Predicting match outcomes in association football using team ratings and player ratings

Halvard Arntzen [1] **and Lars Magnus Hvattum** [1]

[1] Faculty of Logistics, Molde University College, Norway

Address for correspondence: Lars Magnus Hvattum, Faculty of Logistics, Molde University College, P. O. Box 2110, N-6402 Molde, Norway.

E-mail: hvattum@himolde.no.

Phone: (+47) 71 21 42 23.

Fax: NA.

Abstract: The main goal of this paper is to compare the performance of team ratings and individual player ratings when trying to forecast match outcomes in association football. The well-known Elo rating system is used to calculate team ratings, whereas a variant of plus-minus ratings is used to rate individual players. For prediction purposes, two covariates are introduced. The first represents the pre-match difference in Elo-ratings of the two teams competing, while the second is the average difference in individual ratings for the players in the starting line-ups of the two teams. Two different statistical models are used to generate forecasts. The first type is an ordered logit regression (OLR) model that directly outputs probabilities for each of the three possible match outcomes, namely home win, draw, and away win. The second type is based on competing risk modelling, and involves the estimation

of scoring rates for the two competing teams. These scoring rates are used to derive match outcome probabilities using discrete event simulation. Both types of models can be used to generate pre-game forecasts, whereas the competing risk models can also be used for in-game predictions. Computational experiments indicate that there is no statistical difference in the prediction quality for pre-game forecasts between the OLR models and the competing risk models. It is also found that team ratings and player ratings perform about equally well when predicting match outcomes. However, forecasts made when using both team ratings and player ratings as covariates are significantly better than those based on only one of the ratings.

Key words: Elo rating; competing risk; ordered logit regression; plus-minus rating; survival analysis

1 Introduction

Association football is one of the most popular sports, with a huge fan base, attracting the attention of media and entertainment platforms. Forecasts of match outcomes can provide useful inputs to decision makers within the sport, as well as valuable information to pundits and journalists. In particular, in-game predictions can be useful both for entertainment purposes and for the participants in a match that needs to make risk assessments. However, few models for making in-game predictions have been evaluated in the scientific literature.

When predicting match outcomes, it is common to first evaluate the relative playing strength of the teams involved in the match. There is a multitude of alternative

methods in the scientific literature that discusses how to rate teams based on their observed performance. However, a more recent line of research is to use the increasing amount of data available to directly rate individual players, which then allows team ratings to be derived based on the players involved in the match. Existing research has to a very little extent tried to evaluate whether player ratings are more informative than direct team ratings when forecasting match outcomes. In particular, although the use of player ratings may involve more information in the predictions, such as taking into account the actual starting line-ups for each team, it may also be the case that the player ratings are more noisy than the direct team ratings, and therefore will not lead to improved predictions.

This paper makes the following contributions: First, it compares two types of statistical models for forecasting match outcomes. One of these, an ordered logit regression (OLR) model, is known from the literature to be a reasonable method for generating pre-game forecasts. The second is a novel competing risk model, based on survival analysis, which can also be used for in-game predictions. The paper evaluates these two models in terms of their ability to generate pre-game forecasts, and also evaluates two versions of the competing risk model in terms of generating in-game forecasts. Second, the paper compares the use of individual player ratings and direct team ratings as a basis to forecast match outcomes.

The remainder of this paper is structured as follows. Section 2 provides a review of relevant literature. Then, Section 3 describes the experimental setup. This involves a description of the data used for calculations, and the two methods for predicting match results are explained: the OLR model and the novel competing risk model. The system for deriving team ratings, and the system for deriving player

ratings are described in the supplementary material for this manuscript (available at http://www.statmod.org/smij/archive.html). Section 4 contains the results and discussions of the key findings, while Section 5 presents some concluding remarks.

2 Extant literature

This section provides a short overview of literature relevant to the comparison of team ratings and player ratings as predictors of match outcomes in association football. The section has three parts, first discussing literature on team ratings in association football, then looking at literature on player ratings in association football, and finally discussing the most relevant prediction models employed when trying to predict match outcomes on the basis of ratings.

2.1 Team ratings

Stefani and Pollard (2007) presented a survey of rating systems for different forms of football. Two systems were discussed for association football: the system applied to calculate official FIFA ratings at the time, and a system based on Elo ratings, originally devised for rating chess players as described by Elo (1978). A different adaptation of the Elo system to association football was suggested by Hvattum and Arntzen (2010), where it was shown that taking the winning margin into account, in what was labeled a goal-based Elo rating, helped to improve the predictive power of the ratings.

Lasek et al. (2013) found that a version of Elo ratings had better predictive capabilities than several other rating systems when applied to international association

football matches. The other rating systems included the FIFA ratings at the time, least squares ratings, network based ratings, and Markovian ratings. Van Eetvelde and Ley (2019) described the FIFA ranking systems for men and women, least squares ratings, Thurstone-Mosteller and Bradley-Terry models, and Elo ratings, but without discussing the predictive abilities of these rating systems for association football. Wunderlich and Memmert (2018) showed that the goal-based Elo ratings of Hvattum and Arntzen (2010) could be improved by updating the ratings based on the observed pre-game odds from the betting market, rather than the actual result of the match. However, a weakness from that study, which is evident from the supporting information, is that the Elo ratings appear not to have been initialized using a proper bootstrapping procedure.

Constantinou and Fenton (2013) developed an alternative to Elo ratings called pi-ratings, where teams are assigned to both a home rating and an away rating. Used on data from the English Premier League over five seasons, the proposed pi-rating performed better than Elo ratings. Van Haaren and Davis (2015) used both Elo ratings and pi-ratings in the prediction of final tables of domestic leagues. While Elo ratings were found to perform particularly well, they noted that pi-ratings excelled when only match results from the previous season were available.

Robberechts and Davis (2019) adapted a rating system called the offense defense model to association football. This system creates two strength measures for each team, representing offensive and defensive strength, respectively. The authors compared this rating system to goal-based Elo ratings, and found that the Elo ratings worked better when predicting matches from a large data set. They also found that the Elo ratings outperformed predictions based on an extended version of pi-ratings as

presented by Constantinou (2019). Recently, Lasek (2019) proposed several promising rating systems for teams based on the update step from stochastic gradient descent as applied to OLR and Poisson models, respectively.

2.2 Player ratings

The first rating system for soccer players appears to have been proposed by McHale et al. (2012). Their rating system is based on six subindices. The first subindex requires relatively detailed data from each soccer match, regarding such elements as passes, tackles, crosses, dribbles, block, clearances, and yellow and red cards. The other five subindices are based on the number of minutes played (in two different ways), the number of goals scored, the number of assists, and the number of clean sheets, respectively. The final rating is a weighted sum of the six subindices. For several years this rating system was published online as the official rating system of the English Premiere League.

