International Journal of Forecasting 29 (2013) 122–130

Contents lists available at SciVerse ScienceDirect

International Journal of Forecasting

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

Forecasting exact scores in National Football League games

Rose D. Baker, Ian G. McHale [*]

Centre for Operations Management, Management Science and Statistics, Salford Business School, University of Salford, UK

ARTICLE INFO

Keywords:
Sports
Predictions
Point processes

ABSTRACT

The paper presents a point process model for predicting exact end-of-match scores in the premier league of American football, the National Football League. The hazards of scoring are allowed to vary with team statistics from previous games and/or the bookmaker point spread and over-under. The model is used to generate out-of-sample forecasts, which are evaluated using several criteria, including a Kelly betting strategy. In predicting the results of games, the model is marginally outperformed by the betting market. However, when it is used to forecast exact scores, the model proves to do at least as well as the market.

1. Introduction

Understandably, forecasting in sport is most often concerned with forecasting win/lose results. However, in some circumstances it is of interest to model exact scores, from both a practical and a modeling perspective. For example, bets can be placed on either the result of a soccer match (home win, draw or away win) or the exact score (0–0, 1–0, 0–1, 2–1, etc.), and hence a bookmaker (or bettor) is interested in a model for exact scores. Developing exact score models for soccer is relatively simple, due to the nature of the scoring system used. The scoring system in soccer means that models of exact scores take advantage of the relatively well-behaved counts that are goals, and, as such, are based around Poisson-type regression models (see, for example, Dixon & Coles, 1997, and McHale & Scarf, 2011). However, some sports are not scored using such a simple system.

One such example is American Football, since, in the National Football League (NFL), for example, there are five ways to score: an unconverted touchdown (6 points), a touchdown with a one point conversion (7 points), a touchdown with a two point conversion (8 points), a safety (2 points) and a field goal (3 points). The high relative frequency of seven point converted touchdowns and of three point field goals results in a rather peculiar distribution of each team's points in games. Fig. 1 shows a bar plot of the home and away teams' scores in 2128 NFL games between 2001 and 2008, and the peculiar shape depicts a distribution that is indeed non-standard, as it is clear that $3x+7y,\ x, y \in 0, 1, 2, \ldots$ combinations are observed more frequently than other scores.

In the literature, exact score forecasts for NFL games have not previously been the focus of forecasting exercises. This is most probably due to two factors. First, given that the most widely used application of forecasting models for sports is in the analysis of betting market efficiency, the exact score is not of primary interest, as markets (books) have traditionally been offered on the point spread (the difference in the points scored by each team) and the over-under (the sum of all points scored in the game), rather than the exact scores. However, with the advent of betting exchanges such as Betfair, new markets are available for a wide variety of bets: from the number of touchdowns scored in the game to the time of the next scoring play. As such, models that reveal more information about a game than just the result are in demand. Second, as we have already seen, the complex scoring system results in scorelines that are difficult to model using standard modeling approaches, and as such, statisticians may have shied away from a task that was, until recently, of little practical use.

Given that the most convenient, and probably toughest, test of a forecasting model in sport is to compare its predictive power with that of the betting market, previous

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

0169-2070/$ – see front matter © 2012 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2012.07.002

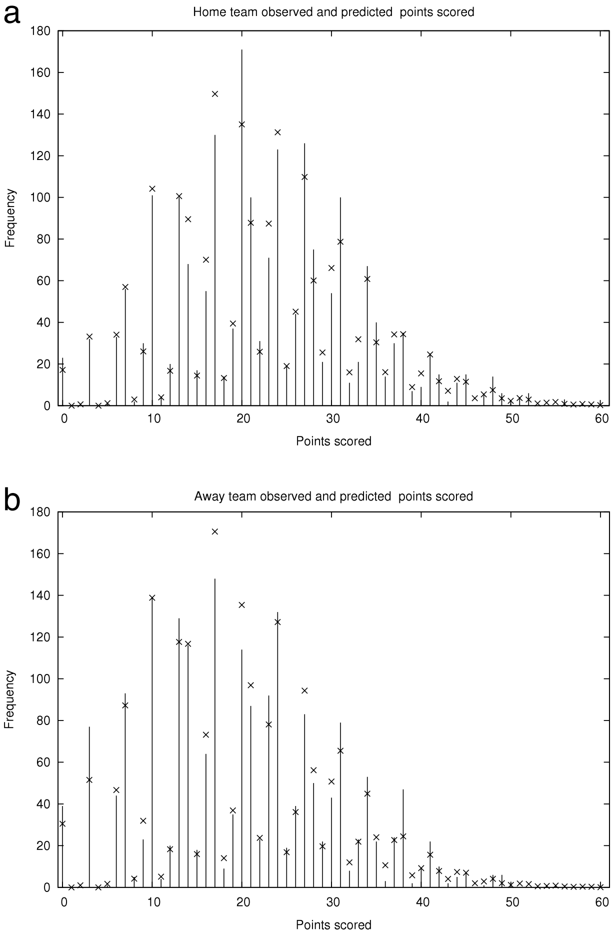

Fig. 1. Observed and predicted frequencies of points scored for (a) home teams and (b) away teams. The solid lines and the ×s represent the observed frequencies and those predicted from our model described below, respectively.

work has focused on either predicting the winning team in a match or forecasting the spread. Harville (1980) presented a linear model of what he calls the *score*, but which is actually the difference of the two teams in points, i.e. the spread, using the differences in points in past games as covariates. Several authors have also used linear models to forecast the difference in points; see, for example, Sauer, Brajer, Ferris, and Marr (1988) and Zuber, Gandar, and Bowers (1985). Stern (1991) bridged the gap between forecasting the spread and forecasting match winners by investigating the distribution of the margin of victory minus the point spread, and, having found that a normal approximation was acceptable, used his result to estimate the probability of a game being won. In a departure from the linear model, Glickman and Stern (1998) used a state space model to estimate the difference in the points scored by the two teams in a game. Cain, Law, and Peel (2000) take the non-normality of forecast errors into account in models of the difference in points, and find, as have most studies, that the over-under and point spread betting markets are efficient, in that it is not possible to obtain positive returns.

