



Using ELO ratings for match result prediction in association football

Lars Magnus Hvattum^{a,*}, Halvard Arntzen^b

^a *Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Norway*

^b *Molde University College, Norway*

Abstract

Sports betting markets are becoming increasingly competitive. These markets are of interest when testing new ideas for quantitative prediction models. This paper examines the value of assigning ratings to teams based on their past performance in order to predict match results in association football. The ELO rating system is used to derive covariates that are then used in ordered logit regression models. In order to make informed statements about the relative merit of the ELO-based predictions compared to those from a set of six benchmark prediction methods, both economic and statistical measures are used. The results of large-scale computational experiments are presented.

© 2009 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

Keywords: Sports forecasting; Loss function; Evaluating forecasts; Rating; Ordered logit

1. Introduction

Operations research (OR) methodologies have been applied to several areas of sports, such as the scheduling of fixtures and sports officials, the making of decisions with respect to tactics and strategies during sporting events, and the forecasting of results (Wright, 2009). In this paper we address the issues related to match result forecasting in association football. This area of application for forecasting methodologies is particularly interesting due to the easy availability of

data and the increase in the number of available adversaries (bookmakers) with whom money can be gambled. There is now a big betting market for association football, with a large turn-over, capturing the interest of many casual observers.

It seems evident that match results in association football are governed partly by chance and partly by skill, as was suggested by Hill (1974). This indicates that forecasting methods could be used to provide decision support for both bookmakers and gamblers by trying to discover how chance and skill are related, while attempting to determine which factors can describe the skills involved, and which influence the role of chance.

* Corresponding author.

E-mail addresses: Lars.M.Hvattum@iot.ntnu.no
(L.M. Hvattum), Halvard.Arntzen@himolde.no (H. Arntzen).

There has previously been a fair amount of research in statistical modeling and forecasting in relation to association football. Maher (1982) introduced attack and defense parameters for each team and developed a model where the goals scored by the home and away teams follow Poisson distributions. However, the model could not be used to predict scores or results for future matches. Dixon and Coles (1997) used a similar approach and developed a model that was able to generate match outcome probabilities. Rue and Salvesen (2000) added the assumption that the attack and defense parameters are time-varying, and used Bayesian methods to update the estimates of them. Markov chain Monte Carlo iterative simulation techniques were used to make inferences. A computationally less demanding procedure for updating parameter estimates was later developed by Crowder, Dixon, Ledford, and Robinson (2002).

While the above studies focused on predicting the number of goals scored and then deriving the match result probabilities, it is also possible to model the match results directly. An ordered probit regression model with several types of explanatory variables was presented for this purpose by Goddard and Asimakopoulou (2004). The same modeling approach was also used by Kuypers (2000). Goddard (2005) compared models for forecasting numbers of goals and match results, and found them to yield similar performances in terms of their ability to predict match results. Forrest, Goddard, and Simmons (2005) used one of the models from Goddard (2005): an ordered probit model including a ‘match results’ dependent variable, covariates based on lagged match results, and other relevant information. Odds from bookmakers were tested against the model, and it was found both that the predictive power of these odds improved over time in the data set examined, and that the odds-setters probably used information which was not available to the probit model.

The main focus in this paper is on investigating the use of ELO ratings to create covariates for match result prediction models. The ELO rating system was initially developed for assessing the strength of chess players (Elo, 1978), but has been widely adopted in various other sports, including association football (Buchdahl, 2003). Similar rating systems (often inspired by the ELO system, but usually under different names, such as *power scores*) are also currently employed by some betting services. However, the use of

ELO ratings for predicting match results in association football does not appear to have been extensively tested and evaluated using scientific methods.

Various other rating systems have been examined in the forecasting literature. For example, Graham and Stott (2008) applied an ordered probit model with one parameter for each team to model and estimate individual team strength. This led to a type of rating, but the parameters were not dynamically updated after each match, as would be the case for ELO ratings. Knorr-Held (2000) focused on creating ratings *per se*, but did not examine their use for making predictions. Boulier and Stekler (2003) and Boulier, Stekler, and Amundson (2006) evaluated predictions based on power scores and an associated ranking for matches in the National Football League. Boulier and Stekler (1999) found that computer-made rankings are useful predictors in professional tennis, while Clarke and Dyte (2000) fitted ATP rating differences to a logistic model and used the resultant probabilities in simulations.

In Section 2 we give a detailed description of the ELO rating system, and discuss how it can be adapted for use with association football. Section 3 first shows how the ELO ratings are used within an ordered logit regression model to produce match result predictions, then describes a set of alternative prediction methods that will serve as benchmarks against which the ELO based predictions will be tested. The tests used to differentiate between the ELO based methods and the benchmarks are presented in Section 4, and the results are described in Section 5, after which concluding remarks follow in Section 6.

