Aalto University
School of Business

# Forecasting football match results - A study on modeling principles and efficiency of fixed-odds betting markets in football

Quantitative Methods of Economics

Master's thesis

Olli Heino

Ville Sillanpää

2013

Department of Information and Service Economy
Aalto University
School of Business

AALTO UNIVERSITY SCHOOL OF BUSINESS
Department of Information and Service Economy
Master's Thesis
Ville Sillanpää & Olli Heino



# ABSTRACT

**Objectives of the Study**

This thesis is about the statistical forecasting of (European) football match results. More specifically, the purpose of this thesis is to assess how a statistical forecast model that uses only publicly available information fares against public market odds in forecasting football match outcomes.

**Academic background and methodology**

The forecasting of sports results has been widely researched because it provides important insight into how betting markets operate. Football and betting associated with it has been the most popular topic because of the global popularity of the sport and because the betting markets associated with it capture large annual turnover.

In spite of research by numerous authors, there is still room for improvement in terms of developing more accurate forecast models. Therefore, we contribute to existing literature by developing a regression model for forecasting football results. We assess the model's performance with forecast accuracy measurements and betting simulations. The principal idea of the model is based on the ELO rating system which assigns relative performance ratings to teams.

**Findings and conclusions**

In terms of accuracy measurements and betting simulations, the model developed in this thesis is able to match or surpass the results of existing statistical models of similar build. The measurements also indicate that the model can on average match the accuracy of the forecasts implied by the publicly quoted odds. However, the model is unable to generate positive betting returns. Together these results indicate that the publicly quoted odds for extensively betted football matches are slightly inefficient, but that this inefficiency does not make statistical betting algorithms consistently profitable. The results also indicate that historical league match results are the most important components of a statistical football forecast model, and that supplementing these components with other data yields only modest improvements to forecast accuracy.

**Keywords**

football results forecasting, ordered logit regression, ELO rating system, betting market efficiency

AALTO-YLIOPISTON KAUPPAKORKEAKOULU
Tieto- ja palvelutalouden laitos
Pro Gradu-tutkielma
Ville Sillanpää & Olli Heino

TIIVISTELMÄ
11.11.2013

# ABSTRAKTI

**Tutkimuksen tavoitteet**

Tutkimuksen tavoitteena on kehittää tilastollinen ennustemalli (eurooppalaisen) jalkapallon ottelutulosten ennustamiseen. Kehitetty malli hyödyntää pelkästään julkisesti saatavilla olevaa informaatiota. Mallin tehokkuutta testataan tutkielmassa julkisten vedonlyöntikertoimien avulla.

**Kirjallisuuskatsaus ja metodologia**

Urheilutapahtumien tulosten ennustamista on tutkittu laajalti, sillä ennustemenetelmien tutkiminen avaa oven vedonlyöntimarkkinoiden toiminnan tutkimiseen. Jalkapallo ja siihen liittyvä vedonlyönti ovat historiallisesti olleet suosittuja tutkimuskohteita, koska jalkapallo on suosittu laji ja koska jalkapalloon liittyvien vedonlyöntimarkkinoiden liikevaihto on todella suuri.

Vaikka tulosten ennustamista ja vedonlyöntiä on tutkittu laajalti, on tilastollisissa ennustemalleissa edelleen paljon kehitettävää. Tässä tutkielmassa kehitetään regressiomalli, jolla ennustetaan jalkapallo-otteluiden lopputuloksia. Mallin suorituskykyä mitataan ennustevirhemittareilla sekä vedonlyöntisimulaatioilla. Mallin perusideana on joukkueiden suhteellisia tasoeroja mallintava ELO-pistejärjestelmä.

**Tulokset ja päätelmät**

Tulosten perusteella tutkielmaa varten kehitetty malli on parempi tai yhtä hyvä kuin kirjallisuudessa aiemmin esitellyt samankaltaiset mallit. Tulosten perusteella mallin ennustetarkkuus vastaa internetissä tarjolla olevien kertoimien implikoimaa vedonvälittäjien keskimääräistä ennustustarkkuutta. Tulosten perusteella mallin avulla ei kuitenkaan voi lyödä vetoa voitollisesti. Yhdessä nämä seikat osoittavat, että vedonlyöntimarkkinat voisivat olla tehokkaammat ja että tämä tehottomuus ei kuitenkaan mahdollista voitollista vedonlyöntiä tilastollisin keinoin. Tutkielman tulokset osoittavat myös, että rakentamamme mallin tärkein osa on liigaotteluiden otteluhistoria, ja että muun kuin sen käyttäminen ennustamisessa ei paranna mallin ennustustarkkuutta merkittävästi.

**Avainsanat**

jalkapallo-otteluiden tulosten ennustaminen, regressiomalli, ELO-pistejärjestelmä, vedonlyöntimarkkinoiden tehokkuus

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# 1 INTRODUCTION

This thesis is about statistical forecasting of football match results. More specifically, the purpose of this thesis is to assess how a statistical forecast model that uses only publicly available information fares against market odds in forecasting football match outcomes. Effectiveness of the model built in this thesis is assessed with forecast accuracy measurements and betting simulations. The data used to conduct the empirical tests of this thesis consists of match results from the English Premier League (EPL) as well as of several other related datasets.

## 1.1 Motivation for research

Forecasting of sports results has been widely researched because it can provide theoretically and practically important insight into how betting markets operate. The motivation to research betting markets is therefore twofold: On one hand, academic audience is interested in the topic because forecast models provide an instrument for testing the efficiency of betting markets. On the other hand, more practically oriented audience finds the topic interesting because the forecast models can be combined with betting rules to create betting algorithms, which in turn can possibly be used to earn positive returns from sports betting. Professional football has been the sport of choice in this kind of research mainly because of its global popularity and because the betting markets associated with it capture large annual turnover (Hvattum & Arntzen, 2010, 461). Hence academic contributions to this subject are interesting to large audiences and they can have a significant impact on an industry with multi-million turnover.

In spite of extensive literature on the subject, the methodologies for forecasting match outcomes in professional football are still relatively underdeveloped in the academic literature. Notable recent research in this field includes Hvattum & Arntzen's work (2010) where ratings were first assigned to teams based on their historical performance, and then used in forecasting. The system that the authors used in their work, the ELO rating system, was originally developed by Arpad Elo (1979) for the purpose of measuring skill levels between

chess players. As another notable development, Goddard & Asimakopoulos (2004) have discovered that factors other than recent league performance also contribute to football team's future performance, and should thus be considered in forecasting. However, as reported by – for example – Hvattum & Arntzen (2010), none of the statistical models discussed in the literature have been able to consistently outperform markets.

This thesis builds upon these recent developments by constructing an ELO rating -based forecasting model. The model developed in this thesis is based on the rating-assigning technique introduced by Hvattum & Arntzen (2010). In this thesis we expand the authors' core idea by complementing the model in two ways: First, we extend the original variable specification to consider team-specific home advantage. Second, we extend the forecast model with variables that capture effects not directly associated with the recent league match history of a team. The resulting model should be robust in its use of historical results as well as comprehensive in its way of accounting for effects that influence future match performance indirectly.

## 1.2 Research objectives

In order to specify how our model build contributes to existing literature, we have broken the main task of this thesis into three research questions. These research questions are detailed in the following paragraphs.

> ***Question 1****: Is it possible to improve Hvattum & Arntzen's (2010) ELO rating –based variable by introducing team-specific home advantage into it?*

The first research question relates to the improvement we propose to Hvattum & Arntzen's (2010) model build. On one hand, the existence of home advantage in league football is well-documented (Nevill & Holder, 1999). On the other hand, there is also evidence that some part of home advantage is team-specific (Clarke & Norman, 1995, 515-516); effectively meaning that team's performance level at the team's home games is consistently different from its performance level at away games. As Hvattum & Arntzen's (2010) model does not model team-specific home advantage in any way, we propose a modified version of the model, which

accounts for team-specific home advantage explicitly. Thus the answer to our first research question is obtained by building the adjusted model and comparing its performance with the performance of the original model.

> ***Question 2****: Should other information than direct league match history be used in forecasting?*

The second research question relates to combining the modeling ideas introduced by Goddard & Asimakopoulos (2004) to the modeling ideas introduced by Hvattum & Arntzen (2010). On one hand, Goddard & Asimakopoulos have showed that the information not directly related to team's historical league performance is relevant to forecasting team's future league performance. On the other hand, Hvattum & Arntzen (2010) have introduced an ELO rating based algorithm for encoding team's performance level from the historical match results. As the ELO rating based encoding is shown to be superior to the encoding algorithm introduced by Goddard & Asimakopoulos (Hvattum & Arntzen 2010, 469), it would be interesting to see whether this ELO based model's performance could be enhanced by complementing the ELO rating based model build with variables that model the indirect effects in the spirit of Goddard & Asimakopoulos's (2004, 56) work. Thus, this research question is answered by constructing a forecast model with multiple variables and comparing its performance with the performance of the univariate ELO rating based model.