Paralleling the literature on forecasting the difference in points, many papers have concentrated on the problem of forecasting the match winner. For example, Boulier and Stekler (2003) used a variety of approaches to estimate the probability of victory. They compared the forecasting performances of probit models based on the power scores provided by *The New York Times* with the forecasting performances of the betting market and the sports editor's predictions. The betting market prevailed, providing the best predictions.

Regardless of the type of prediction (probability of victory or margin of victory) and of the information used in the model (experts' opinions, past match statistics and/or power scores), outperforming forecasts based on information from the betting market has, for the most part, proven elusive. Such findings suggest that the football betting market is, in the main, efficient, in that it is not possible to obtain abnormal returns. For a review of the literature on the efficiency of betting markets, see, for example, Sauer (1998). For a review of the football forecasting literature, see Stekler, Sendor, and Verlander (2010).

To the best of our knowledge, in neither the statistics nor the economics literature has a paper yet addressed the problem of forecasting exact end-of-match scores in football. In the next section we describe a birth process model for exact end-of-match scores in the NFL. In

Section 3 we describe the data used here, and in Section 4 we present the model fit and some diagnostics. We include a comparison of forecasts based on our model with those of the betting market. Some conclusions are given in Section 5.

2. A model for exact scores in football

The five ways of scoring in the NFL for each team lead to ten different types of scores occurring during a game. Thus, the process of scoring is a birth process, with ten different types of birth. We can model this birth process as a continuous-time Markov process. Each team then has a 'hazard' of making each type of score, and this hazard can vary as the state of play changes. This type of model has previously been used in soccer, for which scoring is a birth process with just two types of birth, goals for the home and away teams (see Dixon & Robinson, 1998). One can model the intensity of each type of score as a function of the time into the game, the current state of play (defined as the numbers of each type of score the two teams have accrued, that is, a 10-tuple of scores), and exogenous variables such as whether a team is playing at home or away, or the previous performance of each team.

Standard birth process models rely on having information on the timings of each score during the game (that is, the minute a touchdown occurred, for example). We do not have this information, and so modeling becomes more complex. To clarify, consider the simpler scoring system in which teams can only score touchdowns with a one point conversion. Fig. 2 shows two routes by which each team might score three touchdowns each (for a final scoreline of 21–21). There are in fact $\binom{6}{3} = 20$ routes in total. If the timings of the scoring events, and hence the particular route by which the final score came about, are known, then calculating the contribution to the likelihood function of the match is straightforward. However, if the timings are unknown, as is the case here, we must compute the probability of arriving at the final score for every single route (permutation) possible. In this simplified example, there were only two ways to score, resulting in a problem in two dimensions that we could illustrate easily. However, in reality, we have five ways of scoring for each team, resulting in a problem in ten dimensions. The complexity of the model and the computation power required therefore increase dramatically.

Before we describe our solution to this issue and the resulting likelihood computation and estimation, we turn our attention to specifying the hazards of scoring in more detail.

2.1. Hazard functions

Two hazard functions, λ_{ijk} and ρ_{ijk}, for home and away teams respectively, are needed, giving the hazard of a type k score ($k = 1$ corresponds to the most common type of score, a touchdown with a one point conversion, while $k = 2, 3, 4$ and 5 correspond to field goals, touchdowns with no conversion, touchdowns with a two point conversion, and safeties, respectively) when the home team has i points and the away team has j points. The form of hazard for the home team was:

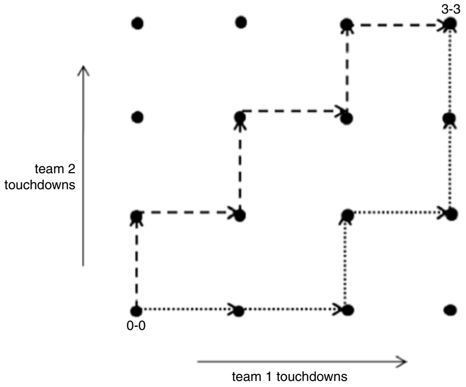

Fig. 2. Example routes to two teams scoring three touchdowns with one point conversion each. The final score for the match is 21–21.

$$\lambda_{ijk} = b_h^{\delta} \exp\left\{\left(\sum_{l=1}^{p} \beta_l x_l\right) + c_k + c_h + \gamma_1 i + \gamma_2 j\right\}, \qquad (1)$$

where the β_l are coefficients of exogenous variables x_l, in practice predictors of attacking and defensive strengths derived from previous games played by the team and the opposition respectively; c_k is a constant; and c_h is a home-team advantage coefficient. The coefficient c_k can be interpreted in much the same way as an intercept term in a regression: as a base rate of scoring intensity for score type k when all other variables are set equal to 0. When bookmaker predictions are included, b_h is the predicted number of points scored by the home team, and experimentation has led us to use $\hat{\delta} \simeq 0.95$. We first experimented with b_h being included inside the exponential and raised to a power, but estimation suggested that this power was zero, i.e., the term was a logarithm, and we therefore moved it outside the exponential. For model specifications not using the bookmaker predictions, $b_h = 1$.