2. The Elo rating system

We will briefly describe both the basic ELO rating system and a modification thereof. Based on the results of a set of preceding matches, each team can be assigned an ELO rating as a measure of the team’s current strength. Let ℓ_0^H and ℓ_0^A be the current ratings, at the start of a match, of the home and away teams, respectively. Define a score system in which a win gives a score of 1, a draw a score of 0.5, and a loss a score of 0. The ELO calculations then assume that, on average, for the match in question, the home and away

teams should score γ^H and γ^A respectively, where

$$\gamma^H = \frac{1}{1 + c^{(\ell_0^A - \ell_0^H)/d}} \quad \text{and}$$

$$\gamma^A = 1 - \gamma^H = \frac{1}{1 + c^{(\ell_0^H - \ell_0^A)/d}}.$$

Now, let the actual score for the home team be given by

$$\alpha^H = \begin{cases} 1 & \text{if the home team won,} \\ 0.5 & \text{if the match was drawn, or} \\ 0 & \text{otherwise.} \end{cases}$$

The actual score for the away team is then $\alpha^A = 1 - \alpha^H$. The ELO ratings are updated after the match, and the new rating for the home team is

$$\ell_1^H = \ell_0^H + k(\alpha^H - \gamma^H). \quad (1)$$

The new rating ℓ_1^A for the away team is calculated in the same way. The rating follows the team, and is updated after each match.

Note that the computation of ELO ratings after a specific set of matches requires some initial ratings to be provided for each team. The ratings cannot be expected to be reliable indicators of strength until a sufficient number of past match results have been taken into account.

We will refer to the rating method described above as the *basic* ELO rating. An interesting variant of the basic ELO rating that can be used in association football is to allow the rating update coefficient k to depend on the goal difference, thus rewarding a 3–0 win more strongly than a 2–1 win. One possibility is to use Eq. (1) with k replaced by the expression

$$k = k_0(1 + \delta)^\lambda,$$

with δ being the absolute goal difference, and taking $k_0 > 0$ and $\lambda > 0$ as fixed parameters. This approach will be referred to as the *goal based* ELO rating.

There are three parameters in the basic ELO rating method, c , d , and k . While c and d can be interpreted as setting an appropriate scale for the ratings, some care is needed in choosing k . If the value of k is too low, a team's rating will not change quickly enough when it improves its performance, while if the value of k is too high, the rating will be unreliable, changing too much from one match to another. In the goal based ELO rating, there are four parameters, with k being

replaced by k_0 and λ . We discuss the settings for these parameters in Section 5.

3. ELO based prediction methods and benchmarks

In this section we first outline two prediction methods based on the use of ELO ratings, then describe six benchmark prediction methods that are used for assessing the performance of the ELO based methods.

3.1. ELO based prediction methods

In order to use the ELO ratings for making match result predictions, we make use of an ordered logit regression model (Greene, 1999). The basic ELO scheme outlined in Section 2 can be used with the ordered logit model as follows. An initial set of matches is used to compute initial ELO ratings for all teams in a league. Then a second set of matches is used to estimate the parameters of the ordered logit model, using as a single covariate the rating difference in favor of the home team,

$$x = \ell_0^H - \ell_0^A.$$

For future matches we then have a natural prediction method obtained by assigning the corresponding probability generated by the ordered logit model for each outcome of a match, with x being equal to the rating difference prior to the match. The ratings used will always be the most recently updated ones. The method also allows for periodic updates of the regression parameters, so as to always utilize as much data as possible. This prediction method will be referred to as **ELO_b**.

A very similar method, labeled **ELO_g**, is obtained if we just replace the basic ELO rating with the goal based ELO rating in the above scheme. The basic methods described here use no other covariates than the ELO rating differences, but one can easily extend these methods by including other covariates.

3.2. Benchmark prediction methods

Our main focus in this paper is on rating-based predictions of match results. We now present six other prediction methods that will serve mainly as benchmarks for comparisons with the ELO based methods.

These benchmarks are very different in terms of how they utilize the available information.

The first benchmark prediction method is labeled **UNI**. This method completely ignores any available information and gives a uniform distribution on the possible outcomes of a match, giving a predicted probability of $\frac{1}{3}$ for each outcome (home win, draw, away win).

The second method is called **FRQ**, and uses the available information about past matches in a very simple way. It takes the observed frequency of each outcome into account when making predictions, and thus takes the probability of a home win as equal to the observed frequency of home wins in previous matches. These two prediction methods, **UNI** and **FRQ**, are probably the most naïve alternatives available for predicting outcomes of football matches. Thus, one should interpret these as giving lower bounds on the quality of reasonable predictions.

