Available online at www.sciencedirect.com

SCIENCE DIRECT®

International Journal of Forecasting 21 (2005) 331–340

www.elsevier.com/locate/ijforecast

Regression models for forecasting goals and match results in association football

John Goddard*

Department of Economics, University of Wales Swansea, Singleton Park, Swansea SA2 8PP, UK

Abstract

In the previous literature, two approaches have been used to model match outcomes in association football (soccer): first, modelling the goals scored and conceded by each team; and second, modelling win–draw–lose match results directly. There have been no previous attempts to compare the forecasting performance of these two types of model. This paper aims to fill this gap. Bivariate Poisson regression is used to estimate forecasting models for goals scored and conceded. Ordered probit regression is used to estimate forecasting models for match results. Both types of models are estimated using the same 25-year data set on English league football match outcomes. The best forecasting performance is achieved using a 'hybrid' specification, in which goals-based team performance covariates are used to forecast win–draw–lose match results. However, the differences between the forecasting performance of models based on goals data and models based on results data appear to be relatively small.

Keywords: Bivariate Poisson; Ordered probit; Football match results

1. Introduction

There are two distinct strands of empirical literature on modelling the outcomes of matches in association football (soccer). The first approach, favoured by most applied statisticians, involves modelling the number of goals scored and conceded directly. Forecasts of win–draw–lose match results can be derived indirectly, by aggregating the estimated probabilities assigned to appropriate permutations of goals scored and conceded by the two teams. The second approach, developed recently by a number of applied econometricians, involves modelling win–draw–lose results directly, using discrete choice regression models such as ordered probit or logit.

Although these two strands in the literature have developed independently, it is natural to ask which approach, if any, is superior for forecasting. Accurate forecasting matters, not only to statisticians for whom uncovering regularities in empirical goal scoring or match results data sets is a relevant and interesting exercise in its own right, but also to bookmakers and bettors who may have a significant financial interest

* Tel.: +44 1792 205678x4835; fax: +44 1792 295872.
E-mail address: j.a.goddard@swan.ac.uk.

0169-2070/$ - see front matter © 2004 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2004.08.002

in being able to assess accurately prior probabilities for football match outcomes. Previous contributions to the literature on both sides have avoided drawing direct comparisons between goals- and results-based models, perhaps because there has been considerable diversity in both the techniques employed, and the scope of the data sets examined. This article draws direct comparisons between the forecasting performance of goals- and results-based models. The comparisons are facilitated by estimating both types of model over data sets identical in every respect other than their focus on goals or results, respectively.

A win–draw–lose match results data set is effectively 'nested' within a goals data set: the result of a match is established from the goals scored by the two teams, but the result on its own does not indicate the number of goals scored. Any direct comparison between the forecasting capabilities of the two types of models must be based on forecasts of results: while goals-based models can forecast goals and results, results-based models forecast results only. One possible prior is that a goals-based model should outperform a results-based model, because the former draws on a more extensive data set than the latter. On the other hand, given that league points are awarded for results and not goals (with goals relevant only for separating teams with equal points totals), insofar as they affect points totals, wins of 1–0 or 6–3 are of equal worth. What is crucial is which (if either) team won; the number of goals scored and conceded are incidental. Consequently, goals data might contain more noise than results data. Moreover, it is clear from the literature on modelling goal scoring (see Section 2) that the choice of distributional assumptions and the treatment of the problem of interdependence between the goals scored by the two teams are difficult issues, which have been addressed in various ways by different researchers. In contrast, the use of discrete choice regression appears to be relatively uncontroversial in the literature on modelling results. Accordingly, a results-based model might be expected to outperform a goals-based model, on the grounds that model selection and specification issues are more straightforward.

The rest of this paper is structured as follows. Section 2 provides a brief review of the previous literature on forecasting match outcomes in association football. Section 3 specifies the bivariate Poisson and ordered probit regressions that are used to model goals and results in this paper. Section 4 describes the data set and presents the estimation results. Section 5 compares the forecasting performance of the models developed in the previous sections. Section 6 concludes.

2. Literature review

In an early contribution to the literature on modelling goal scoring, Maher (1982) uses univariate and bivariate Poisson distributions, with means reflecting the attacking and defensive capabilities of the two teams. Attacking and defensive parameters for each team are estimated ex post (after a complete set of data for each season have been collected), but the model is not capable of predicting scores or results for individual matches ex ante (before the match is played). A tendency to underestimate the proportion of draws when using univariate distributions is attributed to interdependence between the goals scored by the home and away teams, and is corrected using a bivariate Poisson specification.[1]

Dixon and Coles (1997) develop a forecasting model capable of generating ex ante probabilities for goals and match results. Home and away team goals follow univariate Poisson distributions, and for low-scoring matches an ad hoc adjustment to the probabilities corrects for interdependence. Dixon and Pope (2004) compare probabilistic forecasts obtained from the Dixon–Coles model with probabilities inferred from UK bookmakers' prices for fixed-odds betting. Using a forecasting approach similar to that of Dixon-Coles, Rue and Salvesen (2000) allow the attacking and defensive parameters for all teams to vary randomly over time. The estimates of these parameters are updated as new data on match outcomes are obtained. Markov chain Monte Carlo iterative simulation techniques are used for inference. Crowder, Dixon, Ledford, and Robinson (2002) propose a procedure for updating the team

[1] In the present data set, for 29,562 matches played between the 1986–1987 and 2000–2001 seasons (inclusive), Pearson's correlation between the goals scored by the home and away teams is +0.0199. The null hypothesis that the true correlation is zero is rejected at the 1% level.

strength parameters that is computationally less demanding.

The impact of specific factors on match outcomes has been considered in a number of studies. Barnett and Hilditch (1993) investigate whether artificial playing surfaces, adopted by a few teams during the 1980s and 1990s, conferred an additional home-team advantage. Ridder, Cramer, and Hopstaken (1994) show that player dismissals have a negative effect on the match outcome for the teams concerned. Clarke and Norman (1995) quantify the effect of home advantage on match outcomes. Dixon and Robinson (1998) investigate variations in the scoring rates of the home and away teams during the course of a match. The scoring rates at any time are partly dependent on the time elapsed, and partly on which (if either) team is leading at the time. Using data on the 1998 international World Cup tournament, Dyte and Clarke (2000) examine the relationship between a set of pre-tournament national team rankings, and the teams' actual performance in the tournament.

