International Journal of Forecasting 33 (2017) 458–466

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

A bivariate Weibull count model for forecasting association football scores

Georgi Boshnakov[a], Tarak Kharrat[a,b], Ian G. McHale[b,*]

[a] School of Mathematics, University of Manchester, UK

[b] Centre for Sports Business, Salford Business School, University of Salford, UK

ARTICLE INFO

Keywords:
Betting
Calibration
Copula
Counting process
Soccer
Weibull

ABSTRACT

The paper presents a model for forecasting association football scores. The model uses a Weibull inter-arrival-times-based count process and a copula to produce a bivariate distribution of the numbers of goals scored by the home and away teams in a match. We test it against a variety of alternatives, including the simpler Poisson distribution-based model and an independent version of our model. The out-of-sample performance of our methodology is illustrated using, first, calibration curves, then a Kelly-type betting strategy that is applied to the pre-match win/draw/loss market and to the over–under 2.5 goals market. The new model provides an improved fit to the data relative to previous models, and results in positive returns to betting.

1. Introduction

Since the seminal paper by Maher (1982), a considerable amount of effort has been invested in modelling the probability distribution of scores in association football. Maher's model assumes that the numbers of goals scored by each team in a football match follow independent Poisson processes, and that the rates at which the two teams can expect to score goals are functions of their respective abilities in attack and defence. Subsequent efforts have enhanced the Maher model in a variety of directions. Dixon and Coles (1997) make two enhancements to Maher's model: first, they allow for dependence between the goals scored by the two teams; and second, they address the dynamic nature of teams' abilities by using a time-decay function in the likelihood, so that more recent results affect a team's estimated strength parameters more than results further in the past. Rue and Salvesen (2000) address the dynamic nature of teams' abilities in a Bayesian framework, as does Owen (2011). Karlis and Ntzoufras (2003) use a bivariate Poisson model with diagonal inflation, so that the probabilities of draw scores are calibrated better than with the simple independent Poisson model. Most recently, Koopman and Lit (2015) use a state space model to allow team strengths to vary stochastically with time.

These models all assume that the basic scoring pattern in football follows a (time-homogeneous) Poisson process. This assumption may be made more out of convenience than for any other reason, since there are surprisingly few natural alternatives, other than the negative binomial distribution.

Here, we propose the use of a count process that is derived when the inter-arrival times are assumed to follow an independent and identically distributed Weibull distribution. We refer to this model as the Weibull count distribution, and the form of the distribution for the count process generated by Weibull inter-arrival times was not known until recently. However, McShane, Adrian, Bradlow, and Fader (2008) derived this distribution, meaning that a new, more general, count process model can now be

* Corresponding author.
E-mail address: i.mchale@salford.ac.uk (I.G. McHale).

http://dx.doi.org/10.1016/j.ijforecast.2016.11.006

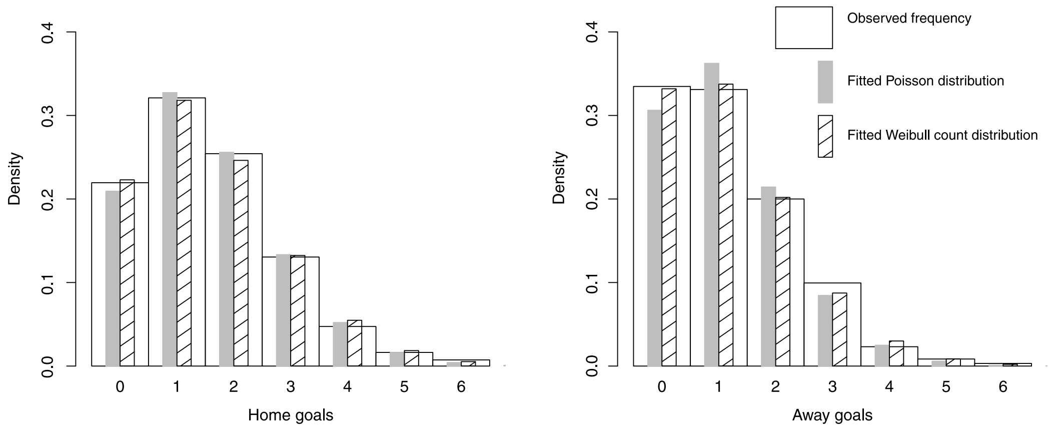

Fig. 1. Histograms of home (left) and away (right) goals with the fitted Poisson and Weibull count models. The estimated parameters for the home team are $\lambda_H = 1.50\ (0.04)$, $c_H = 1.56\ (0.03)$, while those for the away team are $\lambda_A = 1.10\ (0.03)$ and $c_A = 0.85\ (0.04)$, where the figures in parentheses are standard errors.

adopted. In addition to using a Weibull count model, we also allow for dependence between the goals scored by the two teams by employing a copula to generate a bivariate distribution allowing for positive or negative dependence.