The inclusion of γ_1 and γ_2, the current score variables, is intended to take into account any changes to the intensity of scoring for different current scores. For example, if a home team has scored a lot of points in the current game, the hazard of scoring again might be expected to increase, since it is likely to be a high scoring match. In our case, as a consequence of not knowing the timings of scoring events, we do not know what the current score actually is at any point in time during a match. However, because we are forecasting end-of-match scores, and estimating the probability of every possible ordering of the scoring events occurring, we can include the covariates i and j to obtain estimates of their influence on the hazards. As such, this model could be used to predict end-of-match scores at any point during a match, although we do not do so here.

Correspondingly, for the away team,

$$\rho_{ijk} = b_a^{\delta} \exp\left\{\left(\sum_{l=1}^{p} \beta_l x_l\right) + c_k + \gamma_1 j + \gamma_2 i\right\}. \qquad (2)$$

This is a proportional hazards model.

In order to reduce the correlations between fitted values of model parameters, the most common score, a touchdown with a 1-point conversion, was chosen as the 'base' method of scoring, with parameter c_1, and the c_k for other types of scores were computed as $c_k = c_1 + p_k$, where the p_k are now the parameters (we use the definition of k given above). This reparameterization also causes the iterations to converge faster. This leaves us with the following parameters to be estimated: $c_1, p_2, p_3, p_4, p_5, c_h, \gamma_1$ and γ_2, as well as the βs, the coefficients on the exogenous variables.

Of course, it would be possible to increase the complexity of the model, for example by making the hazards depend on other aspects of the state of play than the numbers of points scored thus far by each team. Also, the coefficients of the exogenous variables could vary with k, the type of score. However, as we will see, this simpler model performs well.

Given this model for the hazards, the probability of any final state can be found using the Chapman–Kolmogorov forward equation. The notation unavoidably becomes cumbersome; first, let $\mathbf{v}$ denote the 10-tuple defining the state of play (a vector of the number of each type of score accrued by each of the two teams), and let $P_{\mathbf{v}}$ be the probability of the 10-tuple score occurring. Letting n_k be the number of points accruing from the kth type of score, we now rewrite the home and away hazards of a type k score from a 10-tuple $\mathbf{v}$ as $\alpha_{\mathbf{v}}^{(k)}$, $\beta_{\mathbf{v}}^{(k)}$ respectively, and denote the state that is $\mathbf{v}$ except that the kth component of the tuple is decreased by n_k as $\mathbf{v}$, $i_k - n_k$.

Then

$$dP_{\mathbf{v}}(t)/dt = \sum_{k=1}^{5} \{\alpha_{\mathbf{v},i_k-n_k}^{(k)} P_{\mathbf{v},i_k-n_k}(t) + \beta_{\mathbf{v},i_{k+5}-n_k}^{(k)} P_{\mathbf{v},i_{k+5}-n_k}(t) - (\alpha_{\mathbf{v}}^{(k)} + \beta_{\mathbf{v}}^{(k)}) P_{\mathbf{v}}(t)\}, \qquad (3)$$

where any probabilities corresponding to impossible (negative) scores are zero.

It is straightforward to solve such equations analytically for low score final states; e.g., the probability of a zero–zero final score is

$$P = \exp(-a), \qquad (4)$$

where a is the sum of the 10 hazard functions of scoring from a zero–zero state of play. The probability that the home team makes one type k score and the other scores nothing is

$$Q = \frac{\gamma_k \{\exp(-a) - \exp(-b)\}}{b - a}, \qquad (5)$$

where γ_k is the hazard of the home team scoring from a nil–nil state of play, and b is the sum of the 10 hazards of scoring from the final state of play.

Analytic solutions for high scores are tedious to derive, although this could be mechanized using an algebra package. Another problem with using analytical solutions to compute the likelihood function is that the numerator and denominator in Eq. (5) can be very small or zero if $a \simeq b$. To avoid overflows, the computer program would have to check for this and find the solution using L'Hospital's rule. These special cases make analytical solutions unattractive as a means of computing the log-likelihood.

2.2. The likelihood function

Having specified the functional form of the hazards for each score type for both the home and away teams, in order to make predictions, we must fit the model to data. This was done by the method of maximum likelihood, and the computation of the likelihood function, although non-trivial, could be done quickly enough to make the use of this method feasible.

In order to solve Eq. (3) and calculate the log-likelihood of a set of final scores, we follow Tijms (2003, p. 166). Imagine all of the possible 10-tuples being enumerated and the probability of the state of play at time t written as a vector $\mathbf{P}$. Then, we can write the Chapman–Kolmogorov forward differential equations as

$$d\mathbf{P}(t)/dt = \mathbf{GP}, \qquad (6)$$

where $\mathbf{G}$ is the transition matrix. The Taylor series expansion of each element of $\mathbf{P}$, with substitution of Eq. (6), gives

$$\mathbf{P}(t) = \left\{ \sum_{n=0}^{\infty} t^n \mathbf{G}^n / n! \right\} \mathbf{P}(0) = \exp(\mathbf{G}t)\mathbf{P}(0). \qquad (7)$$

The likelihood of observing the outcomes of n games is the product $L = \prod_{i=1}^{n} P_{k(i),i}(t)$, where the outcome of game i is $k(i)$.

If we knew the sequence of scores and their timings, we could fit this same model. What would change is that the likelihood function, the probability of the final score, would be a product of conditional probabilities, with each being the probability of the next score occurring at some known time Δt after the previous one. The final probability would be the probability of no score occurring between the final score and the end of the match. Since we do not know the scoring sequence, the likelihood function must be summed over all possible ways that the final score could have arisen.

