Stockholms universitet

Is the oddsmarket for soccer efficient?

Carl Ringqvist

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2013:9
Matematisk statistik
September 2013

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2013:9**
http://www.math.su.se

Is the oddsmarket for soccer efficient?

Carl Ringqvist*

September 2013

Abstract

This report aims to investigate to some extent whether or not the odds-market for soccer-game outcomes is efficient. In order to try this hypothesis, we build a model for predicting the probability of each outcome alternative of a soccer game, in the sense of win, loss or draw. We base the survey on the complete history of the English Premier League, including a range of quantitative information for each game. We construct our own explanatory variables for our model over this quantitative information, a procedure requiring selection of several parameters. We will try to find the "optimal" set of parameters by optimizing a likelihood function. After choosing parameters for variable construction, we perform a statistical analysis over these, and settle with a model based on a rating for each of the opposing teams. The model performs well in predicting outcome probabilities, but fails to generate statistically significant profits, by which draw a tentative conclusion the odds-market in question is information efficient at least with respect to quantitative information. Also, some interesting findings regarding the variable construction procedure are noted.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail:carl.ringqvist@gmail.com . Supervisor: Mikael Petersson.

Acknowledgements

I would like to thank my supervisor for this thesis, Mikael Pettersson, my friends and colleagues Sebastian Rosengren and Carl Adam Briggert, and Jan-Olov Persson, thesis coordinator, for support and valuable consultation.

Contents

1 Introduction

During recent years the betting industry has seen a surge in turnover, mainly due to the development of internet gambling. As a result, the industry has experienced vastly increased competition. Whereas bookmakers historically could sell odds completely out of tune with the probability of the underlying event, thus easily managing the risks involved, they nowadays are in increasing need of correctly estimating the "true" underlying probabilities, enabling them to compete in the odds-market while still turning profit. This environment, one can imagine, would be more prone to sometimes "slip over the edge", i.e producing odds that actually are in statistical favor for the buyer. Further impetus are added to this notion by a possible conflicting interest among the bookmakers against setting the "correct" odds: if amounts pour in disproportionally on alternatives, and this is thoroughgoing in the market, the bookmakers might actually jointly alter the odds away from the "correct" ones in order to hedge their risk. This strategy would secure profit for the bookmakers whatever the outcome of the game, at cost of statistical profit, and would favor the buyer.

In this report we are going to investigate whether or not the odds-market for soccer is indeed inefficient, i.e if there exist opportunities to make consistent profit from identifying and taking position on wrongly priced odds. More specifically, we are going to look at the complete history of the English Premier League, through analyzing a dataset consisting of results of all games ever played with adherent information such as goals, corners, shots on target etcetera. Although the conclusions obviously will be directly applicable only to the Premier League, it should not be too bold an assumption it is somewhat representative for the professional leagues in general.

We are going to construct measures of the teams' strength with rating algorithms and by listing historic performance quantities for each game, that will allow us to estimate a probability for the outcome of each game through fitting a Multinomial Logistic Regression. With the outcome of a game is here meant precisely whether it is a win for the home team, a loss for the home team or a draw. The results are thus not applicable to odds-markets for other types of bets (eg goal difference, half time results). Moreover, our analysis does not account for the qualitative side of a game analysis, by which one can conclude that whatever the result here, wrongly priced odds might still exist and be observable by a qualitative analyst, even in the Premier League on the same type of bets here considered.

We are going to assume basic knowledge of the dynamics of the Multinomial Logistic Regression, an intermediate level of probability theory and familiarity with likelihood theory and adherent concepts such as the "maximum likelihood principle".

2 Definition of terms and some basic Odds theory

2.1 Odds markets, fair odds and implied probability

The betting markets we will consider works in the following way. A player places a bet on a certain event to occur. If the event occur the player win his betted amount times the odds the bookmaker has marketed for this event, otherwise the bookmaker keeps the betted amount in full.

We will refer to "fair" or "correct" odds as odds resulting in zero expected profit for the bookmaker. One might of course indulge in a discussion treating the existence of a probability for a soccer game outcome, which in turn quickly intrudes on philosophical territory treating the existence of human choice and other not so easily answered questions. Needless to say, we are simply assuming their existence.

It is a trifle task showing the fair odds for any binary event is the inverted probability of the event. Let V_e be the stochastic profit/loss resulting from placing a bet on the binary event e, X be the fair odds, p_e the probability of the event, and B the size of the bet. Then

$$E[V_e] = p_e(X-1)B - (1-p_e)B = 0 \Leftrightarrow X = \frac{1}{p_e} \Leftrightarrow p_e = \frac{1}{X}$$

Each of the outcome alternatives considered in this report (see below) can be seen as a binary events in isolation, why the fair odds for each of the three outcome alternatives must be precisely the inversion of their respective probabilities.

We will refer to the "implied" probability of an odds as the inverted odds. This is the probability of an event given the odds is fair; a concept that is going to be important throughout.

2.2 Variables and measures

Throughout this report, we are going to refer to certain clusters of explanatory variables as **measures**. For example, the teams' past performance with respect to goals scored and let in will be referred to as the "goal measure", although it consists of four variables for each game: the goals scored and let in by the home team, and the goals scored and let in by the visiting team. Together they comprise a measure of the probable outcome of the game. The reason for this nomenclature is the clusters intuitive connection, the analysis will treat the measures essentially as individual variables.

3 Data overview

The data on which the analysis in this report is based comprises the complete history of the Premier League: season 93/94 to present (2012/2013). An observation is a game, the outcome being one of the following:

H= Home team wins

D= Draw

A= Away team wins

A range of variables are observed for each game. For the seasons 93/94-01/02, the following variables occur in the data:

HT= Name of the home team

AT= Name of the away team

HG= Home team goals scored at full time

AG= Away team goals scored at full time

For the rest of the seasons, the additional variables occur in the data:

HST= Home Team Shots on Target

AST= Away Team Shots on Target

HC= Home Team Corners

AC= Away Team Corners

In addition. a range of odds, for each of the outcome alternatives, are listed for different betting sites (detailed later).

3.1 Historic goal, average points, corner and shots on target

For each game, measures of the opposing teams pre-match strength will be listed in the form of historic points, goals, corners and shots on target. The precise algorithm for obtaining these measures reads:

1. Find the n previous games played by each of the opposing teams.

2. For each of the opposing teams, list the scores/corners/shots made during the previous n games, and the scores/corners/shots made against during the previous n games. For points, list the average points obtained during the previous n games. This yields four positive whole numbers per game for scores/corners/shots, and two real positive numbers per game for points. These will be referred to as:

HomeS= historic Home Goals Scored
HomeA= historic Home Goals Allowed
AwayS= historic Away Goals Scored
AwayA= historic Away Goals Allowed
} comprising the **Goal measure**

SHomS= historic Home Shots on target Score
SHomA= historic Home Shots on target Allowed
SAwaS= historic Away Shots on target Score
SAwaA= historic Away Shots on target Allowed
} comprising the **Shots measure**

CHomS= historic Home Corners Score
CHomA= historic Home Corners Allowed
CAwaS= historic Away Corners Score
CAwaA= historic Away Corners Allowed
} comprising the **Corners measure**

PontH= historic average Home Points
PontA= historic average Away Points
} comprising the **Points measure**

3.2 ELO rating

For each game, a rating will be listed for each of the opposing teams. The rating used is a variant of the ELO rating *(Elo,1978)*, used originally for rating chess players.

Here, one assume that the situation of team i opposing team j can be well described by two stochastic variables (S.V:s) representing each of the teams' strengths, call them S_i and S_j. The event that team i wins over team j is in this model framework expressed through the event $[S_i > S_j]$, a probability that is easy to calculate for most distributions. The S.V:s are made dependent on the ratings, so that strong rating implies a larger probability of the adherent S.V coming out larger than any other S.V., than if the rating is weak. The possibility of a draw is here neglected, the reason becoming clear later on.

Mathematical details

For a specific game, let team 1 be the team playing home, opposing team 2. Then for i=1,2, j=1,2, let

R_i = Rating for team i prior to the game

S_i = Stochastic strength variable for team i

$$\lambda_i(\text{game}) = \begin{cases} 1 & \text{if team } i \text{ wins the game} \\ 0.5 & \text{if the result is a draw} \\ 0 & \text{if team } i \text{ loses the game} \end{cases}$$

$P_{i,j}$ = The probability that team i wins the game against opponent j

In our variant of the ELO rating, one assume:

$$S_i = \beta R_i + \epsilon_i \quad , \quad \beta \in \mathbf{R}_+ \quad , \quad \epsilon_i \sim EV_1(0,1)$$

where $EV_1(0,1)$ is the type-1 extreme value distribution.

The remarks and definitions above lend themselves to a straightforward calculation:

$$P_{i,j} = P[S_i > S_j] = P[\beta R_i + \epsilon_i > \beta R_j + \epsilon_j] = P[\beta(R_i - R_j) > \epsilon_j - \epsilon_i] = \frac{1}{1+e^{-\beta(R_i - R_j)}}$$

since

$$\epsilon_j - \epsilon_i \sim \text{Logistic(0,1)}$$

and since the the cumulative distribution function for the Logistic(0,1) is given by:

$$\text{F}_{Logistic}(x;0,1) = \frac{1}{1+e^{-x}}$$

After the game, the rating of the opposing teams are updated according to the following algorithm:

$$R_i^{new} = R_i^{old} + K(\lambda_i(\text{game}) - P_{i,j}) \quad , \quad K \in \mathbf{R}$$

In the beginning of the considered period, all teams in the league are assigned the same rating R_I. The algorithm procedure outlined above is performed for each game played in the data, making the ratings move according to performance. For each game, we list the following variables:

$$\left.\begin{array}{l} \text{eloHK} = \text{pre-match ELO rating for the home team} \\ \text{eloAK} = \text{pre-match ELO rating for the away team} \end{array}\right\} \text{comprising the } \textbf{ELO-rating measure}$$

eloPr = The probability for home win according to the ELO algorithm, also later referred to as the **ELO-probability measure**.

The system elegantly captures especially two features desirable in a model of our purpose. First, games played a long time ago tend to have less influence on the current rating than do recently played games, although the information in these old games are accounted for. Second, winning a game against an opponent when the theoretical prior probability ("ELO"-probability that is) of winning is large, grants less an advance in rating, than if winning against an opponent when the prior probability is less. Hence, with right choice of parameters, one ought to make the rating very sensitive to recent performance and accurately describe the team's current form, without neglecting either the fact that performing against strong teams is better than performing against weak teams, nor their long term performance. The measures of historic goals etc mentioned above for example, does not take these facts into consideration.

4 Model selection

The model used for fitting is the Multinomial Logistic Regression. For detail about the dynamics of the this model, see *Hosmer & Lemeshow, 2000*. The ordered logistic regression is left out as an alternative for the analysis, since it comprises essentially a special case of the non-ordered. The benefit of using it over the non-ordered alternative lies mainly in reducing the number of model parameters, but since our observational data is vast, this is not an immediate concern.

All explanatory variables are ordered and multileveled, and expected to explain the probability of a loss or a win in a monotone way (for example, the probability of a team winning a game should steadily increase the larger its goal difference from previous matches). This indicates a suitable situation for using the multinomial logistic approach, as it is very flexible in creating monotone curves. The probability of draw will be expected to take the shape of a 'bell' or a second degree polynomial over an explanatory variable, a form also possible to obtain with the logistic regression. The only serious competitor would be the Multinomial Probit Model. However, this is known to be very similar to the logistic alternative. Moreover, since the logistic model has a more intuitive interpretation (log-odds linear in explanatory variables), it seems to be the more reasonable alternative.

