Forecasting international soccer match results using bivariate discrete distributions

Ian McHale and Phil Scarf

Centre for Operational Research and Applied Statistics,
Salford Business School,
University of Salford,
Salford,
Manchester
M5 4WT, UK.

email i.mchale@salford.ac.uk, p.a.scarf@salford.ac.uk

Salford Business School Working Paper Series

Paper no. 322/06

Forecasting international soccer match results using bivariate discrete distributions

Ian McHale*, Phil Scarf
*Centre for Operational Research and Applied Statistics, The University of Salford, Greater Manchester, M5 4WT, UK. *i.mchale@salford.ac.uk*

Abstract

This paper models international soccer match results for the period 1993 to 2004 using bivariate discrete distributions. The models employed are defined in terms of the marginal distributions and a dependence copula, and are used to describe match scores. This copula representation allows dependence in the bivariate distribution to be modelled in a flexible manner by specifying a suitable family of copula functions and fitting this to the bivariate data using maximum likelihood. A graphical diagnostic for considering the model fit to the dependence structure is presented. This plot requires the discrete data to be "continuised"—this is in the spirit of copula functions since although the bivariate copula is a continuous function on the unit square, it can be used to specify the unique dependence structure in a bivariate discrete distribution which is defined on a lattice. Marginal means are modelled with match covariates, and a forecasting model is developed. Forecasts with this model are compared to a probit model for match outcome, and it is shown that the bivariate discrete distribution forecasts match results slightly better than the probit forecasts. The former have the advantage that exact scores and match outcomes are forecast.

Keywords: football; copula; continuised; negative binomial.

Introduction

Much of the empirical research on soccer has been related to forecasting match outcome. From analysing the efficiency of the betting market on soccer and judging the skills of tipsters at predicting match outcomes (see, for example, Goddard and Asimakopoulos (2004) and Forrest and Simmons (2000)) to assessing the effect of red cards (Ridder et al (1994)), each task requires some sort of forecasting model. The forecasting methodology can broadly be split into two categories: a direct approach and an indirect approach. The direct approach utilizes ordinal regression models such as the probit or logit models to forecast the ordered response variable, win/draw/loss. This method seems to have been favoured by economists, see for example, Koning (2000). The indirect method goes back to Maher (1982) and models the distributions of goals scored by each team, either independently or dependently. With this approach inferences can be made as to the most probable outcome – defined in terms of result or exact score. The indirect method has typically employed the bivariate Poisson model of Griffiths and Milne (1978), or some variation of this model, for example the diagonally inflated bivariate Poisson model of Karlis and Ntzoufras (2003). This method of forecasting seems to have been popular amongst statisticians. Goddard (2005) assesses the relative performance of the direct and indirect forecasting models using a data set based on domestic soccer in England over an 11 year period from 1991 to 2002, and concludes that neither the direct nor the indirect approach clearly outperforms the other in terms of forecasting match outcome. However, the indirect approach does have the advantage that not only is the match outcome forecast, but the plethora of exact scores is assigned probabilities, providing a much richer outcome description. For instance, one is able to identify high scoring games versus low scoring games using the probabilities assigned to each possible score. Such information can be of use to bettors since wagering on there being x or more goals is a popular bet taken at the bookmakers.

To date, the forecasting models have been largely confined to domestic leagues such as that in England, see, for example, Dixon and Coles (1997). This paper considers forecasting soccer played on a higher stage. Competition between national soccer teams may be considered the most popular sporting activity in the world. For example, the 2002 FIFA World Cup final, held in stadiums across Korea and Japan, attracted an international television audience of 28.8 billion across 213 countries, which even exceeds viewing figures for the Olympic Games[1].

[1] Figures taken from FIFA website, www.fifa.com.

One key difference between domestic soccer and national team soccer exists which renders the bivariate models used in the literature to date useless. As a consequence of the bivariate Poisson model of Griffiths and Milne the two dependent variables must have non-negative correlation. This can be seen if one interprets the bivariate Poisson model in terms of three univariate Poisson random variables *Z1, Z2, Z3,* with rates λ_1, λ_2 and λ_3, such that *X=Z1+Z3* and *Y=Z2+Z3*. The joint probability function is given by

$$\Pr(X = x, Y = y) = \exp(-\lambda_1 - \lambda_2 - \lambda_3) \sum_{k=0}^{\min(x,y)} \frac{\lambda_1^{x-k}\lambda_2^{y-k}\lambda_3^{k}}{(x-k)!(y-k)!k!}$$

It is easily shown that $\mathrm{cov}(X,Y) = \lambda_3$ and must therefore be positive since λ_3 is a Poisson rate. For domestic soccer, goals by the two teams display only slight positive or no correlation, so this model may be applicable, whereas for national team soccer, goals are significantly negatively correlated, see Table 1 (all figures are statistically significant at the 99% level). Thus the bivariate Poisson model used in previous studies cannot be employed here and in order to forecast exact scores for national team soccer we seek a bivariate Poisson distribution which not only allows for positively correlated dependent variables, but also permits negative dependence.