The rating system of McHale et al. (2012) mixes two different philosophies for deriving player ratings. Their first subindex follows a bottom-up philosophy to estimate the contribution of players to match outcomes: that is, the contribution of a player is calculated based on each separate contribution made during the match. For instance, if a player performs a successful dribble twenty meters inside the opponent's half, that player will be credited with the average value that such an action has towards the ultimate goal of the team (such as scoring a goal, or winning the match). This approach requires very detailed data from each match used as a basis for calculating the player rating.

An alternative to the bottom-up approach is to create player ratings with a top-down

philosophy. This type of ratings can be calculated when detailed information on actions during a match is unavailable, as it only relies on the knowledge of which players are playing at any time of the match and on when any goals are scored. The advantage of this is that less detailed data is required for the calculation of ratings. However, it also means that it does not make sense to produce ratings from a single match, as there is simply not enough data to differentiate the players involved based on the outcome of the match.

Sæbø and Hvattum (2015) proposed a top-down rating model for soccer players, using a regression model capturing the performance of players relative to their team mates and the opposition. The model was based on similar models, referred to as adjusted plus-minus ratings, from basketball and ice hockey. Each match is split into segments where the players on the pitch are constant. These segments are then used as observations in a linear regression model where the dependent variable corresponds to the observed goal difference in the segment. The model is adjusted for home field advantage and players being sent off, and older observations are down-weighed so that newer observations are more important for the final player ratings derived. To estimate model coefficients, Tikhonov regularization, also known as ridge regression, was used. Sæbø and Hvattum (2019) refined the model slightly by splitting the individual player rating into an individual component and a tournament component. Later, the model was developed further, as presented by Pantuso and Hvattum (2019). Gelade and Hvattum (2020) analyzed the resulting player ratings and found that the additional use of event-based key performance indicators for individual players could only marginally improve predictions formed on the basis of the player ratings.

Sittl and Warnke (2016) created a model similar to an adjusted plus-minus model,

where each observation corresponded to the goal margin of a single player in a single match, weighted for minutes played and with additional covariates for other players on the pitch, a fixed team effect, a fixed coach effect, time varying characteristics such as league and season effects, a home team indicator, the number of dismissals, and both the age and the age squared of the player. Vilain and Kolkovsky (2016) presented another variation of the plus-minus idea. They defined observations as entire matches, as opposed to the segments of Sæbø and Hvattum (2015) or the combination of matches and players of Sittl and Warnke (2016). The major difference, however, is the type of model built. Instead of relying on linear regression, the dependent variable is taken as a categorical variable and an ordered probit regression (OPR) model is built using maximum likelihood estimation. Regularization terms are added to the log-likelihood function, to deal with the issues of collinearity. Another interesting point of the ratings developed by Vilain and Kolkovsky (2016) is that the contribution of each player is split into a defensive and an offensive part.

Schultze and Wellbrock (2018) calculated weighted plus-minus ratings without using a regression model. Instead, the ratings are calculated using a formula that is applied for each minute played of a match. The formula included two novelties: First, bookmaker odds were used to find an expected result for a match, and the ratings reflect how the players contribute to a result relative to this expectation. Second, goals are weighted differently depending on whether they are immediately changing the outcome of a match, for example going from a won to a drawn game, or whether they are simply changing the winning margin. Hvattum (2019) presented a comprehensive overview of literature on plus-minus ratings for association football and other team sports.

Other contributions have provided less complete rating systems, either being able to

rate only a subset of players, or focusing on particular aspects of the game, such as passing. Tiedemann et al. (2010) used data envelopment analysis to evaluate the performance of outfield players in the German 1. Bundesliga. They used playing time as the only input, and considered the number of goals scored, the number of assists, the pass completion ratio, and the ratio of successful tackles as outputs. Duch et al. (2010) created networks based on passes between players as well as shots, and used centrality measures for networks as a means to find the importance of players. Results were obtained for a data set consisting of the 2008 Euro Cup tournament. The authors left out goal keepers from parts of their analysis, which may suggest that the approach works best for outfield players. Brooks et al. (2016) developed a metric to rank soccer players based entirely on passes. Data from the 2012/2013 season of the Spanish top division was used to test the metric, and the resulting player ratings were found by the authors to be consistent with general perceptions of offensive ability. Szczepański and McHale (2016) created a model to assess the passing ability of players. Fitted on data from the 2006/2007 season of the English top division, the model was significantly better at predicting player's completion rates for the next season than just using the previous season's completion rates.

Finally, there are player ratings based on subjective assessments. Peeters (2018) investigated whether subjective player valuations from a popular website were useful to predict outcomes of international matches, and found that the subjective valuations, averaged over the team as a whole, provided better predictions than both the official FIFA ratings and a version of Elo ratings. Cotta et al. (2016) pointed out that a popular video game series includes evaluations of many players for up to 34 attributes, using inputs from scouts hired specifically for this purpose. However, they did not test whether these evaluations were useful for predicting match outcomes.

Kharrat (2016) showed that such ratings from video games can be useful for prediction purposes, but also showed that they could be further improved by incorporating information from the objective rating system developed by McHale et al. (2012). In another test, Lasek (2019) found that aggregated ratings from the same video game series performed better than Elo ratings when predicting match outcomes in domestic leagues.

2.3 Prediction models for match outcomes

Statistical models for match outcomes in association football can be divided into three categories. Match outcomes can either be 1) modelled directly as home win, draw, or away win; 2) represented as the difference between the goals scored by the two teams involved; or 3) represented by the distribution of the number of goals scored by each team, which can then be translated into overall win probabilities. The first and latter category are by far most popular.

Discrete choice models are often used to model match outcomes directly. An early contribution to this field was made by Koning (2000), who presented an OPR model. The representation as the difference between the goals scored was proposed by Karlis and Ntzoufras (2009) in a Bayesian model based on the Skellam distribution. The technique of modelling the goals scored by each team has the longest history, starting with Maher (1982), and with important contributions from Dixon and Coles (1997). The most typical assumption in this type of modelling is that goals scored follow a Poisson distribution.

Goddard (2005) compared OPR models with bivariate Poisson regression models, with either results-based or goal-based covariates. The objective was to predict the

outcome of football matches, encoded as home wins, draws, and away wins. Results showed that the differences between the models were small, but that an OPR model with goal-based covariates dominated the other variants for most seasons in the data. Hvattum and Arntzen (2010) used the difference of Elo ratings for teams as a single covariate in an OLR model, and found that this worked better than the goal-based covariates used by Goddard (2005), at least when applied to relatively small data sets. Robberechts and Davis (2019) compared an OLR model and a bivariate Poisson model when using a single covariate based on Elo ratings, and reported a better prediction quality for the OLR model. Hvattum (2017) highlighted some drawbacks of ordinal regression models in that they fail to properly incorporate information relevant for the prediction of draws. Empirical results suggested that this has little practical consequence, as none of the commonly used covariates in such models seem to directly affect the predicted proportion of draws. It was found, however, that multinomial logit regression models can be used if such covariates are devised.