Two other benchmark models are derived from the work of [Goddard \(2005\)](#), who used ordered probit regressions to create predictions for match results in association football. Rather than using the full number of covariates suggested, we extract the covariates that are based on the results of past matches only, and thus that use the same information as **ELO_b** and **ELO_g**. In the benchmark that we refer to as **GOD_b**, the covariates we use are the 50 result based team performance covariates used in Model 4 of [Goddard \(2005\)](#), while in **GOD_g** we use the 100 goal based team performance covariates used in Model 3 of the same paper. In other words, we exclude some of the covariates from the original models, such as those concerning match significance, cup elimination, geographical distance, and crowd attendance. We apply the remaining covariates within the same ordered logit framework that is used for our ELO rating based prediction methods.

Finally, two alternatives that can be expected to perform vastly better than the naïve methods are those based on the odds offered by a set of bookmakers. These bookmakers attempt to make a living by inducing punters to place bets, and their odds must therefore represent an approximation of the actual probabilities governing the outcome of a given football match. We will utilize the *vox populi* ([Galton, 1907](#)) of bookmakers (albeit exploiting the average rather than the median) to create a fifth benchmark, labeled **AVG**. Consider an upcoming match, and suppose that a number

of bookmakers offer odds for each of the three outcomes: home win, draw, and away win. For each outcome we compute the average odds and then take its inverse. To finally arrive at outcome probabilities, we normalize the inverse average odds, so that the sum of probabilities over the three outcomes equals one. Our last benchmark, **MAX**, is similar to **AVG**, but instead of using the average odds from the bookmakers, it uses the maximum odds for each of home win, draw, and away win.

The different benchmarks introduced here make radically different assumptions about the amount of information available from past history. While **UNI** assumes that no information is available, and **FRQ** assumes that only actual results are recorded (with no attention to which teams are playing, how many goals are scored, or whether there have been injuries, suspensions, transfers, change of managerial staff, etc.), the amount of information implicitly used by **AVG** and **MAX** is potentially enormous. Under the efficient market hypothesis ([Sauer, 1998](#)), the odds given by bookmakers should encompass all available relevant information, as otherwise there could be opportunities to exploit the market and achieve a profit. However, there are potential reasons for a misrepresentation of the odds by the bookmakers, as they may have objectives other than making perfect predictions: for example, [Levitt \(2004\)](#) argues that odds may be adapted to be in line with consumer preferences, while [Forrest and Simmons \(2008\)](#) show that odds may be set to exploit fan sentiment.

Therefore, the average/maximum odds should not be seen as upper bounds on the quality of predictions, but rather as giving predictions that one should try to outdo.

4. Evaluation procedures

Having presented eight different prediction methods in Section 3, it is necessary to have some means of differentiating between them in terms of quality: how well does each of them actually predict match results in association football? In the following we describe some statistical and economic measures for the evaluation of predictions. We then discuss how to make comparisons between the different prediction methods, after simulating their performances on a large set of matches.

4.1. Loss functions

There are several different loss functions that can be used to evaluate prediction methods (Witten & Frank, 2005). Here we will concentrate on quadratic loss and informational loss. In general, the loss is a measure obtained by a comparison between the prediction probabilities and the observed outcome.

The *quadratic* loss, which we will denote as L^2 , is sometimes referred to as the Brier score, due to the fact that it was introduced by Brier (1950) as a way of evaluating weather forecasts.

The *informational* loss, here denoted by L^I , corresponds to the amount of information required to communicate the observed results of matches given the true probabilities of each outcome (Shannon, 1948). Within the association football match result forecasting literature, a measure called the *pseudo-likelihood* statistic was presented by Rue and Salvesen (2000) for a set of matches, and has subsequently been used by others. This statistic is mathematically equivalent to the average informational loss over the same set of matches.

If one is minimizing the expected value of L^2 or L^I , the unique optimal method is the one that uses the true probabilities.

There are other loss functions that are not considered here. For example, the zero-one loss can be useful when comparing prediction methods that only give classifications rather than probabilities, such as newspaper tipsters (Forrest & Simmons, 2000). In this work, however, all of the prediction methods considered yield probabilities as outputs, and can therefore be evaluated using quadratic and informational loss.

4.2. Economic measures

As an alternative to loss functions, various economic measures can be applied to compare different prediction methods. We can simulate betting on past matches, based on the probabilities given by the prediction methods and the odds available from bookmakers. Suppose that a bet is placed on every outcome for which the probabilities indicate a value bet; that is, every time the probability times the maximal odds is greater than one.

We will consider three different money management schemes. In the UNIT BET strategy, we use a

fixed bet size of one unit, which will give either a gain equal to the odds minus one if the bet is a success, or a loss equal to one if the bet is a failure. In the UNIT WIN strategy, we also take the odds into consideration when placing bets, by choosing a stake such that a successful bet will give a gain of one unit. This strategy is less prone to yielding heavy losses from long strings of lost bets with high odds. Finally, assuming that we hold the correct probabilities, an optimal bet size can be set according to the Kelly criterion (Kelly Jr., 1956). In what we label the KELLY strategy, we take into account both the odds and the edge (defining the latter as the product of the probability and the odds, with a higher edge indicating that higher return rates are expected), with a bet size equal to $(op - 1)/(o - 1)$, where p is the prediction probability and o is the odds. For simplicity, when applying the KELLY strategy, we will assume a constant bank roll equal to one unit.