> ***Question 3:*** *Is it possible to build a model that would outperform fixed odds betting markets?*

The third research question is about assessing if the improvements we propose to the existing modeling practices result in forecast accuracy that would outperform the market odds. To our knowledge, no author has so far been able to build a forecast model that would produce forecasts superior to the forecasts produced by bookmakers – or in other words the market. To answer this research question we conduct two types of measurements: first, we compare our best model's forecast accuracy with the accuracy of forecasts produced by the betting markets. After this we conduct betting simulations, whose results help to assess how profitable our model would be if it is used for making betting decisions.

## 1.3 Results and conclusions

Our research yielded conclusive answers to all three research questions presented in part 1.2. These answers, and therefore the results of our research, are summarized in this part.

As an answer to the first research question, our results indicate that modeling team-specific home advantage does improve Hvattum & Arntzen's (2010) ELO rating based build, if certain issues related to cross-tracking home- and away performance can be accounted for. The magnitude of this improvement is, however, quite small, as our forecast accuracy metrics could not uniformly agree on whether the improved model was equally good or superior when compared with Hvattum & Arntzen's (2010) original build. Therefore our results are, to an extent, in line with the results Clarke & Norman (1995, 514) observed in their examination of home advantage in league football: there is evidence on the existence of team-specific home advantage, but statistically the evidence is relatively weak. Based on our results, it seems that modelling for team-specific home advantage is important in forecasting. However, it is not as important as modelling generally for performance across all historical games.

As an answer to the second research question, our results indicate that modeling indirect effects is significantly less important than modeling direct effects. While some of the indirect-effect variables we experimented with showed that the indirect effects do contribute to forecasts, the contribution of these variables is relatively small in comparison to the contribution yielded by match history variables that model for the direct effects. Moreover, our results suggested that some indirect-effect variables were only proxies for the inefficient encoding of the direct match history. Hence, our results on one hand indicate that it is worthwhile to complement the ELO rating based direct-effect variables with some indirect-effect variables, as they contain additional information about teams' future performance. On the other hand, our results also indicate that the improvement these indirect-effect variables bring is relatively small, and thus we cannot unreservedly recommend using them for practical results forecasting or betting purposes.

As an answer to the third research question, our results indicate that our best model is on average able to forecast as accurately as bookmakers forecast. However, according to our

betting simulation results this does not translate into betting that would be profitable on average. Hence, our model is on par with market odds in terms of forecast accuracy, but it needs improvement if it is to be used as a betting tool.

## 1.4 Contribution to existing literature

The largest contribution this thesis makes to the existing literature is the fact that it answers – to an extent – the question raised by Hvattum & Arntzen. At the end of their paper on ELO rating based forecasting the authors raised a question, which asked that would adding variables improve their model build enough to bring the model's forecasts to par with market odds (Hvattum & Arntzen, 2010, 469). As mentioned at the end of part 1.3, our results showed that adding variables does indeed seem to improve the forecast accuracy enough to bring forecast accuracy to par.

However, as mentioned in previous part, our results also indicate that betting with our model would not be profitable. This result raises a question of how our model's (relatively good) forecasts could be used profitably. While we could not answer this question within this study, we were able to point out several topics for future research that could address this. These suggestions for future research are covered in Chapter 6 of this thesis.

Another contribution that this thesis makes is related to the relationship between direct-effect variables and some of the indirect-effect variables. As mentioned in part 1.3, some indirect-effect variables used in our study were found to be – at least partially – proxies for inefficient direct-effect variables. These results contribute to the existing literature as they raise questions about Goddard & Asimakopoulos' (2004) results: as their model has a number of highly significant indirect-effect variables and as their model's direct-effect variables have been shown to be inferior to the ELO rating based variable in terms of forecast accuracy (Hvattum & Arntzen, 2010, 469), it seems that some of the indirect-effect variables presented by Goddard & Asimakopoulos (2004) could be replaced with much less complex direct-effect variables without a significant loss of forecast accuracy.

## 1.5 Limitations of this thesis

The limitations of our thesis are all related to our restricted and relatively small sample size. These limitations are discussed more elaborately in the following paragraphs.

First limitation, brought about by our choice of focusing only on the English Premier League, is the potential inability to generalize our results to the other top-tier football leagues. As the dataset that is used in the evaluation of our model consists of only one league, it is possible that some of our conclusions are only valid within the context of this specific football league. This limitation is, however, alleviated slightly by the fact that market odds for all European top-tier football leagues face similar pressures related to the competition and online availability of odds. Thus, it is somewhat safe to assume that odds for different top-tier leagues behave on average in the same manner as they do in our dataset, although cross-country datasets could be used to verify this empirically.

Second limitation brought about by the dataset is related to the fact that it might not be large enough to reveal differences between forecast models compared in this thesis. As the dataset used in the evaluation consists of only (approximately) 7000 matches, it is possible that a larger dataset could reveal larger differences between different forecast models and thus alter our conclusions about the pecking order of different forecast models.

The third limitation of our study is the fact that the sample only consists of matches that are played on the highest competitive level of English football. As there is some evidence that results of matches in lower-level leagues are more difficult to forecast (Constantinou, 2012, 85), this limitation could make generalizations to lower-level leagues invalid.

The decision of focusing on only a single league was made mainly because including other leagues would have brought about complexities related to modelling indirect effects: First, indirect-effect variables would have had to be slightly different for each league, and thus model specification would have gotten very complicated. Second, deriving the indirect-effect variables was relatively laborious already with one league, and thus introducing more leagues would have expanded the workload beyond the requirements of a master's thesis. And as a

third point, the data to model these indirect effects is not as readily as available as, for example, the match histories of professional football leagues are. Hence data gathering would have been too burdensome considering the time available for writing this thesis. However, our results indicated that indirect effects are not that important in the first place. Therefore it would be interesting to repeat some parts of this study with datasets that consider multiple leagues, but do not consider those indirect-effect variables that are difficult to derive and insignificant for forecasting.

## 1.6 Structure of this thesis

This thesis is divided into six chapters. The chapters following this introductory chapter are 2 – Literature Review, 3 – Data and Research Methods, 4 – Results, 5 – Discussion and Conclusions and 6 – Suggestions for future research.

The next two parts following this introductory chapter are structured as follows. First a literature review is presented. This review draws from the fields relevant to this research by starting from the market efficiency postulations and by ending with the recent developments in football results forecasting. After the review, we describe our research methods and data in Chapter 3. In this chapter we describe the research data, the variables of our model, the details of our regression model estimation, and the methods that are used in the assessment of our model's performance.

The remaining four parts of this thesis are structured as follows. In Chapter 4, the results of our analysis are presented. These results include observations about our model, the results of forecast accuracy measurements and the results of betting simulations. In Chapter 5, these results are discussed in the context of our research questions, and thus answers to the questions are postulated. In Chapter 6, we discuss what new avenues for research our thesis opened, and how exploring them could contribute to existing literature.

# 2 LITERATURE REVIEW

Before we can discuss football results forecasting in more detail, we must put our subject into context. While some might argue that researching the theme is relevant simply because of the sport itself, the theme's economic relevance is mostly derived from its connection to betting markets. In fact, the majority of academic literature on sports forecasting is primarily interested in using forecasting models as tools for studying the dynamics of sports betting. This thesis is no exception to this.

In order to highlight this connection we start our literature review by discussing betting markets and previous research done on them. After this, we can move on to discuss the different modelling practices in more detail.

## 2.1 Betting, betting markets and research on betting market's information efficiency

Betting is an activity of placing wagers on the outcomes of an event whose end results are governed by uncertainty (Encyclopaedia Britannica, 2012a). Opportunities for these wagers are offered by bookmakers in the form of odds (Encyclopaedia Britannica, 2012b). Bookmaker's revenue logic in this activity is based on an overround, a percentage of received bets that is held by the bookmaker instead of distributing it to winners of the bet in question (O'Connor, 2012). We will discuss odds and overround in more detail later in this chapter. In sports betting the wagers are placed different outcomes of a sports event. Here the concept of an outcome can mean anything that can be objectively measured from the sports event: for instance in football bets can be placed on full- and half-time results, first goal scorers or correct scores (OLBG.com, 2013).

Sports betting markets are the forum where bets for sports can be placed, and they represent a major commercial activity. A publicly traded online betting company Bwin.party (2011a) states an estimate by H2 Gambling Capital (H2GC), which predicts that online sports betting will reach an annual turnover of $13.4 billion by 2015, excluding business conducted in the United States. Another part of the same report (Bwin.party, 2011b) also states an estimate by

H2GC, which predicts that sports betting will constitute as 48.1% of all online gambling by 2015. Hence the economic value of sports betting markets is quite significant and will grow to be even more so.

While significant in financial value, betting markets have also attracted a large amount of academic research (Vlastakis et al., 2009, 428). Kyupers (2000, 1353) points out that this is likely due to the way betting markets are set up: bets offered in the betting market require that detailed information on prices and outcomes of the bet and its corresponding events are given. This richness of public and explicit information provides a fertile ground for empirical research.

Our research is concerned with fixed-odds betting on the match results in football. Thus in order to explain the idea behind our study, we continue by describing how fixed-odds betting on football works.