A further difficulty with the bivariate Poisson model of Griffiths and Milne is that the marginal distributions of the two random variables must be Poisson. However, studies have shown that this is not the case for goals in soccer, see, for example, Reep and Benjamin (1968). It is evident in goals data that over-dispersion is present, and the mean and variance of goals scored by each team are not equal. Thus, not only must we use a model which allows for negative dependence structure, we should seek a bivariate model which can have marginal distributions other than that of the Poisson distribution.

A natural way forward is to consider copula functions (Nelsen (2006)) to generate various bivariate dependent discrete distributions, and use these distributions to forecast exact score results, and then infer match outcomes. McHale and Scarf (2006) consider using such functions to generate several bivariate distributions, and fit the models to shots data from the English Premier League.

Thus in this paper we consider new models for indirect modelling of soccer match results; these models can explain the negative dependence that is found in international soccer match data. This paper is structured as follows. Section 2 presents

the models employed here. The data we have obtained and used is then described in Section 3, and some results of a preliminary analysis are given. Section 4 presents the results of a forecasting model based on the bivariate distributions. Finally, some closing remarks are given in Section 5.

1 Bivariate discrete distributions using Copulas

A copula, C, is a multivariate distribution with all univariate marginal distributions being uniformly distributed on the unit interval, [0,1]; hence C is the distribution of a multivariate uniform random vector. For a bivariate distribution F with margins F_1 and F_2, the copula associated with F is a distribution function $C:[0,1]^2 \to [0,1]$ that satisfies

$$F(x,y) = C\{F_1(x), F_2(y)\}, \quad (x,y) \in \Re^2 . \tag{1}$$

The copula $C(u,v)$ itself characterises the dependence between the random variables X and Y with marginal distributions F_1 and F_2. Thus the copula representation (1) resolves the joint distribution into the marginals F_1 and F_2 and the dependence structure C. For more information on copulas see, for example, Nelsen (2006) and Joe (1997).

Copula functions that can model positive and negative dependence may be constructed in a number of ways. The class of extendable Archimedian copulas (copulas for which $C(u,v) = \phi_\kappa\{\phi_\kappa^{-1}(u) + \phi_\kappa^{-1}(v)\}$ for any parametric function ϕ_κ that is convex with $\phi(0) = 1$, $\phi(\infty) = 0$), and $\phi^{-1}(e^{-z})$ being concave in z (Joe, 1997)) are convenient for modelling purposes. The copulas in this class that are employed in this paper are Frank's (F) copula, given by

$$C(u,v) = -\kappa^{-1}\log\{1-(1-e^{-\kappa u})(1-e^{-\kappa v})/(1-e^{-\kappa})\}, \quad (\kappa \in \Re), \tag{2}$$

and Kimeldorf and Sampson's (KS) copula, given by

$$C(u,v) = \max\{(u^{-\kappa} + v^{-\kappa} - 1)^{-1/\kappa}, 0\}, \quad (\kappa \in \Re). \tag{3}$$

Nelson (2006) refers to this latter copula as Clayton's copula. κ is the dependence parameter.

There are a large number of copula families (Nelsen (2006), p.116). The choice of a copula family can be guided by the (dependence) properties of that family. The choice of the copulas we employ here is due to the fact that they are comprehensive, meaning

they can model the full range of correlations possible, i.e. Kendall's τ going from -1 to 1.

The discussion here is not restricted to the case of a bivariate Poisson model, since any discrete distribution can be used as the marginal distributions in the copula to generate a bivariate distribution. For example, a bivariate geometric distribution with Frank's copula is given by

$$F_{XY}(x,y) = \\ -\frac{1}{\kappa}\log[1-\{1-\exp(-\kappa\sum_{i=1}^{x}\mu_1^{-1}(1-\mu_1^{-1})^{i-1})\}\{1-\exp(-\kappa\sum_{i=1}^{y}\mu_2^{-1}(1-\mu_2^{-1})^{i-1})\}]/(1-e^{-\kappa})] \quad (4)$$

where $x, y = 0,1,\ldots$, $0 \le \mu_1, \mu_2 \le 1$, and $-\infty < \kappa < \infty$. This distribution has marginal means μ_1 and μ_2, a dependence parameter κ and negative dependence for $\kappa < 0$. The marginal distributions are independent when $\kappa = 0$. Bivariate Poisson and negative binomial distributions may be obtained in a similar manner by replacing u and v in (2) and (3) by the appropriate marginal distribution functions. Using negative binomial distributions as the marginals results in a flexible bivariate distribution which can capture over or under dispersion in the margins.