Our objective in this paper is to build a model for the goals scored by the two teams in a football match. Our model can be used to construct the probabilities of the score-lines, and hence can be employed in betting market analysis or to study market efficiency, for example.

The computations and the graphs in the paper were done in R (R Core Team, 2016) using the Countr package (Kharrat & Boshnakov, 2016), available from the Comprehensive R Archive Network (CRAN).

The remainder of the paper is structured as follows. Section 2 presents the Weibull count distribution, our bivariate model, and provides our specification for its use when modelling the goals scored by the two teams in a football match. The results from fitting our model to data from the English Premier League are presented in Section 3, whilst the out-of-sample predictive performances, including the results of a simple Kelly-based betting strategy, are described in Section 4. We conclude with some closing remarks in Section 5.

2. A bivariate Weibull count distribution

2.1. The Weibull renewal process

McShane et al. (2008) derive the probability distribution of the number of events occurring by some time t when the inter-arrival times are assumed to be independent and identically distributed Weibull random variables (this process is also known as a Weibull renewal process). They do this by using a Taylor series expansion of the exponential in the Weibull density. They refer to the resulting count process as the '*Weibull count model*', and its probability mass function is given by

$$\Pr(X(t) = x) = \sum_{j=x}^{\infty} \frac{(-1)^{x+j}(\lambda t^c)^j \alpha_j^x}{\Gamma(cj+1)}, \tag{1}$$

where $\alpha_j^0 = \Gamma(cj+1)/\Gamma(j+1)$, $j = 0, 1, 2, \ldots$, and $\alpha_j^{x+1} = \sum_{m=x}^{j-1} \alpha_m^x \Gamma(cj - cm + 1)/\Gamma(j - m + 1)$, for $x = 0, 1, 2, \ldots$, and $j = x + 1, x + 2, x + 3, \ldots$. In Eq. (1), λ is a 'rate' parameter and c is the 'shape' parameter of the distribution, where the observation unit is the match, which we take as having a duration of one time unit. Thus, λ is the scoring rate per match.

The use of the Weibull distribution to model the inter-arrival times allows the hazard $h(t)$ associated with the count process to vary over time. The Weibull hazard is given by

$$h(t) = \lambda c t^{c-1},$$

and may be monotonically increasing for $c > 1$, monotonically decreasing for $c < 1$, or constant (and equal to λ) for $c = 1$. Note that we recover the (time-homogeneous) Poisson process when $c = 1$. It is also interesting to note that this model handles both over-dispersed data (the mean is smaller than the variance; $c < 1$) and under-dispersed data (the mean is larger than the variance; $c > 1$) naturally, whilst the Poisson count distribution ($c = 1$) can only accommodate equi-dispersed data (the mean is equal to the variance).

Despite the somewhat intimidating appearance of Eq. (1), the computations for the Weibull count model can be performed without much trouble. For the usual values of the count (goals) observed in association football ($x \in [0, 10]$), the first 50 terms of the infinite series are sufficient to enable the probabilities to be computed accurately. For the sake of speed, we implemented these computations in C++, though McShane et al. (2008) were able to perform the computations in Microsoft Excel. We validated the computations by retrieving the Poisson case ($c = 1$) and reproducing the analysis conducted by McShane et al. (2008). All of the computations described in this paper can be reproduced using the R (R Core Team, 2016) add-on package Countr (Kharrat & Boshnakov, 2016).

Fig. 1 shows the Weibull count model and the Poisson distribution fitted to the goals scored by the home (left) and away (right) teams in matches played in the English Premier League during the five seasons from 2010–11 to

Table 1
χ^2 goodness-of-fit test statistics for the fitted Weibull count model and the Poisson distribution with home and away goals.

	Home goals			Away goals		
	χ^2	df	p-value	χ^2	df	p-value
Weibull count model	7.65	4	0.10	6.59	4	0.16
Poisson distribution	13.23	5	0.002	23.5	5	0.0002

2014–15. The density histograms of home and away goals are also shown. Eyeballing the fits of the two distributions suggests that the Weibull count model and the Poisson distribution provide similar levels of goodness-of-fit for home goals (although the Weibull count is slightly better, especially for the 0 count), whereas that the Weibull count model is a clear improvement for away goals. The χ^2 goodness-of-fit test statistics for the fitted models shown in Table 1 support this. In fact, they suggest that the Poisson distribution is not adequate for either home or away goals, while the Weibull count model is appropriate.

We now present a bivariate distribution based on the Weibull count model, and include some modifications that can be used for forecasting the results of football matches.

2.2. *Using a copula to generate a bivariate model*

The existence of some sort of dependence between the goals scored by two teams in a football match is widely accepted. However, the exact specification of this dependence is less clear. For example, Dixon and Coles (1997) find evidence of dependence by studying the difference between the empirical joint distribution of goals scored by the two teams and the implied joint distribution under the hypothesis that the two random variables are independent (i.e., the product of the marginal distributions). After confirming that the distributions are not independent, they impose an *ad hoc* correction on their bivariate Poisson distribution. Karlis and Ntzoufras (2003) use a diagonally inflated distribution to account for the fact that draw results are observed more frequently than they would be under the independent bivariate Poisson model (although only positive dependence can be captured by the bivariate Poisson distribution). Here, we follow McHale and Scarf (2011), who choose to allow for any potential dependence between the goals scored by the two teams through the use of a copula to 'glue' together the two marginal distributions of goals scored.