The data will throughout the report be divided into two blocks, according to which variables they contain. As outlined above, a range of additional variables occur from 02/03 and onward, and this year just so happens to divide the material in approximately two equal halves. These halves of the data will be referred to as section 1 and section 2 respectively. The coefficients for the variables contained in section 1 will be estimated over this period (see below). The data for this period will be referred to as data1. The variables contained in section 2 will be estimated over the first half of section 2. The data for this period will be referred to as data2. We will build two models. One based on data1, and tested over section 2. Another based on data2, and tested over the second half of the section 2. The model that performs the best can then be estimated over the complete material (if possible) for future estimations. We will however not undertake such a procedure in this report.

5 Variable parameter selection

This statistical investigation differers somewhat from the traditional in that not only the model contain parameters, but also the variables constructed to comprise its explanatory ground. This section explain the procedure for choosing such parameters in an optimal way.

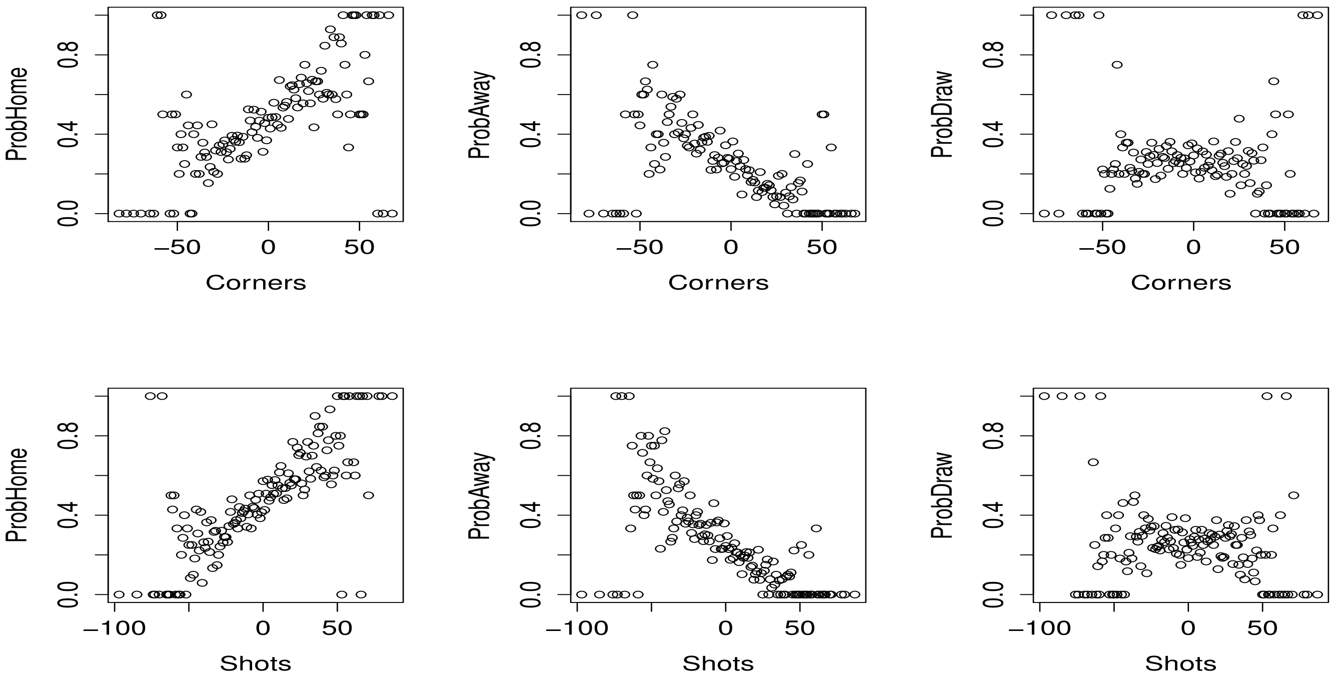
5.1 Choosing the right n for the historic goals, point averages, corners and shots on target

It seems plausible that using different n (number of historic games to include) for the calculation of the mentioned measures of historic points, goals, corners and shots on target; could yield significantly different results in terms of how well they explain data. The maximum likelihood principle lends a numerical estimation of the optimal n for each of the four measures. Since the number of games in a season is limited, one only have a finite amount of n to consider. Moreover, many candidates for an n even within the range of the number of seasonal games is inappropriate, since the larger n the more of the initial games will not get assigned a value; they wont have n previous games played in the season. Estimating over different seasons seem suboptimal, as changes in form can be significant in between. We will therefore content ourselves with calculating and comparing Likelihoods for n between one and twenty, for each of the four measures. Games that cannot be assigned a measure (games played early in a season), are deleted from the material. However, deleting these games for each n individually would result in there being different number of observations underlying the calculation of the likelihood. This would of course undermine the credibility of comparing the Likelihoods, since generally, a material with more observations tend to have a smaller Likelihood than material based on fewer observations. Therefore, for all n, the games not assigned a measure for $n = 20$ are deleted before calculating a likelihood, rendering all likelihoods to be calculated from the same number of observations.

The likelihood will be calculated fitting the Multinomial Logistic Regression model over data1, one measure at a time. One might consider trying optimal combinations of different n for the different measures, fitting the model on several, but this would be inappropriate of two reasons. One, it would be numerically heavy, the number of combinations to try using any approach quickly getting unmanageable. Second, if one later on in the analysis decides to drop one of the measures from the model, the reason for choosing the n:s of the remaining measures has disappeared. One would have to re-optimize every time such a decision was made.

The Likelihood Optimization procedure results in a rather surprising result: the Likelihoods are, in general, getting bigger the larger the n, for all four measures. This holds when optimizing both over data1 and data2. Choosing $n = 20$ would however be suboptimal, as it would significantly reduce the number of games possible to play on during a season (each

Figure 1: Observed frequencies over variables



team only plays around 38 games). Here, $n = 6$ is chosen, for all four measures. Plots seem to confirm this being a good choice, and it seems like a reasonable tradeoff between the number of games it renders possible for betting, and the likelihoods behavior. To illustrate the degree of explanatory power of these quantities, a "difference measure" is plotted against the observed fraction of home wins, draws and home losses, for each of the four measures. The "Goaldifference", for example, for a specific game i is calculated through:

$$GD_i = [HomeS_i - HomeA_i] - [AwayS_i - AwayA_i]$$

where GD is Goaldifference. The other difference measures are calculated in the same way, except the points measure where the measure is simply the difference between the average historic points for the the opposing teams.

An interesting fact to observe is that the Shots on target and Corners measures (plotted in Figure 1) seems to explain data just as good as the more intuitively relevant goals and points measures. Graphs of the point and goal measure can be observed in appendix, Figure 14. All graphs are for n=6.

5.2 Choosing the parameters for the ELO algorithm

Like previously, the suitability of different parameter values are evaluated using the likelihood they generate over data1. The beginning rating R_I is set to 0 for each team. We allow for a slight modification of the ELO-rating presented before. Instead of fixing K for all match results, the values of K are set differently in the updating algorithm depending on the goal difference occurred. For example, if the result is draw, K could be set to 1, if the winning team wins by one goal over the other team, K could be set to 2, etc. We optimize over different combinations of β and set of K:s. The set of K:s are chosen to be different integer values for all goal differences up to four, at which it is set constant. The number four is chosen with the motivation that only 54 out of the 2892 observations in data1 have a goal difference of more than four. Very little flexibility hence ought to be lost due to fixing the K for results above four. One can even argue that there is a certain limit at which winning with additional goals does not add any information regarding the teams' strengths, as the losing team would tend to 'give up'. A goal difference of four is very hard to redeem in a game of football, indicating this limit is a reasonable one.

Now consider the function we are trying to optimize: the maximal likelihood obtained by fitting a multinomal logistic regression over the ratings, that in turn depend on the parameter values chosen for the ELO- updating algorithm:

$$\mathcal{L} = L(\mathbf{eloHK}(\beta, K_0, K_1, K_2, K_3, K_4), \mathbf{eloAK}(\beta, K_0, K_1, K_2, K_3, K_4))$$

Contemplating the structure of this function (vectors denoted in bold), one is quickly deterred from analyzing it even in the most abstract way. For one, each of the ELO ratings in the data are functions of the previous ratings and the outcomes, rendering the vector valued functions

$\mathbf{eloHK}(\beta, K_0, K_1, K_2, K_3, K_4) =$
$= \{eloHK_1(\beta, K_0, K_1, K_2, K_3, K_4), \; eloHK_2(\beta, K_0, K_1, K_2, K_3, K_4),,, \; eloHK_n(\beta, K_0, K_1, K_2, K_3, K_4)\}$

$\mathbf{eloAK}(\beta, K_0, K_1, K_2, K_3, K_4) =$
$= \{eloAK_1(\beta, K_0, K_1, K_2, K_3, K_4), \; eloAK_2(\beta, K_0, K_1, K_2, K_3, K_4),,, \; eloAK_n(\beta, K_0, K_1, K_2, K_3, K_4)\}$

to rather be of the joint form

$$\{\boldsymbol{f_1}(R_I, \beta, K_0, K_1, K_2, K_3, K_4), \; \boldsymbol{f_2}(\boldsymbol{f_1}(R_I, \beta, K_0, K_1, K_2, K_3, K_4)),,, \; \boldsymbol{f_n}(\boldsymbol{f_{n-1}}(....(\boldsymbol{f_1}(R_I, \beta, K_0, K_1, K_2, K_3, K_4))...)\}$$

where $\boldsymbol{f_i}$ is a vector valued function (dim=2) returning pre-match ELO-ratings for the home and away team for game i. Add to this the complexity brought by the fact the coefficients in the Multinomial Logistic Regression fit have to be optimized for a given alternative of ELO parameters in order to obtain the Likelihood-functions' value. We hence abstain from attempts of abstract reasoning about the structure of this function.

The optimization procedure outlined here finds the global maximum of the function in case it exists and given it is the only local maximum. This is not a very likely event, but the procedure would still make sense since it would produce 'a good alternative' even if not the optimal alternative; or the best alternative one could hope to obtain with our computer power.

In the first optimization, β ranges from zero to one with intervals of 0.1, and the different K:s ranges from zero to fifteen with intervals of 1. Combinations such that K for a larger goal difference is larger than K for a smaller is tried out. The choice of ranges are motivated by that they cover values suggested as suitable in previous literature on the subject *(Hvattum Arntzen, 2010).* The procedure yields β=0.1 and $\mathbf{K} = \{K_0, K_1, K_2, K_3, K_4\} = \{1, 1, 2, 3, 5\}$. The procedure is repeated with new ranges including the obtained values, with finer intervals than previously (since computing power limitations does not allow to run over the original span of values with finer division). The ranges are chosen so that they cover the two values closest to the obtained values in the previous optimization. For example, in the second optimization, β ranges from 0 to 0.2, with 0.02 as intervals, $K1$ ranges from 0 to 2 with 0.2 intervals, etc. The procedure is iterated several times. If the values obtained along the way lies on the boundaries of the ranges chosen, we redo the same optimization with slightly extended ranges for these parameters. Continuing in this fashion, we get

β=0.06
K={1, 1, 2.1, 3.2, 4.1}

with an accuracy of 0.01 for β and 0.1 for the other variables. This will be our final choice of parameters, and we list the pre-match ELO rating for each game and team as explanatory variables. For listing the ELO-ratings for data2, a similar optimization procedure is undertaken, but over the data comprising data1 and data2. The result is

β=0.08
K={1.1, 1.1, 1.9, 2.9, 4.1}

with the same degree of accuracy as before; rather comforting as it is not at all very different from the previous set of values. It speaks of a certain stability in this approach of seeking to explain data. Plots over the rating development of a few teams over time can be observed in the appendix, Figure 15.