The parameterisation for the negative binomial distribution used here is given by

$$f(x;\mu,\sigma) = \frac{\mu^x}{x!}.\frac{\Gamma(\sigma+x)}{\Gamma(\sigma)(\mu+\sigma)^x}.\left(1+\frac{\mu}{\sigma}\right)^{-\sigma}$$

where μ is the mean and σ is a scale parameter.

We have fitted the models using geometric marginals but the results suggest this model is not appropriate for this data. We thus consider four bivariate distributions generated from two copulas and two pairs of marginal distributions, denoted by F_{PP}, F_{nbnb}, KS_{PP} and KS_{nbnb}, representing Frank's copula with Poisson marginals, Frank's copula with negative binomial marginals, Kimeldorf and Sampson's copula with Poisson marginals and Kimeldorf and Sampson's copula with negative binomial marginals, respectively.

Given a bivariate discrete distribution specified in terms of marginal distributions $F_1(x;\boldsymbol{\theta}_1)$ and $F_2(y;\boldsymbol{\theta}_2)$ and copula $C(u,v;\kappa)$, the likelihood function for the parameters $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\kappa)$ given a datum (x_i,y_i) is

$$L\{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\kappa),(x_i,y_i)\} = \Pr(X=x_i, Y=y_i) = C\{F_1(x_i),F_2(y_i)\} \\ -C\{F_1(x_i-1),F_2(y_i)\} - C\{F_1(x_i),F_2(y_i-1)\} + C\{F_1(x_i-1),F_2(y_i-1)\}.$$

For sample data (x_i, y_i), $i = 1,...,n$, the log-likelihood, $\sum_i \log L\{(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \kappa), (x_i, y_i)\}$, may be maximised in a standard way. This approach therefore opens to model fitting a wide and rich field of discrete bivariate distributions with general dependence structure.

2 Data

We have collected a total of 8,735 international soccer results from two main sources. The data for the period 1993-2001 were obtained from the archive of International Soccer Results[2] and the data for the period 2001-2004 were obtained from the RSSSF archive[3]. Data on the FIFA world rankings were collected from the FIFA website[4] for each month during the years 1993-2004.

In addition to the results and rankings data, we collected geographical coordinates to be used to calculate approximate distances travelled by teams and/or fans to games. These data were collected from various internet sources.

Thus for each match we have the goals scored by team 1 and the goals scored by team 2. Since some games are held on neutral territory it is not appropriate to define team 1 as the home team, as is done in domestic league studies. Here team 1 is simply the team whose name comes first alphabetically.

Table 1: Correlation for goals scored by each team in a match

Correlation	Corr(g_1,g_2)
Pearson	-0.194
Kendall's Tau	-0.154
Spearman Rho	-0.185

As already mentioned, goals for and goals against for matches involving national soccer teams have significant negative correlation, see Table 1. All correlations shown in Table 1 are negative and statistically significant at the 99% level. This result contrasts with the corresponding result for domestic soccer. For example, McHale and Scarf (2006) say that the positive correlation between goals scored and conceded for

[2] http://www.staff.city.ac.uk/r.j.gerrard/football/aifr21_1.htm

[3] http://www.RSSSF.com

[4] http://www.fifa.com

matches played in the English Premier League from 2001 to 2005 is not statistically significant. One may ask why such a contrast exists between domestic league and national team soccer. One explanation we offer is that of the difference in competitive balance between the two forms of the game. For domestic soccer, the teams are separated into divisions, with the top division having the twenty (or so) top teams in the country, the second division having the next best twenty teams and so on. The vast majority of matches are between the twenty teams from each division all playing each other, and thus the matches are closely competed. Thus, the two relatively equally matched teams will tend to score similar numbers of goals, with the total number of goals scored being dependent more on the tempo and style of the match, rather than the relative quality of the teams. National team soccer on the other hand is organised quite differently. The majority of the matches are qualifying games for major tournaments, and teams are given a seeding so as to avoid two top teams meeting in the early stages of a competition. The resulting effect may be that a large proportion of the games have low competitive balance, causing a negative relationship between the two teams' scores.

Table 2 summarises the fits of each of the four bivariate models for shots, with parameter estimates, the corresponding standard errors, the log-likelihood (LL) and the Akaike Information Criterion (AIC) all being shown. The model parameters have been estimated by maximising the log-likelihood, using various routines written in R (R Development Core Team, 2005).