Some models that are used to predict match outcomes can produce team ratings as a part of the estimation process. Ley et al. (2019) investigated a family of such models, and found that a bivariate Poisson model with one strength parameter per team provided the best predictions. Boshnakov et al. (2017) presented results indicating that a Weibull count model provided a better fit than a Poisson distribution. For these models the strength parameters, which correspond to ratings, are assumed to be static. The models emphasize more recent results by down-weighting observations according to their age, as first suggested by Dixon and Coles (1997). Other models allow ratings to change dynamically, so that there is a strength parameter for each team and each time instant. Rue and Salvesen (2000) proposed a Bayesian model that allowed time-dependent skill estimates for teams in a league, which was also the

case for Crowder et al. (2002). Owen (2011) showed that calculating dynamic ratings using a dynamic generalized linear model provided better predictions of future results than a non-dynamic form of the model. Koopman and Lit (2019) compared models with static ratings and models with dynamic ratings for each of the three categories of match outcomes: using a bivariate Poisson model for goals scored by each team, a model based on the Skellam distribution for the difference in goals scored, and an OPR model for the match outcome. They concluded that dynamic models with time-varying parameters show better forecasting performance than models with static parameters.

Recently, the use of machine learning methods outside of statistical regression models has increased. Schauberger and Groll (2018) applied random forests to predict matches in international association football tournaments. They considered the FIFA ranking of teams, but no other ratings as such. Later, Groll et al. (2019) incorporated ratings from the bivariate Poisson model investigated by Ley et al. (2019) in a hybrid random forest method yielding improved predictions. Baboota and Kaur (2019) found that gradient boosting outperformed random forests in terms of prediction quality for matches from the English Premier League. Some of the features included in their models came from player ratings of the video game series FIFA.

The models discussed above all consider a match as an atomic unit: a prediction is made when the match starts, and the result is observed at the end. Few models have been made that incorporate the dynamics of a match itself, taking into account the timing of the goals scored or other events that can happen during a match to influence its outcome. Dixon and Robinson (1998) applied techniques from survival analysis to analyse scoring rates of teams as a function of the number of minutes played of

a match. They devised a birth-process model where scoring rates are allowed to vary based on both the amount of time played and the current score. The model simultaneously estimates both attack and defense parameters for each team in the data set. Volf (2009) also considered that scoring rates are varying during a match, and used a semi-parametric multiplicative regression model for the scoring intensities. The model was applied to data from matches of the 2006 World Championship.

Titman et al. (2015) presented a multivariate counting process formulation, where both goals and bookings are assumed to follow a Weibull distribution. In their model, relative team ability was derived from match outcome odds provided by several bookmakers. The modelling framework appears flexible and suitable for realtime prediction of match outcomes, as well as goal differences and bookings. Robberechts et al. (2019) considered a Bayesian model for in-game win probabilities, with Elo ratings to represent team strengths. Although the consideration of the timing of events within matches could in principle provide better pre-game forecasting methods than simply relying on the observed final result, the literature on association football match outcome forecasting does not currently present any direct comparisons to support such a claim.

3 Experimental setup

This section first describes the available data and how this is used in the experiments. Then, an overview is given of the team ratings calculated using the Elo system and the player ratings calculated using an adjusted plus-minus system. Two methods for predicting match outcomes are presented next: an OLR model and a competing risk

model. Finally, the evaluation criteria used to assess the resulting prediction models are explained.

3.1 Data

The data used in the reported experiments consist of matches from the four divisions of the English league system: Premier League, Championship, League One, and League Two. In addition, matches from the English League Cup are included. Matches from ten seasons, from 2009/2010 to 2018/2019 are considered, and each match has information about the date of playing, the teams involved, the players that started the match, and the time for each of the goals scored as well as any substitutions made and red cards given. Due to data quality, some matches are discarded. In addition, the data set does not include all matches from the earliest rounds of the League Cup, and all matches played on neutral ground are removed. This results in a total of 21,129 matches.

The matches are split into different sets. Seasons 2009/2010 through 2013/2014 are used only for initial calculations of ratings. Then, seasons 2014/2015 through 2016/2017 provide initial observations for the statistical models, where ratings are updated before each new day with matches played. Finally, seasons 2017/2018 and 2018/2019 are used to evaluate the predictions created by the statistical models. Before each day of matches, ratings are updated and the statistical models are re-estimated using all current observations.

All the data used in the following are available from various online sources. Kharrat (2016) discussed how to obtain the type of data required. While the data are easily found online, there can be challenges with respect to cleaning data or combining

different data sources.

3.2 Ratings

The team ratings examined in this paper are based on the rating system of Elo (1978) for chess players, with adjustments for association football by Hvattum and Arntzen (2010). The considered ratings for individual players are based on the plus-minus ratings developed for association football by Sæbø and Hvattum (2015, 2019) and later refined by Pantuso and Hvattum (2019) and analyzed by Gelade and Hvattum (2020).

Details of the calculation of these ratings are given in supplementary material for this manuscript (available at `http://www.statmod.org/smij/archive.html`). The two rating systems are used to generate two covariates: x_{Elo} as the difference between the team rating of the home team and the away team, and x_{PM} as the difference between the average ratings of players in the starting line-up of the home team and the away team.

3.3 Ordered logit regression

The first model used to generate predictions in this work is an OLR model. Dobson and Goddard (2001) presented an early description of such a model in the context of association football, while Greene (2012) provides a general exposition of the technique. The outcome of a football match is encoded as an ordinal dependent variable, with $y = 1$ representing a home win, $y = 2$ representing a draw, and $y = 3$ representing an away win. Given a vector of covariates x, the probability of outcome

$j \in \{1, 2, 3\}$ is stated as

$$\pi_j(x) = F(-\theta_j - \beta x) - F(-\theta_{j-1} - \beta x) \tag{3.1}$$

where $F(z)$ is taken as the logit link function

$$F(z) = \frac{1}{1 + e^{-z}}$$

and where $\theta_0 = \infty$, $\theta_3 = -\infty$, $F(-\infty) = 0$, and $F(\infty) = 1$. The parameters that must be estimated are the coefficients of the covariates, β, as well as θ_1 and θ_2 which are used to differentiate between the three ordinal values of the dependent variable.