We also run the same optimization procedure for the probability generated by the ELO rating in the updating algorithm as an explanatory variable. This probability would have nothing to do with the "true" probability of an outcome, since the draw outcome is not even considered as part of the event space here. However, it seams reasonable that scaling the measure between 0 and 1 would be favorable for modeling the true probability. Moreover it incorporates the information contained in the two pre-match ELO ratings into one variable. Parameter combinations for this variable is therefore optimized for investigation, both for data1 and data1+data2 as previously, and the respective ELO probabilities for the respective combination outcomes listed for each of the datasets. The result is, with the same accuracy as previously for the respective parameters,

data1: $\beta = 0.07$
$\mathbf{K} = \{1.05, 1.05, 2.2, 2.8, 4.45\}$

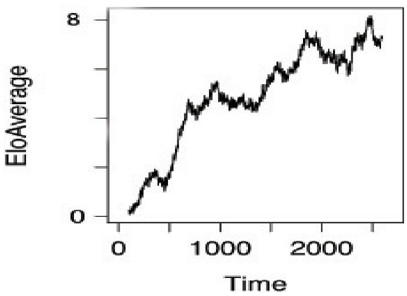
data1+data2:
$\beta = 0.06$
$\mathbf{K} = \{1, 1.02, 2, 2.5, 4\}$

These results are further adding to the notion it might exist an somewhat optimal set of combinations for explaining data of our type, as the results are very similar internally but also to the previous results.

5.3 ELO inflation

We wish the ELO ratings not to be prone to inflation, in order for the estimated coefficients to be stable over time. Although the rating system is constructed so that the sum of a static number of teams' ratings are the same over time (the ELO rating update algorithm keeps the sum of two teams' ratings unchanged for any event), new teams enter the Premier League with a beginning rate of 0, while teams with lower ratings tend to be the ones excluded over time. Thus bad teams may tend to loose points to the system and then leave it, causing the average rating to inflate. The graph in Figure 2 depicts a moving average of the ratings registered over time, confirming suspicions. However, when using a model like the one under construction, one would make use of frequent updates of the coefficient estimates as games are being played and new observations made available. As this would compensate for the rather moderate degree of inflation present, we choose to not go further in an analysis to treat the system for this flaw.

Figure 2: Moving average of the ELO-ratings over time



6 Building model 1

6.1 Overview

The explanatory variables here consists of the following:

HomeS - Goals made by the home team during the previous six games
HomeA- Goals allowed by the home team during the previous six games
AwayS - Goals made by the away team during the previous six games
AwayA - Goals allowed by the away team during the previous six games

PontH - Average points scored by the home team during the previous six games
PontA - Average points scored by the away team during the previous six games

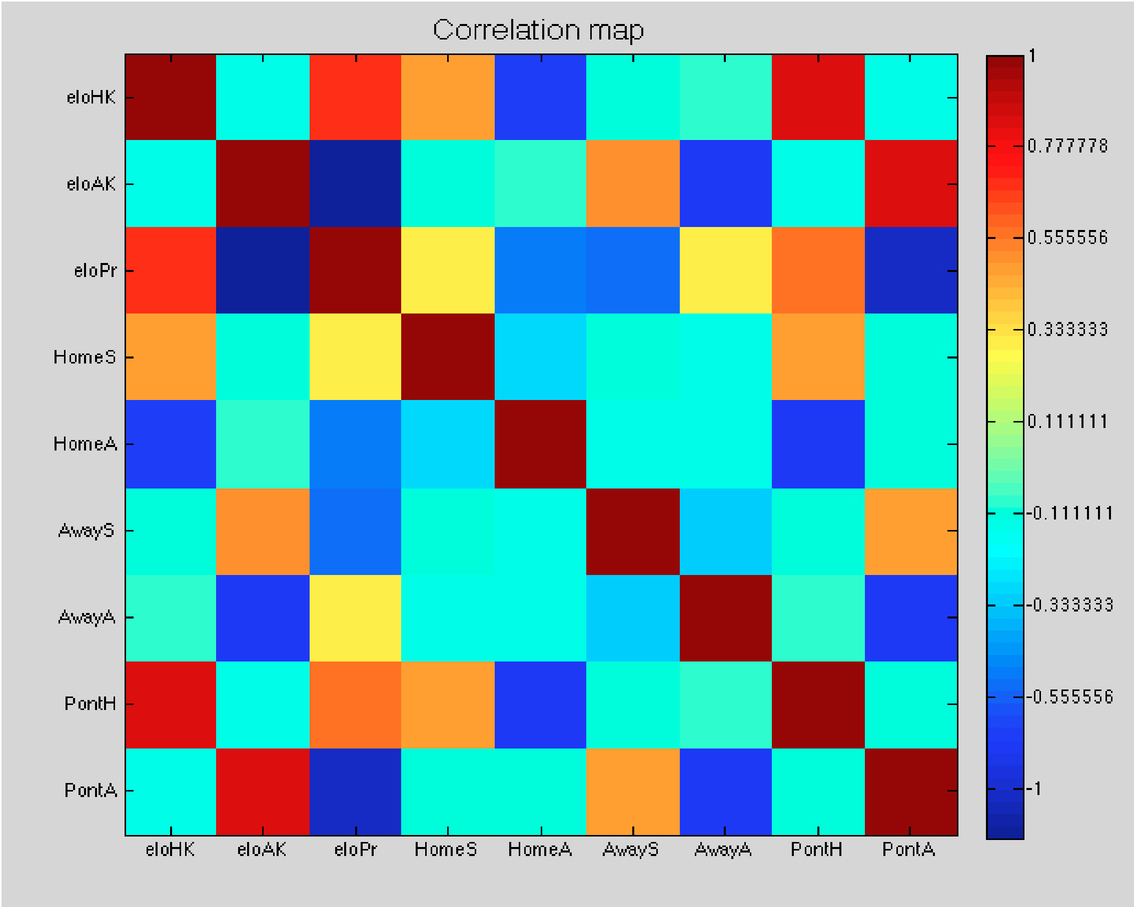
eloHK - Pre-match ELO rating for the home team
eloAK - Pre-match ELO for the away team

eloPr - The "probability" generated in the ELO algorithm for a game

6.2 Univariate analysis

A univariate analysis of each of the explanatory measures confirms that they cannot be excluded individually; they all have p-values of less than 0.01, below any reasonable significance level. Note that when referring to univariate analysis, we are in this report going to mean analyzing the variables by measure. By a p-value for a measure, we mean the joint p-value for the variables comprising the measure (for example, when we speak of the p-value for the points measure, we mean the joint p-value for PontH and PontA). As stated, the analysis is performed in this manner out of intuitive reasons. It does make little sense including just one or a few of the goal measures without the others. Also, we need a way to manage the rather large number of variables, especially when it comes to optimization procedures. The analysis hence starts from the multivariate situation.

Figure 3: Colored map illustrating the correlation between all explanatory variables



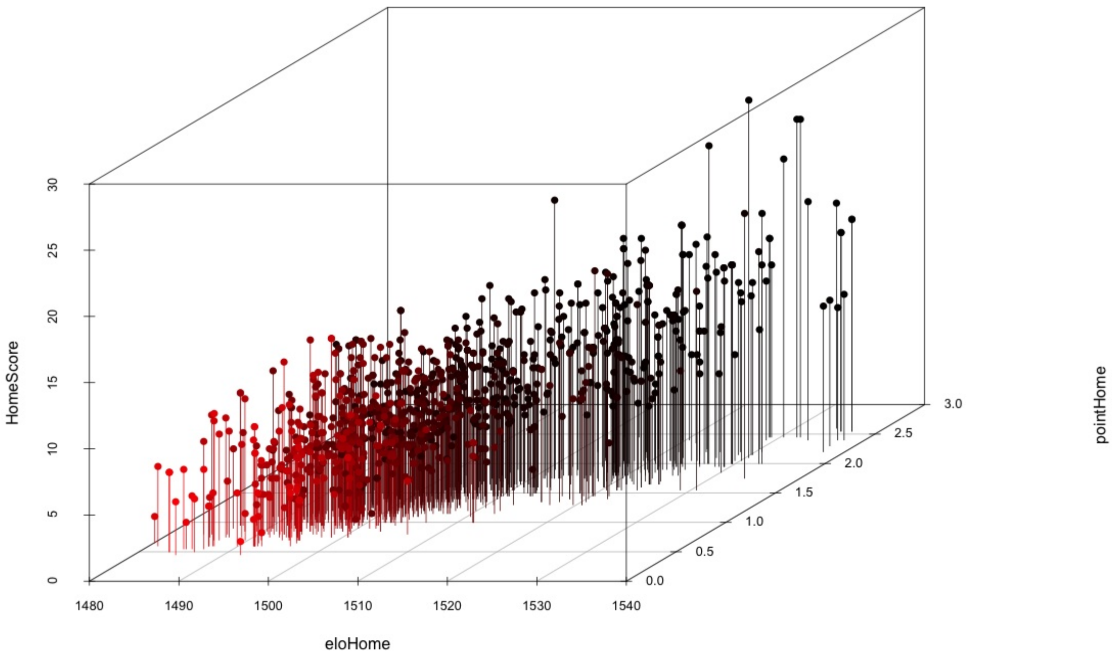
6.3 Correlation analysis

The following picture illustrates the correlation between all the explanatory variables:

There lies a positive surprise in the figure: the goal measures seem to be only moderately correlated with the other measures, indicating at least not all four measures contain essentially the same information. The point measure on the other hand seem to be very correlated with the ELO measure, so that one might expect one of the sorts to be excluded in an analysis.

In Figure 4 a 3D scatterplot can be observed, plotting 1000 elements of eloHK against PontH against HomeS, chosen at random from the complete number of observations (figures of this graph "viewed" from each direction, i.e 2D scatters of the included variables, can be found in Appendix, Figure 16,17,18). The graphs confirm the notion conveyed by the correlation map: a strong correlation between the ELO measures and the point measures seem to be present, a weaker otherwise. Also the graphs give one the impression that no notable non-linear dependence is present in the data. Examination of plots over the other variables similar to those shown below, does not convey a different picture, as they look essentially as one could expect (strong correlation between the point measure and ELO measures, weak between the goal measure and the others, non notable non-linear dependence). We therefore omit any analysis investigating non-linear dependence.

Figure 4: Correlation graph of variables measuring home team strength



6.4 Construction

We begin by performing a forward, backward and stepwise selection method on the data. We use p-values of the coefficients as criterion for inclusion/exclusion. Since we begin with a quite extensive number of parameters, the p-value is set to 0.1, i.e quite restrictive. Literature suggests up to 0.25 *(Hosmer & Lemeshow, 2000, p. 95)* in order to not miss variables that together might add extensive information without doing so individually. As this is probably not the case of our data (the variables are shown to very well explain the data on their own), we can afford to put on some pressure on with a restrictive p-value criterion. The result is summarized in table1.