### *2.1.1* Fixed odds betting on football outcomes

Fixed-odds betting on match results is the most popular form of betting in professional football (Forrest et al., 2005, 552). In this form of betting the bookmakers determine the odds for each possible outcome of the match: home team win, draw or away team win. The odds are offered to bettors approximately a week before the match starts (Webber et al., 2007, 2). Curiously, bookmakers rarely adjust the odds during the week leading into the match (Forrest et al., 2005, 552) even though new information on the upcoming match or the distribution of betting volumes might justify such action. This differs from conventional betting common in – for example – horse racing where odds are often adjusted freely up until the start of the event (Forrest et al., 2005, 552). Given this circumstance the odds offered on professional football matches can be considered to be fixed, and hence the concept is called "fixed odds betting".

Due to their static nature fixed odds provide a way to examine the subjective probabilities the bookmakers estimate for the match outcomes (Hvattum & Arntzen, 2010, 463). This becomes apparent when we transform the odds into probabilities by normalizing them. More specifically, the bookmaker's estimate for match outcome's probability is approximately $\frac{\frac{1}{a_i}}{r}$,

where $a_i$ is the decimal odd[1] offered on the outcome $i$, and $r = \sum \frac{1}{a_i}$ is the sum of inversed odds for all outcomes. For example, suppose that a bookmaker offers an odd of 2.45 for home win, 3.35 for draw and 2.75 for away win. To calculate $r$ we sum the inverses of the odds together: $r = \frac{1}{2,45} + \frac{1}{3,35} + \frac{1}{2,75} = 1,07$. Then we can use the formula to approximate the bookmaker's subjective probabilities. The approximation yields values $0.38$, $0.28$ and $0.34$ for home win, draw and away win respectively.

Given this, we also observe that the overround of these odds is given by $1 - \frac{1}{r}$. As mentioned earlier, the overround represents the bookmaker's take in the odds, and it is present in all publicly offered fixed odds (Webber et al., 2007, 2). Size of the overround varies among bookmakers and also as a function of time, and it represents the bookmaker's costs of providing the odds (Makropoulou & Markellos, 2011, 522). In addition to covering for operating costs, it is commonly assumed in the literature that that the overround also covers for the cost of uncertainty associated with not adjusting the odds between the initial publication and start of the match (Makropoulou and Markellos, 2011, 521).

Now that we have shed light on how fixed-odds betting in football works and how bookmaker's probability estimates can be derived from fixed odds, we can continue to discuss the market efficiency of fixed-odds markets. This topic is important as it provides the theoretical justification for our study. In the next part, we will introduce the concept of market efficiency and further discuss how it is related to our research.

### *2.1.2* Information efficiency of a market

The concept of market's informational efficiency was first introduced by Fama in the context of capital markets (1970). By definition, a capital market is said to be efficient if the security prices on the market always fully reflect all relevant information about the traded securities (Fama, 1970, 383). In other words, in information-efficient markets securities are always

[1] In United Kingdom, they are often quoted as fractions. For example, a decimal odd of 2.50 would be 6/4 when quoted in fractions (6/4 + 1 = 2.50).

traded at the "right price", so that abnormal returns cannot be achieved. Definition of the right price, however, varies in literature. Sharpe (1964) approaches the concept by explaining that expected returns of a security should compensate for the variance – or in other words risk – of the returns. Hence, any returns above this compensation are abnormal.[2]

Three different forms of market efficiency can be distinguished: weak-form efficiency, semi-strong form efficiency and strong form efficiency. Markets are said to be weak-form efficient as long as consistent abnormally large returns cannot be achieved by using historical price and volume data in trading. Semi-strong form efficiency requires that all relevant public information is quickly incorporated into security prices, and hence the opportunities for abnormal returns disappear quickly. Strong form efficiency requires that all public and private information is incorporated in the security prices and hence no one can consistently achieve abnormal returns in the market. (Fama, 1970)

In essence, the question of market's informational efficiency is a question of whether information, be it historical, public or private can be used to achieve consistent abnormal returns in the given market. Next, we will discuss how this concept relates to fixed-odds betting.

### *2.1.3* Information efficiency of fixed-odds betting markets

Kyupers (2000, 1355) connects the concept of information efficiency to betting markets by noting that bookmaking business is an information market in addition to being a service. He further argues (2000, 1355) that returns on a wager are abnormal if they exceed the bookmaker overround. For example, let us consider again the odd presented in 2.1.1. Let us assume that the observed 7% overround would be constant in all fixed-odds bets offered on a betting market. By Kyupers' definition, the returns would be abnormal if they were consistently larger than this -7%. As the overround is always positive, this definition implicates that one can observe inefficiency in the market without observing opportunities for

[2] The concept of abnormal returns is widely covered in the scientific literature, and hence the definition given here is likely to be slightly outdated. The definition given here, however, suffices for this thesis, as its only purpose is to illustrate the concept of market efficiency.

profit. This may sound obscure, as from the bettor's perspective a return of -7% does not sound abnormally large. Observing inefficiency in fixed-odds betting has, however, interesting theoretical and practical implications. These become apparent when one considers how the odds are formulated.

As mentioned in 2.1.1, fixed odds can be seen as the approximations of probability distributions estimated by the bookmaker. Given that the practice is to let the odds remain unchanged for approximately a week before the match, any errors in the probability estimations pose a significant financial risk for the bookmaker (Forrest et al., 2005, 552). This is because in the case of a mispriced odd bettors have a whole week to exploit the error and thus cause severe financial damages to the bookmaker. Because of this interaction, bookmakers offering fixed odds on outcomes should be under heavy pressure to estimate match outcome probabilities as correctly as possible (Forrest et al., 2005, 552). This interpretation explains why even negative abnormal returns are relevant: if a bettor is consistently able to earn above the bookmakers' overrounds, it means that she is consistently able to estimate the outcome probabilities more accurately than bookmakers are. While she may or may not achieve positive returns in doing this, she is still making better judgments of the match outcomes than the bookmakers are. As bookmakers' probability estimations are the primary pricing mechanism of this particular market, then any systematic errors in their probability estimations directly implicate that the pricing mechanism of the fixed-odds market is not perhaps working as well as it could.

While the concept of a flawed pricing mechanism in betting markets is interesting from an academic point of view, one might argue that it does not bear any practical implications. This is because bookmakers can compensate for the estimation error with a large overround and thus restrain opportunities for making profit consistently. In addition, bookmakers can place restrictions on betting to further inhibit exploitation: instead of letting bettors wager on single matches the bookmakers can force bettors to bet in bundles of three to five bets, thus making wagering more capital intensive (Kyupers, 2000, 1361) and making the exploitation of errors in the individual odds more difficult (Forrest et al., 2005, 552). However, during the recent years both of these compensation mechanisms have gotten more difficult to use. Study by

Hvattum & Arntzen (2010, 465) shows that the average overround on the English football odds has fallen from 7.56% to 2.50 % between 2000 and 2008. Forrest et al. (2005, 552) also report that by 2003 all British bookmakers had abandoned the practice of forcing bundled bets. These developments are due to increased competition, caused by the rise of online bookmaking (Paton et al., 2002). In addition to stiffening competition between the traditional bookmakers, the emergence of betting exchanges has started to threaten the entire business model of bookmaking (Makropoulou & Markellos, 2011, 522). Betting exchanges are online services where bettors wanting to bet on a particular outcome are paired with other bettors wanting to bet against the outcome (Makropoulou & Markellos, 2011, 522). As this business model does not require the service provider to estimate any odds, the operating costs are significantly smaller and thus betting can be facilitated with much lower margins (Smith et al., 2006, 673). In addition, betting exchanges transfer the information uncertainty completely to bettors (Makropoulou & Markellos, 2011, 522), and thus betting exchanges can offer the same service as the bookmakers but without the added risk premium. Both internal competition and the external threat of betting exchanges build up to the same conclusion: as high overrounds and bet bundling can no longer be used as a cushion, bookmakers should have very strong incentives to produce accurate estimations of football match outcomes. Therefore observing inefficiency in fixed odds is relevant, as bookmaking business largely depends on the high degree of efficiency and inefficiencies can potentially result in large financial consequences.

Given the theoretical and practical significance as well as the convenience of the research setting, it is no wonder that the efficiency of betting markets has been subject to extensive amounts of research. The way in which the research is carried out brings us the justification of our thesis: if bookmakers' ability to forecast results is a measure of market's efficiency, then developing forecast models and using them in betting is a way to assess whether a particular betting market shows signs of inefficiency.

In order to illustrate how these experiments are carried out, we take a look at the previous research on market efficiency in fixed-odds betting. As different types of experimental setups can be conveniently categorized by Fama's three-form classification, we also take turns to discuss the research on weak-, semi-strong- and strong-form efficiency separately.

#### 2.1.3.1 Previous research on weak-form efficiency in fixed-odds betting

Kyupers (2000) formulated a setup for testing the assumption of weak form efficiency in fixed odds betting markets. In this test, the odds formulated by bookmakers are compared with the actual average probabilities of match outcomes (Kyupers, 2000, 1358). If probability distributions across the entire market differ significantly and systematically from the actual average outcome probabilities, then one can conclude that markets are not weak-form efficient as these inconsistencies would be easy to exploit in betting (Kyupers, 2000, 1358). To illustrate this, consider a bettor who observes that historically bookmakers have systematically underestimated the chance of a home team win. After making this observation, the bettor starts to bet always for the home team. If this strategy yields returns larger than average overround, then we have a scenario where only historical information from betting has been used to achieve abnormal returns. Thus in this scenario we could conclude that betting markets in question show the signs of weak form inefficiency.