An additional way of examining economic measures is to perform the same betting while putting restrictions on the required edge. A winning method would be expected to have a larger return on bets when enforcing a stricter limit on the required edge. By increasing the required edge (upward from one), a decreasing number of bets are placed (tending toward zero as the required edge increases), but the returns per unit staked should increase if the prediction method produces better probabilities than the market.

4.3. Statistical tests

The loss functions and the returns on bets can be calculated for each individual match, but a reliable evaluation of a prediction method requires that the calculations are repeated for a large number of events, reporting the observed average performance.

A simple method for comparing two prediction methods, based on a loss function, can be described as follows. For each match, compute the losses for each of the methods. A matched pairs t -test can then be applied to test for significant differences in the theoretical mean loss for the two methods. The test requires the two methods to be applied to the same matches.

In the alternative situation, when the average loss is obtained using different sets of matches for each prediction method, one should be very careful about using the average loss to compare the methods. Since the average predictability of matches typically varies,

both between league systems and over time within a given system, comparing the average losses may not be meaningful.

When testing for statistical significance between methods based on their economic performances, some additional problems arise. What we want to compare is the returns per unit staked, which we will call the return on bets. When staking one unit per bet, it is easy to test for significant differences between returns on bets for different methods based on the observed averages and standard deviations, but matched pairs t -tests are not possible, since the methods may place bets on different matches and different outcomes.

5. Computational testing

In this section we report the results from our computational testing. First, the data used in the tests are described. Second, the calibration of the ELO rating based methods is discussed. Third, the main results, where we attempt to assess the quality of predictions made by the ELO based methods relative to those of the benchmark methods, are presented. In particular, we will focus on the **ELO_g** method, which seems to be the best of the methods which are built around ELO ratings.

5.1. Data

Nowadays, large amounts of data on the results of past football matches are readily available. Here, we have focused on the top four divisions of the English league system (currently named Premiership, Championship, League One, and League Two). Match results are obtained for 14 seasons, from 93/94 to 07/08, covering a total of 30,524 matches.¹ The data also include odds which were collected from various bookmakers during the course of the last eight seasons, for a total of 16,288 matches. Due to the increasing availability of bookmakers, the odds data are based on a variable number of bookmakers and are collected from various sources. Table 1 shows how the competitiveness of the football betting market has increased during the eight seasons in question. The table shows, both for each

Table 1

Expected returns on bets, calculated as the sum of the inverse of the best odds over all outcomes, based on 16,015 matches.

| Season | Return |
|-----------|--------|
| 2000/2001 | 0.9244 |
| 2001/2002 | 0.9324 |
| 2002/2003 | 0.9312 |
| 2003/2004 | 0.9323 |
| 2004/2005 | 0.9378 |
| 2005/2006 | 0.9686 |
| 2006/2007 | 0.9664 |
| 2007/2008 | 0.9750 |
| Avg. | 0.9456 |
| St.dev. | 0.0242 |

season and on average, the expected returns on bets, based on the best odds available. That is, it reports how much one would expect to collect for each unit bet if placing bets completely at random. The odds data for some of the matches indicate arbitrage opportunities. However, these matches are not included in Table 1, which is based on the remaining 16,015 matches.

5.2. Calibration

Recall from Section 2 that there are three parameters in the basic ELO rating method, c , d , and k , and four in the goal based ELO method, c , d , k_0 , and λ . The two first parameters of each, c and d , serve only to set a scale for the ratings, and we use $c = 10$ and $d = 400$. Alternative values of c and d give identical rating systems, but one has to find a suitable coefficient k for the accompanying rating adjustment procedure. Fig. 1 shows how the average quadratic loss depends on the values of k_0 and λ . Similar results were also observed for the informational loss. The basic ELO rating corresponds to setting $\lambda = 0$, and thus we can determine suitable parameter values for both **ELO_b** and **ELO_g** based on the figure. In the following we use $k = 20$ in **ELO_b**, and $k_0 = 10$ and $\lambda = 1$ in **ELO_g**, which gives $k = 10(1 + \delta)$.

As an illustration of how the ELO ratings of each team evolve over time, Fig. 2 shows the goal based ELO ratings of four selected teams in the English league system, from the start of the 95/96 season until the end of the 07/08 season. Manchester United won in 93/94, came second in 94/95, and was the top rated team by the start of the 95/96 season. The highest

¹ The data used here were collected from <http://www.football-data.co.uk>.