One limitation of the model is that the hazards of scoring are assumed to depend only on the state-of-play, and do not depend directly on time. However, if $\mathbf{G} = \mathbf{G}(u)$, the solution is

$$\mathbf{P}(t) = \exp \left\{ \int_0^t \mathbf{G}(u) du \right\} \mathbf{P}(0).$$

Hence, if each of the 5 hazards of scoring scales arbitrarily (and differently) with time, we regain Eq. (7) to a multiplicative constant, which is absorbed into the unknown constants c_k. Hence, the model is more general than it seems at first sight. This is a strength, but shows that caution would be needed if the model were to be used to predict scores partway through a game.

Thus, to evaluate Eq. (7), we need to compute the exponential of a matrix, and there are a number of ways of doing this. The most widely used method is 'scaling and squaring' (Al-Mohy & Higham, 2009; Higham, 2005). Ashi, Cummings, and Matthews (2009) review the various methods, and conclude that the scaling and squaring, and matrix diagonalization methods are the best.

However, here we seek only the one element of the exponential matrix that corresponds to the transition between the initial and final states, and the matrix is sparse. Hence, a different methodology that exploits these facts is used, and we describe it in the Appendix.

3. Data and model specification

We obtained data from the historical betting archive at sportsinsights.com for the eight seasons of NFL games between 2001 and 2008, which includes information on 2128 games. The information included the names of the home and away teams, the points scored by each team, the playing surface, and various game statistics such as rushing and passing yards for each team and how each team's score was obtained (number of touchdowns, field goals, safeties, etc.).

In addition to the game data, the dataset also included the point spread, *PS*, and the over-under, *OU*, for each game, as given by Pinnacle Sportsbook. The point spread (or betting line) can be thought of as the betting market's estimate of the difference between the points scored by each team. If a gambler bets on the favorite, he or she wins the bet if the favorite wins by more than the point spread. If a gambler bets on the underdog, he or she wins the bet if the underdog either wins, or loses by less than the point spread. If the difference in points is equal to the point spread, the gambler receives his wager back and the bet is effectively cancelled (known as a 'push'). Both teams are offered at fixed odds of 10/11 on the point spread market, so that the bookmaker includes some protection in the odds offered, the 'vigorish' (the vig). In this case, the vig is 0.045.

We should note that the evidence suggests that bookmakers do not construct the point spread as a prediction of the difference in points. Rather, Levitt (2004) demonstrates that the bookmakers take positions themselves and rely on their superior ability to predict game outcomes and a knowledge of bettor preferences, e.g. for local teams, to attain higher returns than would be the case if they 'cleared' the market (i.e., accepted an equal volume of money on either side of the point spread, so as to lock in a risk-free return, as defined by the vig). Despite Levitt's findings, there is very little evidence of any inefficiency in the betting market on the NFL (see, for example, Dare & Holland, 2004). In addition to the economics literature on the betting market, Stern (1991) showed that modeling the score difference of a game to have a mean equal to the point spread is empirically justifiable—a further testament to the efficiency of the betting market on the NFL.

Betting on the over-under is analogous to the point-spread market, where the bettor can place money on more or fewer points being scored in the game than the figure identified by the bookmaker; that is, more or less than *OU*. The over and under are also offered at fixed odds of 10/11.

We present two specifications of our model, and in each we let the intensity of scoring (the hazards) depend on covariates that represent the strengths of the two teams' offenses and defenses, which include past scoring records and past game statistics. In our second specification, in addition to specific team-based statistics, we also include the bookmakers' thoughts on the outcome of the game. Following Cain et al. (2000), for example, we recalibrate the point spread and over-under to produce a more interpretable specification. Specifically, $b_h = (PS + OU)/2$ gives an estimate of the number of points the bookmaker expects the home team to score and $b_a = (OU - PS)/2$ gives an estimate of the number of points the bookmaker expects the away team to score. In our second specification, then, b_h and b_a are used as covariates in Eqs. (1) and (2).

The model takes the two teams' offensive abilities into account using exponentially weighted averages of the team's record of previous: *points scored, first downs, rushing yards/attempt, passing yards/attempt, turnover differential, passing interceptions, percentage previous wins* and *minutes in possession*. Similarly, we include variables to represent the opposition's defensive ability using exponentially weighted averages of the opposition's *times opposition sacked, rushing yards conceded, passing yards conceded, opposition fumbles, penalties conceded* and *points conceded*. The choice of which variables to experiment with in the model is based on the findings of Zuber et al. (1985), and experimentation with the variables included in the data set.

The variables in our model are similar to the statistics used to construct power scores. Boulier and Stekler (2003) investigate the use of power scores themselves as covariates in a probit model for predicting the winner in a game. However, we use the raw, disaggregated data to build a model, rather than using the power scores themselves.

4. Results

We fitted the model to the data described above so as to provide genuine out-of-sample forecasts. The procedure adopted was as follows. We first found the best model for the five seasons from 2001–2005. We then used the first week of the 2006 season to generate the past game covariates, and produced forecasts for week 2 (we experimented with using more than one week of data for producing forecasts, but the results were largely the same, so we chose to forecast as many games as possible). The covariates based on the weighted averages of various past game statistics were updated after each week of games. At the end of each season, we refitted the model and repeated the forecasting procedure for the following season's games. As such, we do not produce forecasts for the first week of each season. One might expect that such a procedure would need to 'warm up', as the weighted averages of the various past game statistics would be less reliable in the first few weeks of the season, due to the small sample size. Furthermore, a team may have a tough (easy) start to a season, playing against strong (weak) teams, which would result in a further distortion of the past game statistics in the early stages of a season. However, as we shall see, the model performs well compared to the betting market, even with this simplification.