Assume that data consists of n observations and that the dependent variable y takes values from $\{1, 2, 3\}$. Let d_{ij} be indicator variables such that $d_{ij} = 1$ if observation i provides $y_i = j$. The likelihood function can now be written as

$$L = \prod_{i=1}^{n} \prod_{j=1}^{3} \pi_j(x_i)^{d_{ij}}$$

which in the case of the logit link function leads to a log-likelihood of

$$l(\beta, \theta) = \sum_{i=1}^{n} \sum_{j=1}^{3} d_{ij} \ln(F(-\theta_j - \beta x_i) - F(-\theta_{j-1} - \beta x_i)).$$

Given a data set of observations, maximum likelihood estimation can be used to find the parameters of the model. As the likelihood function is convex, Newton's method can be applied to maximize the likelihood once the gradient and Hessian of the function has been derived. Equation (3.1) can then be used to directly predict

match outcomes. In this paper, two covariates are considered in the OLR model: x_{Elo} and x_{PM}, reflecting the difference in team ratings and player ratings for the two teams involved in a match.

3.4 Competing risk

The second model used to generate predictions in this work is based on survival analysis, and has similarities to the framework presented by Titman et al. (2015). The model as presented here was derived from the work of Kalbfleisch and Prentice (2002). Similar to Poisson regression models, the goal is to find scoring rates of teams as a function of selected covariates. This is achieved by making observations of the time until goals are scored.

In particular, assume that an association football match is being observed. The match is split into intervals, each terminated either by the occurrence of some event (e.g. home goal, away goal, home team red card) or by the end of match. We consider each interval a separate *observation*. In line with the terminology of competing risk models, we say that the observation is ended by a *cause* or by reaching the end of the game. If an observation is terminated because the match ends, thus not having an observed cause, the observation is *censored*.

The random variable giving the cause for ending an observation is denoted by C, and the random variable giving the duration until the cause is T. This means that one will either observe a cause $C = c$ after a duration $T = t$, or the observation will end providing the single information that $T > t$, where t is the length of the observation. The probability distributions for C and T are generally depending on a vector of covariates x, specific to the observation.

In the following we outline the ideas for deducing a likelihood function. Further details are given by Kalbfleisch and Prentice (2002). In competing risk modeling, the scoring rates correspond to cause-specific hazard rates,

$$\lambda_c(t,x) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P[\ t < T < t + \Delta t, C = c \mid T > t, x\ ]\ .$$

The overall hazard rate is

$$\lambda(t,x) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P[\ t < T < t + \Delta t \mid T > t, x\ ] = \sum_c \lambda_c(t,x)\ ,$$

assuming two causes cannot occur simultaneously. Let $S(t,x) = P[\ T > t \mid x\ ]$ and let $f(t,x) = -S'(t,x)$ be the density of T for given x. Under reasonable regularity assumptions, we get

$$\lambda(t,x) = \frac{f(t,x)}{S(t,x)} = -\frac{S'(t,x)}{S(t,x)}\ .$$

Integrate this identity over $[0,t]$, then exponentiate, to get

$$\ln(S(t,x)) = -\int_0^t \lambda(s,t)ds\ ,$$

$$S(t,x) = \exp(-\int_0^t \lambda(s,t)ds) = \prod_c e^{-\Lambda_c(t,x)}\ , \tag{3.2}$$

where

$$\Lambda_c(t,x) = \int_0^t \lambda_c(s,x)ds.$$

Additionally, we will consider the *cause-specific* density $f_c(t,x)$ representing component c of the joint density of (C,T). We then have

$$f_c(t,x) = \lim_{\Delta t \to 0} P[\ t < T < t + \Delta t, C = c \mid x\ ] = \lambda_c(t,x)S(t,x)\ , \tag{3.3}$$

where the latter identity is obtained by conditioning on $T > t$.

Now, turning to the likelihood, assume that the data consists of n observations $O_i = (t_i, d_i, c_i, x_i)$, where t_i is the observed duration, $d_i = 1$ if a cause was observed at time t_i and $d_i = 0$ otherwise, $c_i \in \{1, \ldots, m\}$ is the cause observed or $c_i = 0$ if $d_i = 0$, and x_i is the covariates vector. If $d_i = 1$ the contribution to the likelihood is $f_{c_i}(t_i, x_i)$. If $d_i = 0$ we only know that $T > t_i$, so the contribution to the likelihood is $S(t_i, x_i)$. From (3.2) and (3.3) the likelihood function can be written

$$L = \prod_{i=1}^{n} \lambda_{c_i}(t_i, x_i)^{d_i} \prod_{c=1}^{m} e^{-\Lambda_c(t_i, x_i)} \, .$$

By introducing indicator variables d_{ic}, such that $d_{ic} = 1$ if $c_i = c$ and $d_{ic} = 0$ otherwise, the log-likelihood can be written as

$$l = \ln(L) = \sum_{c=1}^{m} \sum_{i=1}^{n} \left(\, d_{ic} \ln(\lambda_c(t_i, x_i)) - \Lambda_c(t_i, x_i) \, \right) .$$

In the particular model derived here, the time until a cause is assumed to be exponentially distributed. This is consistent with the body of literature on forecasting of match results in association football using Poisson regression, although other distributions are possible, such as the Weibull distribution used by Boshnakov et al. (2017). The competing risk model further assumes that the causes are independent, which is sometimes contested in the setting of association football. However, Groll et al. (2015) argued that the independence assumption can be warranted when appropriate covariates are included. The assumption of exponentially distributed times until a

cause leads to the parametric form

$$\lambda_c(t, x) = \alpha_c e^{x\beta_c}$$

where x is a row vector of values for the covariates, β_c is a column vector with coefficients of the covariates for cause c, and α_c is a parameter for cause c. In general, the hazard rates may depend on the time t, but this is not the case when using an exponential distribution.

In the particular case of using an exponential distribution, the log-likelihood becomes

$$l(\alpha, \beta) = \sum_{c=1}^{m} \sum_{i=1}^{n} \left(d_{ic} \ln(\alpha_c) + d_{ic} x_i \beta_c - t_i \alpha_c e^{x_i \beta_c} \right).$$

As for the OLR model, the likelihood function is convex, and Newton's method can be used to estimate the coefficients. While the OLR model directly provides match outcome predictions, the competing risk model only estimates scoring rates as functions of the covariates. To obtain match outcome predictions, discrete event simulation is used. To this end, the inverse transformation method is used to sample the time to the next event, and the corresponding cause is drawn randomly based on the relative hazard rates for the different causes. Each simulation provides one possible result for a match, and to estimate a probability distribution over all possible results, the simulation is repeated 30,000 times per match.

Two variants of the competing risk model are investigated in this paper. In the simple variant, only scoring rates for the two teams are estimated, using only two covariates: x_{Elo} and x_{PM}. An extended model additionally includes red cards for either team as

causes, thus also estimating the rates at which the two teams accrue red cards. In this latter model, covariates are included to represent the number of red cards already given out to the respective teams, and also to reflect the current game state, that is, whether either of the teams is in the lead.