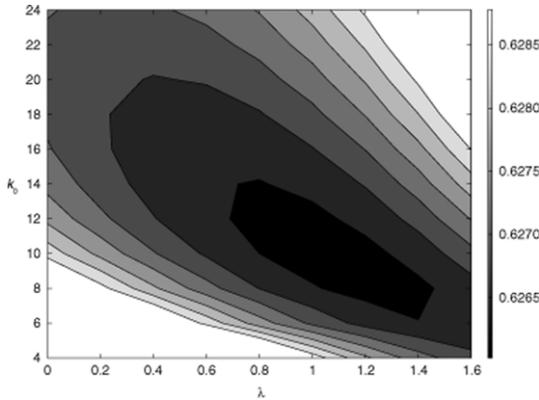


Fig. 1. Observed average quadratic loss when calibrating the two free parameters of the ELO_g method, k_0 and λ .

rating observed was for Chelsea, which had a rating of 719 points at the middle of January 2006. Leeds had a notoriously troublesome period, starting after the 01/02 season, when the team ended up in the fifth place of the Premier League. In the 07/08 season, Leeds played in League One, the third level of the league system, for the first time in the history of the team.

5.3. Main results

The data from the 93/94 to 07/08 seasons are used as follows. The first two seasons, 93/94 and 94/95, are used for initial calculations of ratings (ELO_b , ELO_g), historic frequencies (FRQ), and team performances (GOD_b , GOD_g). The next five seasons, from 95/96 to 99/00, are used for estimating the initial parameters for the ordered logit regression models (ELO_b , ELO_g , GOD_b , GOD_g). Finally, the remaining eight seasons, from 00/01 to 07/08, are used for the actual testing. During these seasons various calculations are required to keep the methods up to date, such as updating the ELO ratings after each match and updating the team performance measures used in GOD_b and GOD_g . The regression parameters used by the logit models and the frequency estimates used by FRQ are updated between seasons. Further illustrating the ELO_g method, Fig. 3 shows the home win, draw, and away win probabilities as a function of rating differences, as given by the ordered logit model estimated on the basis of data up to the end of the 06/07 season. The home win and away win probabilities are equal when the rating dif-

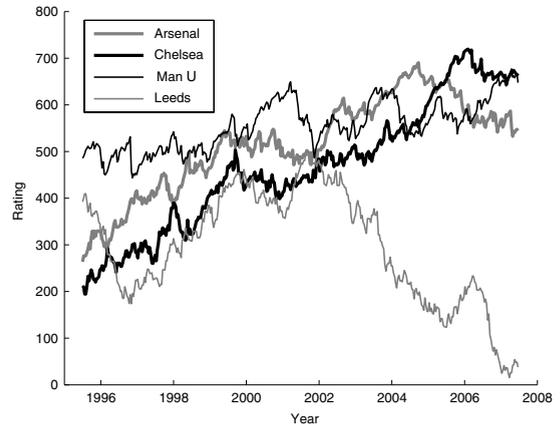


Fig. 2. Rating development for four selected teams in English association football.

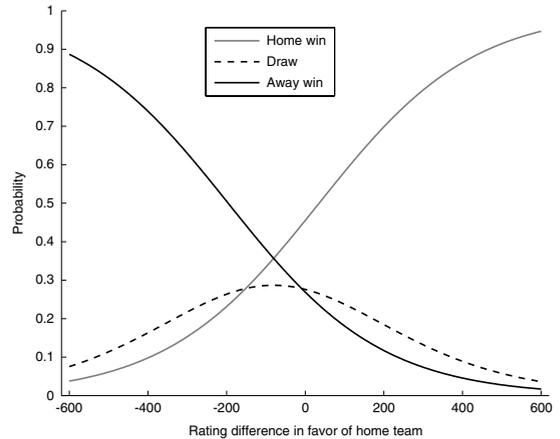


Fig. 3. Logit model probabilities of a home win, draw, and away win as functions of the rating difference.

ference is about -80 , thus giving a measure of the average home team advantage in terms of rating points.

In Tables 2–4 we show the results of placing bets based on the outcome predictions made by the different methods. From the total of 16,288 matches with odds data, we discard the betting results if either the odds indicate arbitrage opportunities or any of the prediction methods fail to produce outcome probabilities. This leaves a total of 14,927 matches. We can see that some methods may fail to provide probabilities for some matches, given that ELO_b will fail on any match involving a team that played below level four of the league system in the preceding season, and AVG will fail on any match where odds are not available.