Recently, several applied econometricians have used discrete choice regression models to model win–draw–lose match results directly, rather than indirectly through goals scored and conceded.[2] Forrest and Simmons (2000a, 2000b) investigate the predictive quality of newspaper tipsters' results forecasts, and the performance of the pools panel in providing hypothetical results for postponed matches. Koning (2000) estimates a model to describe a set of match results ex post, as part of a broader analysis of changes in competitive balance in Dutch football. Audas, Dobson, and Goddard (2002) examine whether managerial change has any short-term impact on subsequent team performance measured by results. Goddard and Asimakopoulos (2004) and Kuypers (2000) estimate ex ante forecasting models to investigate the efficiency of prices quoted by high-street bookmakers for fixed-odds betting on results. Cain, Law, and Peel (2000) and Dixon and Pope (2004) investigate similar issues for fixed-odds betting on goals.

3. Regression models for goal scoring and match results

Section 3 specifies the bivariate Poisson and ordered probit regression models used in this paper to produce forecasts of football match outcomes. Full definitions of the covariates used in these models are given in Appendix A. Summary definitions are as follows.

$F^{d}_{i,y,s}$, $A^{d}_{i,y,s}$=Average numbers of goals scored and conceded by team i, indexed by period prior to current match (y), season (s) and division (d).
$P^{d}_{i,y,s}$=Team i's average recent results (on a scale of 1=win, 0.5=draw, 0=loss), indexed as above.
$S^{\mathrm{H}}_{i,m}$, $C^{\mathrm{H}}_{i,m}$, $S^{\mathrm{A}}_{i,n}$, $C^{\mathrm{A}}_{i,n}$=Goals scored and conceded in mth and nth most recent home and away matches by team i.
$R^{\mathrm{H}}_{i,m}$, $R^{\mathrm{A}}_{i,n}$=Results of team i's mth and nth most recent home and away matches.
SIGH$_{i,j}$=dummy variable identifying matches important for championship, promotion or relegation outcomes for home team i but not for away team j.
SIGA$_{i,j}$=as above, for matches important for away team j but not for home team i.
CUP$_{i}$=1 if team i is eliminated from the FA Cup; 0 otherwise.
DIST$_{i,j}$=natural logarithm of the geographical distance between the grounds of teams i and j.
AP$_{i,s}$=residual from regression of team i's average home attendance on league position, indexed by season (s).

For convenience, the following notation is used below to refer to groups of covariates:

Home team attack, away team defence goals covariates: $\mathbf{x_A}$=$\{F^{d}_{i,y,s}, A^{d}_{j,y,s}, S^{\mathrm{H}}_{i,m}, S^{\mathrm{A}}_{i,n}, C^{\mathrm{H}}_{j,n}, C^{\mathrm{A}}_{j,m}\}$;
Home team defence, away team attack goals covariates: $\mathbf{x_B}$=$\{A^{d}_{i,y,s}, F^{d}_{j,y,s}, C^{\mathrm{H}}_{i,m}, C^{\mathrm{A}}_{i,n}, S^{\mathrm{H}}_{j,n}, S^{\mathrm{A}}_{j,m}\}$;
Home and away team results covariates: $\mathbf{x_C}$=$\{P^{d}_{i,y,s}, P^{d}_{j,y,s}, R^{\mathrm{H}}_{i,m}, R^{\mathrm{A}}_{i,n}, R^{\mathrm{H}}_{j,n}, R^{\mathrm{A}}_{j,m}\}$;
Other covariates: $\mathbf{z}$=$\{$ SIGH$_{i,j}$, SIGA$_{i,j}$, DIST$_{i,j}$, CUP$_{i}$, CUP$_{j}$, AP$_{i,s}$, AP$_{j,s}\}$.

[2] Karlis and Ntzoufras (2003) discuss a third modelling approach, which involves modelling the difference between the scores of the two teams using the Poisson difference distribution. This approach, which is not pursued in the present paper, falls midway between the goals-based and results-based models that are reported below. In a goals-based model the (bivariate) dependent variables are S and C, the goals scored and conceded by the home team. Possible values taken by S and C are $\{0, 1, 2, \ldots\}$. In a difference-based model the (univariate) dependent variable is S–C. Possible values taken by S–C are $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$. In results-based models, the (univariate) dependent variable has only three possible categories, which correspond to S–C<0, S–C=0 and S–C>0, respectively.

For a bivariate Poisson regression with goals scored and conceded as outcomes, $S^{\mathrm{H}}_{i,0}$(=$C^{\mathrm{A}}_{j,0}$) and $C^{\mathrm{H}}_{i,0}$ (=$S^{\mathrm{A}}_{j,0}$) denote the goals scored and conceded by home team i in the current match, respectively. Following Holgate (1964), the bivariate Poisson joint probability function for $S^{\mathrm{H}}_{i,0}$ and $C^{\mathrm{H}}_{i,0}$ takes the form:

$$P\left(S^{\mathrm{H}}_{i,0}=s,\ C^{\mathrm{H}}_{i,0}=c\right)=\exp\left(-\lambda_{1,i,j}-\lambda_{2,i,j}+\lambda_{3,i,j}\right)\sum_{k=0}^{\min(s,c)}\lambda^{s-k}_{1,i,j}\lambda^{c-k}_{2,i,j}\lambda^{k}_{3,i,j}/\{(s-k)!(c-k)!k!\}$$

This joint probability function can be interpreted as the product of three univariate Poisson probability functions with means $\lambda_{1,i,j}-\lambda_{3,i,j}$, $\lambda_{2,i,j}-\lambda_{3,i,j}$ and $\lambda_{3,i,j}$, respectively. $\lambda_{1,i,j}$, the expected number of goals scored by the home team, depends on covariates reflecting the propensities of home team i to score and away team j to concede goals. Similarly, $\lambda_{2,i,j}$, the expected number of goals scored by the away team, depends on covariates reflecting the propensities of away team j to score and home team i to concede goals. Two specifications for $\lambda_{1,i,j}$ and $\lambda_{2,i,j}$ (Models 1 and 2) are considered. Both include a set of lagged team performance covariates which differ between the two specifications, and a set of other covariates common to both specifications. In Model 1, the performance covariates are based on lagged goals data: $\lambda_{1,i,j}$=$f_{11}(\mathbf{x_A},\ \mathbf{z})$, $\lambda_{2,i,j}$=$f_{12}(\mathbf{x_B},\ \mathbf{z})$. In Model 2, the corresponding covariates are based on lagged results data: $\lambda_{1,i,j}$=$f_{21}(\mathbf{x_C},\ \mathbf{z})$, $\lambda_{2,i,j}$=$f_{22}(\mathbf{x_C},\ \mathbf{z})$. f_{11}, f_{12}, f_{21} and f_{22} are linear functions. Model 1 is a 'pure' goals model, with lagged goals covariates used to model goals. Model 2 is a 'hybrid' model, with lagged results covariates used to model goals. Finally, in both models, $\lambda_{3,i,j}=\eta\sqrt{\lambda_{1,i,j}\lambda_{2,i,j}}$ is the covariance between $S^{\mathrm{H}}_{i,0}$ and $C^{\mathrm{H}}_{i,0}$. The prior is η>0.