A copula, C, is a multivariate distribution for which all univariate marginal distributions are distributed uniformly on the unit interval, $[0, 1]$. In other words, C is the distribution of a multivariate uniform random vector. The power of a copula approach for modelling the dependence comes from the theorem of Sklar (1973) that the joint cumulative distribution function F of any pair of random variables (Y_1, Y_2) may be written in the form

$$F(y_1, y_2) = C(F_1(y_1), F_2(y_2)), \quad (y_1, y_2) \in \Re^2,$$

where F_1 and F_2 are their marginal cumulative distribution functions and C is a copula.

Here, we want to join two marginal distributions, and there are a plethora of copulas to choose from in this bivariate case. Initially, we experimented with the Gaussian copula and Frank's copula, but Frank's copula tended to provide a superior fit to the data according to both the AIC and the BIC. Of course, it also has the advantage of allowing for the full range of dependence, meaning that the correlation can range from -1 to 1. The Frank copula is given by

$$C(u, v) = -\frac{1}{\kappa} \ln\left(1 + \frac{(e^{-\kappa u} - 1)(e^{-\kappa v} - 1)}{e^{-\kappa} - 1}\right),$$

where $\kappa \in \Re$ is the dependence parameter.

We construct our bivariate Weibull count model by using the Weibull count probability mass function given in Eq. (1) to calculate the cumulative distribution functions, $F_1(y_1; \lambda_1, c_1)$ and $F_2(y_2; \lambda_2, c_2)$. Using the copula $C(u, v; \kappa)$ to glue these marginals together, the likelihood function for the parameter vector $(\lambda_1, c_1, \lambda_2, c_2, \kappa)$ given the ith pair of observations (y_{1i}, y_{2i}) is then

$$\begin{aligned}
&\mathcal{L}(\lambda_1, c_1, \lambda_2, c_2, \kappa; y_{1i}, y_{2i}) \\
&\quad = \Pr(Y_1 = y_{1i}, Y_2 = y_{2i}) \\
&\quad = C(F_1(y_{1i}), F_2(y_{2i})) - C(F_1(y_{1i} - 1), F_2(y_{2i})) \\
&\qquad - C(F_1(y_{1i}), F_2(y_{2i} - 1)) \\
&\qquad + C(F_1(y_{1i} - 1), F_2(y_{2i} - 1)). \qquad (2)
\end{aligned}$$

The log-likelihood, $\ell(\lambda_1, c_1, \lambda_2, c_2, \kappa; \mathbf{y}_1, \mathbf{y}_2) = \sum_{i=1}^{n} \log \mathcal{L}$, can be maximised for a sample of n football matches using standard numerical optimization routines.

We note that Frank's copula nests the independence case ($\kappa = 0$), so that a test of whether κ is equal to zero is equivalent to testing the assumption of independence.

2.3. *A model for goals*

Since Maher (1982) first provided his specification for modelling the goals scored by the two teams in a football match, many researchers have followed suit. Adapting the specification to our bivariate model is simple due to the presence of the rate parameter λ in the distribution function given in Eq. (1). Following Maher (1982), we let the rate parameter for home team i playing against away team j be

$$\log(\lambda_i) = \alpha_i + \beta_j + \gamma,$$

where α_i is the attack strength (the higher the value of α, the stronger the attack) of team i, β_j is the defence strength (the smaller the value of β, the stronger the defence) of team j, and γ is a home advantage parameter. The away team's scoring rate is given by

$$\log(\lambda_j) = \alpha_j + \beta_i.$$

The above model is static, in that the team strength parameters are not allowed to vary with time. This is not

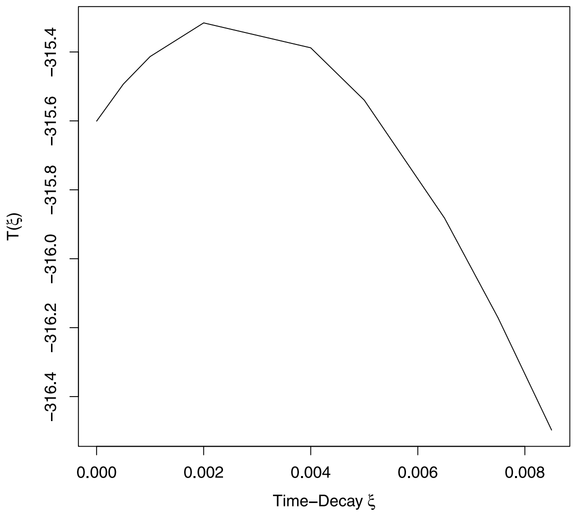

Fig. 2. Objective function.

a good approximation. For example, Baker and McHale (2015) present a time-varying model for team strengths and find that the strengths do indeed vary over time as, for example, teams buy and sell players, injuries occur, and runs of good and bad form happen. The forecasting literature has seen two approaches to allowing time-varying team strengths.