Table 1: Result of selection procedures

Backward	eloHome, eloAway, eloProb, HomeScore, AwayAllowed
Forward	eloProb
Stepwise	eloProb

Two facts are stated. First, the point measures does seem redundant, as they are excluded from all end results. This is in line with what to expect from the correlation analysis, where we established a strong correlation between the points and ELO measures. Second, the ELO measures perform the strongest, indicating at least either eloProb or the individual ratings should be part of the final model. The rest of the analysis will therefore focus on models with the point measures excluded and some or all of the ELO measures included.

Since the models surviving this filtering are of a rather manageable size, a manual assessment will constitute the analysis. We will compare seven different models, covering alternatives for intuitive models along with the model given by the backward elimination . They will be compared through different measures of their strengths, being:
- AIC
- Stability: will the coefficients be approximately the same when estimating over section 2? The strength in this respect will be measured by adding up the absolute value of the percentage changes of the coefficients, and dividing by the number of explanatory variables.
- Predictive power: A predictive AIC value is listed for each of the models, that is; one calculate the Likelihood of section 2 using the coefficients estimated from data1. This measures how well future data is explained by the model. Also, plotted in Figure 6 are the estimated probabilities against their corresponding frequencies as they occur in section2. The predictive intervals are set to 0.05. Thus, all outcomes with a predicted probability of between 0 and 0.05, 0.05 and 1, etc, are pooled, and the frequency at which they actually occur calculated.
- Economic measure: which model will generate most money (least loss) when playing on section 2, and which model will

have the clearest trend of doing so (least risk)? The strengths in this respect will be measured by simply plotting the development of capital invested in the model in the beginning of section 2 against time (Figure 5), when playing on the maximal odds of the ones provided by the bookmakers. That is, for each game and alternative, we place 1 unit on the maximum of the odds among those higher than our own estimated fair ones (if any at all).

The results from the analysis are shown in Table 2.

Table 2: Model comparison

Model	AIC	Predictive AIC	Stability
eloHome, eloAway	5868.3	6381.377	1.007
eloHome, eloAway, Goals	5874.393	6414.846	1.152
eloProb	5868.477	6392.664	0.979
eloProb, Goals	5874.74	6428.138	0.658
eloHome, eloAway, eloProb	5865.121	6403.445	0.652
eloHome, eloAway, eloProb, Goals	5870.656	6437.703	0.794
Backward elimination result	5864.108	6428.406	0.609

Figure 5: Profit/Loss curves using different models

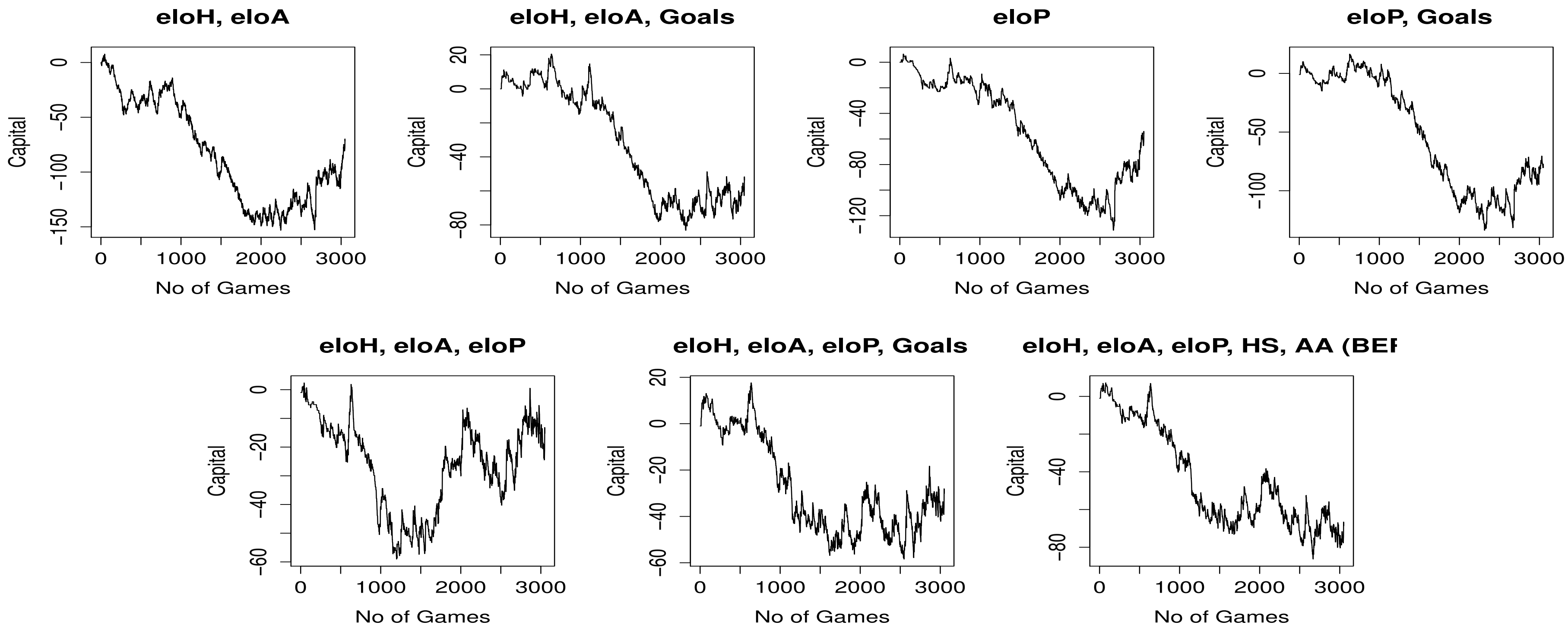


Figure 6: Observed frequencies over estimated probabilities for different models

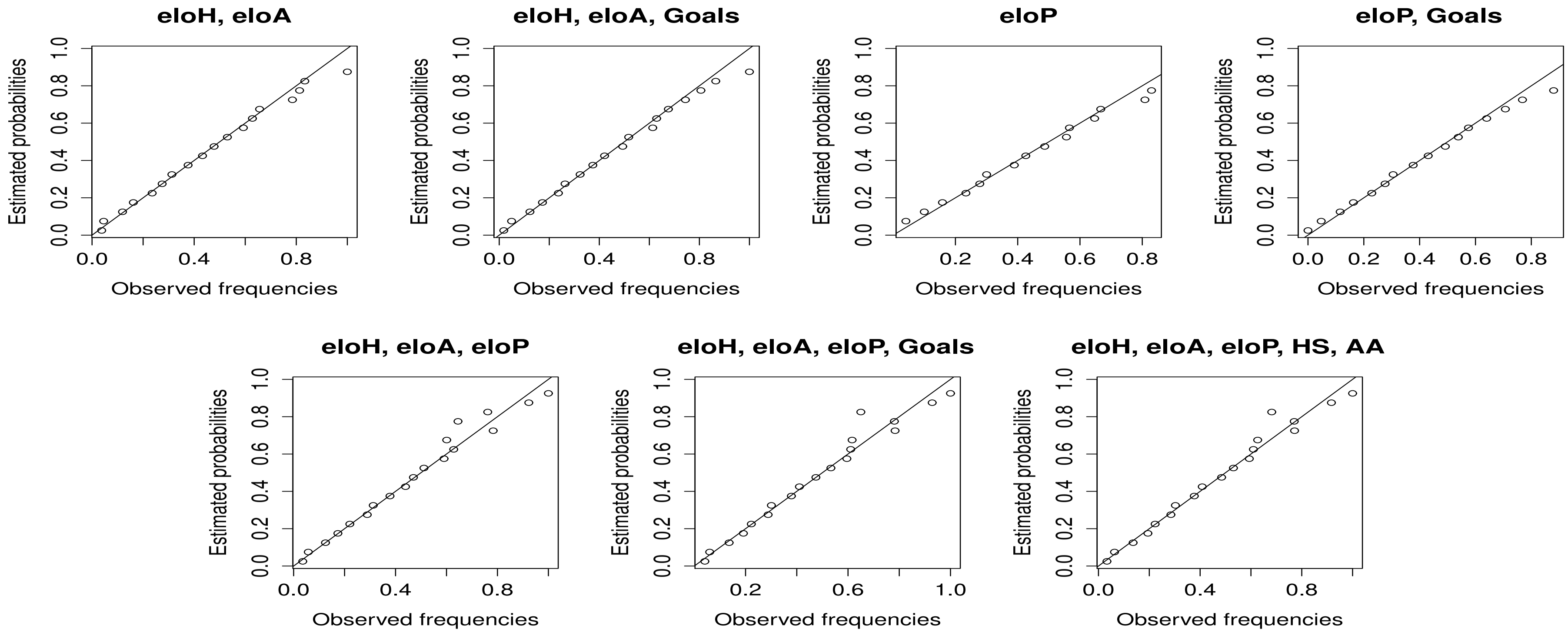


Table 3: Coefficients for model eloHome, eloAway

Coefficients :

	Estimate	Std. Error	t-value	Pr(>|t|)	
A:(intercept)	-0.6434946	0.0511625	-12.5775	< 2.2e-16	***
D:(intercept)	-0.4715169	0.0471666	-9.9968	< 2.2e-16	***
A:eloHome	-0.0661225	0.0064980	-10.1759	< 2.2e-16	***
D:eloHome	-0.0405393	0.0059202	-6.8476	7.511e-12	***
A:eloAway	0.0704762	0.0062159	11.3380	< 2.2e-16	***
D:eloAway	0.0353858	0.0060087	5.8891	3.884e-09	***

6.5 Choosing a model

The model with the lowest AIC value and best stability measure, the backward elimination result (BER), seems rather counter intuitive as it encompasses two goal measures without consideration of the other two. Its predictive AIC also supply us with a reason for excluding it as a choice, as it is highly unsatisfactory. The BER is only moderately better than its closest follower with respect to the AIC: the model encompassing all ELO measures. Its profit/loss curve is quite preferable among the observed ones, and it has the second best measure of stability and third best predictive AIC. The model even sustains a hypothesis test of the type $eloPr = 0$ against $eloPr \neq 0$. There are however some disturbing facts about this model. First, the profit/loss curves should how tempting it now may be not serve as a primarily source for decision making, as it is influenced not only by model accuracy, but also of how the odds are set. This makes the profit curves hard to evaluate distinguished from luck. Moreover, including all three ELO measures does not make much sense from an intuitive point of view. The individual ratings and the ELO-Probability are very correlated (see correlation map), as they are essentially just two ways of measuring the exact same information. One can of course imagine the information in the ratings are best extracted through this setup, but it seems rather ad hoc. At last, even though it has the third best predictive AIC, it is still performing considerably lesser than the first and second model in this respect, and inspecting the frequency graphs, one indeed sees that it does not perform very well in predicting high probabilities. Thus we drop the eloProb measure and obtain the model consisting of only the individual ratings. This one has the third best AIC value, its graph of frequencies is among the top performers, it has the best stability measure, and most importantly, the predictive AIC is considerably stronger than even its closest follower, the eloProb model. The eloProb model has essentially the same AIC and Stability measure but lesser predictive AIC, and is hence excluded in a comparison. Since the purpose of our model is to predict outcomes, the predictive AIC should be emphasized in the decision-making, and so the model consisting solely of the ratings will constitute our choice of model for the first data set. It is simple and intuitive, encompassing a lot of information in the explanatory variables rather than through a complicated model setup. The remaining models including the goal measures are all excluded as alternatives, as none of them passes a hypothesis test of the type

$\boldsymbol{H_0} : \boldsymbol{\beta_{goals}} = \mathbf{0}$ against $\boldsymbol{H_0} : \boldsymbol{\beta_{goals}} \neq \mathbf{0}$.