Using Italian Serie A match results data, Karlis and Ntzoufras (2003) identify a tendency for the bivariate Poisson regression to underestimate slightly the probabilities for low-scoring draws. Accordingly, they suggest a modification that involves including additional parameters to inflate the relevant draw probabilities, and deflate the other probabilities.[3] In the present case, likelihood-ratio tests indicate that additional parameters inflating the probabilities of 0–0 and 1–1 draws are significant in both Models 1 and 2. Accordingly, the adjusted bivariate probability function used to estimate Models 1 and 2, which contains two additional parameters π and θ, is as follows:

$$\begin{aligned}\tilde{P}\left(S^{\mathrm{H}}_{i,0}=s,C^{\mathrm{H}}_{i,0}=c\right)&=(1-\pi)P\left(S^{\mathrm{H}}_{i,0}=s,C^{\mathrm{H}}_{i,0}=c\right)+\pi\theta && \text{for } \{s,c\}=\{0,0\};\\ \tilde{P}\left(S^{\mathrm{H}}_{i,0}=s,C^{\mathrm{H}}_{i,0}=c\right)&=(1-\pi)P\left(S^{\mathrm{H}}_{i,0}=s,C^{\mathrm{H}}_{i,0}=c\right)+\pi(1-\theta) && \text{for } \{s,c\}=\{1,1\};\\ \tilde{P}\left(S^{\mathrm{H}}_{i,0}=s,C^{\mathrm{H}}_{i,0}=c\right)&=(1-\pi)P\left(S^{\mathrm{H}}_{i,0}=s,C^{\mathrm{H}}_{i,0}=c\right) && \text{for } \{s,c\}\neq\{0,0\},\{1,1\}\end{aligned}$$

For an ordered probit regression model with win–draw–lose match results as outcomes, the result of the match between teams i and j, denoted $R^{\mathrm{H}}_{i,0}$ (=1−$R^{\mathrm{A}}_{j,0}$), depends on the unobserved or latent variable $y^{*}_{i,j}$ and a Normal independent and identically distributed disturbance term, $\varepsilon_{i,j}$, as follows:

$$\begin{array}{lll}\text{Homewin}: & R^{\mathrm{H}}_{i,0}=1 & \text{if } \mu_2<y^{*}_{i,j}+\varepsilon_{i,j}\\ \text{Draw}: & R^{\mathrm{H}}_{i,0}=0.5 & \text{if } \mu_1<y^{*}_{i,j}+\varepsilon_{i,j}<\mu_2\\ \text{Away win}: & R^{\mathrm{H}}_{i,0}=0 & \text{if } y^{*}_{i,j}+\varepsilon_{i,j}<\mu_1\end{array}$$

There are two alternative specifications for $y^{*}_{i,j}$. Model 3 is a 'hybrid' model, with lagged goals covariates used to model results: $y^{*}_{i,j}$=$f_3(\mathbf{x_A},\ \mathbf{x_B},\ \mathbf{z})$. Model 4 is a 'pure' results model, with lagged results covariates used to model results: $y^{*}_{i,j}$=$f_4(\mathbf{x_C},\ \mathbf{z})$. f_3 and f_4 are linear functions. Specifications similar to Model 4, with some variations in the selection of covariates, are reported by Audas et al. (2002) and Goddard and Asimakopoulos (2004).[4]

[3] As noted above, Dixon and Coles (1997) apply a similar adjustment to their probabilities for low-scoring match outcomes, obtained from univariate Poisson distributions.

[4] The assumption of Normality for the disturbance term in the ordered probit specification does not appear problematic. For example, in the version of Model 4 reported in Table 1 (below), Glewwe (1997) LM test fails to reject the null hypothesis that $\varepsilon_{i,j}$ is Normal tested against the alternative that $\varepsilon_{i,j}$ follows some other member of the Pearson family of distributions. The test fails to reject on all other (unreported) versions of Models 3 and 4 that are used to compile Table 3 (below).

4. Data and estimation results

Models 1 to 4 are used to generate four sets of win–draw–lose match results forecasts for matches played in the Premier League (PL) and the three divisions of the Football League (FL), for each of the 10 seasons from 1992–1993 to 2001–2002 inclusive. Below, Football League Divisions One to Three are abbreviated FLD1 to FLD3. The forecasts take the form of ex ante home win, draw and away win probabilities. The four sets of forecasts for each season are obtained from versions of the models estimated using data for the preceding 15 seasons. Accordingly, the estimation period for the models used to obtain the forecasts for 1992–1993 is 1977–1978 to 1991–1992; the estimation period for 2001–2002 forecasts is 1986–1987 to 2000–2001; and so on. Preliminary experimentation suggested that extending the estimation period up to about 15 seasons produced tangible benefits in terms of improved forecasting accuracy, but beyond 15 seasons there was little or no further gain. All data on match outcomes are obtained from various editions of *Rothmans Football Yearbook*.[5]

Table 1 reports the estimated versions of Models 1 and 4 based on data for seasons 1986–1987 to 2000–2001 (inclusive). To conserve space, the 'hybrid' Models 2 and 3 are not reported. The contribution to Models 1 and 4 of each set of covariates is now considered.

In Model 1, the average goals scored and goals conceded variables $F^d_{i,y,s}$ and $A^d_{i,y,s}$ (for team i and their counterparts for team j), calculated over the 24 months prior to the current match, are the main indicators of attacking and defensive capability, or team quality. The indexing of these variables allows for separate contributions to the team quality measures from goals scored and conceded in matches played: 0–12 months (y=0) or 12–24 months (y=1) before the current match; within the current season (s=0) or previous season (s=1) or two seasons ago (s=2); and in the team's current division (d=0) or one (d=±1) or two (d=±2) divisions above or below the current division.