This idea of testing weak-form efficiency has been used and further developed by numerous authors, and even though strict weak-form inefficiency has not been observed, the results by numerous authors have found evidence that systematic biases exist in odds. In his research Kyupers (2000) failed to identify the signs of weak-form efficiency in odds for the top four English football leagues from seasons 1993-1994 and 1994-1995, but his findings suggested that bookmakers capitalize on the bettors' irrationality by offering biased odds (Kyupers, 2000, 1362).

Since then biased odds have been discovered in numerous datasets of fixed-odds for football results. Perhaps the most famous of these observed biases is the favorite-longshot bias. Originally observed in horse racetrack betting by Ali (1977), the bias is a type of systematic error where favorites tend to win more often than what the odds would imply (Cain et al., 2000, 25). Cain et al. (2000) discovered that this bias existed in 1991-1992 odds of English football leagues. By comparing returns between betting rules that bet for favorites and underdogs respectively, Cain et al. (2000) were able to conclude that the returns produced by these two strategies were significantly different in favor of the strategy that placed bets on

favorites. Based on data from 2002 to 2004 Vlastakis et al. (2009) discovered that the favourite-longshot -bias also exists in the odds of several large internet-based football bookmakers who offer odds on European football. Makropoulou & Markellos (2011, 529) claim that the favourite-longshot bias exists because uninformed bettors have a tendency to overbet underdogs. As a response to this phenomenon, bookmakers then simply adjust by creating an appropriate bias into the odds. Study by Preston and Baratta (1948) speaks in favour of this hypothesis, as it has showed that in experimental conditions people often overbet low probability events and underbet high probability events.

Other biases observed in fixed-odds football betting are related to misestimating home advantage. Vlastakis et al. (2009) showed that odds from 2002-2004 on European football systematically overestimated the winning probabilities of away favourites in comparison to home favourites. While a formal theoretical explanation for this phenomenon is yet to be formulated, it is likely that similar reasoning to the one in favourite-longshot bias could be applicable: uninformed bettors are most likely aware of the home advantage, and thus it is likely that existence of it guides the decision making of uninformed bettors.

#### 2.1.3.2 Previous research on semi strong form efficiency in fixed-odds betting

As stated earlier, semi-strong efficiency in capital markets requires that all public and relevant information regarding the security is built into the price of the security. A parallel to fixed-odds betting can once again be drawn: any publicly available information that can affect the outcome of a football match should be reflected in the match's odds, and thus by using this information in betting no abnormal returns should be achieved if markets are semi-strong efficient (Kyupers, 2000, 1354). Naturally, the best method to test this is to find a way to estimate the probabilities for match outcomes from the publicly available information, and then use these estimations in betting (Kyupers, 2000, 1359). If abnormal returns can be achieved by betting in this manner, then one can conclude that markets are not semi-strong efficient. Another way to assess semi-strong efficiency is to normalize fixed-odds into probabilities and then interpret them as bookmaker's estimations of match outcomes (Hvattum

& Arntzen, 2010, 463). Then by comparing forecast accuracy of these estimations with the ones provided by a competing model, conclusions about semi-strong efficiency can be drawn.

Results different authors have reported suggest that the efficiency of fixed-odds betting markets has improved during the last fifteen years. Kyupers (2000) used an unspecified, historical results based, regression model in conjunction with a simple probability-ratio based betting rule and found the signs of inefficiency in odds for English football for season 1994-1995. His forecast model managed to outperform odds in over 50% of the cases, and thus his betting rule yielded positive and abnormal returns (Kyupers, 2000, 1361). Kyupers, however, notes that on the subsequent season the bookmakers imposed bundling restrictions on bets (Kyupers, 2000, 1362), thus suggesting that the bookmakers themselves observed this inefficiency as well. Goddard & Asimakopoulos (2004) tested the efficiency of English football odds from 1999-2000 and 2000-2001 seasons with a regression model. The model used historical match results along with information about distance, FA-cup success and relative match importance, and like in Kyupers' work (2000) it was used in conjunction with a simple probability-ratio based betting rule (Goddard & Asimakopoulos, 2004). The authors did not explicitly report findings about abnormal returns, but instead they noted that their model yielded positive returns if bets were placed only on matches in April and May (Goddard & Asimakopoulos, 2004, 63). These results would point towards some level of semi-strong inefficiency, but the consistency of these results should be verified by using data from more than two seasons. The fact that their model did not outperform odds consistently also suggests that bookmakers have improved their forecasting since the 1994-1995 season originally tested by Kyupers (2000).

Further improvement in betting markets' efficiency is observed by Forrest et al. (2005), who tested odds for English football from season 1998-1999 to season 2002-2003 with a model similar to the one used by Goddard & Asimakopoulos (2004). After contrasting their model to the historical odds with a range of heuristic tests, the authors conclude that bookmakers' forecast accuracy is superior to the one given by their model, and that the accuracy has risen during the sample period. The authors hypothesize that the rise in the efficiency is due to the enhanced use of information: subjective judgments about information – for example the

injuries of the key players – are employed in forecasting and thus higher forecast accuracy is observed (Forrest et al., 2005, 563). Milliner et al. (2009, 90) also note that bookmakers often employ panels of experts that together make these subjective judgments.

Hvattum & Arntzen (2010) developed a model that used team ratings as a basis of forecasting. In their model, the team ratings were derived from historical match results (Hvattum & Arntzen, 2010). Although their model seemed to yield better forecasting accuracy than several other models discussed in the literature, their model failed to outperform bookmakers' forecast accuracy between seasons 2001-2008 (Hvattum & Arntzen, 2010).

Given these results, it seems that in recent years bookmakers have started to outperform academics in results forecasting. Hence based on the recent research, we can say that fixed-odds betting markets have started to show the signs of semi-strong efficiency. These results also suggest that using historical data as the sole predictor of future match outcomes is not enough to produce forecasts that would outperform fixed-odds betting markets.

#### 2.1.3.3 Previous research on strong form efficiency in fixed-odds betting

Strong-form efficiency of capital markets requires that all relevant information, both public and private, is reflected in security prices. The betting market analogy for this is a situation, where all information relevant to match outcomes – be it public or private – is built into the odds. This information could be, for example, non-disclosed information about recent player injuries. Strong-form efficiency in fixed-odds betting markets is relatively hard to test, as it is often difficult to determine which part of inefficiency in odds is attributable to publicly available information and which part is attributable to non-public information. Makropoulou & Markellos (2011) further argue that due to extensive media coverage and regulation it is very unlikely that odds for professional football would even have significant amounts of insider information available for use.

Because both empirical testing and theoretical reasoning around this topic would warrant another thesis completely, we will not discuss strong form efficiency further in this work. For

those interested in the topic, refer to work done by Shin (1993) and Cain et al. (2003), who have done research on strong form efficiency in horse track betting and football respectively.

### *2.1.4* Conclusions on betting market efficiency's relation to football forecasting

Our literature review has so far revealed that forecasting football results is intrinsically connected with the research on betting market efficiency. In fact all semi strong form tests on the efficiency employ a forecast model and a betting rule in order to detect inefficiencies, and thus abnormal returns from fixed-odds.

The development of these forecasts models is also relevant to the development of fixed-odds betting markets. It seems that the popularity of online betting through various online bookmakers and betting exchanges has exposed bookmaking business to increased competition. This in turn has forced the bookmakers to forecast better, as it is no longer possible to compensate for inaccuracy with very large overrounds. On the other hand, the internet has also exposed bookmakers to larger audiences of informed bettors, thus placing further emphasis on developing better forecasting methods. Evidence for these claims is provided by the fact that academics have not been able to find the definitive signs of weak- or semi-strong form efficiency: while some forecast methods proposed in scientific papers have been able to reveal inefficiencies here and there, the results speak more in favour of semi-strong efficient markets than in favour of inefficiency.

Two distinct features of forecast models rise to prominence in the recent research on betting markets. On one hand, it seems that in order to even try to achieve abnormal returns, the employed forecast model should use historical match information as richly as possible. On the other hand, this information about historical matches should be supplemented with relevant auxiliary data that captures effects not directly related to match history. These principles, combined with a suitable betting rule, seem to be the best starting points for building a forecast model for fixed-odds betting. As the purpose of this thesis is to build an efficient forecast model, these findings are very relevant to this thesis.

However, in order to specify a forecast model we have to examine the modelling practices themselves in more detail. First step in this examination is to look at the different factors that are likely to influence the outcome of a football match. A review of literature on these factors is done in the next parts of this chapter.

## 2.2 Factors to be considered in football results forecasting

In practice, the forecasting of association football results is about finding proxies for effects that affect match outcomes. The underlying assumption of this approach is that different kinds of publicly available information act as a proxy for teams' future performance level.

The most obvious method for this kind of modelling is to infer the future performance level directly from the historical match information, and hence assume that the historical performance explains future performance. While this causality is a good starting point for modelling, the thinking should not stop there. As already touched upon in 2.1.4, according to the literature on football results forecasting, one reasonable way to forecast future results is to use the information from past match results with other public information relevant to match outcomes. Hence, in all likelihood a team's performance level implied by historical matches should not be the only factor to be used in results forecasting.