In Figure 3 a summary table is shown for the chosen final model. Similar summary models can be observed in the appendix for a couple of other models, Table 6 and Table 7.

6.6 Model behavior

Having settled with a model, we want to investigate how it behaves intuitively with respect to movements in our explanatory variables. For example, what happens to the predicted probabilities when fixing eloAway while increasing eloHome? Obviously we want the probability for a home win to increase, and the probability for away win to decrease. But what about the draw alternative? The most intuitively appealing scenario would be for the probability for draw to increase whenever eloHome moves towards eloAway, as the teams are becoming more "equal", while declining otherwise, as the teams are becoming more "unequal". Lets investigate how the chosen model behaves:

Let

$I_1, I_2, D_1, D_2, A_1, A_2 \in \mathbf{R}$,

eloHome=x
eloAway=y

The probabilities as predicted by the Multinomial Logistic Regression are of the form:

$P[H] = \frac{1}{1+exp(I_1+D_1x+D_2y)+exp(I_2+A_1x+A_2y)}$

$P[D] = \frac{exp(I_1+D_1x+D_2y)}{1+exp(I_1+D_1x+D_2y)+exp(I_2+A_1x+A_2y)}$

$P[A] = \frac{exp(I_2+A_1x+A_2y)}{1+exp(I_1+D_1x+D_2y)+exp(I_2+A_1x+A_2y)}$

One sees straight away that with $D_1, A_1 < 0$, the probability for home win increases for any fix eloAway when increasing eloHome. Consider now the draw alternative. Call the function calculating its probability z. Using the quotient rule for derivatives, we get, after some simplifications:

$\frac{dz}{dx} = \frac{D_1exp(I_1+D_1x+D_2y)+exp(I_1+D_1x+D_2y)exp(I_2+A_1x+A_2y)(D_1-A_1)}{\lambda}; \lambda \in \mathbf{R}_+$

Setting the derivative to zero gives:

$D_1 + exp(I_2 + A_1x + A_2y)(D_1 - A_1) = 0 \Leftrightarrow \frac{D_1}{A_1-D_1} = exp(I_2 + A_1x + A_2y) \Leftrightarrow$

$\frac{ln(\frac{D_1}{A_1-D_1})-I_2}{A_1} - \frac{A_2}{A_1}y = x$

With our estimates we get:

$-16.69 + 1.066y = x$

Furthermore, straightforward calculations gives:

$\lim_{x\to\infty} z(x, y) = \lim_{x\to-\infty} z(x, y) = 0$

for any given y. We know the multinomial functions are nonnegative and limited upwards by 1. These facts speaks of a bell shaped curve in x for any y, with the crest in three dimensions running over just mentioned line in the x,y-plane. These facts are in tune with our intuitive preferences, except for the rather low intercept. A natural explanation for this result is however found in the home team advantage: one can imagine the teams being "most" equal when the teams has a an ELO-difference of approximately 16.69 units and the lesser team plays at home, enjoying the advantage of doing so. Taking his into consideration, the function behaves essentially intuitively in the sense outlined above. Doing the same calculations for the away win probability function yields a negative sign within the log-function for calculating zero derivative points, implying no local extreme-value points present. We conclude the function must be monotone, also in line with our intuitive preferences, as we would like the probability for away win to decrease for any given y, when increasing x (see the graphs for confirmation the monotonicity is in the "right" direction). We plot the graphs with x and y ranging from -100 to 100 (by far covering all observed values of ELO-ratings), to be seen in Figure 7 and Figure 8. Our calculations show nothing unexpected occurs outside these ranges. A similar analysis for eloAway yields nothing unexpected, it follows our intuitive preferences equally well.

Figure 7: Draw probability function, the y-axis straight into the picture, the x-axis straight to the right

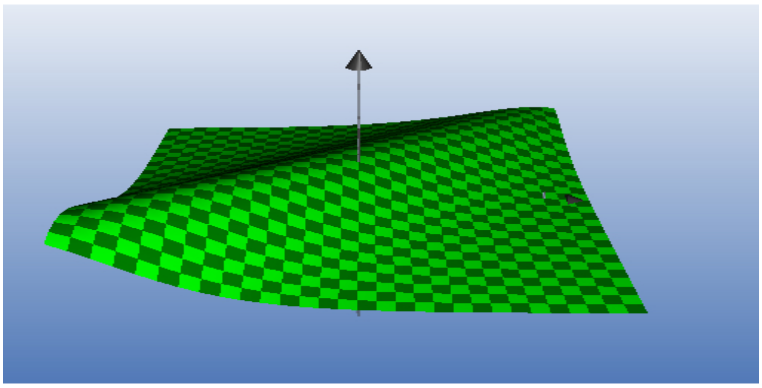
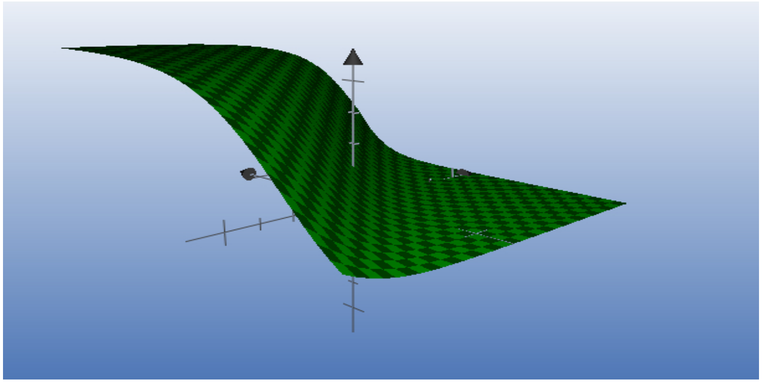


Figure 8: Away probability function, the y-axis being visible left, the x-axis straight in and right



7 Building model 2

7.1 Overview

The explanatory variables in data2 are:

HomeSco - Goals made by the home team during the previous six games
HomeAll- Goals allowed by the home team during the previous six games
AwaySco - Goals made by the away team during the previous six games
AwayAll - Goals allowed by the away team during the previous six games

PH - Average points scored by the home team during the previous six games
PA - Average points scored by the away team during the previous six games

eloH - Pre-match ELO rating for the home team
eloA - Pre-match ELO for the away team

eloProb - The "probability" generated in the ELO algorithm for a game

Shots measure:
HomeScoS - Shots on target made by the home team during the previous six games
HomeAllS - Shots on target allowed by the home team during the previous six games
AwayScoS - Shots on target made by the away team during the previous six games
AwayAllS - Shots on target allowed by the away team during the previous six games

Corners measure:
HomeScoC - Corners made by the home team during the previous six games
HomeAllC - Corners allowed by the home team during the previous six games
AwayScoC - Corners made by the away team during the previous six games
AwayAllC - Corners allowed by the away team during the previous six games

Odds for Home, Away and Draw, for 6 different bookmakers, quoted the day before the games, being:
- BET365
- Gamebookers, (GB)
- Interwetten (IW)
- Ladbrokes (LB)
- Sportingbet (SB)
- William Hill (WH)

7.2 Univariate analysis

Like before, we would like to reduce the rather vast number of variables before proceeding, but like before, a quick univariate analysis (univariate by measures like in model 1, odds in groups after bookmaker) of each of the explanatory measures confirms that they cannot be excluded individually; they all have p-values of less than 0.01, below any reasonable significance level. The analysis hence has to start from the multivariate situation.

7.3 The odds

It is a hard question whether to include the odds in the model or not. Since we would like to use it to discover odds that are priced wrong, this intuitively seems like a bad idea: when an odds is wrongly priced, a model would estimate a wrong probability, which would then be used to assess whether the same odds is wrong, giving an inconclusive result. But there is more here than first meets the eye. In order to understand the dynamics, it is instructive to imagine two scenarios.

First assume a very large part of the odds are corresponding to "correct" probabilities. The model would then, for given odds, estimate a probability very close to the one implied by the odds, i.e the "true" probability. Therefore, if the odds was wrongly priced i.e the implied probability wrong, the model would predict a wrong probability, making the model useless in detecting a wrongly priced odds, even if ever so suitable for predicting probabilities for the outcomes in general.

In the second scenario, the odds does not infer the correct probabilities, but very well explain the correct probabilities, that is; they are highly suitable as explanatory variables in a model. This is the most likely scenario: betting sites ought to set bets so that their implied probabilities are highly correlated with the 'correct' ones, being slightly higher in order for them to make profit. This would imply that a model over the odds in general would return a slightly lower estimated probability than the implied. If then an odds is wrongly priced in favor of the player, the implied probability would be lower than the real, and hence the estimated probability even lower, making a player demand disproportionally high odds, probably causing him to miss the opportunity. The only case were a strategy based on the odds could actually generate a statistical profit would be if the odds for a certain interval or so where systematically too highly priced. A model would then be able to estimate the correct probabilities AND spot wrongly priced odds. Given the assumed expertise of the bookmakers in setting the odds, this event seems highly unlikely however; they would systematically lose money for odds set in this interval.

In Figure 9, the implied probabilities are plotted over the observed frequencies of events, for intervals of 0.05 in implied probabilities, for each of the six bookmakers. They seem to depict the second scenario. Especially, none of the sites seem to underestimate their probabilities in any systematic way. This implies odds should not be included in our model, since although they certainly very well explain the data, they are unable to spot the wrongly priced odds. However, this line of reasoning only holds for the case when all odds in question are from the same provider. Odds might still be suitable in assessing odds from different providers. An approach to assess this question is to build models based on odds from one provider, and try to detect wrongly priced ones from another. Three such graphs are shown below in Figure 10, building models on the odds from one site and, like previously, playing on the maximal of the odds provided by the others. Instead of using the odds directly as explanatory variables, we use the inverted odds, i.e their implied probabilities. As outlined by *Rosengren, 2012* this should be a better format of the information in the odds for applying the Multinomial Logistic Regression model (the relationship becomes linear rather than non-linear).

As seen, the models seem unable to correctly identify wrongly priced odds. This probably depends on the odds being highly correlated, see correlation map, Figure 11. If an odds is wrongly priced at one provider, it would tend to be so at the others too. Thus the games the algorithm do find are probably in general the same it would find if it played autonomously (a model on Interwetten odds playing on Interwetten odds for example). These games are according to previous reasoning not correctly spotted wrongly priced odds, instead they result out of imperfections in the mathematical structure of the model. The relationships are clearly linear between the implied probabilities and the frequencies (inspecting the graphs), but the model is not assuming perfect linearity, making it flawed in certain intervals.

7.4 The historic goals, points, corners and shots on target

With the results from the previous analysis as motivation, we choose to exclude the goal measure and the point measure, as they were not at all among the serious competitors for our choice of model. With the odds excluded as well, we are left with the ELO variables, the shots-on-target- and corners measures. According to the result we got from our optimization regarding the n-parameter involved in the historic measures, we choose the n for both the corner and the shots-on-target measure to 6. Thus we begin the multivariate analysis with the three ELO-variables, with the parameters as suggested by the parameter optimization over data2, the four corners-variables and the four shots on target-variables with $n = 6$.

7.5 Correlation analysis

Like before, we begin with a correlation analysis of our set of variables. Again, we draw the correlation map, seen in Figure 12.