We fitted several specifications of the model, but, for the sake of brevity, present the results from just two models: a model based purely on past game statistics, which we refer to as *model* 1 and which is our main model; and a minimum AIC estimator, which includes the bookmaker's expected points for each team, referred to as *model* 2. The parameter estimates and model fit summaries are shown in Table 1. We show four measures of goodness-of-fit: the log-likelihood per game modeled,

Table 1
Model summaries.

Parameter	Model 1	Model 2
$c_1(TD+1)$	-0.0825	-1.9053^{***}
$p_2(\textit{field goal})$	-0.4066^{***}	-0.3943^{***}
$p_3(TD+0)$	-3.0126^{***}	-3.0226^{***}
$p_4(TD+2)$	-3.4959^{***}	-3.5174^{***}
$p_5(\textit{safety})$	-4.3911^{***}	-4.3142^{***}
$c_h(\textit{home advantage})$	0.1464^{***}	
$\gamma_x(\textit{team points so far})$	-0.0163^{***}	-0.0187^{***}
$\gamma_2(\textit{opposition team points so far})$	-0.0027^{***}	-0.0027^{***}
βs		
Expected points (from PS and OU)		0.9549^{***}
Previous points scored	0.0080^{***}	
Previous first downs	-0.0011	
Previous rushing yards/attempt	0.0338^{***}	0.0191
Passing yards/attempt	0.0219	
Times opposition sacked	0.0083	
Turnover differential	0.0027	
Passing interceptions	0.0075	-0.0273^{*}
Percentage previous wins	0.0011^{***}	
Minutes in possession	-0.0041	
Opposition previous points conceded	0.0130^{***}	
Opposition rushing yards conceded	0.0339^{**}	
Opposition passing yards conceded	0.0316^{***}	
Opposition fumbles won	0.0221^{*}	
Penalties conceded	-0.0083	
$-\ell$/game	7.402	7.888
Brier score	0.2235	0.2149
MAD	7.9913	7.7021
% correct	63.600	66.9169

* Statistically significant at the 10% level.
** Statistically significant at the 5% level.
*** Statistically significant at the 1% level.

the Brier Score, the mean absolute deviation (MAD) for the predicted points, and the percentage of game results predicted correctly.

Our main concern here is with the forecasting ability of the models, and so we comment only briefly on the parameter estimates in Table 1. The values of the estimated coefficients in Table 1 are a little difficult to interpret. However, it is interesting to note that the coefficients of the expected points are not exactly equal to 1. This may be a sign that the market forecasts contain a small bias, as was found by Dare, Gandar, Zuber, and Pavlik (2005), although we note that the values are not statistically significantly different from 1.

4.1. Forecast accuracy: predicting game results

Our first and simplest measure of goodness-of-fit is to compare the percentages of game results which are predicted correctly by the models with some benchmarks. In order to get an idea of what percentage of results a model might be expected to predict correctly, we note that in our data set, the observed home win percentage is 57.3% and the bookmaker's point spread predicted the winner correctly in 65.7% of games. We further note that Boulier and Stekler's model (Boulier & Stekler, 2003) predicted the winner correctly in 61% of games, based on power scores.

Model 1 predicts the winner of the game correctly in 63.6% of games (a hit rate higher than that of Boulier & Stekler, 2003, but lower than that attained from the point spread).

Model 2 (column 2 of Table 1) achieves a hit rate of 66.9%, a marginally higher percentage than when using only the bookmaker's point spread. The slight improvement over using only the bookmaker's point spread is thought to be a consequence of the expected points being made up of two pieces of information, *PS* and *OU*. An interesting finding from this model is that the home advantage parameter has lost statistical significance. We expect this to be because the bookmakers (and the market) are aware of the home advantage and have taken it fully into account in the *PS*. As such, the home advantage is included in the model indirectly, despite the home advantage parameter having been dropped.

We also fitted a logistic regression model with a logit link (not presented here), where the dependent variable is a binary variable indicating whether the home team won or not, and the covariates included only the weighted averages of past game statistics, not the bookmaker odds. This model predicts the winner correctly in 62.5% of the games, which is a little lower than for our model based only on past game statistics. A McNemar test reveals that the difference between the performances of our model and the logistic regression model is not statistically significant.

4.2. Forecast accuracy: predicting exact scores

Although it is still relevant, the assessment of our models' success in predicting game results is not of primary interest here. We have a model that can predict exact scores, and as such, calculating the percentage of results predicted correctly is judging only a fraction of what the model is capable of.

Assessing the goodness-of-fit for the forecast exact scores is not straightforward. The mean absolute deviation (MAD) reported in Table 1 is a good starting point. As for the percentage of correct predictions, our model including the bookmaker's odds appears to just outperform the model based only on past game statistics, and the MAICE model provides a slight improvement again. Using the *OU* and *PS* to predict exact scores results in a MAD of 7.836, which is again slightly better than our model.

Fig. 1 shows the observed and predicted frequencies of the points scored for the home team (Fig. 1(a)) and the away team (Fig. 1(b)) for our model 1. A visual inspection appears to show a more than satisfactory replication of the observed frequencies. More formally, χ^2 tests of the goodness-of-fit for the home and away distributions result in test statistics of 34.65 with 24 degrees of freedom and 21.93 with 22 degrees of freedom, respectively, and therefore, at a significance level of 0.10, we do not reject the null hypotheses that the observed and predicted distributions of home and away scores are the same.