First, one can build a model in which the team strengths are assumed to vary stochastically, as per Crowder, Dixon, Ledford, and Robinson (2002), Koopman and Lit (2015) or Owen (2011). Second, one can adopt a time decay factor in the likelihood function, so that more recent matches have greater weight in estimating the team strength parameters than matches that were further in the past. This implicitly assumes that each team's recent results are indicative of the team's changing strength—which seems an intuitive and reasonable assumption to make. This approach was first presented by Dixon and Coles (1997), and is how we deal with the time-varying nature of team strengths here.

Using an exponential weighting function, our 'pseudo-likelihood', $\tilde{\mathcal{L}}$, at time t is given by

$$\tilde{\mathcal{L}}_t(\kappa, c, \alpha, \beta) = \prod_{k \in A_t} e^{-\xi(t-t_k)} \mathcal{L},$$

where t_k is the time when match k was played, $A_t = \{k : t_k < t\}$ is the subset of all matches played up to, but not including, time t, and $\mathcal{L}$ is as given in Eq. (2). The parameter ξ cannot be estimated by maximising the likelihood. Dixon and Coles (1997) select the value of ξ that maximises

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

where $\delta_k^H = 1$ if match k is a home win (else $\delta_k^H = 0$), and p_k^H, p_k^A and p_k^D are the maximum likelihood estimates of a home win, an away win and a draw, respectively. We modify this approach because our model may be 'wasted' testing it against only the 1X2 (home win, draw, away win) market. Rather, we choose ξ to maximise the following objective function, which includes the over–under 2.5 goals market:

$$T(\xi) = \sum_{k=1}^{N} (\delta_k^H \log p_k^H + \delta_k^A \log p_k^A + \delta_k^D \log p_k^D + \gamma_k^{O2.5} \log p_k^{O2.5} + \gamma_k^{U2.5} \log p_k^{U2.5}), \quad (3)$$

where $\gamma_k^{O2.5} = 1$ if there are more than 2.5 goals in match k and $\gamma_k^{U2.5} = 1$ if there are fewer than 2.5 goals in match k, and $p_k^{O2.5}$ and $p_k^{U2.5}$ are the model-implied probabilities of there being more than or fewer than 2.5 goals in match k. Fig. 2 shows a plot of Eq. (3) as ξ varies from 0 to 0.008, with a clear maximum at $\xi = 0.002$; thus, this is the value that we use throughout. Following Dixon and Coles (1997), we use a half-week as the unit of time. They estimate a value of 0.0065, which is of the same order of magnitude as our estimated value of ξ.

3. Results

We obtained data on the results of matches in the English Premier League for the ten seasons from 2006/07 to 2015/16, inclusive, from www.football-data.co.uk. The data also include pre-match odds from several bookmakers for both the 1X2 (win, draw, lose) market and the over–under 2.5 goals market. We examine the goodness-of-fit of the models to these data using several methods: the log-likelihood and Akaike information criterion (AIC) are used for in-sample goodness-of-fit diagnostics, whilst returns to betting and calibration curves are used to examine the models' out-of-sample predictive abilities. First, though, we take a look at the estimated team strengths.

3.1. Estimated team strengths

As we describe below, we fitted the model to rolling windows of four-and-a-half seasons. This section reports the results at the end of the 2015/16 season. The estimated distribution parameters c in Eq. (1) are $c_H = 1.050$ (0.0416) for the home team and $c_A = 0.9831$ (0.0482) for the away team (with standard errors in parentheses). The estimated value of the dependence parameter in the copula is $\kappa = -0.4561$ (0.1961), so that the value of Kendall's τ is -0.050, suggesting a negative (and statistically significant) dependence. Lastly, the estimated home advantage is $\gamma = 0.2948$ (0.0503). The standard errors have been computed under the assumption of asymptotic normality and using numerical estimates of the gradient and the hessian. As McShane et al. (2008) pointed out, this gives reasonably close results compared to a bootstrap approach.

Table 2 shows the estimated team strength parameters, α (attack) and β (defence), for our model. As expected, teams with fewer observations have larger standard errors for the estimated strengths. These teams have not played in the league every year due to promotion or relegation (after each season, three teams are relegated and three other teams are promoted into the Premier League from the Championship). For example, the Blackburn Rovers and Wolverhampton Wanderers have the two largest standard errors on their estimated attack strengths.

Table 2
Estimated team strength parameters, based on the last four-and-a-half seasons' matches.