Table 2

Average values for returns on simulated betting on 14,927 matches from the English league system, obtained by seven different prediction methods using a unit bet size.

| | #BETS | HWR | DR | AWR | AROB | L | U |
|------------------------|--------|-------|-------|-------|-------|-------|-------|
| UNI | 27 290 | 0.942 | 0.937 | 0.931 | 0.935 | 0.915 | 0.955 |
| FRQ | 16 892 | 0.942 | 0.903 | 0.920 | 0.928 | 0.903 | 0.953 |
| GOD_b | 13 644 | 0.935 | 0.940 | 0.886 | 0.912 | 0.885 | 0.939 |
| GOD_g | 14 161 | 0.935 | 0.951 | 0.953 | 0.945 | 0.919 | 0.972 |
| ELO_b | 12 142 | 0.956 | 0.937 | 0.891 | 0.930 | 0.902 | 0.958 |
| ELO_g | 12 152 | 0.972 | 0.928 | 0.902 | 0.939 | 0.911 | 0.967 |
| AVG | 5 594 | 0.953 | 0.915 | 0.960 | 0.954 | 0.904 | 1.005 |

Table 3

Average bet size (BS) and total return on bets (TROB) based on simulated betting on 14,927 matches from the English league system, obtained from seven different prediction methods, using three different bet sizing strategies.

| | #BETS | UNIT BET | | UNIT WIN | | KELLY | |
|------------------------|--------|----------|-------|----------|-------|-------|-------|
| | | BS | TROB | BS | TROB | BS | TROB |
| UNI | 27 290 | 1.000 | 0.935 | 0.371 | 0.940 | 0.086 | 0.922 |
| FRQ | 16 892 | 1.000 | 0.928 | 0.448 | 0.938 | 0.090 | 0.915 |
| GOD_b | 13 644 | 1.000 | 0.912 | 0.505 | 0.920 | 0.052 | 0.898 |
| GOD_g | 14 161 | 1.000 | 0.945 | 0.535 | 0.947 | 0.056 | 0.927 |
| ELO_b | 12 142 | 1.000 | 0.930 | 0.559 | 0.943 | 0.041 | 0.924 |
| ELO_g | 12 152 | 1.000 | 0.939 | 0.568 | 0.954 | 0.040 | 0.940 |
| AVG | 5 594 | 1.000 | 0.954 | 0.318 | 0.967 | 0.009 | 0.946 |

Table 4

Average bet size (BS) and total return on bets (TROB), based on simulated betting on 14,927 matches from the English league system, obtained by **ELO_g** when requiring different levels of edge. Here, the edge is defined as the predicted probability times the best available odds.

| EDGE | #BETS | UNIT BET | | UNIT WIN | | KELLY | |
|------|--------|----------|-------|----------|-------|-------|-------|
| | | BS | TROB | BS | TROB | BS | TROB |
| 1.0 | 12 152 | 1.000 | 0.939 | 0.568 | 0.955 | 0.040 | 0.937 |
| 1.1 | 3 673 | 1.000 | 0.935 | 0.433 | 0.964 | 0.079 | 0.944 |
| 1.2 | 1 389 | 1.000 | 0.924 | 0.298 | 0.955 | 0.091 | 0.910 |
| 1.3 | 662 | 1.000 | 0.776 | 0.236 | 0.908 | 0.098 | 0.847 |
| 1.4 | 325 | 1.000 | 0.521 | 0.190 | 0.706 | 0.100 | 0.651 |
| 1.5 | 173 | 1.000 | 0.412 | 0.170 | 0.637 | 0.105 | 0.567 |
| 1.6 | 87 | 1.000 | 0.110 | 0.142 | 0.216 | 0.102 | 0.193 |

Table 2 shows a first attempt to use returns on bets to differentiate between the benchmark prediction methods and the ELO based methods. For each method, using the UNIT BET strategy, the table shows the number of bets made (#BETS), the return on bets for home wins (HWR), draws (DR), and away wins (AWR), as well as the average return on bets (AROB). In addition, it shows 95% confidence intervals [L, U] for the return on bets for each method. The number of bets can be larger than the number of matches, since

bets can be placed on more than one outcome for a single match. The confidence intervals are all overlapping. Thus, in terms of returns on unit sized bets, no method can be said to be superior to any other method based on two-sample *t*-tests, despite the large number of bets made. The confidence intervals further show that it is very unlikely that any of the methods are profitable on average.

In Table 3, we present the results from three different betting strategies (UNIT BET, UNIT WIN, and

KELLY). The overall pattern is that the UNIT WIN strategy seems to give higher returns than UNIT BET. Since the former will place smaller bets when the odds are high, this result could be interpreted as evidence in support of a favorite-longshot bias (Cain, Law, & Peel, 2000), which means that returns are higher when betting on favorites than when betting on underdogs. We can also see that the KELLY strategy does not fare well in terms of total returns on bets. Since this strategy sets the size of the bets assuming that the true probabilities are available, this indicates that neither of the predictions produced are competitive when compared to the market. Interestingly, AVG places a total of 5594 bets over the 14,927 matches. In some betting markets, the average odds (or spreads) provide better predictions than the maximum odds, as was illustrated by Paton and Vaughan Williams (2005), who showed that the average mid-point of all spreads in a spread betting market produced better predictions than the mid-points of spreads offered by market outliers. This was used to construct a simple betting strategy with statistically significant profits. However, our findings so far indicate that MAX produces better predictions than AVG for match results in association football.