4.3. Forecasting accuracy: comparison with the betting market

Perhaps the sternest test of a forecasting model in sports is to use it in practice to bet. Such an analysis has sometimes been used in the study of market efficiency, see for example Cain et al. (2000). Here, we do not concentrate on using our model to examine betting market efficiency, but rather use the betting market as a goodness-of-fit test for our model. Here we use model 1, a model which does not include bookmaker's information, to inform a betting strategy, purely to assess the forecast accuracy.

There are many betting strategies available to bettors, see for example Grant and Johnstone (2010). Here, we adopt a betting strategy based on the Kelly Criterion (see Kelly, 1956, or for a more thorough exposition, MacLean, Thorp, & Ziemba, 2011). The Kelly betting strategy is derived by supposing that a bettor has a logarithmic utility function for wealth. Then suppose that the bettor estimates the probability of an event (e.g., a team beating the spread) as p, whilst the bookmaker offers odds of b (where $1/(b+1)$ can be interpreted loosely as the bookmaker's implied probability of the event occurring). When betting on the over-under or point spread market in the United States, b is fixed at odds of 10/11. The expected utility (using a logarithmic utility function) from betting a fraction f of the bettor's wealth on this wager is given by $E\{u(f)\} = pln(1+bf) + (1-p)\ln(1-f)$. Using a little elementary calculus to find the maximum expected utility with respect to the fraction of the bettor's wealth wagered, f, gives the Kelly betting formula for the optimal fraction to bet:

$$f = \frac{(b+1)\,p-1}{b}. \tag{8}$$

More recent papers have discussed variations of Kelly betting strategies; for example, Kadane (2011) discusses fractional Kelly betting, whereby the fraction of the bettor's wealth that is invested, f, is deflated by a factor k, resulting in a less aggressive betting strategy. When $k = 0.5$, the resulting betting strategy is known as half-Kelly betting. Here, we adopted fractional Kelly betting, and found that using $k = 0.25$ maximized the overall return; i.e., we used quarter-Kelly betting.

In order to implement Kelly betting here, we need to estimate the probability, p; that is, the probability of the total points being more (or less) than the over-under. We do this using Monte–Carlo simulations of each match as follows. With the two teams' scores set to zero and the time elapsed also at zero, we first calculate the hazard functions given in Eqs. (1) and (2) and sum them to give the hazard of any type of score occurring. Next, we generate an exponential random number to give the time of the next event (a score). A second random number is used to simulate which type of score occurs, where the probability of a type i score, $p_i = hazard(i)/\sum_j hazard(j)$, where $hazard(i)$ is the hazard of event i occurring.

Accumulating points for the two teams, we continue this process until the time elapsed is greater than 1. We performed 10,000 simulations for each match, and from these we were able to estimate the probabilities of the total points being more or less than the over-under.

Our betting strategy for the over-under market is as follows. Let p_{under} (p_{over}) be our estimated probability of the total points being below (above) the over-under quoted by the bookmaker. If $p_{under} > 11/21$, then we placed a bet of size $0.25 \times f \times$ *current wealth* on the under. Similarly, if $p_{over} > 11/21$, then we placed a bet on the over. Wealth was accumulated during the season as bets were placed.

Table 2 summarizes the performance of model 1 (a model based purely on past game information) when the fractional Kelly betting strategy described above is used for betting in the over-under market. The returns shown are the average rate of arithmetic returns, given by $ARR = final\ wealth^{(1/n)} - 1$, where n is the total number of bets placed.

The odds offered by the bookmaker are fixed at 10/11, so that betting randomly is expected to result in a return of −0.045 on any one bet. Attaining a return higher than −0.045 would violate strong form efficiency, whilst a return higher than 0 would violate weak form efficiency. For our purposes, we consider a return of −0.045 or higher as evidence that the model is performing as well as the market. Examining the year-on-year returns enables us to identify whether any one year contributed to the positive return, whilst disaggregating by week of the season allows us to identify whether the model needs to warm up.

Looking at returns on the over-under market, the average rate of arithmetic return is positive, at 0.07% across the 568 bets placed. It appears our model does well when it is used to bet in weeks 2–9 of the season, achieving a positive return of 0.16%. Indeed, the largest return on the over-under market (0.038%) is observed in weeks 2–9 of 2008.

There is no obvious pattern in returns when disaggregated by week, suggesting that the model does not need to "warm-up" in relation to the bookmakers.

We adopt the same betting strategy for the point spread market. Table 3 gives the corresponding performance summary for the point spread market.

Table 2
Average rate of arithmetic return from fractional Kelly betting on the over-under market. *n* is the number of bets placed.

Year	All weeks		$2 \leq$ week ≤ 9		Week > 9	
	n	Return	*n*	Return	*n*	Return
2006	182	0.0014	84	−0.0008	98	0.0033
2007	208	0.0017	88	0.0019	120	0.0015
2008	178	−0.0011	79	0.0038	99	−0.0051
All	568	0.0007	251	0.0016	317	0.00002

Table 3
Average rate of arithmetic return from fractional Kelly betting on the point spread market. *n* is the number of bets placed.

Year	All weeks		$2 \leq$ week ≤ 9		Week > 9	
	n	Return	*n*	Return	*n*	Return
2006	168	0.0029	77	0.0079	91	−0.0012
2007	195	0.0021	89	0.0088	106	−0.0035
2008	193	0.00005	86	0.0022	107	−0.0017
All	556	0.0016	252	0.0062	304	−0.0022

The results for the point spread market are similar to those on the over-under market. In general, returns are positive but small, with an overall compound return of 0.01% over the 556 bets placed. However, the results differ when disaggregated by week. Our model does considerably better in weeks 2–9 than later on in the season. In other words, it appears that the model does not need to "warm-up"; rather, the market warms up.