Team	α (s.e.)	β (s.e.)
Arsenal	0.392 (0.076)	0.241 (0.103)
Aston Villa	−0.272 (0.101)	−0.281 (0.084)
Blackburn Rovers	−0.138 (0.316)	−0.422 (0.253)
Bolton Wanderers	−0.152 (0.310)	−0.379 (0.263)
Bournemouth	0.014 (0.156)	−0.283 (0.139)
Burnley	−0.470 (0.213)	−0.053 (0.165)
Cardiff City	−0.344 (0.221)	−0.423 (0.156)
Chelsea	0.371 (0.077)	0.196 (0.102)
Crystal Palace	−0.129 (0.107)	0.011 (0.105)
Everton	0.183 (0.083)	0.039 (0.095)
Fulham	−0.020 (0.129)	−0.373 (0.113)
Hull City	−0.261 (0.144)	−0.063 (0.127)
Leicester City	0.264 (0.104)	0.096 (0.125)
Liverpool	0.401 (0.076)	−0.013 (0.094)
Manchester City	0.550 (0.072)	0.242 (0.104)
Manchester United	0.312 (0.079)	0.230 (0.102)
Newcastle United	−0.058 (0.093)	−0.256 (0.085)
Norwich City	−0.201 (0.111)	−0.241 (0.095)
Queens Park Rangers	−0.173 (0.132)	−0.314 (0.110)
Reading	−0.103 (0.211)	−0.387 (0.172)
Southampton	0.165 (0.087)	0.120 (0.101)
Stoke City	−0.097 (0.093)	−0.013 (0.093)
Sunderland	−0.123 (0.094)	−0.128 (0.088)
Swansea City	0.024 (0.089)	−0.059 (0.091)
Tottenham Hotspur	0.302 (0.079)	0.106 (0.098)
Watford	−0.135 (0.167)	0.012 (0.156)
West Bromwich Albion	−0.111 (0.094)	−0.068 (0.091)
West Ham United	0.093 (0.089)	−0.019 (0.096)
Wigan Athletic	−0.001 (0.168)	−0.310 (0.153)
Wolverhampton Wanderers	−0.279 (0.338)	−0.682 (0.228)

Larger αs indicate a stronger attack, smaller βs indicate a stronger defence.

3.2. In-sample model diagnostics

We compare the fit of our main model with the performances of three other models, an independent Poisson model, an independent Weibull count model and a Frank copula-induced bivariate Poisson model. Using these three models as benchmarks enables us to determine the cause of any out-performance (for example, an improvement in goodness-of-fit may come from either modelling the dependence structure using a copula or modelling the counts using a Weibull count model rather than a Poisson distribution).

Table 3 shows the log-likelihood, the number of model parameters and the AIC for each of the four models under consideration. Although the copula-induced bivariate Weibull count model has more parameters, it is the best-fitting model based on the AIC. It is worth noting that the change from Poisson to a Weibull count distribution improves the AIC by approximately 6–10 units, while the change from independence to copula-induced dependence improves the AIC by approximately 12–16 units. As such, it looks like the overall improvement in our model comes from both the copula-based dependence and the use of the Weibull count distribution.

4. Out-of-sample performance

To test the model out-of-sample, we fit the model to rolling windows of four-and-a-half seasons (1710 matches), and predict the following week's results for each fit. The first four-and-a-half season window begins at the start of the 2006/07 season and ends half way through the 2009/10 season. Having made predictions for the following week's games, we move forward one week and refit the model to take account of the latest round of results. We repeat this until the last round of games in the 2009/10 season have been predicted. We then wait for half of the next season to be completed so that there are plenty of data available with which to obtain reasonable parameter estimates for the newly promoted teams and teams with large turnovers in playing staff. This process is repeated for each of the five seasons, resulting in a total of 1140 games (six $\times$ half seasons) for which out-of-sample forecasts were generated and on which bets could be placed. The four-and-a-half season window length seems a good compromise between the desire to use more data for model fitting and keeping the model useful for prediction.

4.1. Calibration curves for the bivariate Weibull count model

Calibration can be seen intuitively as a way of visualising how often a model is right or wrong. In fact, a perfectly calibrated model *knows* how often it is right or wrong: when it predicts an event with 80% confidence, the event will occur 80% of the time. Whilst perfect accuracy is probably an unachievable goal for football forecasting models, theoretically perfect calibration is a more realistic target, since, in principle, a model with imperfect accuracy could still be perfectly calibrated. Although the notion of calibration is popular in quantitative finance, it has not yet been investigated in the sports forecasting literature (to the best of our knowledge).

Table 3
Comparison of the four models for football scores, fitted (in-sample) to the Premier League data.

	Log-likelihood	Number of parameters	AIC
Copula Weibull count model	−3250.00	64	6628.00
Copula Poison model	−3257.09	62	6638.19
Independent Weibull count model	−3258.99	63	6643.98
Independent Poisson model	−3264.00	61	6650.00

This section evaluates the calibration of the bivariate Weibull count model's posterior prediction distribution directly using the 1140 matches in our out-of-sample period. For each event forecasted, we visualise the model's performance graphically by plotting the *calibration curves* (also known as *reliability plots*). We now briefly describe the estimation of the calibration curves in football.