Table 2: Summary results for bivariate models fitted to shots data

Parameter	F_{PP}	KS_{PP}	F_{nbnb}	KS_{nbnb}
κ	-1.103 (0.002)	-0.094 (0.000)	-1.327 (0.061)	-0.187 (0.013)
$\text{Log}(\mu_1)$ (s_H)	0.304 (0.002)	0.295 (0.000)	0.321 (0.012)	0.320 (0.012)
$\text{Log}(\mu_2)$ (s_A)	0.336 (0.003)	0.335 (0.000)	0.350 (0.011)	0.348 (0.011)
$\text{Log}(\sigma_1)$			0.794 (0.039)	0.803 (0.044)
$\text{Log}(\sigma_2)$			0.901 (0.043)	0.908 (0.034)
LL	-28921.547	-28938.478	-27841.132	-27919.585
AIC	57837.094	57870.955	55672.265	55829.171

Standard errors are shown in parenthesis to 3 decimal places.

It is important to note that the reference team is not taken to be the home team (since in an international tournament match there may be no home team), and thus any difference in the parameter estimates for scores labelled 1 and 2 is merely a consequence of the way we have assigned the reference team, team 1.

INSERT FIGURE 1 HERE.

Table 3: Bivariate frequency table for international soccer match results data

		T2 goals							
		0	1	2	3	4	5	>5	Total
T1 goals	0	0.088	0.090	0.068	0.038	0.019	0.011	0.016	0.331
	1	0.088	0.104	0.066	0.028	0.013	0.005	0.003	0.308
	2	0.063	0.064	0.042	0.015	0.006	0.002	0.001	0.194
	3	0.033	0.028	0.015	0.006	0.003	0.001	0.000	0.086
	4	0.020	0.014	0.005	0.002	0.000	0.000	0.000	0.041
	5	0.010	0.006	0.002	0.000	0.000	0.000	0.000	0.018
	>5	0.013	0.006	0.002	0.000	0.000	0.000	0.000	0.022
	Total	0.314	0.312	0.201	0.090	0.041	0.020	0.021	1

A scatter plot of the goals for and goals against the reference team for each match is given in Figure 1. The data have been made continuous for illustrative purposes by adding a uniform U(-0.5,0.5) random number to each of the observations in order to distinguish the individual data points. Alternatively, the data may be presented in a bivariate frequency table, shown in Table 3. "Continuising" the data, figure 1, is a natural way of proceeding when modelling dependence in discrete bivariate distributions with copulas. This is because the copula itself is a continuous function on $[0,1]^2$, while the bivariate probability mass function that involves the copula is only defined on the discrete lattice $[0,1,2,\ldots]^2$. Furthermore, given the "continuised" data, we can now transform the fitted marginal distributions to U(0,1), using the probability integral transformation. For univariate data this is done in percentile plotting in order to consider the appropriateness of the fitted distribution. For bivariate data, transformation of the margins to U(0,1) allows us to consider graphically the dependence structure in the data, albeit the "continuised" version of the data. We can then seek copula functions that possess the structure indicated by such a plot, noting that the copula describes the dependence structure independently of the chosen marginal distributions. Such a "uniform margins" plot for the international goals data is shown in Figure 2. The fitted copula density which we expect to follow the pattern in the "uniform margins" plot is shown in Figure 3. The copula density for Frank's copula is given by

$$\frac{\partial^2 C}{\partial u \partial v} = \frac{\kappa e^{-\kappa u} e^{-\kappa v}}{(1-e^{-\kappa}) - (1-e^{-\kappa u})(1-e^{-\kappa v})} + \frac{\kappa (1-e^{-\kappa u})(1-e^{-\kappa v}) e^{-\kappa v} e^{-\kappa u}}{\left[(1-e^{-\kappa}) - (1-e^{-\kappa u})(1-e^{-\kappa v})\right]^2}$$

INSERT FIGURE 2 HERE.

INSERT FIGURE 3 HERE.

Figure 2 appears to have three clusters of points arranged on the diagonal running from the top left to the bottom right. The top left and bottom right clusters represent games where one team has scored a high number of goals and the opposition has scored few goals (one-sided matches with distinctly low competitive balance). The central cluster, which seems to be stretched along the opposite diagonal, represents close matches where the two teams have scored similar numbers of goals (matches with high competitive balance). Figure 3 closely mimics the raised diagonal evident in Figure 2. This suggests that Frank's copula may well be appropriate when modelling the dependence structure present in the data. One shortcoming of Frank's copula in modelling this data is that it cannot reproduce the "three peaks" evident in Figure 2. A possible way forward would be to explore multi-parameter copulas that can replicate such nuances, although such a copula family has not been identified as yet.

3 Regression model for forecasting exact scores

We now use covariates in a link function for the mean term of the marginal distributions to produce a regression model which can be used to forecast exact scores. Such regression models have been estimated for each copula/marginal distribution pair, and again we find that the F_{nbnb} model provides the best fit to our data.

The covariates used in the final model are: a home dummy variable for each team, set equal to 0 if the team is not at home and 1 if they are; a neutral dummy variable for each team, set equal to 0 if the team is not playing at a neutral ground and 1 if they are; the FIFA world ranking of each team as at the date of the game; and past goals scored and conceded for the 8 most recent matches for both teams.