In both the point spread and over-under markets, the returns obtained from a Kelly Criterion based betting strategy are statistically significantly higher than betting at random would achieve. As such, economists would conclude that this demonstrates a violation of strong form efficiency (that the expected return from any betting strategy is the same). From a forecasting perspective, this can be considered a testament to the performance of the model.

5. Conclusions

Modeling exact scores in sports that have more than one type of score has not been studied in the literature before. In this paper, a model for forecasting end-of-match exact scores in NFL games has been presented. The model is of interest, not only as an exercise in statistical modeling and computation, but also, given the ever-expanding number of markets available on betting exchanges, forecasting exact scores is of relevance from a practical point of view too. Using a simple set of covariates based on past game statistics, the model performs as well as bookmakers in predicting game results and exact scores. Generating superior (better than average) returns is a testament to the forecasting approach adopted, and, perhaps more importantly, has implications for the literature on betting market efficiency. Follow-on studies may concentrate on investigating the economic implications, such as betting market efficiency, of our model further. Indeed, unlike the point spread market, the over-under market on NFL has previously received little attention from economists, and our model enables both markets to be evaluated.

In addition to using the model for forecasting the final score prior to kickoff, the model can also be used to forecast whilst the game is in progress. In-play markets are readily available on betting exchanges, offering 'traders' the opportunity to bet on a wide variety of outcomes, e.g., the total number of touchdowns, or whether there will be a safety. Our model can be used to estimate the probabilities of such events. In practice, the Black–Scholes option pricing model is often used 'in reverse'. The prices of frequently-traded options are fed into the model, and the 'implied volatility' is calculated. The model, with these implied volatilities, is then used to price less-frequently traded options (e.g., Williams, 2006). Our model could also be used in this way. Once calibrated using the bookmaker's score predictions, it would then enable predictions of the probability of any outcome on which bets could be made.

Although we do not do it here, our model could be used to predict scores and outcomes during a game. Lastly, we note that this type of model can easily be adapted to forecasting in other sports, such as rugby union and rugby league, or basketball.

Acknowledgments

We are grateful to Brad Humphreys for providing some data and to our student, Asem Zarin, and his sister, for helping to collect additional data. We also thank the anonymous reviewers and the associate editor for their suggestions, which have helped to improve the paper from its original incarnation.

Appendix A. Computation

Computations were done using a purpose-written fortran95 program that called the NAG (Numerical Algorithms Group) library to invoke a function minimizer.

The first way of speeding up the computation is to adopt the method of uniformization discussed by Tijms (2003), who gives a probabilistic interpretation. Uniformization solves the problem that, in order to build up a negative exponential such as that in Eq. (5), the Taylor series terms alternate in signs. The improvement means that the number of terms to be summed is reduced greatly.

Write ν as the largest value of $\sum_{k=1}^{5}(\alpha_{\mathbf{v}}^{(k)} + \beta_{\mathbf{v}}^{(k)})$, and define $\mathbf{S} = \mathbf{G} + \nu\mathbf{I}$, where $\mathbf{I}$ is the unit matrix. Then, all elements of $\mathbf{S}$ are positive. We can compute $\exp\{\mathbf{S}(t)\}$ and then write

$$\exp(\mathbf{G}t)\mathbf{P}(0) = \exp(-\nu t)\exp\{\mathbf{S}(t)\}\mathbf{P}(0) = \sum_{n=0}^{\infty} t^n \mathbf{R}_n$$

as $\exp(-\nu\mathbf{I}) = \exp(-\nu)\mathbf{I}$. The terms in the Taylor series expansion are then positive, and the final probability is multiplied by $\exp(-\nu t)$.

The expansion is simplified slightly by setting $t = 1$ at the game end. Each term contributing to the vector $\mathbf{P}(1)$ can be obtained from the previous vector $\mathbf{R}_{n-1}$, as $\mathbf{R}_n = \mathbf{G}\mathbf{R}_{n-1}/n$. Initially, $\mathbf{R}_0 = \mathbf{P}(0)$.

The second methodology we adopt in order to speed up the computation takes advantage of the fact that the matrix $\mathbf{G}$ is sparse; from Eq. (3), each row of $\mathbf{G}$ has at most 11 non-zero elements, so that only 11 multiplications are needed per element of $\mathbf{R}_n$. Hence, the computation of $\mathbf{R}_n$ is not done by formally multiplying a matrix by a vector, but simply by accumulating the required 11 terms.

All of the elements of $\mathbf{R}_n$ that must be computed *en route* to the required element are stored in a vector. We needed 10 nested 'do-loops' to allow the state probability contributions $\mathbf{R}_n$ to be incremented. As each term in the exponential expansion allows the probability of states with a score that is one higher (of any type) to be non-zero, for the nth term, the upper limit of the do-loop for a score $n_k^{(i)}$ of type k for team i need be no higher than the minimum of $n_k^{(i)}$ and n. Also, updating $\mathbf{R}_n$ can be skipped if $\sum_{i=1}^{2}\sum_{k=1}^{5} n_k^{(i)} > n$. Implementing these considerations saves a lot of computing time. In all, 11 nested do-loops are required, but this is not a problem in modern Fortran.