Positive coefficients on $F^d_{i,y,s}$ and $A^d_{i,y,s}$ are expected. It is assumed that team i's propensity to score is captured by its average scoring rate over the previous 12 months, $F^0_{i,0,0}+\sum_{d=-1}^{+1}F^d_{i,0,1}$, and its average scoring rate between 12 and 24 months ago, $\sum_{d=-1}^{+1}F^d_{i,1,1}+\sum_{d=-2}^{+2}F^d_{i,1,2}$. The individual components of these sums make separate contributions to the attacking and defensive capability measures. For example, if the current-season scoring rate is a better indicator of the team's current attacking capability than the previous-season scoring rate in the same division within the same 12-month period, the coefficient on $F^0_{i,0,0}$ should (and does) exceed the coefficient on $F^0_{i,0,1}$. The covariates $A^d_{i,y,s}$, $F^d_{j,y,s}$ and $A^d_{j,y,s}$ all make similar contributions to Model 1, and the average 'points per match' (or win ratio) covariates $P^d_{i,y,s}$ and $P^d_{j,y,s}$ play an equivalent role in Model 4.

The estimated coefficients on all of these variables are predominantly correctly signed and well defined. Preliminary experimentation indicated that the coefficients on $\{F^d_{i,y,s}, A^d_{i,y,s}, F^d_{j,y,s}, A^d_{j,y,s}\}$ and $\{P^d_{i,y,s}, P^d_{j,y,s}\}$ were highly significant for y=0,1 (data from matches played 0–12 months and 12–24 months before the current match); but not for y=2 (24–36 months before the current match).

The variables $\{F^d_{i,y,s}, A^d_{j,y,s}, F^d_{j,y,s}, A^d_{j,y,s}\}$ play a similar role in Model 1 to the time-varying parameters reflecting attacking and defensive propensities in the models of Crowder et al. (2002), Dixon and Coles (1997) and Rue and Salvesen (2000). Table 2 illustrates the contribution of these covariates, using data for one arbitrarily chosen FLD1 team, Millwall. Prior to the 2001–2002 season, Millwall enjoyed two successful seasons in FLD2, finishing fifth in 1999–2000 and first in 2000–2001. As the 2001–2002 season unwinds over 23 home matches (col (1)), the evolution of the covariates contributing to Millwall's 'attack' coefficient (col (8)) is tracked in cols (2) to (7). For example, data from matches played in 2001–2002 accumulate (col (2)), while data from 1999–2000 drop out of the relevant data set (col (6)). In the event, Millwall enjoyed another successful season in 2001–2002, finishing fourth in FLD1. Although Millwall's win ratios in 2001–2002 and 1999–2000

[5] Occasional errors in the data recorded in *Rothmans* were identified by comparing final league tables for each season compiled from the recorded data against actual league tables, and noting anomalies. The errors were then eliminated by cross-checking the recorded match outcomes for teams with anomalies against data obtained from other (mainly Web-based) databases.