3.5 Evaluation

To evaluate the predictions produced by different models, different metrics can be used. Two popular choices for evaluating match outcome forecasts include the informational loss and the quadratic loss, as described by Witten et al. (2011). Let p_j be the probability for outcome j of a match, with $j \in \{1, 2, 3\}$ as there are three possible match results, and let $d_j = 1$ if the match ended with outcome j and $d_j = 0$ otherwise. The informational loss can then be stated as

$$L^I = -\log_2 \left(\sum_{j=1}^{3} d_j p_j \right),$$

whereas the quadratic loss can be stated as

$$L^Q = \sum_{j=1}^{3} (d_j - p_j)^2.$$

These loss functions are also known as the ignorance score and the Brier score, respectively. Wheatcroft (2019) shows that, while both are proper scores, they can differ in terms of how quickly they are able to identify the best forecasting systems. He also reports on computational experiments indicating that both scores outperform the ranked probability score, which is also commonly used for evaluating match outcome forecasts in association football.

To apply the loss functions, their values are calculated for each model for each match of the 2017/2018 and 2018/2019 seasons. Then, the average loss is considered over all matches, or over specific subsets of matches, to indicate the performance of a prediction model. To determine whether a given set of forecasts outperforms another set, paired t-tests can be applied under the null-hypothesis that the loss of predictions is identical for the two sets of forecasts.

A secondary method to evaluate the resulting prediction models is to look at whether the resulting regression coefficients are reasonable and different from zero with statistical significance. To this end, for the main models, the regression coefficients and their standard errors are presented, together with the corresponding p-values. The latter are calculated using normal distributions, based on asymptotic likelihood theory described by Greene (2012).

4 Results and discussion

This section first reports on the top ranked teams and players based on the rating systems used. Then, we present tests using the Elo ratings for teams and the plus-minus ratings for individual players to predict match outcomes. To this end, both the OLR model and the competing risk model are used. Finally, a more elaborate competing risk model is analyzed, also taking a look at its use for in-game predictions.

4.1 Top teams and players

To illustrate the ability of the team ratings and the player ratings to differentiate between teams and players of unequal strength, lists of the top ten ranked teams and

players are shown in Table 1 and Table 2, respectively. As the lists are created using data from the English league system and the English League Cup only, spanning seasons 2009/2010 through 2018/2019, the lists are dominated by teams and players from the best performing teams of the Premier League in the 2018/2019 season.

Table 1: Top ten highest rated teams according to Elo ratings at the end of the 2018/2019 season.

Rank	Team	Elo
1	Manchester City	750.1
2	Liverpool	694.2
3	Tottenham Hotspur	546.4
4	Chelsea	514.0
5	Arsenal	484.8
6	Manchester United	477.2
7	Everton	414.4
8	Leicester City	383.5
9	Crystal Palace	365.2
10	Wolverhampton Wanderers	360.0

Table 2: Top ten highest rated players according to plus-minus ratings at the end of the 2018/2019 season.

Rank	Player	Team	Minutes	Rating
1	Ederson	Man. City	6,870	0.341
2	Kyle Andrew Walker	Man. City	28,299	0.328
3	Aymeric Laporte	Man. City	3,873	0.328
4	Leroy Sane	Man. City	6,746	0.308
5	Bernardo Silva	Man. City	5,024	0.306
6	Benjamin Mendy	Man. City	1,314	0.300
7	Ilkay Gundogan	Man. City	4,973	0.291
8	Kevin de Bruyne	Man. City	10,458	0.288
9	Sadio Mane	Liverpool	12,960	0.288
10	Gabriel Jesus	Man. City	4,036	0.279

4.2 Team ratings versus player ratings

This test considers the two types of prediction models, the OLR model and the competing risk model. For each, three combinations of covariates are tested: either

using only the Elo ratings with x_{Elo}, using only the plus-minus ratings with x_{PM}, or using both of them simultaneously. To compare the six resulting variants, we use the informational loss L^I and the quadratic loss L^2.

Table 3: Informational and quadratic loss over $n = 4225$ matches using two different prediction models and three different sets of covariates.

	OLR		Competing risk	
Covariates	L^I	L^2	L^I	L^2
Elo	1.4781	0.6150	1.4782	0.6151
PM	1.4760	0.6141	1.4757	0.6139
Elo, PM	1.4721	0.6122	1.4721	0.6121

Overall results are shown in Table 3. The informational loss and the quadratic loss are very similar for both prediction models: Using a two-sided paired t-test, it can be concluded that the results for the OLR models and the competing risk models are not statistically different, for either of the three sets of covariates. Comparing the use of team ratings and player ratings, it can be observed that for both models and both evaluation metrics, the player ratings provide better results. However, these differences are not statistically significant for any combination of model and metric.

The best results are seemingly obtained when using both Elo ratings and plus-minus ratings as separate covariates in the same model. The t-tests confirm this: the differences between the predictions using only Elo ratings and the models using both ratings are statistically significant with p-values below 0.001, for both informational and quadratic loss. The improvements observed when comparing the use of both ratings and using only the plus-minus ratings are also statistically significant, but only at a level of 0.05 and not at a level of 0.01. The differences in p-values in these latter tests are perhaps an indication that plus-minus ratings perform somewhat better than Elo ratings.

To provide a point of reference for the loss values in Table 3, odds data were downloaded from https://www.football-data.co.uk. Pre-game match odds were converted into probabilities by taking the inverse of the decimal odds and then removing overround using the basic normalization described by Štrumbelj (2014). The odds data do not include cup matches and have a few missing data points, so the following comparison was made on the remaining $n = 4016$ matches when considering the bookmaker Pinnacle and their odds provided some time before the match (PS), as well as their closing lines (PSC). The informational loss of PS and PSC is 1.467 and 1.462, respectively, whereas the quadratic loss is and 0.610 and 0.609. The corresponding values for the OLR model with Elo and PM on the same data set are 1.479 and 0.616, and both are higher than those of PSC with p-values below 0.005.

Table 4 presents the regression coefficients from the OLR model, estimated using all seasons of data. Negative β coefficients imply that the probability of the home team winning increases when the value of the corresponding covariate increases. Table 5 similarly shows coefficients for the competing risk model, where $c = 1$ corresponds to home team goals and $c = 2$ corresponds to away team goals. For $c = 1$, positive β coefficients means that a higher value of the covariate increases the scoring rate of the home team, and for $c = 2$, negative β coefficients means higher values of the covariate decreases the scoring rate of the away team. All coefficients for both models are statistically different from 0 with p-values less than 0.00001.

Table 4: Regression coefficients for the OLR model estimated on data from all seasons.

Variable	Coefficient	Std. error	p-value
θ_1	−0.302	0.020	0.000
θ_2	0.870	0.022	0.000
β_{Elo}	−0.700	0.082	0.000
β_{PM}	−0.002	0.000	0.000

Table 5: Regression coefficients for the competing risk model estimated on data from all seasons.