To understand what factors should be considered in results forecast models, we next discuss factors that previous literature has shown to have an effect on team's future performance in a soccer league. After this discussion, we are able to compare the underlying assumptions of different modelling approaches in terms of their feasibility.

### *2.2.1* Effect of home advantage on team's performance

Courneya & Carron (1992, 13) define home advantage as *"the consistent finding that home teams win over 50% of the games played under a balanced home and away schedule"*. During the past decades, numerous authors have shown that home advantage exists in almost all sports where balanced home and away schedule, or 'round robin' schedule, is used (Jamieson, 2010). A brief look into the match result distributions presented in Figure 2.1 also reveals that home advantage is relevant in modern European football.

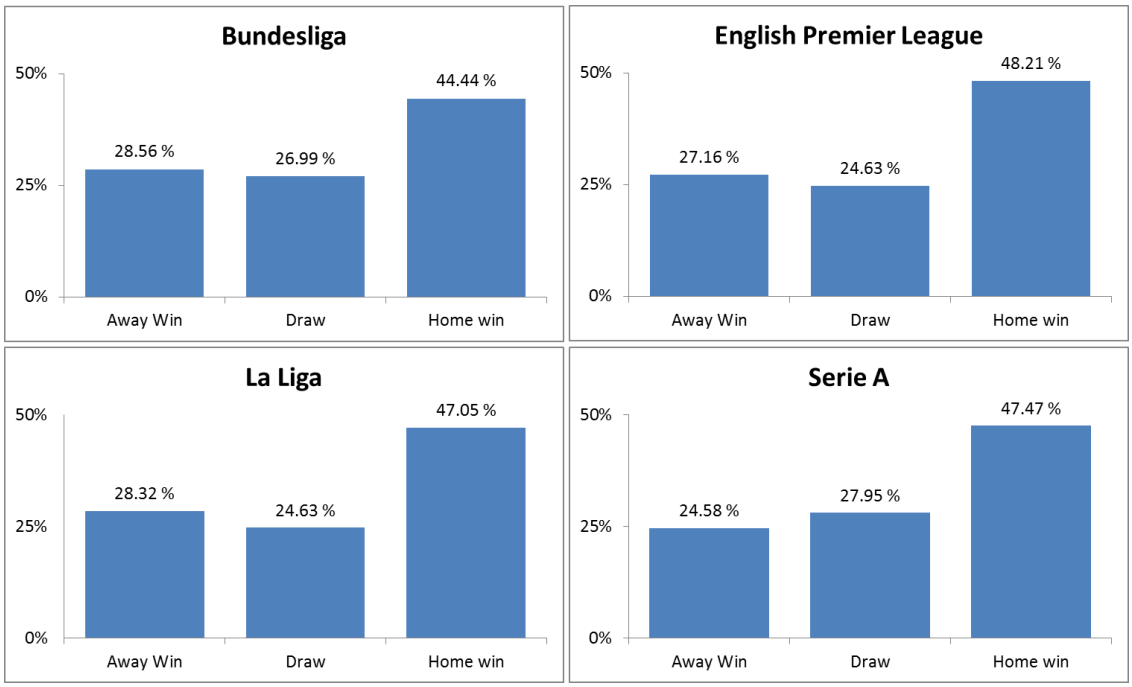


**Figure 2.1 – Distribution of Match Outcomes in Four European Professional Football Leagues from 2005 to 2010, adopted from Football by the Numbers (2011).**

As shown in Figure 2.1, games are won more often – by a large margin – at home than at away in the top European football leagues. This observation further illustrates the notion that playing at home stadium provides a major advantage in football. Literature offers many explanations for why this home advantage exists. Nevill & Holder (1999, 221) categorize these explanations into four factor classes: learning factors, rule factors, travel factors and crowd factors.

The first two of these factors have been largely ruled out in the context of football. Studies on learning factors have tried to determine whether home team's learned familiarity with the environment is a source of home advantage. These studies have shown that teams gain very little benefit from being more familiar with the environment (Nevill & Holder, 1999, 221). Studies on rule factors have in turn tried to explain how much the rules of a sport contribute to home advantage. While rule-based contribution has been found in sports like ice hockey, rule factors have not been found to contribute to home advantage in football (Nevill & Holder, 1999).

As for travelling factors, several authors have hypothesized that they could explain home advantage (Nevill & Holder, 1999). However, studies on home advantage in baseball (Schwartz & Barsky, 1977) and football (Pollard, 1986) present strong evidence against this hypothesis. Based on statistical analysis on large data sets, travel factors were concluded to be insignificant in both studies. Nevill & Holder (1999, 221) also state that the largest argument against the influence of travel factors is the fact that large home advantages are observed even in sports leagues where the distances between home stadiums are relatively small.

With the first three factors ruled out, the crowd effect remains perhaps the most plausible explanation. In fact, a number of authors hypothesize that home advantage is largely due to them (Neville & Holder, 1999). Most notably Schwartz and Barsky (1977) hypothesize that the crowd factor exists because of the social support the home audience provides to the home team. Greer (1983) in turn claims in his study on basketball that the home crowd can also intimidate the referee to be biased towards the home team, and thus give home advantage through unfair penalties. Nevill & Holder (1999, 211) argue that this kind of 'refereeing edge' can be very crucial in sports, as only a few crucial decisions can already alter the course of a match significantly. Several authors have explored Greer's idea further. Nevill et al. (1999) back his notion of the crowd-induced referee bias empirically in their study. The conclusions of the study state that in an experimental setting, semi-professional football players, qualified referees and football coaches were more likely to penalize on a foul, if the decision of penalizing was made with a large crowd noise in the background (Nevill et al., 1999). This suggests that the crowd can affect refereeing outcomes. Empirical study by Buraimo et al. (2009, 431) also reports the evidence of home team favoritism in the English Premier League and the German Bundesliga. Boyko et al. (2007) were also able to establish a connection between the crowd effect and the refereeing bias, although they note that the proneness to home team favoritism varies across individual referees greatly.

Given the previous research, it seems that home advantage can be mainly attributed to the crowd-induced home team favoritism exhibited by referees. How could this effect then be measured? On this theme, the literature points us towards two separate topics.

The first topic considers whether it makes sense to measure home advantage as a team-specific attribute or as a league-specific attribute. If we consider the root cause of home advantage, the crowd, this question is essentially about whether the fans of certain teams are able to cause more refereeing bias than the fans of other teams. Clarke and Norman (1995, 514) suggest that there could be this kind of team-effect in home advantage, effectively meaning that home advantage varies across teams within a league. Results on their statistical testing were not, however, highly conclusive (Clarke & Norman, 1995, 514). Goddard & Asimakopoulos (2004, 56) touch the issue by noting that in their forecast model the recent home results of a team were better predictors of future home matches than the recent away results. A similar conclusion was made about away matches as well (Goddard & Asimakopoulos, 2004). Based on these two studies, it would then seem that there is at least some evidence of the team-effect in home advantage, and hence it seems to make sense to model home and away performance separately.

The second topic considers whether there are any pairwise crowd effects. In other words, we need to consider whether home advantage is a team-pair specific in addition to being team- or league specific. Clarke and Norman (1995, 515-516) found a pairwise effect related to the geographical distances between different football teams in Britain. The authors concluded that home advantage seems to grow as a function of distance between teams (1995, 516). Hence, the authors claim that home teams enjoy larger home advantage when the away team's home stadium is very far away. Goddard & Asimakopoulos (2004, 56) give an explanation for this by suggesting that the distance effect exists in the English football due to travel weariness and due to the intensity of local encounters. The first part of the explanation will not hold if we consider the other literature on the travel factors (which we discussed already). The latter argument about the greater intensity of local matches does, however, seem quite valid: in local matches, or derbies as they are referred to in football, it can be so that the crowd effect is less pronounced as the distribution of fans between home and away teams is likely to be less skewed towards home team. And if this is true, then it is likely that at greater distances the crowd effect is more pronounced because large masses of away team fans are not likely to mobilize for an encounter of their favourite team hundreds of kilometres away from their

home stadium. Findings by Forrest & Simmons (2002, 238) speak in favour of this reasoning as their study on forecasting match attendance showed that the distance negatively affects match attendance in English association football.

### *2.2.2* Other factors

In addition to home advantage, several other factors have been discussed in the literature. More specifically, three broad topics can be recognized: the injuries of key players, the relative differences in the importance of the match, and the effects of external cup competitions.

Drawer & Fuller (2002) studied what kind of effect the injuries of individual players have on the performance the entire football team. Be measuring team quality as a sum of player quality scores from a sample of five seasons of the English Premier League, the authors were able to measure the effect individual players have on their team's performance (Drawer & Fuller, 2002). The measurements made by the authors showed that the injuries of key players were significant in explaining team's performance (Drawer & Fuller, 2002). While these results have not been refuted per se, we were unable to find any documentation about using this kind of information in statistical results forecasting. In their work on time series forecasting Webby & O'Connor (1996, 94) refer to this kind of information as 'broken-leg cues' and further suggest that information like this is judgemental by nature. Hence, they imply that this kind of information should not be considered in statistical forecasting at all. Forrest et al. (2005, 562) approach the issue in a similar manner by hypothesizing that bookmakers use the information about player injuries to make subjective adjustments to statistical forecasts. Forrest & Simmons (2000) also label information like this as contextual, and thus reject the possibility of using it in statistical forecasting. These kinds of views are dominant in the football forecasting literature most likely because statistical inferring from the injury information is quite complex. On one hand, it is often not easy to say how much each player contributes to each team's performance. Hence, it is hard to say anything general about the influence of an individual player, when analysed datasets cover more than ten years' worth of fixtures. On the other hand, it could also be that timely and relevant information about player injuries is so expensive to obtain that it is not relevant to consider it in betting-related forecasting.