Consider a binary probabilistic prediction problem, which consists of binary labels and probabilistic predictions for them. Each instance has a *ground-truth label* $y \in \{0, 1\}$ and an associated *predicted probability* $q \in [0, 1]$ generated by the model, where q represents the model's posterior probability of the instance having a positive label ($y = 1$). The calibration curve is simply a plot of the label frequency, $\mathsf{P}(y = 1|q)$, versus the predicted probability. However, computing $\mathsf{P}(y = 1|q)$ requires an infinite amount of data, and therefore approximation methods are needed for performing the calibration analysis. We follow Tukey's (1961) approach here, and divide the prediction space by 'halves': we split the data into upper and lower halves, then split those halves, then split the extreme halves recursively. Relative to equal-width binning, this allows a visual inspection of the tail behaviour without too many graphical elements being devoted to the bulk of the data. A perfectly calibrated curve would coincide with the $y = x$ line, so that the empirical frequency of an event equalled the model's estimated probability. When the curve lies above the diagonal, the model is *pessimistic*, in that it underestimates the probability of the event occurring; and when it is below the diagonal, the model is *optimistic*, in that it overestimates the probability of the event occurring.

The calibration curves for predicting the home win, draw and away win outcomes in the 1X2 market are shown in Fig. 3. Overall, it appears that the model is 'well-calibrated': the points lie near the $y = x$ line.

4.2. Betting performance

We now test all four models against the betting market. There is a vast array of work in the economics literature on the efficiency of the betting market for football, and, on the whole, there is agreement that the market is efficient, in that it is not possible to accrue 'superior' returns (see for example Snowberg & Wolfers, 2010). Thus, comparing the probabilities implied in the betting market with those produced by the model is a simple but informative guide to the model's effectiveness.

Our betting simulation is *out-of-sample*: team strengths are estimated using results prior to the match to be bet on. As a consequence of the efficient markets hypothesis, we would consider a return that is near the market over-round to be evidence that a model is working well. We use the average odds available on two markets: the 1X2 (home win, draw, away win) market, and the over–under 2.5 goals market. During the last half of each of the ten seasons of data, the average over-rounds on the two markets were 5.5% and 6.0%, respectively. By testing our model against the over–under market, we are gaining an understanding of the model's performance in predicting what it was designed to forecast: goals. If we were to test the model against only the 1X2 market, we would be disregarding the main output from the model: the probabilities of each and every possible scoreline.

Our investment strategy is based on the Kelly criterion (Kelly, 1956). The Kelly criterion arose from a desire to maximise the long-run log-utility, and results in an investment strategy where the bettor invests a fraction f of his overall wealth

$$f = \frac{(b+1)p - 1}{b},$$

where p is the bettor's estimate of the probability of an event (e.g., the home team winning the game), and b is the (fractional) odds offered by the bookmaker (where $1/(b+1)$ can be interpreted loosely as the bookmaker's implied probability of the event occurring).

We allow a maximum of one unit per bet and use the Kelly criterion to decide what fraction of our one unit is staked. Effectively, we reset our bankroll to one after each bet. An additional 'protection' was also introduced: we restrict ourselves to 'quality bets', when the expected value of any bet is greater than some threshold. For each game, there are five possible events to bet on: home win, draw, away win, over 2.5 goals and under 2.5 goals. For event type A, we only bet if

$$EV(A) = \Pr(A) \times Odds(A) - 1 > t,$$

where t is the threshold parameter. In order to choose a relevant value of t, i.e., a value that is a good compromise between betting too much (and losing) and placing a reasonable number of bets, we use our predictions for week 20–21 of every season in our testing set (10 matches $\times$ 2 weeks $\times$ 6 seasons $=$ 120 games), as described before. A value of $t = 0.038$ was obtained and used to bet on the remaining unseen 1020 games. Table 4 shows the *out-of-sample* returns to 1X2 betting on these matches and Table 5 shows the *out-of-sample* returns to betting on the over–under 2.5 goals market for the same matches.

In both cases, the copula Weibull count model produces the highest returns: 21.2% in the case of the 1X2 market and 15.5% in the over–under market. Reading from the bottom row of Table 4 to the top row shows how improvements in the returns are gained by, first, adopting the copula bivariate model rather than the independent model, and second, using the Weibull count distribution instead of the Poisson distribution.

