Statist. Soc. A*, **131**, 581–585.
- Reep, C., Pollard, R. and Benjamin, B. (1971) Skill and chance in ball games. *J. R. Statist. Soc. A*, **134**, 623–629.
- 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.
- Smith, J. Q. (1981) The multiparameter steady model. *J. R. Statist. Soc. B*, **43**, 256–260.
- Williams, T. (1992) *Football Club Directory*. Exeter: Hamsworth Active.
- (1993) *Football Club Directory*. Exeter: Hamsworth Active.
- (1994) *Football Club Directory*. Exeter: Hamsworth Active.