Variable	Coefficient	Std. error	p-value
α_1	0.015	0.000	0.000
$\beta_{1,Elo}$	0.236	0.035	0.000
$\beta_{1,PM}$	0.001	0.000	0.000
α_2	0.012	0.000	0.000
$\beta_{2,Elo}$	−0.273	0.039	0.000
$\beta_{2,PM}$	−0.001	0.000	0.000

The results above show that team ratings and player ratings are complementary, even though both types of ratings work well as a separate means to determine match outcome probabilities. One may also conclude that both OLR models and competing risk models are equally suitable for generating pre-game match outcome probabilities, or at least the differences are very small. It can also be seen that both loss measures, informational and quadratic, lead to the same evaluations when comparing the different models and covariates.

4.3 Variations of competing risk model

While the OLR model is only suitable for pre-game predictions, the competing risk model can take into account the current situation in an on-going match and through simulations obtain an updated prediction for the final outcome. Although the basic competing risk model, with x_{PM} and x_{Elo} as the only covariates, can be used for in-game predictions, an extended model is also considered. The extended model includes additional covariates, as well as two additional causes, with $c = 3$ corresponding to an observation being terminated with a red card for the home team and $c = 4$ to an observation terminated with a red card for the away team. The additional covariates are indicator variables for the current goal difference: x_{DG} to indicate an equal number

Table 6: Regression coefficients for the extended competing risk model estimated on data from all seasons.

Variable	Coeff.	Std. err.	p-value
α_1	0.014	0.000	0.000
$\beta_{1,Elo}$	0.216	0.034	0.000
$\beta_{1,PM}$	0.001	0.000	0.000
$\beta_{1,DG}$	0.212	0.028	0.000
$\beta_{1,H+1}$	0.106	0.023	0.000
$\beta_{1,H+2}$	0.132	0.028	0.000
$\beta_{1,A+1}$	0.164	0.025	0.000
$\beta_{1,A+2}$	0.307	0.035	0.000
$\beta_{1,HR}$	−0.321	0.066	0.000
$\beta_{1,AR}$	0.567	0.039	0.000
α_2	0.011	0.000	0.000
$\beta_{2,Elo}$	−0.299	0.038	0.000
$\beta_{2,PM}$	−0.001	0.000	0.000
$\beta_{2,DG}$	0.228	0.031	0.000
$\beta_{2,H+1}$	0.208	0.025	0.000
$\beta_{2,H+2}$	0.295	0.033	0.000
$\beta_{2,A+1}$	0.118	0.027	0.000
$\beta_{2,A+2}$	0.196	0.037	0.000
$\beta_{2,HR}$	0.722	0.046	0.000
$\beta_{2,AR}$	−0.491	0.068	0.000
...			

Variable	Coeff.	Std. err.	p-value
...			
α_3	0.000	0.000	0.000
$\beta_{3,Elo}$	0.000	0.164	0.798
$\beta_{3,PM}$	−0.001	0.001	0.306
$\beta_{3,DG}$	0.967	0.132	0.000
$\beta_{3,H+1}$	0.668	0.119	0.000
$\beta_{3,H+2}$	0.213	0.174	0.376
$\beta_{3,A+1}$	1.034	0.114	0.000
$\beta_{3,A+2}$	1.130	0.147	0.000
$\beta_{3,HR}$	0.072	0.229	0.759
$\beta_{3,AR}$	0.896	0.158	0.000
α_4	0.001	0.000	0.000
$\beta_{4,Elo}$	0.025	0.139	0.786
$\beta_{4,PM}$	0.000	0.000	0.734
$\beta_{4,DG}$	0.829	0.109	0.000
$\beta_{4,H+1}$	0.860	0.090	0.000
$\beta_{4,H+2}$	0.775	0.113	0.000
$\beta_{4,A+1}$	0.396	0.108	0.001
$\beta_{4,A+2}$	0.197	0.166	0.397
$\beta_{4,HR}$	0.933	0.149	0.000
$\beta_{4,AR}$	0.294	0.159	0.145

of goals to each team, but with goals scored, x_{H+1} to indicate that the home team leads by one goal, x_{H+2} to indicate that the home team leads by two or more goals, x_{A+1} to indicate that the away team leads by one goal, and x_{A+2} to indicate that the away team leads by two or more goals. There are also indicator variables for whether or not the home team has received any red cards, x_{HR}, and similarly for the away team x_{AR}.

Table 6 shows the regression coefficients for the extended competing risk model. The covariates that are included in both the extended model and the simpler model shown in Table 5 obtain similar coefficients. The other regression coefficients show that the scoring rates for both teams increase when the match is no longer goalless. They

also show that the scoring rates increase if the other team has any players sent off, and decreases for the team with a red card. For the causes related to red cards, it is clear that the coefficients for the rating covariates are not statistically significant. The coefficients for the red card covariates, x_{HR} and x_{AR}, indicate that the rate at which a team is penalized with red cards increases when the other team has received a red card. Furthermore, it seems that the rates for red cards increase when the match is no longer goalless, although not all of the corresponding regression coefficients are statistically significant.

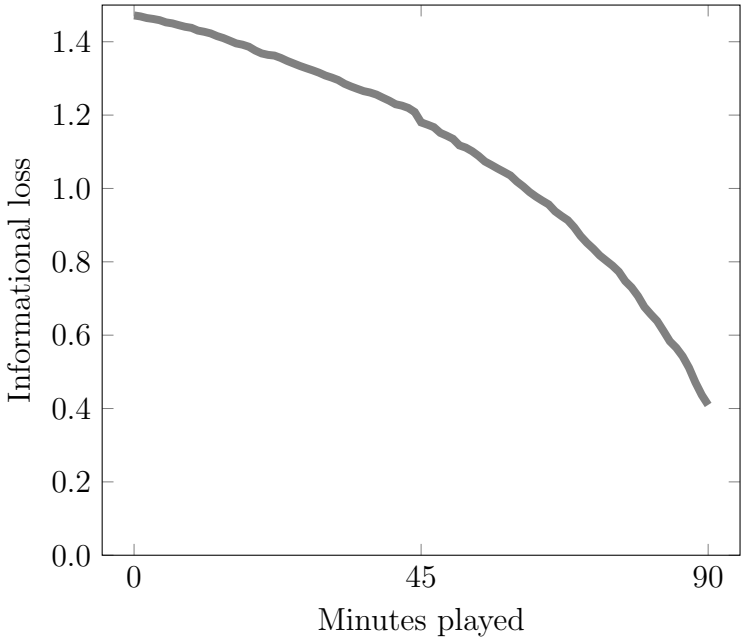

Figure 1: Informational loss for the extended competing risk model.