Table 1
Estimation results: Models 1 and 4, 1986–1987 to 2000–2001 seasons

Model 1: bivariate Poisson regression for goals								Model 4: ordered probit regression for match results			
Dep. var.=$\lambda_{1,i,j}$ (mean home goals)				Dep. var.=$\lambda_{2,i,j}$ (mean away goals)				Dep. var.=$y^{*}_{i,j}$ (latent variable)			
Cov.	Coeff.	Cov.	Coeff.	Cov.	Coeff.	Cov.	Coeff.	Cov.	Coeff.	Cov.	Coeff.
$F^{0}_{i,0,0}$	0.442***	$A^{0}_{j,0,0}$	0.410***	$F^{0}_{j,0,0}$	0.309***	$A^{0}_{i,0,0}$	0.303***	$P^{0}_{i,0,0}$	1.732***	$P^{0}_{j,0,0}$	−1.240***
$F^{+1}_{i,0,1}$	0.534***	$A^{+1}_{j,0,1}$	0.289***	$F^{+1}_{j,0,1}$	0.368***	$A^{+1}_{i,0,1}$	0.271***	$P^{+1}_{i,0,1}$	1.858***	$P^{+1}_{j,0,1}$	−1.516***
$F^{0}_{i,0,1}$	0.345***	$A^{0}_{j,0,1}$	0.386***	$F^{0}_{j,0,1}$	0.227***	$A^{0}_{i,0,1}$	0.368***	$P^{0}_{i,0,1}$	1.202***	$P^{0}_{j,0,1}$	−0.915***
$F^{-1}_{i,0,1}$	0.287***	$A^{-1}_{j,0,1}$	0.488***	$F^{-1}_{j,0,1}$	0.163***	$A^{-1}_{i,0,1}$	0.578***	$P^{-1}_{i,0,1}$	0.874***	$P^{-1}_{j,0,1}$	−0.578***
$F^{+1}_{i,1,1}$	0.178**	$A^{+1}_{j,1,1}$	0.102**	$F^{+1}_{j,1,1}$	0.081	$A^{+1}_{i,1,1}$	0.100***	$P^{+1}_{i,1,1}$	0.859***	$P^{+1}_{j,1,1}$	−0.728***
$F^{0}_{i,1,1}$	0.166***	$A^{0}_{j,1,1}$	0.114**	$F^{0}_{j,1,1}$	0.044	$A^{0}_{i,1,1}$	0.122***	$P^{0}_{i,1,1}$	0.719***	$P^{0}_{j,1,1}$	−0.541***
$F^{-1}_{i,1,1}$	0.104**	$A^{-1}_{j,1,1}$	0.251***	$F^{-1}_{j,1,1}$	0.016	$A^{-1}_{i,1,1}$	0.182***	$P^{-1}_{i,1,1}$	0.507***	$P^{-1}_{j,1,1}$	−0.316***
$F^{+2}_{i,1,2}$	−0.019	$A^{+2}_{j,1,1}$	0.094	$F^{+2}_{j,1,2}$	0.120	$A^{+2}_{i,1,2}$	0.022	$P^{+2}_{i,1,2}$	−0.012	$P^{+2}_{j,1,2}$	−0.301
$F^{+1}_{i,1,2}$	0.255***	$A^{+1}_{j,1,2}$	0.082*	$F^{+1}_{j,1,2}$	0.138**	$A^{+1}_{i,1,2}$	0.133***	$P^{+1}_{i,1,2}$	0.566***	$P^{+1}_{j,1,2}$	−0.466**
$F^{0}_{i,1,2}$	0.186***	$A^{0}_{j,1,2}$	0.141***	$F^{0}_{j,1,2}$	0.121***	$A^{0}_{i,1,2}$	0.124***	$P^{0}_{i,1,2}$	0.494***	$P^{0}_{j,1,2}$	−0.310**
$F^{-1}_{i,1,2}$	0.143***	$A^{-1}_{j,1,2}$	0.265***	$F^{-1}_{j,1,2}$	0.088**	$A^{-1}_{i,1,2}$	0.125**	$P^{-1}_{i,1,2}$	0.449***	$P^{-1}_{j,1,2}$	−0.125
$F^{-2}_{i,1,2}$	−0.016	$A^{-2}_{j,1,2}$	0.208	$F^{-2}_{j,1,2}$	0.154	$A^{-2}_{i,1,2}$	0.096	$P^{-2}_{i,1,2}$	−0.057	$P^{-2}_{j,1,2}$	−0.415**
$S^{\mathrm{H}}_{i,1}$	0.023***	$C^{\mathrm{H}}_{j,1}$	0.010	$S^{\mathrm{A}}_{j,1}$	0.018***	$C^{\mathrm{A}}_{i,1}$	0.010*	$R^{\mathrm{H}}_{i,1}$	0.012	$R^{\mathrm{H}}_{j,1}$	−0.020**
$S^{\mathrm{H}}_{i,2}$	0.008	$C^{\mathrm{H}}_{j,2}$	0.014*	$S^{\mathrm{A}}_{j,2}$	0.006	$C^{\mathrm{A}}_{i,2}$	0.011**	$R^{\mathrm{H}}_{i,2}$	0.008	$R^{\mathrm{H}}_{j,2}$	−0.018**
$S^{\mathrm{H}}_{i,3}$	0.012*	$C^{\mathrm{H}}_{j,3}$	0.019***	$S^{\mathrm{A}}_{j,3}$	0.006	$C^{\mathrm{A}}_{i,3}$	0.008	$R^{\mathrm{H}}_{i,3}$	0.036***	$R^{\mathrm{H}}_{j,3}$	−0.013
$S^{\mathrm{H}}_{i,4}$	0.004	$C^{\mathrm{H}}_{j,4}$	0.008	$S^{\mathrm{A}}_{j,4}$	0.000	$C^{\mathrm{A}}_{i,4}$	0.005	$R^{\mathrm{H}}_{i,4}$	−0.004	$R^{\mathrm{H}}_{j,4}$	−0.008
$S^{\mathrm{H}}_{i,5}$	0.013**	$C^{\mathrm{A}}_{j,1}$	0.017***	$S^{\mathrm{A}}_{j,5}$	0.002	$C^{\mathrm{H}}_{i,1}$	0.009	$R^{\mathrm{H}}_{i,5}$	0.006	$R^{\mathrm{A}}_{j,1}$	−0.016**
$S^{\mathrm{H}}_{i,6}$	0.009	$C^{\mathrm{A}}_{j,2}$	0.015**	$S^{\mathrm{A}}_{j,6}$	0.006	$C^{\mathrm{H}}_{i,2}$	0.013**	$R^{\mathrm{H}}_{i,6}$	−0.010	$R^{\mathrm{A}}_{j,2}$	−0.013
$S^{\mathrm{H}}_{i,7}$	0.005	$C^{\mathrm{A}}_{j,3}$	0.002	$S^{\mathrm{A}}_{j,7}$	0.007	$C^{\mathrm{H}}_{i,3}$	0.015**	$R^{\mathrm{H}}_{i,7}$	0.005	$R^{\mathrm{A}}_{j,3}$	−0.019**
$S^{\mathrm{H}}_{i,8}$	0.004	$C^{\mathrm{A}}_{j,4}$	0.006	$S^{\mathrm{A}}_{j,8}$	0.016***	$C^{\mathrm{H}}_{i,4}$	0.009	$R^{\mathrm{H}}_{i,8}$	−0.004	$R^{\mathrm{A}}_{j,4}$	−0.021**
$S^{\mathrm{H}}_{i,9}$	0.016***	$C^{\mathrm{A}}_{j,5}$	0.016**	$S^{\mathrm{A}}_{j,9}$	0.002	$C^{\mathrm{H}}_{i,5}$	−0.000	$R^{\mathrm{H}}_{i,9}$	0.009	$R^{\mathrm{A}}_{j,5}$	−0.015*
$S^{\mathrm{A}}_{i,1}$	0.012*	$C^{\mathrm{A}}_{j,6}$	0.003	$S^{\mathrm{H}}_{j,1}$	0.006	$C^{\mathrm{H}}_{i,6}$	−0.014**	$R^{\mathrm{A}}_{i,1}$	0.007	$R^{\mathrm{A}}_{j,6}$	0.001
$S^{\mathrm{A}}_{i,2}$	0.004	$C^{\mathrm{A}}_{j,7}$	0.004	$S^{\mathrm{H}}_{j,2}$	0.014***	$C^{\mathrm{H}}_{i,7}$	0.007	$R^{\mathrm{A}}_{i,2}$	0.019**	$R^{\mathrm{A}}_{j,7}$	−0.008
$S^{\mathrm{A}}_{i,3}$	0.021***	$C^{\mathrm{A}}_{j,8}$	0.003	$S^{\mathrm{H}}_{j,3}$	0.006	$C^{\mathrm{H}}_{i,8}$	−0.005	$R^{\mathrm{A}}_{i,3}$	0.021**	$R^{\mathrm{A}}_{j,8}$	−0.012
$S^{\mathrm{A}}_{i,4}$	−0.011	$C^{\mathrm{A}}_{j,9}$	0.004	$S^{\mathrm{H}}_{j,4}$	0.002	$C^{\mathrm{H}}_{i,9}$	0.003	$R^{\mathrm{A}}_{i,4}$	−0.008	$R^{\mathrm{A}}_{j,9}$	−0.026***
$\mathrm{SIGH}_{i,j}$	0.076***	$\Delta\mathrm{AP}_{i,1}$	0.104***	$\mathrm{SIGH}_{i,j}$	−0.058**	$\Delta\mathrm{AP}_{i,1}$	−0.187***	$\mathrm{SIGH}_{i,j}$	0.151***	$\Delta\mathrm{AP}_{i,1}$	0.193***
$\mathrm{SIGA}_{i,j}$	−0.000	$\mathrm{AP}_{i,2}$	0.138***	$\mathrm{SIGA}_{i,j}$	0.060**	$\mathrm{AP}_{i,2}$	−0.106***	$\mathrm{SIGA}_{i,j}$	−0.060*	$\mathrm{AP}_{i,2}$	0.140***
CUP_{i}	−0.138***	$\Delta\mathrm{AP}_{j,1}$	−0.153***	CUP_{i}	0.055**	$\Delta\mathrm{AP}_{j,1}$	0.114***	CUP_{i}	−0.112***	$\Delta\mathrm{AP}_{j,1}$	−0.186***
CUP_{j}	0.044*	$\mathrm{AP}_{j,2}$	−0.177***	CUP_{j}	−0.058***	$\mathrm{AP}_{j,2}$	0.086***	CUP_{j}	0.064***	$\mathrm{AP}_{j,2}$	−0.165***
$\mathrm{DIST}_{i,j}$	0.046***	Const.	−0.415***	$\mathrm{DIST}_{i,j}$	−0.035***	Const.	−0.000	$\mathrm{DIST}_{i,j}$	0.056***		