Studies by Jennett (1984) and Peel & Thomas (1988) have identified that championship, promotion and relegation issues are important determinants of the match attendance. Goddard & Asimakopoulos (2004, 56) argue that these factors are also important determinants of the match results in association football. More specifically, they hypothesize that if there is a significant difference in the match importance between the teams, then the team to whom the match is more important has an edge (Goddard & Asimakopoulos, 2004, 56). If we consider how a round-robin association football season is organized, this argument makes sense. At the end of each season, there most likely are situations where some mid-table teams do not have championship-, promotion- or relegation-related incentives for the last couple of matches, as they no longer can affect their league table position with the remaining games. Meanwhile, these 'indifferent' teams could be paired against teams to whom the same incentives are much higher due to the fact that winning or losing these same matches can affect whether they – for example – get relegated after the season. The authors tested this hypothesis by labeling a match to be significant for a team if "*it is still possible (before the match is played) for the team in question to win the championship or be promoted or relegated, assuming that all other teams currently in contention for the same outcome take one point on average from each of their remaining fixtures"* (Goddard & Asimakopoulos, 2004, 56). The authors further reported that the variables derived with this criterion were consistent with the initial hypothesis of the match significance affecting the results positively, and thus the factor was found to be useful in predicting future match results (Goddard & Asimakopoulos, 2004, 56).

As football teams often play in cup competitions during the regular league seasons, it is also important to investigate how these external cups affect the performance in the league. One interpretation for their effect is founded on the financial incentives cups offer. For example, for English Premier League teams the two European Cup competitions – the Champions League and the Europa League – are major sources of revenue (Deloitte, 2012) and thus the teams should be very interested to perform well in them. In the context of the Premier League, the FA Cup is not as significant in terms of the direct revenue (The FA, 2012a), but as its winner qualifies for the Europa League, English Premier League clubs should have strong

financial incentives to fare well in it as well. As these cup competitions represent an extra strain on the team, it is possible that team's involvement in them could distract league efforts (Goddard & Asimakopoulos, 2004, 56). Hence, if argued from the financial point of view, it is possible that the cup involvement could predict decreased league performance. Contrary to this view, it is also possible to argue that the success in external cups provides a significant morale boost for the team on the league side as well. Therefore, cup participation would imply an increase in the league performance, and consequently a cup exit would imply a decrease in the league performance (Goddard & Asimakopoulos, 2004, 56). Goddard & Asimakopoulos (2004) tested which one of these interpretations would prevail in the case of the FA Cup's effect on the top four English football leagues. Their statistical testing provided evidence in favor of the latter interpretation: based on their results, it seems that a cup involvement has a positive effect on the league performance (Goddard & Asimakopoulos, 2004, 56). As for other effects of external competitions, Constantinou et al. (2012, 123) also account for a match fatigue in their model, and thus their formulation implicitly assumes that the extra fixtures brought about by external competitions could hinder the league performance of a team through straining the fixture schedule.

### *2.2.3* Concluding remarks on factors relevant to forecasting

The literature on football forecasting recognizes six different factors that have been shown to have an effect on match outcomes: team performance inferred from the historical match information, team-specific home advantage inferred from the historical home and away match information, pairwise home advantage inferred from distances, relative match importance inferred from the league table related incentives, cup effects inferred from the external cup schedules and the effect of injuries inferred (subjectively) from the player injury data. Table 2.1 presents these factors and summarizes their implications for modelling.

| Factor | Description | How to model? |
|---|---|---|
| Team performance level | How well the team has recently performed | Inferred from historical match information |
| Team-specific home advantage | How well the team has recently performed home in comparison to away performance | Inferred from differences in historical home- and away match information |
| Pairwise home advantage | How home advantage manifests in specific fixtures | Inferred from distance between home stadiums |
| Relative match importance | How differences in match importance affect match outcome | Inferred from league table position |
| Cup effects | How participation in external cups affects league match outcomes | Inferred from external cup schedules and historical cup results |
| Player injuries | How injuries of key players affect match outcomes | Inferred subjectively from teams' injury news |

**Table 2.1 – Summary of the factors affecting match outcomes discussed in football forecasting literature**

As we can see from Table 2.1, various sources of public information have been recognized in the literature as relevant for football results forecasting: In addition to using information about historical league matches, relevant conclusions can be drawn from various other sources of information as well.

However, in spite of numerous sources of information, historical league match information is by far the most significant predictor of future match results. This is because it carries information about team performance as well as about team-specific home advantage, both of which have been shown by many authors to be very significant in results forecasting.[3]

Given that multiple authors from varying disciplines have researched the topic, it is not surprising that there are many very different approaches for incorporating historical match data into a forecast model. In order to determine which approach is the most suitable, we next

[3] See for example Goddard & Asimakopoulos (2004), Hvattum & Arntzen (2010), Dixon & Coles (1998).

discuss the different methods previous authors have used to incorporate historical league match information into a statistical forecast model.

## 2.3 Different methods of forecasting with historical league match information

To structure the discussion about the way in which historical match information is used by different authors, this part of the literature review is broken down into three parts. In the first part, we compare two fundamentally different approaches to results forecasting: goal scoring process modelling and historical results modelling. In the second part, we then discuss the way team rankings can be used for forecasting purposes. In the third part, we summarize our findings about the methods discussed.

### *2.3.1* Differences and similarities between the goal scoring process method and the historical match results method

Two distinct approaches to football results forecasting can be identified in the literature (Goddard, 2005, 331). The first approach uses match history data to model goal scoring processes and thus forecasts match results indirectly by forecasting the amount of goals scored. The second approach forecasts future match results from historical match results (Goddard, 2005, 331). In the following paragraphs, we discuss both of the modelling approaches in order to see the differences and similarities between the two approaches.

Forecasting methods based on goal scoring processes have been developed during the past thirty years. Maher (1982) was the first author to model goal scoring processes in football. In his model Maher (1982) uses attack and defence parameters of opposing teams to derive univariate Poisson distributions for the goal scoring processes of both home and away teams separately. Dixon and Coles (1998) developed Maher's idea into a forecasting model. In this formulation probabilities for different final scores are derived from two Poisson distributions: one for the home team and the other for the away team (Dixon & Coles, 1997, 270). Formally put, $A_{kl} \sim Poisson(\psi_k \omega_l \upsilon)$ and $B_{kl} \sim Poisson(\psi_l \omega_k \upsilon)$, where $A_{kl}$ and $B_{kl}$ are the numbers of goals scored by home team $k$ and away team $l$, as a function of the corresponding attack and defence parameters $\psi$ and $\omega$ and a home advantage adjustment parameter $\upsilon$ (Dixon & Coles,

1997, 270). In this formulation, a log-likelihood function is then used to produce the inferences (Dixon & Coles, 1997, 271-272).

Dixon & Coles (1997, 268) noted that the dependence between the distributions can cause problems in goal scoring modelling, as their approach implicitly assumes that the goal scoring processes for both teams are independent of each other. Dixon & Coles (1997, 270) corrected for this by adjusting the probability estimation procedure for those final scores, where – based on historical results – the independence was deemed implausible. According to our knowledge, no theories have been presented to explain why some scores can be assumed to be independent and why some others cannot. Dixon & Robinson (1998, 537), however, analysed goal scoring in English football matches and concluded that the dependence between processes seems to be especially strong when favourites have a narrow lead. The evidence presented by Karlis and Ntzoufras' (2003) supports this to an extent, as they modelled Italian Serie A match results and identified a bias in the predictions of low-scoring draws.

The variation of the attack and defence parameters over time also presents a challenge for modelling, as it is reasonable to assume that the attacking and defending capabilities of teams are not constant across seasons, or even during one season. Dixon & Coles (1997, 272) accounted for this dynamics by updating both the attack and defence parameters of teams periodically. Their updating procedure decreased the weight of older observations exponentially, thus placing more emphasis on the more recently observed attacking and defending capabilities (Dixon & Coles, 1997, 272). As with the problem of interdependence, the optimal approach for modelling the time-based variation of parameters is also quite difficult to justify theoretically. Dixon & Coles (1997) justify their weighting procedure by testing which exponential smoothing parameters maximized the model's forecasting accuracy on a historical data set. Other approaches that account for varying attack and defence parameters include Rue and Salvesen's (2000) use of Bayesian inference.