Figure 1 shows the informational loss of the predictions by the extended competing risk model as a function of minutes played in the matches. As expected, the average informational loss decreases as the matches progress towards ninety minutes. As many matches have stoppage time added on the end of the second half, the average loss does not reach 0. Figure 2 shows the difference in informational loss of the simple competing risk model and the extended model. The predictions of the two models

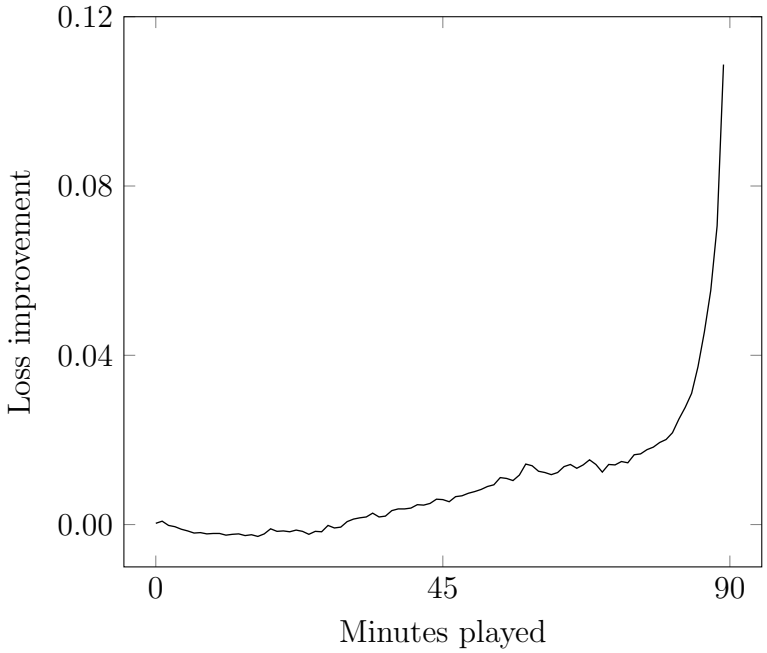

Figure 2: Comparison of informational loss for the extended competing risk model and simple risk model.

are equally good from the start of the match and until about half way through the first half of the match. At that point, it seems that the extended model gradually provides increasingly good predictions compared to the simple model. However, only towards the end of the matches the difference in prediction quality reaches significant proportions.

Figure 3 shows an example of in-game probabilities derived from the extended competing risk model. The match illustrated was played on April 6, 2019, in League Two, between the host Carlisle United and the visitors Bury. The probability of a home win is represented by the black area at the top of the figure, while the probability of an away win is shown in gray at the bottom of the figure. For this match, the rating based covariate values were $x_{Elo} = -85.5, x_{PM} = -0.433$ both indicating an edge for the away team. The initial model probabilities for a home win, draw and away win were respectively $0.323, 0.280$ and 0.397.

At the start of the match, the away team was a slight favourite, according to the model. However, Carlisle scored an early opener and became heavy favorites to win. To turn the events, Bury equalized in minute 8 and then scored their second goal in minute 43. Just before the end of the first half, a Carlisle player was sent off with a red card. At this point the chances of a home win were at a low point. Another equalizer, in minute 50 was followed by a period where the probability of a draw increased gradually, but with Bury being given the best chances of a victory, given their advantage in playing with an extra man. Finally, the home team scored their third goal in minute 89, and went on to win the match.

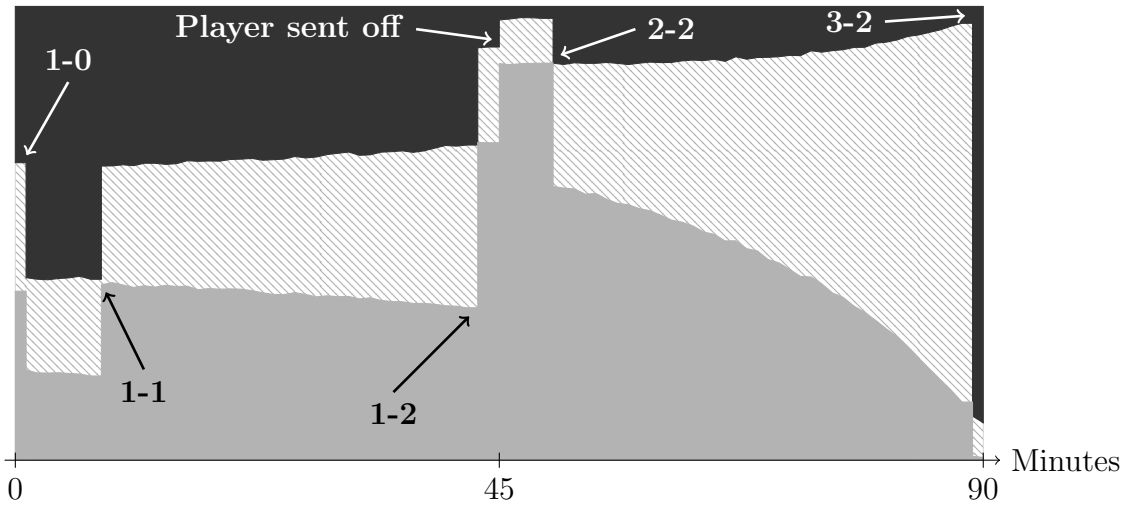

Figure 3: Example of in-game win probabilities from the match between Carlisle United (top) and Bury (bottom) on April 6, 2019, in League Two. The match ended with a 3-2 win for Carlisle, despite playing one man down in the second half.

5 Concluding remarks

Accurate forecasts for match outcomes in association football can provide useful information to decision makers in clubs, to sports fans, and to experts in the media. The scientific literature contains numerous contributions presenting models to create

pre-game forecasts, but only few contributions discuss models for generating forecasts taking into account real-time information while the match is being played. A common theme in the models is the inclusion of a team strength measure, while a more recent development is the attempt to model team strength indirectly by first assigning ratings to individual players.

This paper evaluated two types of models for generating match outcome forecasts: ordered logit regression (OLR) models and competing risk models. Both types of models perform equally well in terms of generating pre-game forecasts. Forecasts generated based on team ratings are of similar quality as forecasts generated based on player ratings. However, forecasts based on using both types of ratings outperform forecasts made based on either team or player ratings only.

Competing risk models can be used to generate in-game forecasts. An extended competing risk model that estimates rates for both goals and red cards provides the best in-game predictions, although its pre-game forecasts are on par with a simpler competing risk model that only estimates scoring rates. Thus, if only pre-game forecasts are needed, the most parsimonious competing risk model is sufficient. As the competing risk models examined are based on the assumption that the time between goals is exponentially distributed, it follows that a standard Poisson regression model would have an equivalent performance as the simple competing risk variant.

Acknowledgements

The authors wish to thank an associate editor and two anonymous reviewers for their comments to the initial version of this paper.

References

Baboota, R. and Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, **35**, 741–755.

Boshnakov, G., Kharrat, T., and McHale, I. (2017). A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, **33**, 458–466.

Brooks, J., Kerr, M., and Guttag, J. (2016). Developing a data-driven player ranking in soccer using predictive model weights. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 49–55, New York, NY, USA, 2016. ACM.