The results of attempts to increase betting returns by placing additional requirements on the edge (EDGE) before allowing a bet are illustrated in Table 4. Results are shown for ELO_g , for each of the three money management schemes. It appears that the return is fairly equal when the required minimum edge is set to 1.0 or 1.1. Increasing the required minimum edge to 1.2 and beyond leads to rapidly decreasing returns, along with the expected drop in the number of bets. For the UNIT WIN strategy, the average bet size (BS) decreases as the edge increases, naturally indicating that the bets with high edges are usually those with high odds. The overall interpretation of the results is that when ELO_g indicates a value bet with a high edge, it is usually the result of a ELO_g probability which is too high, rather than poorly set odds.

Moving on to the statistical measures, Table 5 shows the average informational loss (L^1) and quadratic loss (L^2), as well as their standard deviations. Matches where at least one prediction method is unable to provide outcome probabilities are discarded, resulting in loss values which are calculated from 15,181 of the 16,288 matches with odds data. A separate column (p -value) reports the p -values from

matched pairs t -tests that compare ELO_g with each of the other methods. When using loss functions, we are able to differentiate between all pairs of methods at the 0.01 level, except for two pairs: testing AVG versus MAX is only significant at the 0.05 level, and the difference in losses between GOD_b and GOD_g is not significant at all. Compared to any of our tests based on economic measures, the loss functions are clearly better at ranking the methods. This is partly because we can utilize matched pair tests based on the values generated by the loss functions, but the loss functions are also found to work better in a simple confidence interval comparison like that in Table 2.

Table 5 lists the methods in order of increasing prediction accuracy, as measured by the loss functions. The results support earlier studies by showing that the probabilities based on the betting market odds appear to be better approximations of the true probabilities than those provided by statistical methods of the kind employed here.

The methods based on ELO ratings seem to outperform both GOD_b and GOD_g . However, the latter two methods each have a large number of covariates (50 and 100 respectively), many of which are either almost always zero, or heavily correlated with each other. Even though tens of thousands of matches are used to estimate the parameters for the logit models, we suspect that this is not sufficient to avoid substantial amounts of noise in the estimates of GOD_b and GOD_g ; Goddard (2005) reported that the inclusion of more data continued to improve the predictive power until about 15 seasons of data were included. In line with this, we find that the relative losses of both GOD_b and GOD_g decrease as results from more seasons are added. Thus, it may be that the ELO rating is a more efficient way of encoding past results when relatively short time spans are considered.

6. Concluding remarks

We have implemented and tested two ELO rating based prediction methods for association football. The methods use ELO rating differences as covariates in ordered logit regression models. The ELO based methods were compared to six benchmark prediction methods, and they were found to be significantly worse than the two methods based on market odds, but better than all of the other methods, in terms of observed loss.

Table 5

Average values (AVG) and standard deviations (STD) for informational (L^1) and quadratic (L^2) loss on 15,181 matches from the English league system, obtained using eight different prediction methods. The p -values from matched pair t -tests are also reported, comparing each method with ELO_g .

| | L^1 | | | L^2 | | |
|------------------------|--------|--------|------------|--------|--------|------------|
| | AVG | STD | p -value | AVG | STD | p -value |
| UNI | 1.5850 | 0.0000 | 0.0000 | 0.6667 | 0.0000 | 0.0000 |
| FRQ | 1.5443 | 0.3791 | 0.0000 | 0.6469 | 0.1870 | 0.0000 |
| GOD_b | 1.5135 | 0.4824 | 0.0000 | 0.6321 | 0.2371 | 0.0000 |
| GOD_g | 1.5129 | 0.5136 | 0.0000 | 0.6315 | 0.2472 | 0.0000 |
| ELO_b | 1.5018 | 0.5030 | 0.0001 | 0.6266 | 0.2475 | 0.0001 |
| ELO_g | 1.4995 | 0.5060 | NA | 0.6256 | 0.2489 | NA |
| AVG | 1.4917 | 0.4772 | 0.0000 | 0.6219 | 0.2340 | 0.0000 |
| MAX | 1.4910 | 0.4955 | 0.0000 | 0.6217 | 0.2415 | 0.0000 |

We found that statistical loss functions were more efficient than economic measures in differentiating between the various prediction methods. This is in line with the findings of Johnstone (2007), who provided examples of situations where future profits were maximized when the forecaster chose methods on the basis of historical losses, rather than historical profits.