As a contrast to the goal scoring process method, several authors have more recently started to use historical match results to derive forecasts for future matches. According to our knowledge, Kyupers (2000) was the first to employ an ordered probit based regression model,

which used historical match results as the dependent variables to predict future match outcomes (home team win, draw, or home team loss). Kyupers did not, however, give an explicit formulation for his model. Partly based on Kyuper's (2000) research, Goddard and Asimakopoulos (2004) have constructed a frequently cited lagged performance variable model. In the model historical match results are used to derive long and short term performance statistics for each team. These statistics are then used as independent variables in a regression model, whose dependent variable is a future match result.[4] In the authors' research, this model is then used to predict the probabilities of future match outcomes. To illustrate the idea behind historical match results modelling, a brief description of Goddard & Asimakopoulos's (2004) model is given in the following paragraphs.[5]

In Goddard & Asimakopoulos's (2004) formulation the long term performance level of a team is modelled by constructing win-loss ratios from the team's historical match results. These ratios are measured from the match results data across different timespans in order to see how teams' past performance during different time periods explains the future performance on average. This assessment is done by using the win-loss ratios from different timespans as independent variables in a regression model. When the model was applied to data from English football's top 4 divisions for the seasons between 1989 and 1998, the most recent match history seemed to be the most significant in forecasting future results. Moreover, it seemed that the contribution of the past performance to the forecast accuracy decreased as a function of time relative to the forecasted match: The ratios constructed from the on-going season's past matches were estimated to be the most useful in forecasting, the last season's ratios were the second most useful, the ratios from two seasons ago were the third most useful, and the ratios from matches played earlier than two seasons ago were not useful at all (Goddard & Asimakopoulos, 2004, 55).

---

[4] As the match result is a discrete and ordinal variable, the model used in Goddard & Asimakopoulos' (2004) research is an ordered probit regression model, as it produces meaningful interpretation for variables with these properties. We return to details of such models later in Chapter 3 of this thesis.

[5] The description given here only serves as an introduction to the modeling principles. For a more detailed description, see Goddard & Asimakopoulos (2004).

The teams' short-term performance is captured into the model with discrete indicators, which tell how the team has fared in its recent home and away games (win is encoded as 1, draw is encoded as 0.5 and loss is encoded as 0). As these indicators are used as independent variables together with the long-term performance indicators, they should tell how much the most recent match history has weight in relation to the team's historical long-term performance. Like the previously discussed long-term indicators, empirical tests showed that the short-term performance indicators' contribution to the forecast accuracy decreased as a function of time relative to the forecasted match: The most recent results were the most reliable indicators of future performance and the results older than two months rarely contributed to the forecast accuracy at all (Goddard & Asimakopoulos, 2004, 55). It should also be noted that in general these short term indicators contributed less to the forecast accuracy than the long term indicators (Goddard & Asimakopoulos, 2004, 55).

It is also worthwhile to notice that the model design requires that both the long and short term performance indicators are calculated separately for the home and away teams in order for the model to have a meaningful interpretation. Thus the amount of independent variables used in the model is quite large. In the example design reported by the authors, more than 30 statistically significant variables are reported (Goddard & Asimakopoulos, 2004, 55).

By comparing Goddard & Asimakopoulos's design with the one formulated by Cole & Dixon we immediately observe two differences. The first difference is the fact that by using historical results one does not have to worry about the effects of the in-game interdependencies. When using historical match results to derive variables, the modeller only has to assume that the historical match results explain future match results. When modelling goal scoring processes, the modeller also has to assume the independence of goal scoring processes. And since the independency assumption does not seem to hold across all results, adjustments to modelling practices have to be made. As the literature shows, these adjustments are hard to justify theoretically. Thus, one could argue that the historical match result method is more reliable, as in practical circumstances its key assumptions seem to hold better than those of the goal scoring process method. The second difference between the approaches is the fact that the goal scoring process method requires much more complex

modelling. Although the historical match result model introduced by Goddard & Asimakopoulos does introduce several dozens of independent variables, it is still relatively simple to use, as the variables can be constructed with a relatively simple dataset that only has dates, team names and match results recorded. This is not the case in Dixon & Coles' model, as constructing the attack and defence parameters of the teams requires the modeller to record various attacking and defending statistics per each game. Such information is much more cumbersome to collect, and thus it is more expensive to obtain. However, the data used by the goal scoring process method is also richer than the data used by the historical match result method (Goddard, 332, 2005). While the historical match results method is simpler to use, it also neglects large chucks of information. For example, a 1-0 win is treated equally to a 5-0 win in the historical match result method, and thus the information conveyed by the difference in the scores is lost. As the goal scoring process method accounts for these kinds of differences, it could – at least in theory – provide more accurate forecasts (Goddard, 332, 2005). After all, it is very likely that teams that consistently win with a large goal difference and that have superior attack and defence parameters are better than teams that win only narrowly and have much inferior statistics.

While differing in terms of complexity and data richness, we can observe that both of the methods pose the modeller with very similar challenges on other fronts. More specifically, neither of the modelling approaches can theoretically justify the way different historical time periods are weighted when producing the forecasts. For example, the approach by Dixon & Coles (1997) uses historical match data to calibrate how much weight the historical attack and defence parameters from different timespans should have on forecasts. Similarly, Goddard & Asimakopoulos (2004) experimented with different modelling designs to come up with their final specification, and ultimately the weights of different variables were chosen on the basis of fitting the model into a historical dataset of match results. Other authors, who have cited Goddard & Asimakopoulos's and Dixon & Coles' models in their work, have also resolved

their model parameterization issues with more empirical than theoretical reasoning.[6] Given that neither the goal scoring process method nor the historical match result method can theoretically justify the weighting decisions concerning historical data, we can conclude that both modelling approaches always require significant empirical testing for model calibration.

While there are differences and similarities regarding modelling practices, the most important difference between the two approaches is whether either of them is superior in terms of forecast accuracy. Study by Goddard (2005) can provide some answers to this question. Goddard (2005) tested both of the modelling approaches using a data set of four top divisions of English Football from season 1977-1978 to season 2001-2002[7]. In his work, the author observed that the differences in forecast accuracy between the methodologies were quite small. Based on this, it seems that the richness of the data provided by the goal scoring process method does not seem to bring tangible benefits over the rougher match results based estimation method.

Given that the methods based on using historical matches are less burdensome to use and equally efficient in practical applications, it seems to make more sense to prefer the historical match results based method over the goal scoring process method. By doing so a modeller can operate with much simpler model design and data, but still achieve as good results as she would achieve with a more complex model.

### *2.3.2* Using relative performance levels derived from historical results to forecast future results

While Goddard & Asimakopoulos' (2004) lagged performance variable model discussed earlier is the most widely cited approach to forecasting with historical match results, one can also use the same kind of data to estimate relative performance levels between teams, and then use this information in forecasting. Using this idea Hvattum and Arntzen (2010) developed an ordered logit regression model, which used variables based on relative performance levels to

[6] See Forrest et al. (2005), Goddard (2005) and Hvattum & Arntzen (2010), all of whom justify their model calibration with purely empirical and data-mining oriented means.
[7] See Goddard (2005) for a more specific description of how these tests were conducted.

predict future match results. Variables for the model are derived by using the ELO rating system. The ELO rating system was originally introduced by Arpad Elo (1978) and its original purpose was to assess the skill differences between professional chess players. In order to compare this "ELO model" with the lagged performance variable model, we describe the ELO rating system and explain how it can be used in football results forecasting.

In the ELO rating system the match result is classified with a score system where a win gives a score of 1, a draw gives a score of 0.5 and a loss gives a score of 0 for the home team participating in the match. Let $\alpha^H$ and $\alpha^A$ represent the scores of the home and away teams respectively. Therefore,

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

Given the above, the score of the away team is then $\alpha^A = 1 - \alpha^H$. Now suppose that before the match one can produce estimates for the $\alpha^H$ and $\alpha^A$ based on the performance ratings of the home and away teams. Let the score estimates, or expected scores, be $\gamma^H$ and $\gamma^A$. Then let the performance ratings be $\iota^H$ and $\iota^A$. To illustrate how expected scores, actual scores and performance ratings interact in this model, we introduce the following notation:

$$\iota_0^H = the\ performance\ rating\ of\ the\ home\ team\ before\ the\ match,$$
$$\iota_0^A = the\ performance\ rating\ of\ the\ away\ team\ before\ the\ match,$$
$$\iota_1^H = the\ performance\ rating\ of\ the\ home\ team\ after\ the\ match,$$
$$\iota_1^A = the\ performance\ rating\ of\ the\ away\ team\ after\ the\ match,$$
$$k = a\ constant\ scaling\ parameter,$$
$$c = a\ constant\ scaling\ parameter,$$
$$d = a\ constant\ scaling\ parameter.$$

Of the three scaling parameters mentioned above, $c$ and $d$ can be interpreted as setting an appropriate scale for the ratings, whereas $k$ is a parameter that defines the rate of change in the ratings. The values for $c$ and $d$ can be chosen somewhat arbitrarily, but the value for $k$ must

be chosen more carefully for the rating adjustment process to work properly. The calibration process of $k$ is out of the scope of this thesis, but the general idea behind it is briefly discussed at the end of this section.