Other variables tried but not included in the final model were: time interaction terms with the past goals scored and conceded; distance from the capital city of the team's country to the match location.

We regress the two marginal means on covariates, as

$\log\left(\mu_{ij}\right) = x_{ij}\beta_i \quad i = 1, 2, \ldots$ and $j = 1, \ldots, N$,

where i denotes team 1 and team 2, j represents the j^{th} observation, x is a row vector of covariates for the j^{th} observation and β_i is a column vector of regression coefficients to be estimated. N has been reduced to 5,119 from 8,735 because for each match both teams had to have played at least 8 matches, and further an out sample of 982 observations was kept to assess the forecasting performance of the model at a later stage.

It is evident that since labelling of team 1 and team 2 is not related to the ability of each team, then each team's $\boldsymbol{\beta}$ should be the same. If this was not the case, then the model would become asymmetric having unintuitive consequences. For instance, a change of one ranking place would have a different effect on team depending on whether it was team 1 or team 2.

Since the variable we are regressing on is that team's number of goals, it makes intuitive sense that the variables included effect that team's scoring probability. Thus, the covariates used for each team are their own previous record of goals scored, TGSt-i, for $i = 1, \ldots, 8$ and the opposition's previous record of goals conceded, OGCt-i, for $i = 1, \ldots, 8$.

The final model parameter estimates, standard errors and t statistics for the copula model and the independence model are given in Table 4. Also shown is the sample size (n), the log-likelihood (LL) and the Akaike Information Criterion (AIC).

In the model we only use a lag of 8 games to input past scoring records. More lags were tried, however, the forecasting performance, as measured by number of correct results inferred did not change significantly as more lags were added. In addition to this, it is interesting to note that across our data set, on average a national team plays approximately 8 matches in any one calendar year. Thus using 8 lags can be rationalised in this manner – 8 matches corresponds to approximately one season of football.

In comparison to the independent model, $\kappa = 0$, a likelihood ratio test gives a test statistic of 2.04, not significant at the 95% level. However, the AIC suggests the dependence model is marginally better than the independent model.

Table 4: Estimation summary for forecasting model

parameter	Independence model estimates	Copula model estimates
σ	1.712 (0.075)	1.711 (0.075)
Intercept	-0.262 (0.035)	-0.274 (0.035)
Home	0.210 (0.022)	0.211 (0.022)
Neutral	0.125 (0.026)	0.125 (0.026)
Ranking	-0.004 (0.000)	-0.004 (0.000)
TGSt-1	0.026 (0.006)	0.027 (0.006)
TGSt-2	0.026 (0.006)	0.026 (0.006)
TGSt-3	0.024 (0.006)	0.024 (0.006)
TGSt-4	0.016 (0.006)	0.017 (0.006)
TGSt-5	0.017 (0.006)	0.017 (0.006)
TGSt-6	0.003 (0.006)	0.004 (0.006)
TGSt-7	0.007 (0.006)	0.007 (0.006)
OGSt-8	0.018 (0.006)	0.019 (0.006)
OGCt-1	0.063 (0.007)	0.063 (0.007)
OGCt-2	0.051 (0.007)	0.051 (0.007)
OGCt-3	0.043 (0.007)	0.044 (0.007)
OGCt-4	0.04 (0.007)	0.039 (0.007)
OGCt-5	0.036 (0.007)	0.036 (0.007)
OGCt-6	0.036 (0.007)	0.036 (0.007)
OGCt-7	0.033 (0.007)	0.033 (0.007)
OGCt-8	0.046 (0.007)	0.046 (0.007)
κ		-0.163 (0.097)
LL	-15187.499	-15186.479
AIC	30416.998	30416.958
N	5119	5119

The coefficient on ranking is as expected. The FIFA World Rankings lists the best team as 1, with the ranking number increasing as the perceived quality of the team decreases. Thus, one would expect an increasing ranking (poorer team) to score less goals and hence the coefficient is negative.

The estimates for the home and neutral dummy variables have an intuitive interpretation here. The reference game location is away, and hence the coefficient on the neutral dummy variable can be thought of as the effect playing the match on a neutral venue as opposed to away has on expected number of goals. Similarly, the

coefficient on the home dummy variable indicates the added number of goals for a team playing at home compared to playing away. As one would expect, the home dummy coefficient is larger than the neutral dummy coefficient.

One would expect that the coefficients on the time lagged scoring variables would to decay as the lag increases. Indeed, this is the case for both the goals scored and goals conceded variables.

From the model one can estimate probabilities of each score occurring, and thus by summing the probabilities of team 1's score being greater than team 2's score we can infer the probability of team 1 winning. Similarly we can obtain the probability of team 2 winning and of there being a draw.