Constantinou, A. (2019). Dolores: a model that predicts football match outcomes from all over the world. *Machine Learning*, **108**, 49–75.

Constantinou, A. and Fenton, N. (2013). Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, **9**, 37–50.

Cotta, L., Vaz de Melo, P., Benevenuto, F., and Loureiro, A. (2016). Using FIFA soccer video game data for soccer analytics. In *Proc. KDD Workshop on Large-Scale Sports Analytics*, San Francisco, USA, 2016.

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

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

Dixon, M. and Robinson, M. (1998). A birth process model for association football matches. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **47**, 523–538.

Dobson, S. and Goddard, J. (2001). *The Economics of Football*. Cambridge University Press, Cambridge.

Duch, J., Waitzman, J., and Amaral, L. (2010). Quantifying the performance of individual players in a team activity. *PLoS ONE*, **5**(6), e10937. doi: doi:10.1371/journal.pone.0010937.

Elo, A. (1978). *The Rating of Chessplayers, Past and Present*. Arco Publishing, New York.

Gelade, G. and Hvattum, L. (2020). On the relationship between +/− ratings and event-level performance statistics. *Journal of Sports Analytics*. Forthcoming.

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

Greene, W. (2012). *Econometric Analysis*. Pearson, Harlow, England, 7th edition.

Groll, A., Shauberger, G., and Tutz, G. (2015). Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA World Cup 2014. *Journal of Quantitative Analysis in Sports*, **11**, 97–115.

Groll, A., Ley, C., Shauberger, G., and Van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*, **15**, 271–287.

Hvattum, L. (2017). Ordinal versus nominal regression models and the problem of correctly predicting draws in soccer. *International Journal of Computer Science in Sport*, **16**, 50–64.

Hvattum, L. (2019). A comprehensive review of plus-minus ratings for evaluating individual players in team sports. *International Journal of Computer Science in Sport*, **18**, 1–23.

Hvattum, L. and Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, **26**, 460–470.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. Wiley-Interscience, Hoboken, New Jersey, USA.

Karlis, D. and Ntzoufras, I. (2009). Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, **20**, 133–145.

Kharrat, T. (2016). *A journey across football modelling with application to algorithmic trading*. Phd thesis, University of Manchester, Manchester, UK.

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

Koopman, S. and Lit, R. (2019). Forecasting football match results in national league competitions using score-driven time series models. *International Journal of Forecasting*, **35**, 797–809.

Lasek, J. (2019). *New data-driven rating systems for association football*. Phd thesis, Warsaw University of Technology, Warsaw, Poland.

Lasek, J., Szlávik, Z., and Bhulai, S. (2013). The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recognition*, **1**, 27–46.

Ley, C., Van de Wiele, T., and Van Eetvelde, H. (2019). Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. *Statistical Modelling*, **19**, 55–73.

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

McHale, I., Scarf, P., and Folker, D. (2012). On the development of a soccer player performance rating system for the English Premier League. *Interfaces*, **42**, 339—351.

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.

Pantuso, G. and Hvattum, L. (2019). Maximizing performance with an eye on the finances: a chance-constrained model for football transfer market decisions. arXiv:1911.04689.

Peeters, T. (2018). Testing the wisdom of crowds in the field: Transfermarkt valuations and international soccer results. *International Journal of Forecasting*, **34**, 17–29.

Robberechts, P. and Davis, J. (2019). Forecasting the FIFA World Cup – combining result- and goal-based team ability parameters. In Brefeld, U., Davis, J., Van Haaren, J., and Zimmermann, A., editors, *Machine Learning and Data Mining for Sports Analytics*, pages 16–30. Springer, Switzerland.

Robberechts, P., Van Haaren, J., and Davis, J. (2019). Who will win it? An in-game win probability model for football. In *6th Workshop on Machine Learning and Data Mining for Sports Analytics*, page 13, Würzburg, Germany, 2019.

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

Sæbø, O. and Hvattum, L. (2015). Evaluating the efficiency of the association football transfer market using regression based player ratings. In *NIK: Norsk Informatikkonferanse*. Bibsys Open Journal Systems. 12 pages.

Sæbø, O. and Hvattum, L. (2019). Modelling the financial contribution of soccer players to their clubs. *Journal of Sports Analytics*, **5**, 23–34.

Schauberger, G. and Groll, A. (2018). Predicting matches in international football tournaments with random forests. *Statistical Modelling*, **18**, 1–23.

Schultze, S. and Wellbrock, C.-M. (2018). A weighted plus/minus metric for individual soccer player performance. *Journal of Sports Analytics*, **4**, 121–131.

Sittl, R. and Warnke, A. (2016). Competitive balance and assortative matching in the German Bundesliga. Discussion Paper No. 16-058, ZEW Centre for European Economic Research, Mannheim.

Stefani, R. and Pollard, R. (2007). Football rating systems for top-level competition: A critical survey. *Journal of Quantitative Analysis in Sports*, **3**(3), Article 3.

Štrumbelj, E. (2014). On determining probability forecasts from betting odds. *International Journal of Forecasting*, **30**, 934–943.

Szczepański, Ł. and McHale, I. (2016). Beyond completion rate: evaluating the passing ability of footballers. *Journal of the Royal Statistical Society Series A*, **179**, 513–533.

Tiedemann, T., Francksen, T., and Latacz-Lohmann, U. (2010). Assessing the performance of German Bundesliga football players: a non-parametric metafrontier approach. *Central European Journal of Operations Research*, **19**, 571–587.

Titman, A., Costain, D., Pidall, P., and Gregory, K. (2015). Joint modelling of goals and bookings in association football. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **178**, 659–683.

Van Eetvelde, H. and Ley, C. (2019). Ranking methods in soccer. In Kenett, R., Longford, T., Piegorsch, W., and Ruggeri, F., editors, *Wiley StatsRef: Statistics Reference Online*, pages 1–9. Wiley, Hoboken, NJ, USA.

Van Haaren, J. and Davis, J. (2015). Predicting the final league tables of domestic football leagues. In *Proceedings of the 5th international conference on mathematics in sport*, pages 202–207.

Vilain, J.-B. and Kolkovsky, R. (2016). Estimating individual productivity in football. http://econ.sciences-po.fr/sites/default/files/file/jbvilain.pdf. Accessed 2019-08-03.

Volf, P. (2009). A random point process model for the score in sport matches. *IMA Journal of Management Mathematics*, **20**, 121–131.

Wheatcroft, E. (2019). Evaluating probabilistic forecasts of football matches: the case against ranked probability score. arXiv:1908.08980v1.

Witten, I., Frank, E., and Hall, M. (2011). *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, Burlington, MA, USA.

Wunderlich, F. and Memmert, D. (2018). The betting odds rating system: Using soccer forecasts to forecast soccer. *PLoS ONE*, **13**, e0198668.