According to the graphs in Figure 1, there should be some degree of explanatory power in the shots-on-target and corners measures. The map shows that they are not intercorrelated, and surprisingly independent from the ELO-variables. This suggest the new variables contain new information, valuable for building the model of our purpose. Like previously, we investigate scatter plots over the variables against each other in order to detect non-linear dependence. They do not show any notable such, and we abstain from doing any analysis of non-linear dependence.

Figure 9: Frequencies over implied probabilities for the 6 bookmakers

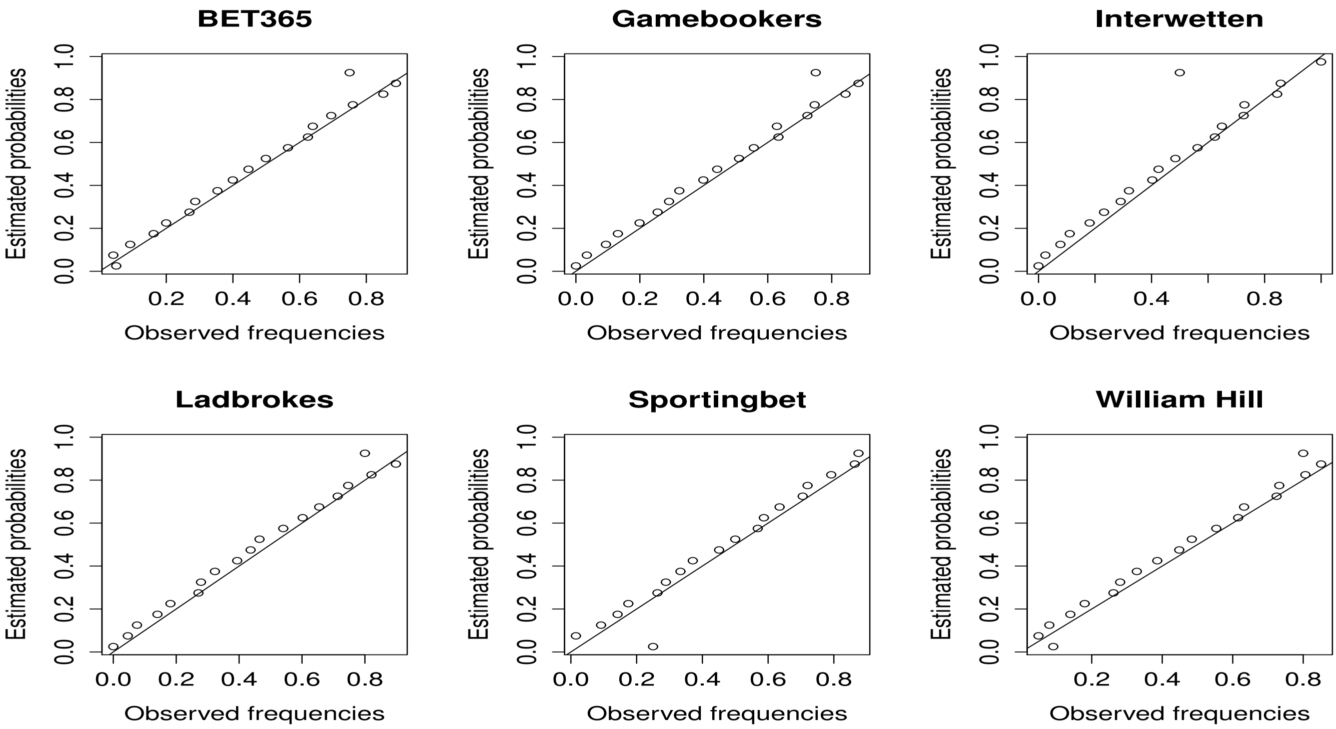


Figure 10: Profit/Loss curves resulting from using models built on odds from a specific bookmaker

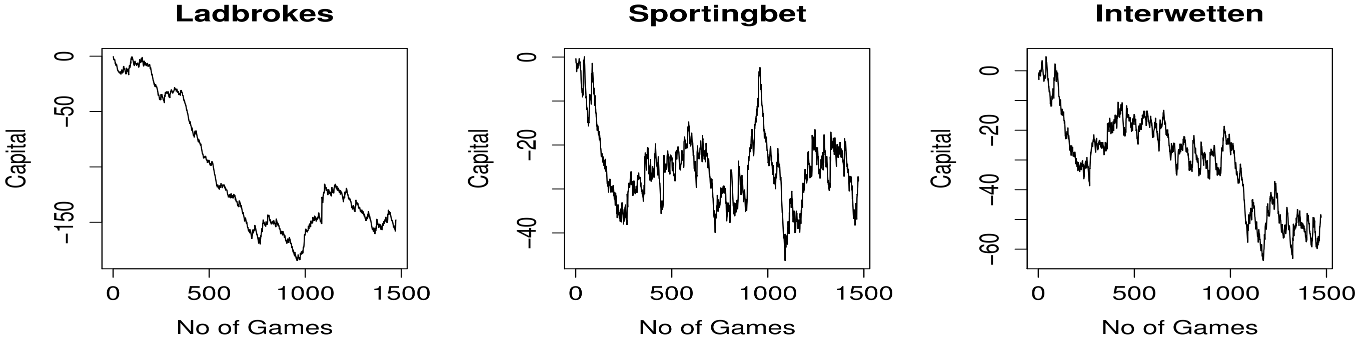


Figure 11: Colored map illustrating the correlation between odds for all alternatives at all bookmakers

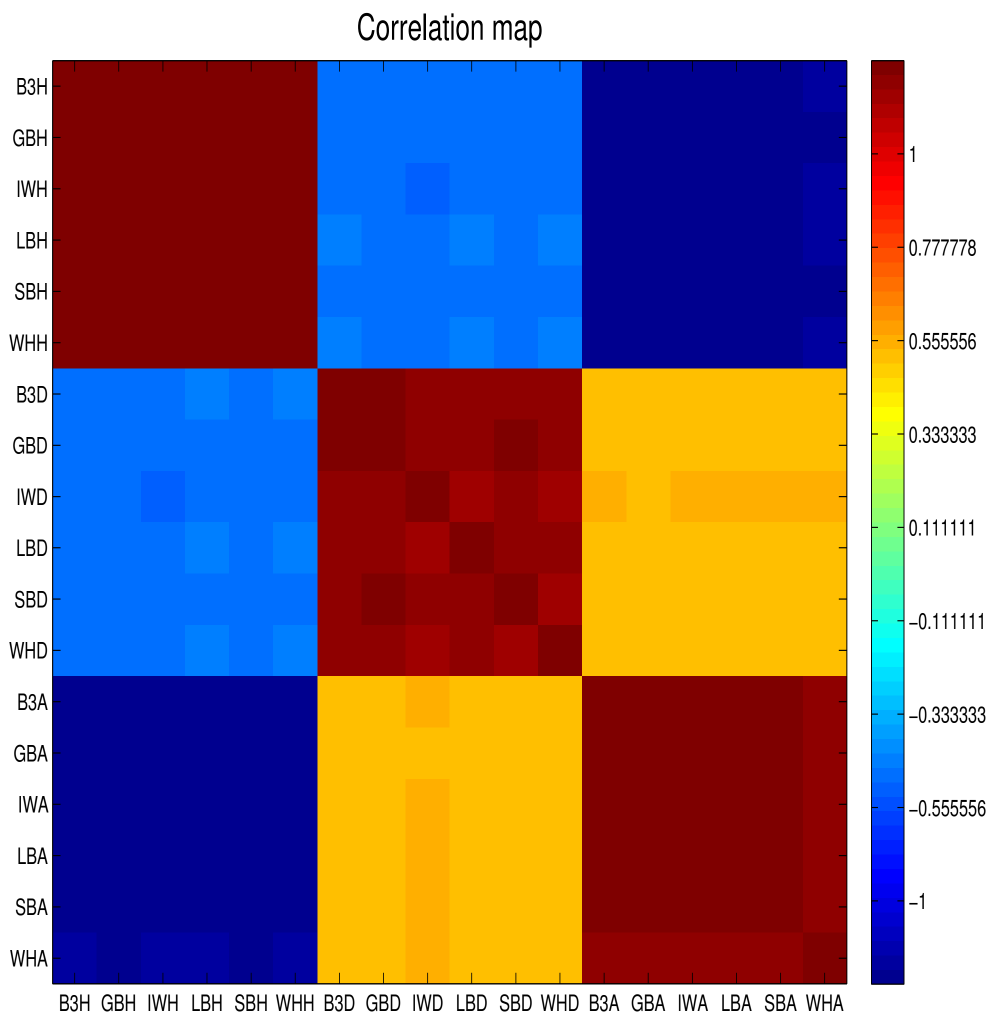
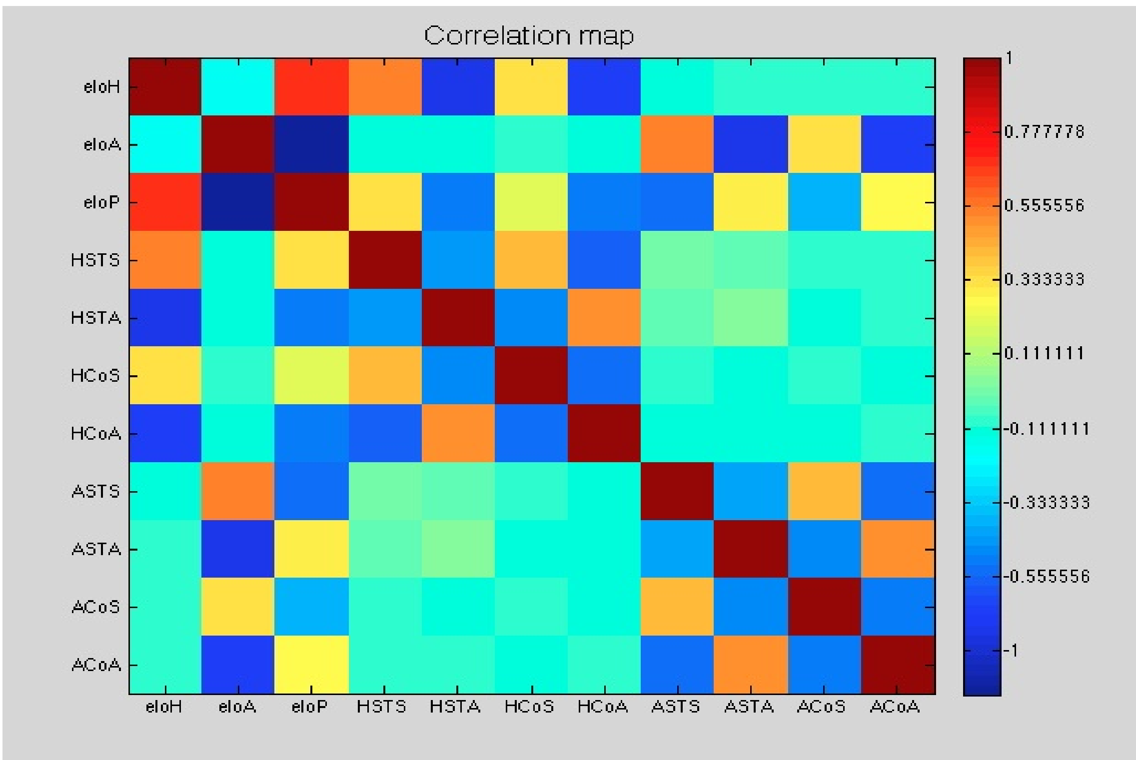


Figure 12: Colored map illustrating the correlation between all explanatory variables of model 2



7.6 Construction

Like before, we begin by doing a forward, backward and stepwise regression on the data. The p-value is set to 0.1; since the variables have been shown to well explain the data on their own, we still afford to put some pressure on the algorithm, even though the variables are not as correlated as before (setting a more conservative p-value results in almost all variables being included). The result is shown in Table 4.