In Fortran, the probabilities cannot be stored in a 10-dimensional array, because the maximum number of dimensions in Fortran is 7. However, it is feasible to use an array P_l, where the home and away team state probability components are stored as a long vector. Suppose that the final score in a game is characterized as n_k, for k being between 1 and 10. Then, any intermediate score m_k can be assigned a unique vector location number

$$l = \sum_{k=1}^{10}\left\{\prod_{i=1}^{k-1}(n_i + 1)m_k\right\}.$$

Here, l is zero for all m_k zero, and its maximum value is $l = \prod_{k=1}^{10}(n_k + 1) - 1$. We are effectively counting in a system where the base differs for each digit. The array size need only be of the order of 2000 (in fact, 2160 for this dataset).

There can be numerical problems of overflow when the function minimizer chooses unsuitable parameter values. Hence, the log-likelihood was set to a large negative number when $\nu > 10^{10}$ or when a term in $\mathbf{R}_n > 10^{200}$. In general, the log-likelihood was computed with sufficient accuracy that a function minimizer relying on function smoothness could be used. The conjugate-gradient and Newton-type minimizers used converge much faster than the Nelder–Mead simplex method, which is used to cope with rougher likelihood surfaces.

Differentiating the log-likelihood analytically with respect to model parameters could be done, but is complicated. Hence, the covariance matrix on fitted model parameters was found from the Hessian obtained using numerical differentiation.

References

Al-Mohy, A. H., & Higham, N. J. (2009). A new scaling and squaring algorithm for the matrix exponential. *SIAM Journal on Matrix Analysis and Applications*, *31*(3), 970–989.

Ashi, H. A., Cummings, L. J., & Matthews, P. C. (2009). Comparison of methods for evaluating functions of a matrix exponential. *Applied Numerical Mathematics*, *59*, 468–486.

Boulier, B. L., & Stekler, H. O. (2003). Predicting the outcomes of National Football League games. *International Journal of Forecasting*, *19*, 257–270.

Cain, M., Law, D., & Peel, D. A. (2000). Testing for statistical and market efficiency when forecast errors are non-normal: the NFL betting market revisited. *Journal of Forecasting*, *19*, 575–586.

Dare, W. H., Gandar, J., Zuber, R., & Pavlik, R. (2005). In search of the source of informed trader information in the college football betting market. *Applied Financial Economics*, *15*, 143–152.

Dare, W. H., & Holland, A. S. (2004). Efficiency in the NFL betting market: modifying and consolidating research methods. *Applied Economics*, *36*, 9–15.

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

Dixon, M. J., & Robinson, M. E. (1998). A birth process model for association football matches. *The Statistician*, *47*(3), 523–538.

Glickman, M. E., & Stern, H. S. (1998). A state-space model for National Football League scores. *Journal of the American Statistical Association*, *93*(441), 25–35.

Grant, A., & Johnstone, D. (2010). Finding profitable forecast combinations using probability scoring rules. *International Journal of Forecasting*, *26*, 498–510.

Harville, D. (1980). Predictions for National Football League games via linear-model methodology. *Journal of the American Statistical Association*, *75*(371), 516–524.

Higham, N. J. (2005). The scaling and squaring method for the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications*, *26*, 1179–1193.

Kadane, J. B. (2011). Partial-Kelly strategies and expected utility: small-edge asymptotics. *Decision Analysis*, *8*, 4–9.

Kelly, J. L., Jr. (1956). A new interpretation of information rate. *IRE Transactions on Information Theory*, *2*, 185–189.

Levitt, S. D. (2004). Why are gambling markets organized so differently from financial markets? *The Economic Journal*, *114*, 223–246.

MacLean, L. C., Thorp, E. O., & Ziemba, W. T. (2011). *The Kelly capital growth investment criterion: theory and practice. Handbook in financial economics series*. World Scientific.

McHale, I. G., & Scarf, P. A. (2011). Modelling the dependence of goals scored by opposing teams in international soccer matches. *Statistical Modelling*, *11*, 199–216.

Sauer, R. D. (1998). The economics of wagering markets. *Journal of Economic Literature*, *36*, 2021–2064.

Sauer, R. D., Brajer, V., Ferris, S. P., & Marr, M. W. (1988). Hold your bets: another look at the efficiency of the market for National Football League games. *Journal of Political Economy*, *96*, 206–213.

Stekler, H. O., Sendor, D., & Verlander, R. (2010). Issues in sports forecasting. *International Journal of Forecasting*, *26*, 606–621.

Stern, H. S. (1991). On the probability of winning a football game. *The American Statistician*, *45*(3), 179–182.

Tijms, H. C. (2003). *A first course in stochastic models*. New York: Wiley.

Williams, R. J. (2006). *Introduction to the mathematics of finance*. USA: American Mathematical Society.

Zuber, R. A., Gandar, J. M., & Bowers, B. D. (1985). Beating the spread: testing the efficiency of the gambling market for National Football League games. *Journal of Political Economy*, *93*, 800–806.

Rose D. Baker is Chair of Statistics at the University of Salford, UK. Having graduated from The University of Cambridge with a degree and Ph.D. in Theoretical Physics, Rose worked as a research scientist interested in particle physics. Rose soon became interested in statistics and has published work in many areas including medical statistics, reliability and sport.

Ian G. McHale is Reader in Business Analytics at the University of Salford, UK. Ian studied Extreme Value Statistics at the University of Manchester to gain his Ph.D., whilst his current research interests include the analysis of gambling markets and statistics in sport. Ian was co-creator of the EA Sports Player Performance Index, the official player ratings system of the English Premier League, and is Chair of the Statistics in Sports Section of the Royal Statistical Society.