Table 5: Ordered probit regression results

parameter	*coefficient*	*t stat*	*parameter*	*coefficient*	*t stat*
home	0.167	8.873	t2gs1	-0.028	-2.350
t1rank	-0.010	-19.397	t2gc1	0.045	3.417
t2rank	0.010	17.908	t2gs2	-0.015	-1.314
t1gs1	-0.006	-0.541	t2gc2	0.010	0.775
t1gc1	-0.045	-3.454	t2gs3	-0.021	-1.841
t1gs2	0.020	1.778	t2gc3	0.027	2.069
t1gc2	-0.029	-2.233	t2gs4	-0.001	-0.069
t1gs3	0.025	2.368	t2gc4	-0.009	-0.693
t1gc3	0.007	0.566	t2gs5	-0.015	-1.386
t1gs4	0.019	1.763	t2gc5	0.017	1.321
t1gc4	-0.017	-1.309	t2gs6	0.023	2.059
t1gs5	0.027	2.502	t2gc6	0.034	2.579
t1gc5	-0.015	-1.179	t2gs7	-0.010	-0.896
t1gs6	-0.015	-1.446	t2gc7	0.000	-0.002
t1gc6	-0.016	-1.261	t2gs8	-0.001	-0.112
t1gs7	0.013	1.238	t2gc8	0.000	0.031
t1gc7	-0.009	-0.700	Cut 1	-0.414	-5.351
t1gs8	0.020	1.880	Cut 2	0.338	4.380
t1gc8	-0.021	-1.655			

To compare the predictive capability with an ordered probit model, we have used the exact same covariates in an ordered probit model to produce the corresponding out of sample forecasts. Table 5 gives the ordered probit regression results for the fitted

model. Due to colinearity, the team 1/team 2 - home/neutral dummy variables have been reduced to one dummy variable, home, set equal to 1 if team one is at home, 0 if the game is played at a neutral venue, and -1 if team 2 is at home.

In order to gauge the forecasting power of the models we produce concordance tables of the forecast results with the actual results for the out sample for both the copula regression model and the ordered probit model. The out sample consists of the 982 matches from 2004. The model forecasts in this table are randomised: that is, forecasted outcomes are generated at random from the win, draw and loss probabilities for the reference team. We do not use maximum probability forecasts—this latter approach has the drawback that it does not consider prediction capability for outcomes which occur infrequently since these would be unlikely to have maximum probability even with reasonably good predictive power. The concordance table for the copula regression forecasts is given in Table 6. Table 7 gives the corresponding table for the ordered probit forecasts.

Table 6: Concordance table for copula regression forecasts and actual results

		Copula regression forecast result				
		t1 win	*draw*	*t2 win*	*Total*	*% correct*
Actual result	*team 1 win*	180	86	102	368	48.9
	draw	102	66	86	254	26.0
	team 2 win	119	77	164	360	45.6
	total	401	229	352	982	41.8

Table 7: Concordance table for ordered probit forecasts and actual results

		ordered probit forecast result				
		t1 win	*draw*	*t2 win*	*Total*	*% correct*
Actual result	*team 1 win*	166	91	111	368	45.1
	draw	90	61	103	254	24.0
	team 2 win	91	97	172	360	47.8
	Total	347	249	386	982	40.6

The copula regression model forecasts 41.8% of the results correctly, whilst the ordered probit model forecasts 40.6% of the results correctly. There is a marginal improvement in the predictive power with the indirect copula model. Clearly there is very little difference in the performance of the two models. However, to forecast exact

scores, one is restricted to the copula model (if choosing between the 2 models considered).

4 Closing remarks

This paper considers novel bivariate distributions for modelling final scores in international soccer matches. Such models are based on copula functions which can consider the dependence between scores in a flexible manner. The negative correlation apparent in the goals for and against in international matches can be explained with such functions. The copula model is first fitted to the bivariate goals data and is shown to explain the dependence structure well. The use of a novel graphical diagnostic technique for visualising the dependence structure in the data has been introduced. Plotting the "uniform margins" plot for "continuised" data and comparing this with the copula density plot provides a useful procedure to gauge the appropriateness of the copula in question. This bivariate uniform margins plot is presented as a diagnostic tool for considering dependence structure in bivariate data. Since scores are necessarily discrete, construction of this plot requires the data to be made continuous by adding U(-1/2,1/2) random numbers to the margins. While this adds to the variability in the data, it does allow comparison of the bivariate data with chosen copula densities. The dependence structure of a bivariate distribution is completely described by the copula and so this plot will be useful for choosing among appropriate families of copula functions.

Next, the mean scores for each team are regressed on predictive covariates to produce a forecasting model of match score. The predictive capability of the model is compared to an ordered probit regression model, with the same covariate set. The two models are compared using an out-sample of 982 matches and the copula model is shown to perform marginally better.