Table 4: Result of selection procedures

Backward (BER)	eloHome, eloAway, AwayScoS, AwayAllS, HomeAllC
Forward (FER)	eloProb, AwayAllS, HomeAllC
Stepwise (SER)	eloProb, AwayAllS, HomeAllC

Like before, the ELO variables seem to perform well. With this fact and the result from the previous analysis as motivation, we again are only going to consider models with at least one of the two ELO measures included. Different from before, here two more variables are part of all end results: AwayAllS and HomeAllC. These are however not going to be part of all models assessed, since they have a more dubious joint interpretation.

We hence perform a manual assessment of suitable models, all including at least one ELO measure. We also investigate the two models chosen by the selection methods above. We draw a similar table to before in Table 5 below.

Table 5: Model comparison

Model	AIC	Predictive AIC	Stability
eloHome, eloAway	3134.13	3262.92	1.088
eloHome, eloAway, Corners	3139.10	3273.88	1.39
eloHome, eloAway, Shots	3130.51	3299.98	1.07
eloHome, eloAway, Corners, Shots	3139.713	3318.226	3.157607
eloProb	3133.857	3263.792	1.575843
eloProb, Corners	3138.164	3272.483	1.577182
eloProb, Shots	3131.44	3296.545	1.385126
eloProb, Corners, Shots	3140.271	3314.704	3.462335
eloHome, eloAway, eloProb	3138.043	3266.906	2.336316
eloHome, eloAway, eloProb, Corners	3143.008	3277.709	2.178776
eloHome, eloAway, eloProb, Shots	3134.331	3304.337	2.180876
eloHome, eloAway, eloProb, Corners, Shots	3143.487	3322.487	3.764928
BER	3131.61	3272.871	1.024276
SER and FER	3133.597	3261.84	1.788705

7.7 Choosing a model

The table points at a rather conclusive result. The final model from our previous analysis, consisting of only the individual ratings, obviously outperform most of the models.

It has the second best predictive AIC, being only a unit or so behind the SER and FER model. The SER also has a slightly better AIC value. It is however not preferable for three reasons. First it is counterintuitive. If one decides a measure contain explanatory power, it seems incoherent and rather ad hoc to include just a subset of the variables comprising the measure. Second, the model is very unstable according to our stability measure, the coefficients changing on average 78 percent. Third, it is not much stronger in the other criteria than other more preferable models.

Our model performs rather well also with respect to the AIC value. Although only at the sixth best place by this criteria, the difference from its superiors are rather small. Moreover, the superiors perform considerably lesser with respect to the predictive AIC, except the model consisting of only the eloProb measure, being only a unit or so behind. However, it is very unstable; the coefficients changing value with approximately 56 percent over the periods.

The model is also third best in the stability measure, behind the BER model and the model consisting of the ratings and the shots measure. This model has a better AIC, but very unsatisfactory predictive AIC, prompting its exclusion. The BER is already excluded above.

Having argued for these exclusions, we are left with the model consisting of the individual ratings as the obvious alternative for our final choice.

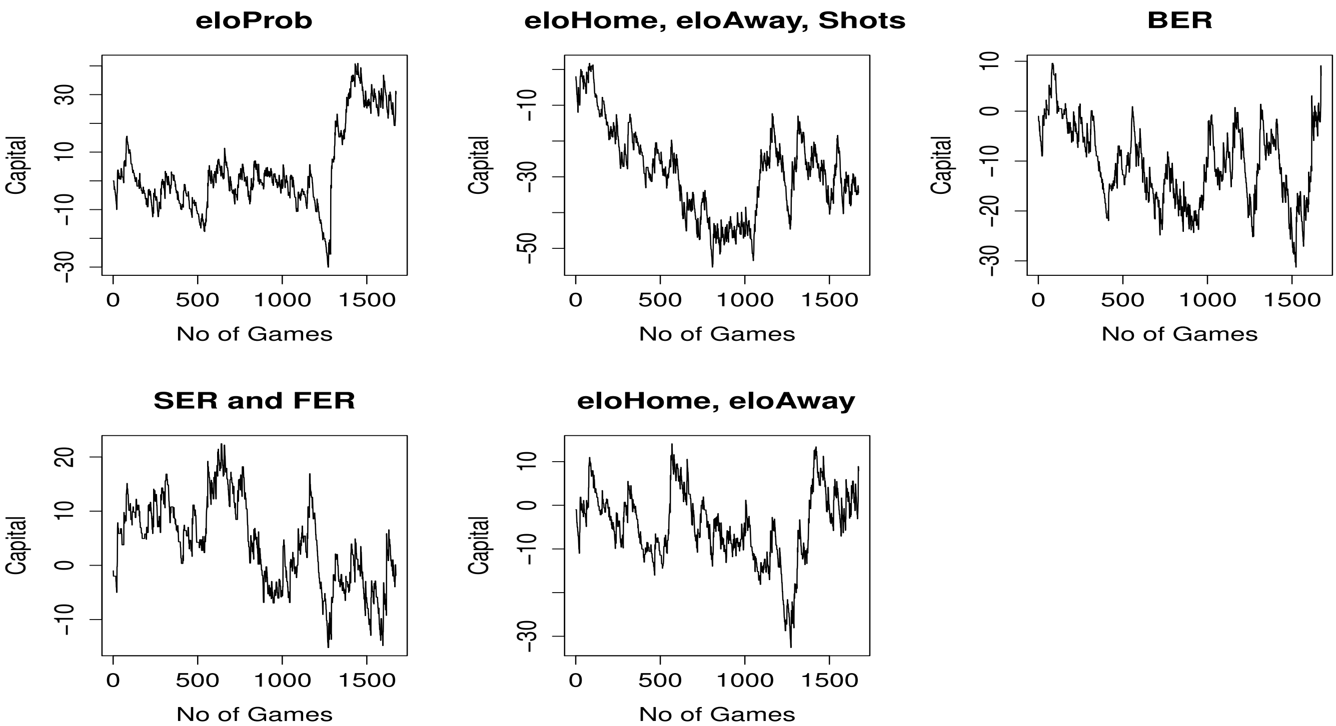
We here omit the frequency graphs for the different models, as they do not show anything unexpected, and since the predictive AIC basically measure the same thing. We do however plot the capital development when playing on some of the more competitive models on the fourth quarter of the data, with coefficients estimated over the third quarter (data2). These are seen in Figure 13. Also a summary table of the model coefficients over data2 is shown in Table 6.

Table 6: Coefficients for model eloHome, eloAway

Coefficients :

	Estimate	Std. Error	t-value	Pr(>|t|)	
A:(intercept)	-0.7287217	0.0697851	-10.4424	< 2.2e-16	***
D:(intercept)	-0.5961230	0.0629282	-9.4731	< 2.2e-16	***
A:eloHome	-0.0800516	0.0079578	-10.0595	< 2.2e-16	***
D:eloHome	-0.0291449	0.0062917	-4.6323	3.617e-06	***
A:eloAway	0.0670934	0.0065235	10.2848	< 2.2e-16	***
D:eloAway	0.0242463	0.0067326	3.6013	0.0003166	***

Figure 13: Profit/Loss curves resulting from using models built on data2



8 Conclusions

As stated in the abstract, concluding efficiency of the odds-market with our findings is in fact somewhat equivalent to no conclusions at all. If we had managed to turn consistent profits with our models, we could indeed conclude inefficiency, whereas now we can only conclude efficiency with respect to our included information and in our analyzed market. We do however feel comfortable in making some extrapolation. First, the information we have taken into account is rather extensive; coming up with quantitative and relevant measures explaining the data in any way our measures do not feels like a hard task. Second, as also stated in the abstract, it seems reasonable that the entire history of the Premier League comprises a fairly good representation of the market situation in general.

Hence, we feel comfortable in believing the soccer odds market (for these type of bets) is quantitatively information efficient, by which we mean the market seems to more or less mirror all available and relevant quantitive information. That is, in order for a more accurate model to be built, one need knowledge of the actual game dynamics and qualitative inputs. The bookmakers are likely to take such information into consideration, given their professional position, which would explain our poor financial results. The profit/loss curves however seem to at least perform rather better using the second model. This is probably due to increasing competition in the bookmaker industry; as the margins have become lower, the odds have become better and hence returning more to any betting strategy applied. This effect is also seen in the profit/loss curves for model1, the losses stabilizing with time.

From a more purely academic viewpoint, we feel this investigation is interesting in another respect. The ELO measures comprises essentially constructed variables, the construction depending heavily on several parameters. It is not a very usual model building procedure estimating such parameters in the way done here, but a very interesting one in terms of model performance. The optimization procedures in a way lets one maximize the extraction of information contained in an already existing explanatory variable. The crude goal measure was excluded from all models, whereas the ELO ratings, being built on the very same numbers, tended to perform very well. This comes at no cost of over-fitting: we indeed enhanced the explanatory power of the model even in terms of prediction, as seen by the predictive AIC value comparisons. When an intuitive interpretation is at hands for constructing explanatory variables over already existing ones, this report indicates that it can indeed be a fruitful approach.

9 Appendix

Figure 14: **Frequencies of occurred events plotted over points and goal differences**:

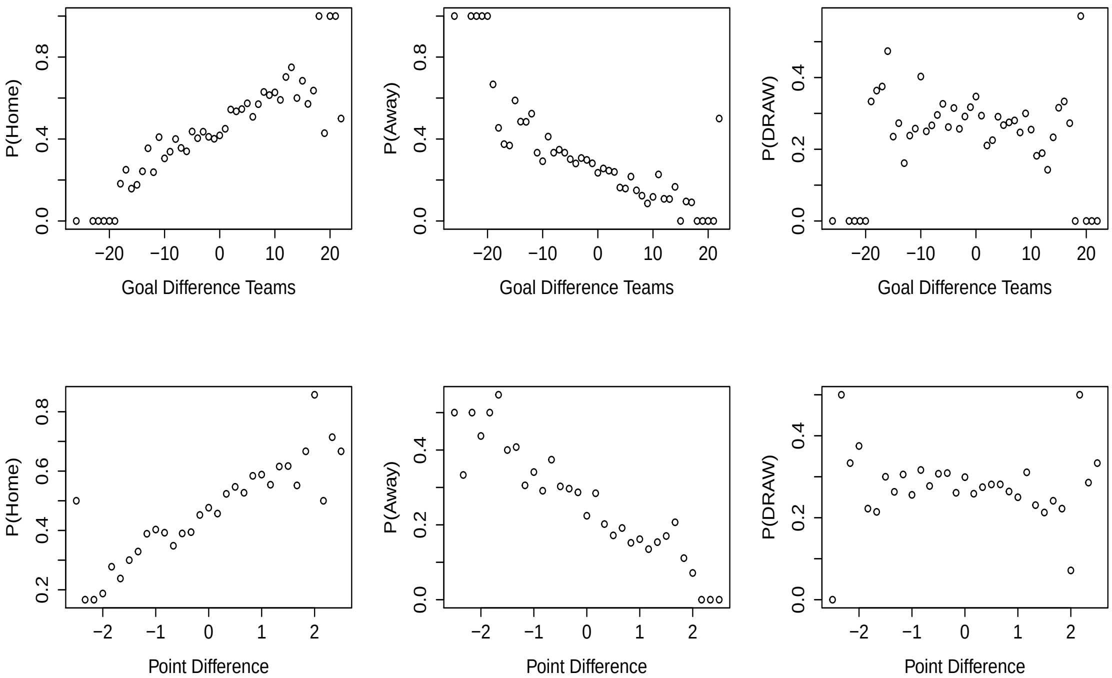


Figure 15: ELO rating development over time

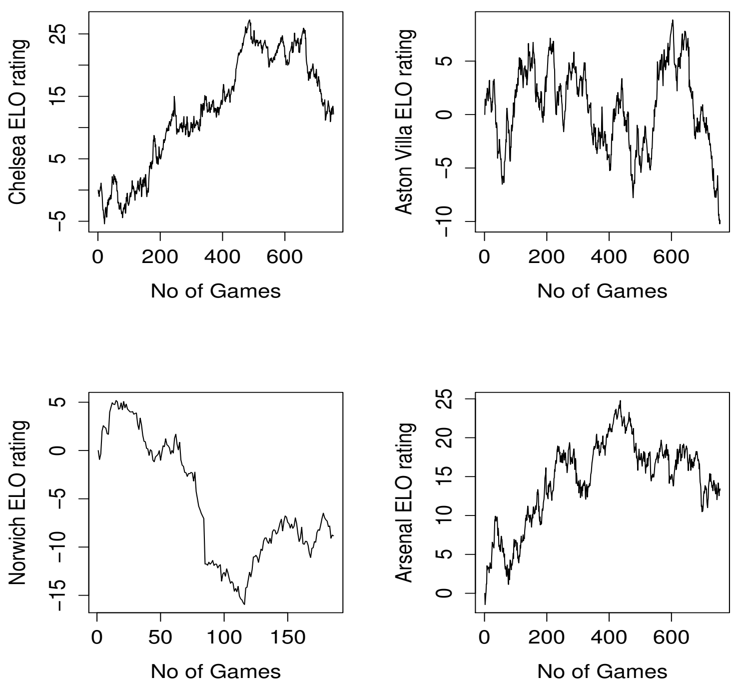


Figure 16:

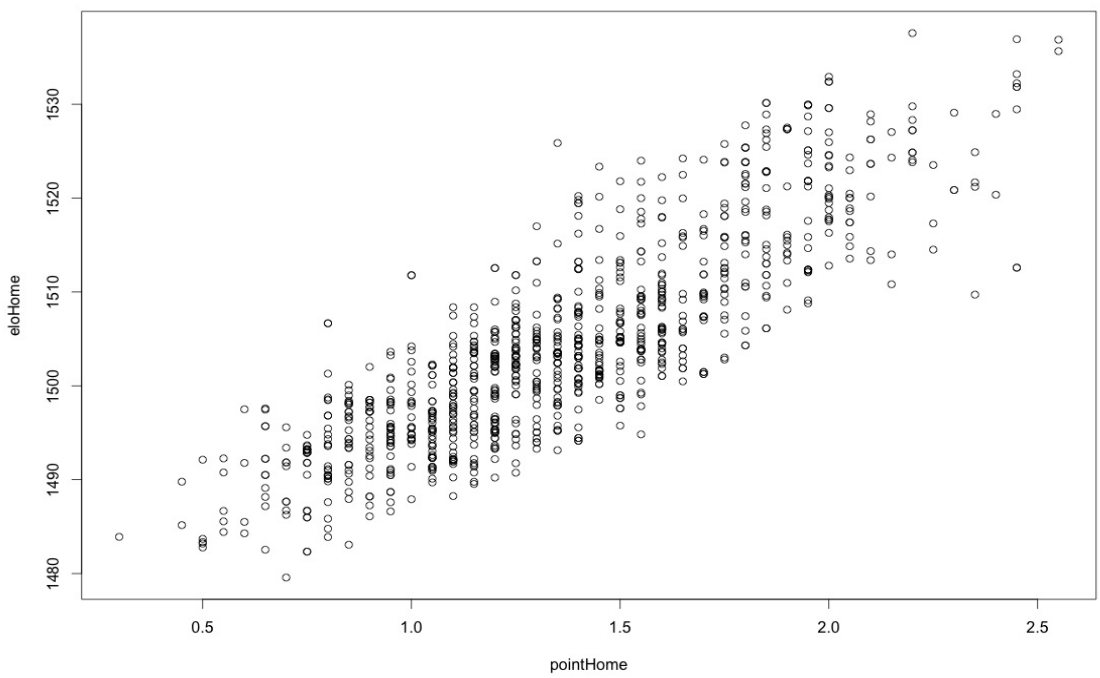


Figure 17:

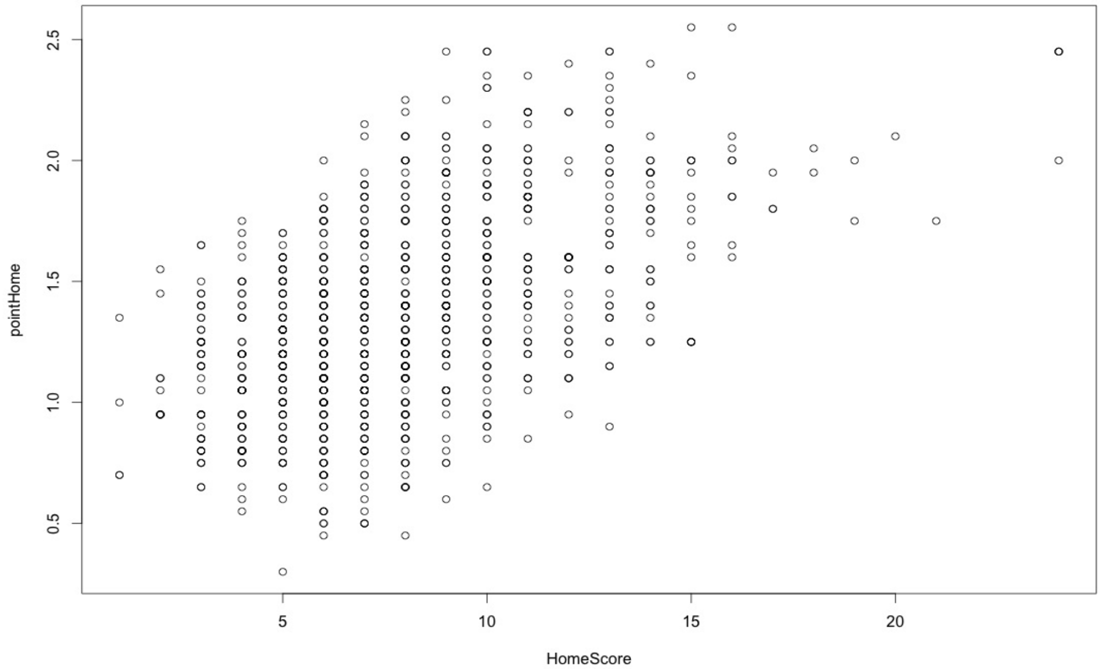


Figure 18:

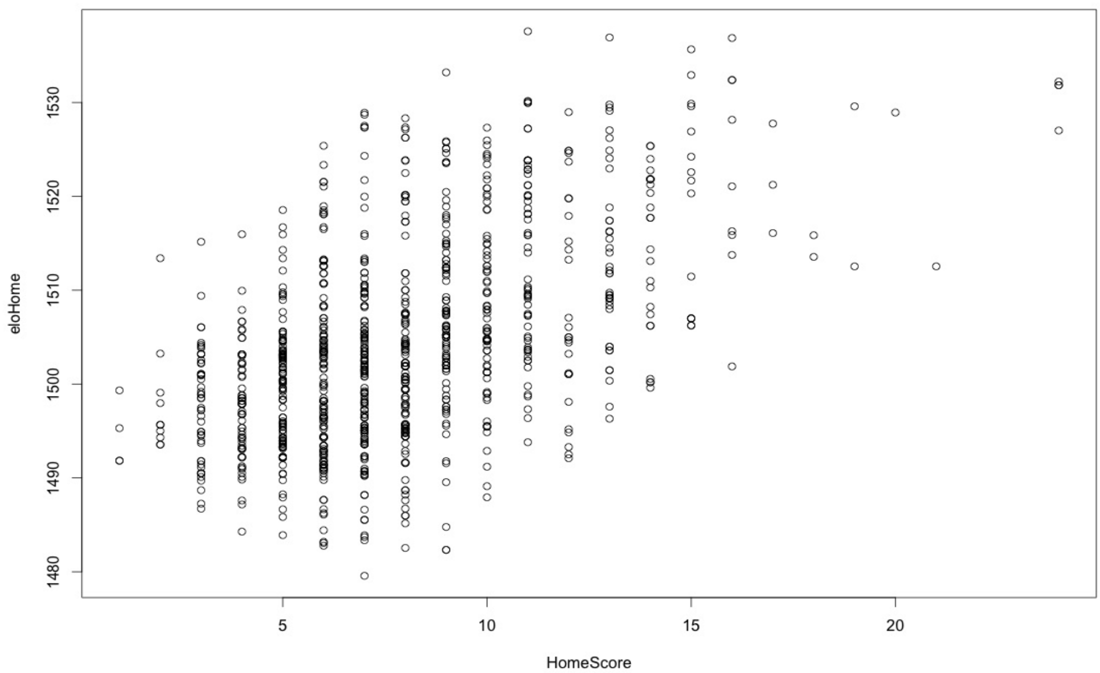


Table 7: Coefficients for model eloHome, eloAway over data1

Coefficients :

	Estimate	Std. Error	t-value	Pr(>|t|)	
A:(intercept)	-2.391920	1.639925	-1.4586	0.14469	
D:(intercept)	1.702809	1.567129	1.0866	0.27722	
A:eloHome	-0.121068	0.052328	-2.3136	0.02069	*
D:eloHome	0.027798	0.049847	0.5577	0.57707	
A:eloAway	0.124736	0.052108	2.3938	0.01668	*
D:eloAway	-0.032902	0.050047	-0.6574	0.51091	
A:eloProb	3.499553	3.276153	1.0682	0.28543	
D:eloProb	-4.350126	3.134975	-1.3876	0.16526	

Table 8: Coefficients for model eloProb over data1

Coefficients :

	Estimate	Std. Error	t-value	Pr(>|t|)	
A:(intercept)	1.50244	0.14307	10.5012	< 2.2e-16	***
D:(intercept)	0.72951	0.14208	5.1344	2.83e-07	***
A:eloProb	-4.28477	0.28604	-14.9796	< 2.2e-16	***
D:eloProb	-2.40258	0.26434	-9.0889	< 2.2e-16	***

References

[1] Arpad Elo (1978), *The rating of chessplayers, Past and Present*, Arco Pub.

[2] David W. Hosmer & Stanley Lemeshow (2000), *Applied Logistic Regression*, John Wiley & Sons

[3] Lars Magnus Hvattum & Halvard Arntzen (2010), *Using ELO ratings for match result prediction in association football*, International Journal of Forecasting

[4] Sebastian Rosengren (2012), *Att med multinomial logistisk regression frklara sannolikheter i fotbolls-matcher*, Department of Mathematics at Stockholm University