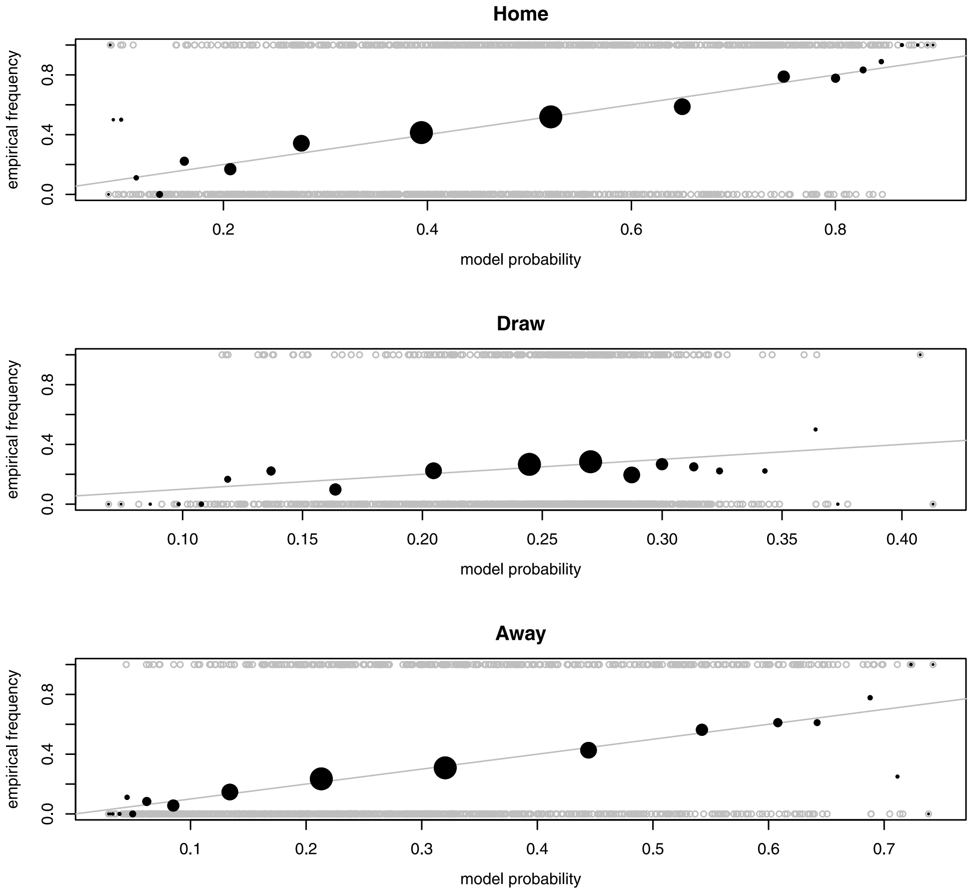

Fig. 3. Calibration curves for predicting outcomes in the 1X2 market. Note that the size of a circle is proportional to the number of observations in that bin.

Table 4
Summary of results when betting on the 1X2 market using a Kelly betting strategy.

Model	Number of bets	Number of winning bets	Gross return	Net return	Total staked	Return on investment
Copula Weibull count	612	312	625.95	13.95	65.80	21.2%
Independent Weibull count	663	278	671.90	8.90	67.42	13.2%
Copula Poisson model	632	265	639.02	7.02	58.01	12.1%
Independent Poisson model	675	291	682.44	7.44	62.52	11.9%

Table 5
Summary of results when betting on the over–under 2.5 goals market using a Kelly betting strategy.

Model	Number of bets	Number of winning bets	Gross return	Net return	Total staked	Return on investment
Copula Weibull count	356	190	361.88	5.88	37.93	15.5%
Independent Weibull count	385	202	388.61	3.61	40.56	8.9%
Copula Poisson model	366	187	369.99	3.99	44.33	9.0%
Independent Poisson model	377	195	381.12	4.12	46.81	8.8%

The profits demonstrated above may be obtained as a consequence of using either an improved model or an improved betting strategy. To illuminate this issue, we also tested the models using a simple one-unit betting strategy: if the expected value of the bet is above the threshold t, we bet one unit. Note that this strategy stakes the same amount irrespective of the value of the expected value (as long as it exceeds the threshold). This resulted in returns of 23.4% from betting on the 1X2 market using the copula Weibull count model. This is higher than the returns obtained using the Kelly betting strategy (which was 21.2%), which suggests that the returns are coming from the model, not necessarily from the betting strategy. Note, however, that we still prefer the use of the Kelly betting strategy because it has smaller confidence intervals and a higher realised Sharpe ratio (0.012 for the Kelly

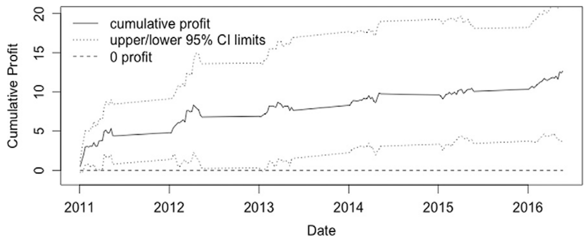

Fig. 4. Cumulative profit from a Kelly betting strategy on the 1X2 market using the bivariate Weibull count model.

strategy versus 0.054 for the unit betting strategy), which makes its returns more 'robust'.