ELO ratings appear to be useful in encoding information on past results. In the case of association football, the single rating difference is a highly significant predictor of match outcomes. This finding can be taken as a justification of the increasingly common use of ELO ratings as a measure of team strength. Still, it remains an open question as to which covariates are needed in an ordered logit regression model in order to create predictions which are on par with the market odds, or even whether this is possible at all.

Acknowledgements

We thank a co-editor, an associate editor, and three anonymous referees for their extensive inputs.

References

- Boulier, B. L., & Stekler, H. O. (1999). Are sports seedings good predictors? An evaluation. *International Journal of Forecasting*, 15, 83–91.
- Boulier, B. L., & Stekler, H. O. (2003). Predicting the outcomes of National Football League games. *International Journal of Forecasting*, 19, 257–270.
- Boulier, B. L., Stekler, H. O., & Amundson, S. (2006). Testing the efficiency of the National Football League betting market. *Applied Economics*, 38, 279–284.
- Brier, G. W. (1950). Verification of weather forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3.
- Buchdahl, J. (2003). *Fixed odds sports betting: Statistical forecasting and risk management*. London: High Stakes.
- Cain, M., Law, D., & Peel, D. (2000). The favourite-longshot bias and market efficiency in UK football betting. *Scottish Journal of Political Economy*, 47, 25–36.
- Clarke, S. R., & Dyte, D. (2000). Using official ratings to simulate major tennis tournaments. *International Transactions in Operational Research*, 7, 585–594.
- Crowder, M., Dixon, M., Ledford, A., & Robinson, M. (2002). Dynamic modelling and prediction of English football league matches for betting. *The Statistician*, 51, 157–168.
- Dixon, M. J., & Coles, S. C. (1997). Modelling association football scores and inefficiencies in the football betting market. *The American Statistician*, 46, 265–280.
- Elo, A. E. (1978). *The rating of chessplayers, past and present*. New York: Arco Publishing.
- Forrest, D., & Simmons, R. (2000). Forecasting sport: The behavior and performance of football tipsters. *International Journal of Forecasting*, 16, 317–331.
- Forrest, D., & Simmons, R. (2008). Sentiment in the betting market on Spanish football. *Applied Economics*, 40, 119–126.
- Forrest, D., Goddard, J., & Simmons, R. (2005). Odds-setters as forecasters: The case of English football. *International Journal of Forecasting*, 21, 551–564.
- Galton, F. (1907). Vox populi. *Nature*, 75, 450–451.
- Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, 21, 331–340.
- Goddard, J., & Asimakopoulos, I. (2004). Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23, 51–66.
- Graham, I., & Stott, H. (2008). Predicting bookmaker odds and efficiency for UK football. *Applied Economics*, 40, 99–109.
- Greene, W. H. (1999). *Econometric analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Hill, I. D. (1974). Association football and statistical inference. *The American Statistician*, 23, 203–208.

- Johnstone, D. (2007). Economic Darwinism: Who has the best probabilities? *Theory and Decision*, 62, 47–96.
- Kelly, J. L., Jr. (1956). A new interpretation of information rate. *The Bell System Technical Journal*, 35, 917–926.
- Knorr-Held, L. (2000). Dynamic rating of sports teams. *The Statistician*, 49, 261–276.
- Kuypers, T. (2000). Information and efficiency: An empirical study of a fixed odds betting market. *Applied Economics*, 32, 1353–1363.
- Levitt, S. D. (2004). Why are gambling markets organized so differently from financial markets? *The Economic Journal*, 114, 223–246.
- Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36, 109–118.
- Paton, D., & Vaughan Williams, L. (2005). Forecasting outcomes in spread betting markets: Can bettors use ‘quarbs’ to beat the book? *Journal of Forecasting*, 24, 139–154.
- Rue, H., & Salvesen, Ø. (2000). Prediction and retrospective analysis of soccer matches in a league. *The Statistician*, 49, 399–418.
- Sauer, R. D. (1998). The economics of wagering markets. *Journal of Economic Literature*, 36, 2021–2064.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423, 623–656.
- Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco, CA: Elsevier.
- Wright, M. B. (2009). 50 years of OR in sport. *Journal of the Operational Research Society*, 60, S161–S168.

Halvard Arntzen is Associate Professor of Applied Mathematics at Molde University College, Norway. His research interests include applied statistics and forecasting, as well as operations research and optimization. He has recently published in journals including *Computers and Operations Research* and the *Journal of Heuristics*.

Lars Magnus Hvattum is a Postdoctoral Fellow at the Norwegian University of Science and Technology, in the Department of Industrial Economics and Technology Management. His main research interests include stochastic combinatorial optimization and heuristic solution methods, primarily for applications within transportation and logistics. He has recently published papers in *Computers and Operations Research*, *INFORMS Journal on Computing*, and the *Journal of Heuristics*.