Estimations of Models 1 and 4 are over 29,562 match observations. In Model 1, $\hat{\eta}$=0.047***, $\hat{\pi}$=0.021***, $\hat{\theta}$=0.337***, ln(L)=−84,728.5. Likelihood ratio (LR) test for significance of the regression: χ^2(120)=2136.8***. In Model 4, $\hat{\mu}_1$=−0.560***, $\hat{\mu}_2$=−0.205***, ln(L)=−30,554.1. LR test: χ^2(59)=1613.1***. Glewwe's (1997) Normality test: χ^2(2)=1.88.

*** Denotes significantly different from zero, 1% level, two-tailed test.

* Denotes 10% level.

** Denotes 5% level.

were similar, the former count for more towards the team quality measure than the latter, because they were recorded in a higher division. Accordingly, Millwall's 'attack' coefficient is progressively upgraded from 0.78 at the start of the 2001–2002 season to 0.84 by the end (col (8)).

Returning to Table 1, the recent goals scored and conceded variables $S^{\mathrm{H}}_{i,m}$, $S^{\mathrm{A}}_{i,n}$, $C^{\mathrm{H}}_{i,m}$ and $C^{\mathrm{A}}_{i,n}$ (and their counterparts for team j) allow for the inclusion in Model 1 of goals data from each team's few most recent matches. $S^{\mathrm{H}}_{i,m}$ is the number of goals scored by team i in its mth most recent home match; $S^{\mathrm{A}}_{i,n}$ is the number of goals scored by team i in its nth most recent away match; $C^{\mathrm{H}}_{i,m}$ and $C^{\mathrm{A}}_{i,n}$ are similarly defined for goals conceded by team i. The recent win–draw–lose match results variables $R^{\mathrm{H}}_{i,m}$ and $R^{\mathrm{A}}_{i,n}$ (and their counterparts for team j) play an equivalent role in Model 4. The possibility of short-term persistence in team performance suggests these variables may have particular relevance in helping

Table 2
Evolution of selected explanatory variables in Model 1: Millwall, 2001–2002 season

Summary playing record (played, won, drew, lost, goals for and against)							
1999–2000	FLD2	P46	W23	D13	L10	F76	A50
2000–2001	FLD2	P46	W28	D9	L9	F89	A38
2001–2002	FLD1	P46	W22	D11	L13	F79	A48

Home match no.	$F^{0}_{i,0,0}$	$F^{-1}_{i,0,1}$	Ave. goals scored, 0–12 months: $F^{0}_{i,0,0}+F^{-1}_{i,0,1}$	$F^{-1}_{i,1,1}$	$F^{-1}_{i,1,2}$	Ave. goals scored, 12–24 months: $F^{-1}_{i,1,1}+F_{i,1,2}{}^{-1}$	$\Sigma\text{coeff}\times F^{d}_{i,y,s}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	0.00	1.93	1.93	0.00	1.67	1.67	0.78
2	0.09	1.80	1.89	0.13	1.58	1.71	0.78
3	0.15	1.74	1.89	0.20	1.48	1.67	0.79
4	0.21	1.70	1.91	0.20	1.48	1.67	0.80
5	0.35	1.41	1.76	0.52	1.37	1.89	0.80
6	0.41	1.37	1.78	0.57	1.33	1.89	0.81
7	0.55	1.23	1.79	0.66	1.28	1.94	0.84
8	0.63	1.24	1.87	0.68	1.17	1.85	0.86
9	0.70	1.17	1.87	0.76	1.04	1.80	0.86
10	0.76	1.11	1.87	0.83	0.96	1.78	0.86
11	0.79	1.06	1.85	0.84	0.93	1.78	0.86
12	0.83	0.87	1.70	1.04	0.89	1.93	0.84
13	0.88	0.85	1.73	1.07	0.84	1.91	0.85
14	1.00	0.73	1.73	1.20	0.78	1.98	0.87
15	1.02	0.67	1.69	1.27	0.76	2.02	0.87
16	1.00	0.64	1.64	1.30	0.75	2.05	0.85
17	1.13	0.56	1.69	1.38	0.64	2.02	0.88
18	1.15	0.50	1.65	1.41	0.63	2.04	0.87
19	1.17	0.46	1.63	1.46	0.50	1.96	0.85
20	1.23	0.40	1.64	1.49	0.43	1.91	0.86
21	1.24	0.39	1.63	1.56	0.33	1.89	0.85
22	1.31	0.35	1.65	1.57	0.30	1.87	0.87
23	1.40	0.13	1.53	1.73	0.17	1.90	0.84

In col (8), 'coeff.' are the relevant coefficients in the equation for $\hat{\lambda}_{1,i,j}$ in Table 1.

predict the outcome of the current match, over and above their contribution to the covariates $\{F^{d}_{i,y,s}, A^{d}_{i,y,s}, F^{d}_{j,y,s}, A^{d}_{j,y,s}\}$ and $\{P^{d}_{i,y,s}, P^{d}_{j,y,s}\}$.

In general, the estimated coefficients on the recent goals and results covariates tend to be rather erratic. Experimentation indicated that data on the home team's recent home performance are more useful as predictors than data on its recent away performance; and similarly, the away team's recent away performance is more useful than its recent home performance. Statistically significant estimated coefficients are obtained for some (but not all) values of $m \leq 9$, and for some $n \leq 4$. Therefore, data are included from the home team's last nine home matches and last four away matches, and from the away team's last four home matches and last nine away matches.

The identification of matches with importance for end-of-season championship, promotion and relegation issues is relevant if match outcomes are affected by incentives: if a match is important for one team and unimportant for the other, the teams may contribute different levels of effort. Of several alternative definitions of importance that were considered, the chosen algorithm produces the match importance dummies with the most explanatory power. A match is important if it is possible (before the match is played) for the team in question to win the championship or be promoted or relegated, assuming all other teams currently in contention for the same end-of-season outcome take one point on average from each of their remaining fixtures. The signs and significance of the estimated coefficients on $\mathrm{SIGH}_{i,j}$ and $\mathrm{SIGA}_{i,j}$ are predominantly consistent with incentive effects as described above.