Before moving away from testing our model against the betting market, we present the evolution of the cumulative returns over the betting period, together with bootstrapped confidence intervals. Using a bootstrap confidence interval allows us to see just how 'robust' the returns we achieve are. We computed the confidence intervals by utilising a bootstrap procedure in which 100 replicate datasets of the same size as the original (1020 games) were generated by sampling individual predicted probabilities-odds pairs, (p_i, O_i), and the eventual outcome, with replacement. The results reported for the (95%) confidence intervals are based on a normal approximation and standard errors estimated from the standard deviation of the returns across those samples.

Fig. 4 shows the cumulative profit on the 1X2 market according to the Kelly betting strategy. The plot for the over–under 2.5 goals market is similar. The flat regions are when we do not bet. The most promising sign here is that the lower bound of the confidence interval is positive.

Obtaining returns that are superior and positive in both the 1X2 and over–under 2.5 markets is interesting for two reasons. First, we take the results as a validation of our model. Second, and more profoundly, there are implications for market efficiency, which may be the subject of future research.

5. Discussion

This paper has studied a new model of bivariate counts for predicting the score distribution in football matches. Our model assumes that the distribution of inter-arrival times for goals is Weibull rather than exponential (the latter is implied when using the Poisson distribution). We induce dependency between the two marginals using a Frank copula, and test this bivariate Weibull count model against three alternative models both in-sample and out-of-sample. The bivariate Weibull count model is found to provide a superior fit to the results data from the English Premier League.

Perhaps the most interesting finding is the fact that our model obtains positive returns in both the 1X2 and over–under 2.5 goals betting markets. However, it is worth noting that although we can 'beat the bookmakers', we are picking and choosing only a relatively small number of matches to bet on. A bookmaker must produce competitive odds for an entire league fixture list, which is a more challenging task and requires an intimate knowledge of the leagues, teams and players involved. Hence, bookmakers in the betting industry continue to rely on a system in which statistical models and traders work in tandem to produce odds. However, the model presented here may prove useful to bookmakers. For example, we believe that bookmakers estimate a supremacy rating for one team over another in a match and an expected number of goals in the match. Bookmakers could use these two variables, which are based on an intimate knowledge of the form of the teams and the players who are actually playing for each team, as inputs to our model. The 1X2 markets will still be priced the same, but a better estimation of the goals distribution will allow more accurate prices to be produced for the subsidiary markets.

In future work, we hope to bridge the gap between the statistical model and traders by incorporating player skills into our model. A knowledge of the identities of the players on each team (and their abilities) should avoid the need to estimate time-varying team strengths, as the key component of the dynamic nature of team strengths is the changing line-up of the team. The use of accurate player ratings as inputs to the model should produce improved forecasts, as it will capture the expected scoring intensities of each team better.

Acknowledgments

We would like to thank the two anonymous reviewers and the Associate Editor for their comments and suggestions, which helped to improve the paper.

References

Baker, R. D., & McHale, I. G. (2015). Time varying ratings in association football: the all-time greatest team is.. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(2), 481–492.

Crowder, M., Dixon, M., Ledford, A., & Robinson, M. (2002). Dynamic modelling and prediction of English football league matches for betting. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *51*(2), 157–168.

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *46*(2), 265–280.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The statistician)*, *52*, 381–393.

Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, *35*(4), 917–926.

Kharrat, T., & Boshnakov, G.N. (2016). Countr: Flexible univariate count models based on renewal processes. R package version 3.2.7.

Koopman, S. J., & Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 167–186.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, *36*(3), 109–118.

McHale, I. G., & Scarf, P. (2011). Modelling the dependence of goals scored by opposing teams in international soccer matches. *Statistical Modelling*, *11*(3), 219–236.

McShane, B., Adrian, M., Bradlow, E. T., & Fader, P. S. (2008). Count models based on Weibull interarrival times. *Journal of Business and Economic Statistics*, *26*(3), 369–378.

Owen, A. (2011). Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics*, *22*, 99–113.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rue, H., & Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The statistician)*, *49*, 399–418.

Sklar, A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika*, *9*(6), 449–460.
Snowberg, E., & Wolfers, J. (2010). Explaining the favorite-longshot bias: is it risk-love or misperceptions? *Journal of Political Economy*, *118*, 723–746.
Tukey, J. W., et al. (1961). Curves as parameters, and touch estimation. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: Contributions to the theory of statistics*. The Regents of the University of California.

Georgi Boshnakov is Lecturer in Statistics at the School of Mathematics, University of Manchester. Georgi's research interests are mainly in time series analysis, statistical computing and dynamic modelling.

Tarak Kharrat is a Research Associate at the Centre for Sports Business at the University of Salford. Prior to studying for his Ph.D. in Statistics at the University of Manchester, Tarak spent time on Wall Street. His research interests are varied, and include prediction modelling in sport, heavy-tailed distributions and state-space modelling.

Ian G. McHale is Professor of Sports Analytics at the Centre for Sports Business at the University of Salford. Ian was founding Chair of the Statistics in Sports Section of the Royal Statistical Society, and serves as an Associate Editor for the *Journal of Quantitative Analysis in Sports*. His research interests include statistics in sport, the analysis of gambling markets and issues related to gambling.