There may be scope to improve on the bivariate dependent score model by considering a wider range of copulas. Some copula families may be parameterised by more than one parameter in order to capture more complex dependence structure in goals for and against, such as the "three peaks" evident in the national team data here. Another extension to the work here would be to model the dependence parameter(s) in the copula function as a function of covariates that indicate the importance of the game or the competitive balance of the match in question.

References

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

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

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

Goddard, J. (2005). Regression models for forecasting goals and match results in association football, *International Journal of Forecasting*, 21, 331-340.

Griffiths, R.C. & Milne, R.K. (1978), A class of bivariate Poisson processes. *Journal of Multivariate Analysis,* 8, 380-395.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.

Karlis, D & Ntzoufras, J. (2003). Analysis of Sports Data Using Bivariate Poisson Models, *The Statistician*, 52, 381-393.

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

McHale, I.G. & Scarf, P.A. (2006). Modelling soccer matches using bivariate discrete distributions. Working paper.

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

Nelsen, R.B. (2006). *An Introduction to Copulas, 2nd Edition*. Springer, New York.

R Development Core Team (2005), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Reep, R. & Benjamin, B. (1968). Skill and chance in Association Football, *Journal of the Royal Statistical Society A*, 131, 581-585.

Ridder, G. & Cramer, J. S. & Hopstaken, P. (1994). Down to Ten: Estimating the Effect of a Red Card in Soccer, *Journal of the American Statistical Association*, Vol. 89, No. 427, pp. 1124-1127.

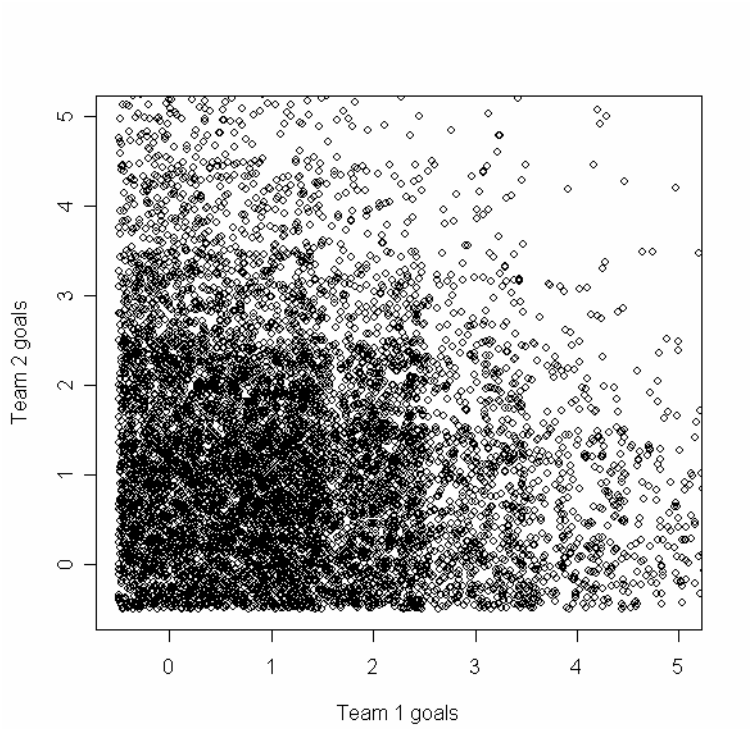

Figure 1: Scatter plot of pseudo-continuous goals data

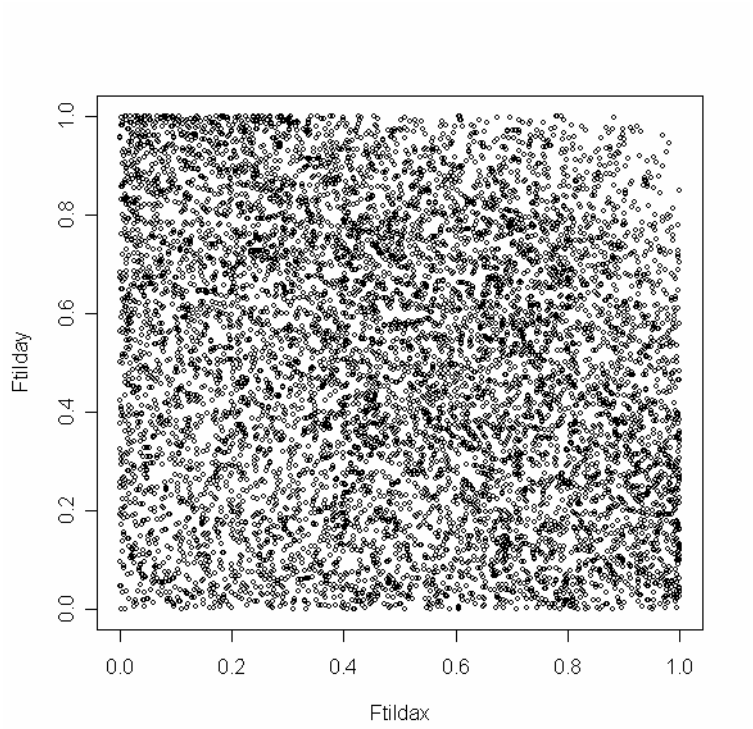

Figure 2: Bivariate uniform margins plot of "continuised" international soccer match score data.