The FA Cup is a sudden-death knockout tournament involving both league and non-league teams. Teams from FLD2 and FLD3 enter the cup in the first round, and teams from the PL and FLD1 in the third round. The final is played at the end of the league season. Early elimination from the cup may have implications for a team's performance in subsequent league matches, though the direction of the effect is ambiguous. A team eliminated from the cup may be able to concentrate its efforts on the league, suggesting an improvement in league performance. Alternatively, elimination may reduce confidence (while progress fosters team spirit), suggesting a decline in league performance. The signs and significance of the estimated coefficients on CUP_i and CUP_j suggest that the second of these two effects dominates.[6]

According to Clarke and Norman (1995), geographical distance is a significant influence on match outcomes: a finding confirmed by the signs and significance of the estimated coefficients on $\mathrm{DIST}_{i,j}$. The greater intensity of competition in local derbies may have some effect in offsetting home advantage, while the psychological or practical difficulties of long-distance travel for teams and supporters may increase home advantage in matches between teams from distant cities. Finally, $\mathrm{AP}_{i,s}$ and $\mathrm{AP}_{j,s}$ are positive for teams that tend to attract higher-than-average home attendances after controlling for league position (and negative in the opposite case). These covariates allow for a 'big team' effect on match outcomes: regardless of the values of other controls, 'big' teams are more likely (and 'small' teams less likely) to win, either through the direct influence of the crowd on the outcome, or because teams with a larger revenue base have more resources to spend on players. To reduce the effect of temporary variation in the attendance–performance relationship, the values of these variables for the two preceding seasons are included in the model. Since the values over successive seasons tend to be highly correlated, $\Delta\mathrm{AP}_{i,1}=\mathrm{AP}_{i,1}-\mathrm{AP}_{i,2}$ (and its counterpart for team j) is used in place of $\mathrm{AP}_{i,1}$ (and $\mathrm{AP}_{j,1}$).

5. Forecasting performance

Using the adjusted bivariate probability functions specified in Section 3 (for Models 1 and 2), the estimated home win, draw and away win probabilities (i.e., win–draw–lose probabilities for the home team) are $\sum_{s>c}\sum_{c}\tilde{P}\left(S^{\mathrm{H}}_{i,0}=s, C^{\mathrm{H}}_{i,0}=c\right)$, $\sum_{s=c}\tilde{P}\left(S^{\mathrm{H}}_{i,0}=s, C^{\mathrm{H}}_{i,0}=c\right)$ and $\sum_{s<c}\sum_{c\geq 1}\tilde{P}\left(S^{\mathrm{H}}_{i,0}=s, C^{\mathrm{H}}_{i,0}=c\right)$, respectively. Using the ordered probit specification (Models 3 and 4), the equivalent probabilities are, $1-\Phi(\hat{\mu}_2-\hat{y}^{*}_{i,j})$, $\Phi(\hat{\mu}_2-\hat{y}^{*}_{i,j})-\Phi(\hat{\mu}_1-\hat{y}^{*}_{i,j})$ and $\Phi(\hat{\mu}_1-\hat{y}^{*}_{i,j})$ respectively, where Φ is the standard Normal distribution function.

As a measure of forecasting performance, Rue and Salvesen (2000) suggest the pseudo-likelihood statistic, equivalent to the geometric mean of the estimated probabilities for the actual results of all matches played during the forecast period. Table 3 reports the values of this statistic for the forecasts obtained from Models 1 to 4, for each of the 10 seasons 1992–1993 to 2001–2002 (inclusive). In terms of numerical magnitude, the differences between the individual results reported in Table 3 appear very small. However, small numerical variations in the pseudo-likelihood statistic can indicate large variations in forecasting capability. For example, a forecasting rule that assigned identical probabilities to all matches, equivalent to the overall proportions of home wins, draws and away wins observed during the 2001–2002 season (46.4%, 26.6% and 27.0%, respectively), would produce a pseudo-likelihood statistic of 0.346 for the same season, only slightly smaller numerically than the values reported in Table 3. For each model, the null hypothesis that all coefficients are zero is rejected at any reasonable significance level by likelihood ratio tests (reported at the foot of Table 1).

[6] As well as the FA Cup, all PL and FL member teams participate in a second domestic tournament, the League Cup. All FLD2 and FLD3 teams also take part in a third tournament, currently known as the LD Vans Trophy. Full-strength line-ups are not always fielded in these tournaments. A few leading teams (between 6 and 10 per season) participate in tournaments at European level. To avoid excessive proliferation of covariates, the impact of participation in these tournaments on league match outcomes is ignored, on the grounds that the other domestic tournaments are widely considered less prestigious than the FA Cup, while the proportion of matches in the entire data set potentially affected by European involvement is very small.

Table 3
Pseudo-likelihood statistics: Models 1 to 4, 1992–1993 to 2001–2002 seasons

	Model 1	Model 2	Model 3	Model 4
1992–1993	**0.35057**	0.34983	0.35034	0.34930
1993–1994	**0.35048**	0.35041	0.34978	0.34959
1994–1995	0.35499	0.35552	**0.35553**	0.35500
1995–1996	0.35026	0.35042	**0.35067**	0.35066
1996–1997	0.34946	0.34882	**0.34965**	0.34853
1997–1998	0.35457	0.35517	**0.35568**	0.35524
1998–1999	0.35294	**0.35379**	0.35317	0.35368
1999–2000	0.35881	0.35912	**0.35949**	0.35920
2000–2001	0.35586	0.35666	0.35656	**0.35670**
2001–2002	0.35950	0.35944	**0.36007**	0.35948
Average	0.35374	0.35392	0.35409	0.35374

Values shown in **bold** are the highest across the four models in each of the 10 seasons.

The results reported in Table 3 suggest there is no difference between the forecasting performance of the two bivariate Poisson regressions. Model 1 outperforms Model 2 in four of the ten seasons and Model 2 outperforms Model 1 in six seasons. It appears the two models cannot be ranked on this basis. There is, however, some difference between the two ordered probit regressions. Model 3 outperforms Model 4 in eight seasons, and Model 4 outperforms Model 3 in two seasons. Across all four specifications, Model 3 is the best performer in six out of ten seasons, Model 1 is the best in two seasons, and Models 2 and 4 are the best in one season each.