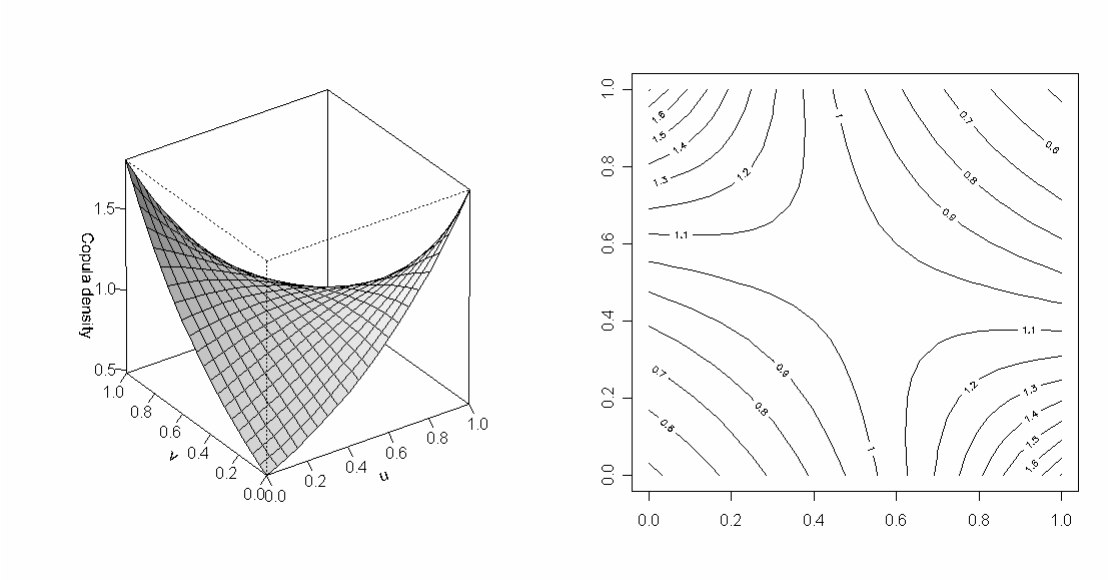

Figure 3: Frank's Copula density with $\kappa = -1.327$

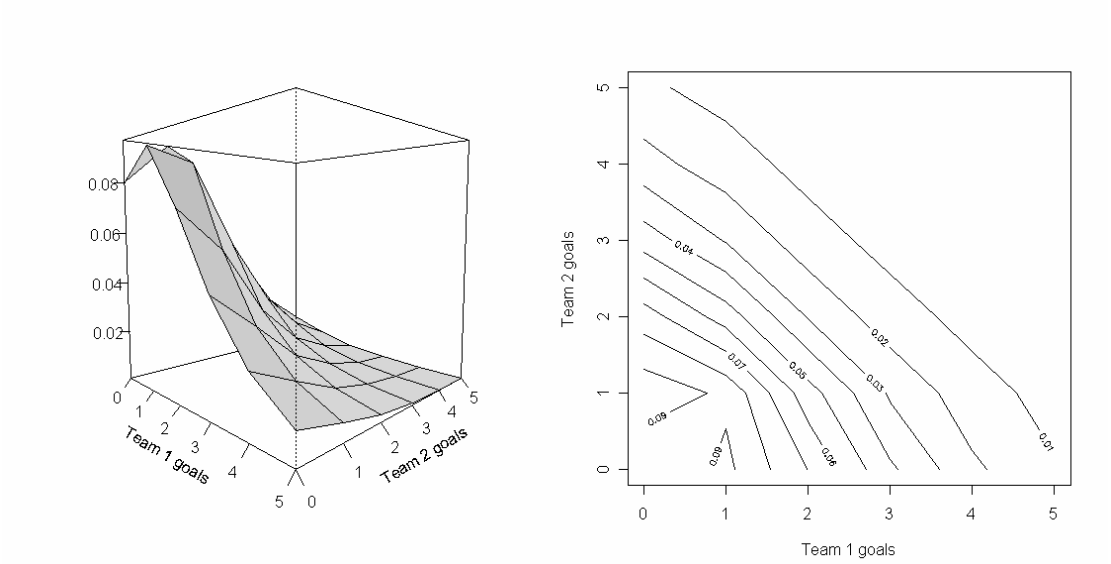

Figure 4: Fitted Frank's copula density, surface plot (left) and contour plot (right)