For the purpose of forecasting win–draw–lose results, the best performance is achieved using the 'hybrid' Model 3, in which an ordered probit regression is used to estimate a specification combining a results-based dependent variable with goals-based lagged performance covariates. No advantage appears to be gained by using the more data-intensive goals-based dependent variable (as in Models 1 and 2), but some advantage is gained by using goals-based team performance covariates rather than the corresponding results-based covariates (as in Model 4). However, while Model 3 dominates the other three specifications on a majority of occasions, it fails to do so on all occasions. Even allowing for the narrow numerical domain of the pseudo-likelihood statistic, the results seem to suggest that the difference in forecasting performance between the four specifications is relatively small.

6. Conclusion

This article has drawn comparisons between the forecasting performance of goals-based and results-based regression models for match outcomes in association football (soccer). The comparisons are facilitated by estimating a set of models over data sets that are identical in all respects apart from their emphasis on goals or results, respectively. Four models are considered, allowing for all possible permutations of goals- and results-based dependent variables, and goals- and results-based lagged performance covariates. Bivariate Poisson and ordered probit regressions are used for estimation. A reasonable prior might be that goals-based models should outperform results-based models, because the former are based on a richer data set. In practice, the best forecasting performance is achieved using a 'hybrid' specification, combining a results-based dependent variable with goals-based lagged performance covariates. No advantage appears to be gained by using the more data-intensive goals-based dependent variable, but some advantage is gained by using goals-based (rather than results-based) lagged performance covariates. However, the difference in forecasting performance between the specifications that have been considered seems to be rather small. This may explain why both goals-based and results-based models have been advocated in the recent applied statistics and econometrics literatures.

Appendix A.

Full definitions of the covariates used in the bivariate Poisson and ordered probit regression models are as follows.

$F^{d}_{i,y,s}=f^{d}_{i,y,s}/n_{i,y}$, where $f^{d}_{i,y,s}$=home team i's total goals scored in matches played 0–12 months (y=0) or 12–24 months (y=1) before current match; within the current season (s=0) or previous season (s=1) or two seasons ago (s=2); in the team's current division (d=0) or one (d=±1) or two (d=±2) divisions above or below the current division; and $n_{i,y}$=i's total matches played 0–12 months (y=0) or 12–24 months (y=1) before current match.

$A^{d}_{i,y,s}=a^{d}_{i,y,s}/n_{i,y}$, where $a^{d}_{i,y,s}$=i's total goals conceded, defined for the same y,s,d as above; $n_{i,y}$ defined as above.

$P^{d}_{i,y,s}=p^{d}_{i,y,s}/n_{i,y}$, where $p^{d}_{i,y,s}$=i's total 'points' score, on a scale of 1=win, 0.5=draw, 0=loss, defined for the same y,s,d as above; $n_{i,y}$ defined as above.

Appendix A (*continued*)

$S^{\mathrm{H}}_{i,m}$=goals scored in mth most recent home match by i.

$C^{\mathrm{H}}_{i,m}$=goals conceded in mth most recent home match by i.

$S^{\mathrm{A}}_{i,n}$, $C^{\mathrm{A}}_{i,n}$=goals scored and conceded in nth most recent away match by i.

$R^{\mathrm{H}}_{i,m}$=result (1=win, 0.5=draw, 0=loss) of i's mth most recent home match.

$R^{\mathrm{A}}_{i,n}$=result of i's nth most recent away match.

$\mathrm{SIGH}_{i,j}$=1 if match is important for championship, promotion or relegation issues for i but not for away team j; 0 otherwise.

$\mathrm{SIGA}_{i,j}$=1 if match is important for j but not for i; 0 otherwise.

CUP_{i}=1 if i is eliminated from the FA Cup; 0 otherwise.

$\mathrm{DIST}_{i,j}$=natural logarithm of the geographical distance between the grounds of i and j.

$\mathrm{AP}_{i,s}$=residual for i from a cross-sectional regression of the log of average home attendance on final league position (defined on a scale of 92 for the PL winner to 1 for the bottom team in FLD3) s seasons before the present season, for s=1,2.

References

Audas, R., Dobson, S., & Goddard, J. (2002). The impact of managerial change on team performance in professional sports. *Journal of Economics and Business*, *54*, 633–650.

Barnett, V., & Hilditch, S. (1993). The effect of an artificial pitch surface on home team performance in football (soccer). *Journal of the Royal Statistical Society. Series A*, *156*, 39–50.

Cain, M., Law, D., & Peel, D. (2000). The favourite-longshot bias and market efficiency in UK football betting. *Scottish Journal of Political Economy*, *47*, 25–36.

Clarke, S. R., & Norman, J. M. (1995). Home ground advantage of individual clubs in English soccer. *Statistician*, *44*, 509–521.

Crowder, M., Dixon, M., Ledford, A., & Robinson, M. (2002). Dynamic modelling and prediction of English Football League matches for betting. *Statistician*, *51*, 157–168.

Dixon, M. J., & Coles, S. C. (1997). Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics*, *46*, 265–280.

Dixon, M. J., & Pope, P. F. (2004). The value of statistical forecasts in the UK association football betting market. *International Journal of Forecasting*, *20*, 686–700.

Dixon, M. J., & Robinson, M. E. (1998). A birth process model for association football matches. *Statistician*, *47*, 523–538.

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

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

Forrest, D., & Simmons, R. (2000b). Making up the results: The work of the football pools panel, 1963–1997. *Statistician*, *49*, 253–260.

Glewwe, P. (1997). A test of the normality assumption in the ordered probit model. *Econometric Reviews*, *16*, 1–19.

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

Holgate, P. (1964). Estimation for the bivariate Poisson distribution. *Biometrika*, *51*, 241–245.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Statistician*, *52*, 381–393.

Koning, R. H. (2000). Balance in competition in Dutch soccer. *Statistician*, *49*, 419–431.

Kuypers, T. (2000). Information and efficiency: An empirical study of a fixed odds betting market. *Applied Economics*, *32*, 1353–1363.

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

Ridder, G., Cramer, J. S., & Hopstaken, P. (1994). Estimating the effect of a red card in soccer. *Journal of the American Statistical Association*, *89*, 1124–1127.

Rue, H., & Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *Statistician*, *49*, 399–418.

Biography: John GODDARD is Professor of Economics at the Department of Economics, University of Wales Swansea. His research interests include the economics of professional team sports, industrial organisation, and the economics of the banking sector. He is co-author with Stephen Dobson of the monograph *The Economics of Football* (Cambridge University Press, 2001).