Now, suppose that the post-match ratings of the home and away teams are a sum of two components: pre-match ratings and the difference between the actual scores and the expected scores scaled with a parameter $k$. More specifically, suppose that the following equations hold:

$$\iota_1^H = \iota_0^H + k(\alpha^H - \gamma^H),$$

$$\iota_1^A = \iota_0^A + k(\alpha^A - \gamma^A).$$

Now, suppose that $\gamma^H$ and $\gamma^A$ are derived from the pre-match ratings with the following formulas:

$$\gamma^H = \frac{1}{1 + c^{\frac{\iota_0^H - \iota_0^A}{d}}},$$

$$\gamma^A = 1 - \gamma^H = \frac{1}{1 + c^{\frac{\iota_0^A - \iota_0^H}{d}}}.$$

Therefore, for any upcoming match the pre-match ratings of the home and away teams, $\iota_0^H$ and $\iota_0^A$, are taken as they are. Then they are used to derive the expected scores, $\gamma^H$ and $\gamma^A$, for the home and away team for the upcoming match. The expected scores thus reflect what kind of actual scores, $\alpha^H$ and $\alpha^A$, the teams will on average get from the upcoming match, given the difference in the pre-match ratings. Then after the match has been played and the values of $\alpha^H$ and $\alpha^A$ have been observed, the post-match ratings can be calculated. As we recall from the equations above, the after-match rating equals the pre-match rating plus the scaled difference of the actual and expected outcome. Then, for the next match of these teams, the adjusted after-match ratings are taken as the new pre-match ratings. Doing this perpetually results in a process, where the ratings of all teams are updated after each match, and where the magnitude of an individual update depends on initial rating difference and the observed result. Figure 2.2

illustrates the idea by showing how team ratings develop as a function of time. In this example scaling parameters are $k = 20$, $c = 10$ and $d = 400$.

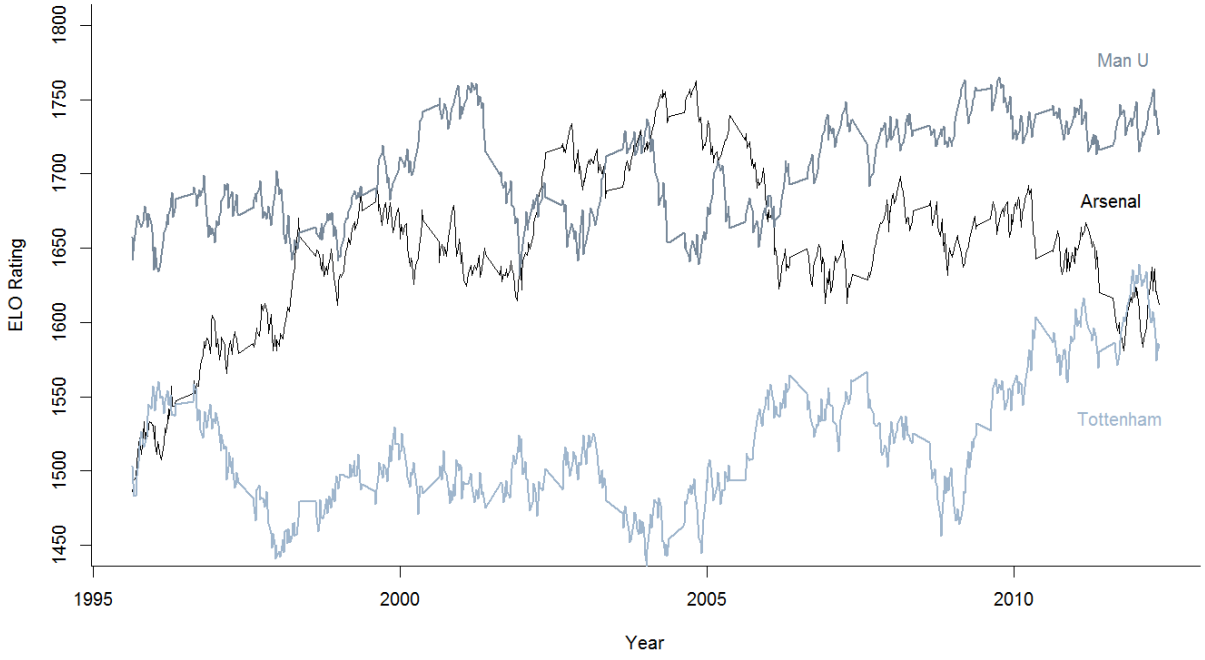


**Figure 2.2– Example of ELO rating development in English Premier League between seasons 1995/1996-2011/2012**

In their model Hvattum & Arntzen (2010) use the pre-match rating difference as an independent variable in an ordered probit regression model, where the dependent variable is $\alpha^H$, the result of a future match. As the forecast model produces probabilities for each possible outcome of $\alpha^H$, the "ELO model" developed by the authors essentially translates skill differences between two teams into probability estimates of how a match between the two teams is likely to end.

As the ELO model uses the same kind of information as the Goddard & Asimakopoulos's (2004) lagged performance variable model described in 2.2.1, it can be considered as a modification of the ideas of Goddard & Asimakopoulos. Given that the principle idea of using historical match results to predict future results is essentially the same in both models, one would expect the differences between the models to be cosmetic. However, when comparing

the properties of the ELO model with the ones of the lagged performance variable model, we observe several quite large differences.

The first difference is the fact that while Goddard & Asimakopoulos's model uses the historical results of an individual team as lagged variables, the ELO model uses a series of historical results to derive a single variable – the rating difference between the two teams participating in the match. This difference means that the underlying assumptions between the two models are slightly different: the lagged performance variable model assumes that a match outcome is determined by comparing two independent match histories directly. The ELO model then in turn assumes that individual match histories can be used to derive a rating difference for each match, and that the difference by itself contains enough information to produce a probability estimate. Because of these different assumptions, the specification of the ELO model is much simpler, as the modeler does not have to worry about the issue of choosing the appropriate number of lagged variables. The information about the team's individual performance history carried by the lagged variables is not, however, lost in the ELO model. On the contrary, it is encapsulated in the ratings of individual teams, as they develop depending on how the teams fare.

The second difference between the models is the potentially richer data of the ELO model. As the lagged performance variable model measures the team quality solely based on match results, it does not take into account the fact that not all wins might be equally good indicators of quality. For example, let us consider a pair of matches: in the first match, team 1 – a clear favorite – prevails over team 2 – a clear underdog. In the second match, team 3 – a mid-table team with a mixed record – prevails over team 4 – another mid-table team with a mixed record. By common sense, the latter win should be more valuable, as the match between the teams 3 and 4 is likely to be tough for team 3, whereas the first match should be easy to win for team 1. The lagged performance variable model does not completely distinguish between these two types of games, as the variables that express the team performance do not directly distinguish results of tougher games from easier ones. Meanwhile, the ELO model would – if appropriately parameterized – recognize that the expected score in the first match is more heavily skewed towards team 1, whereas in the second match the expected score would be

more equal. This richer use of information could potentially lead to the ELO model having a superior forecasting performance over the lagged performance variable model.

The third difference between the models is their consideration for the team-specific home advantage: the lagged performance variable model accounts for the team-specific home advantage via the short-term performance variables. Meanwhile, the ELO model does not account for it at all. In fact, the ELO model outlined here only accounts for "general" home advantage through the fact that home wins are more common, and thus the regression model generally assigns more probability to home wins. Given that there is some evidence on the existence of team-specific home advantage, the lagged performance variable model is richer in this sense as it accounts for it (whereas the ELO model does not).

The calibration of both models is also quite different, even though both models come with the challenge of choosing the appropriate weighting for historical data. The lagged performance variable model is calibrated by selecting a group of time-period discrete variables that produce the best fit for the data. Hence, the choice of appropriate weighting is essentially a question of evaluating different lagged variable structures on the basis of some relative goodness of fit statistic for the regression model. In the ELO model, the historical information is weighted based on parameter $k$, which defines how much the ratings change in each post-match rating adjustment (Hvattum & Arntzen, 462, 2010). The larger the $k$ the faster team ratings develop and hence "forget" older rating adjustments. Hvattum & Arntzen (2010) calibrate $k$ in their model based on a quadratic information loss[8] produced by the model's forecasts in a training sample. Hence their calibration technique is similar to the one Dixon and Coles (1997) used in their goal scoring process model. It is not trivial to say which one of the calibration procedures is more reliable, and perhaps the only way to assess their differences is to assess which modeling approach produces better forecasts altogether.

Hvattum & Arntzen compared the forecasting accuracy of the ELO model with the one of Goddard & Asimakopoulos' model. Based on the comparison with informational loss and

[8] We return to the quadratic loss and other forecast accuracy measurements in Chapter 3 of this thesis.

quadratic loss, the authors concluded that the ELO model produced more accurate estimates for probability distributions. The authors were not, however, able to determine which model would be more profitable in betting. In fact, neither of the models was able to produce abnormal returns when used in betting simulations. The authors also suggested that the differences in forecasting accuracy between the two approaches are likely to become smaller as sample size increases. Hence, they concluded that the ELO ratings are a more efficient way of encoding past results when only short time periods are available to calibrate the model. (Hvattum, & Arntzen, 2010)

### *2.3.3* Concluding remarks on incorporating historical match data into a forecast model

Historical match data is the central component of any football results forecast model. As presented in Table 2.1, league match history can be used to model team performance level as well as team-specific home advantage – depending on what kind of variables are derived from it.

Based on a review of literature at least three different models for incorporating league match history into a forecast model can be distinguished: the goal scoring process model, the lagged performance variable model and the relative performance indicator model (the ELO-model). Figure 2.3 summarizes the properties of the three methods and thus highlights differences between them.

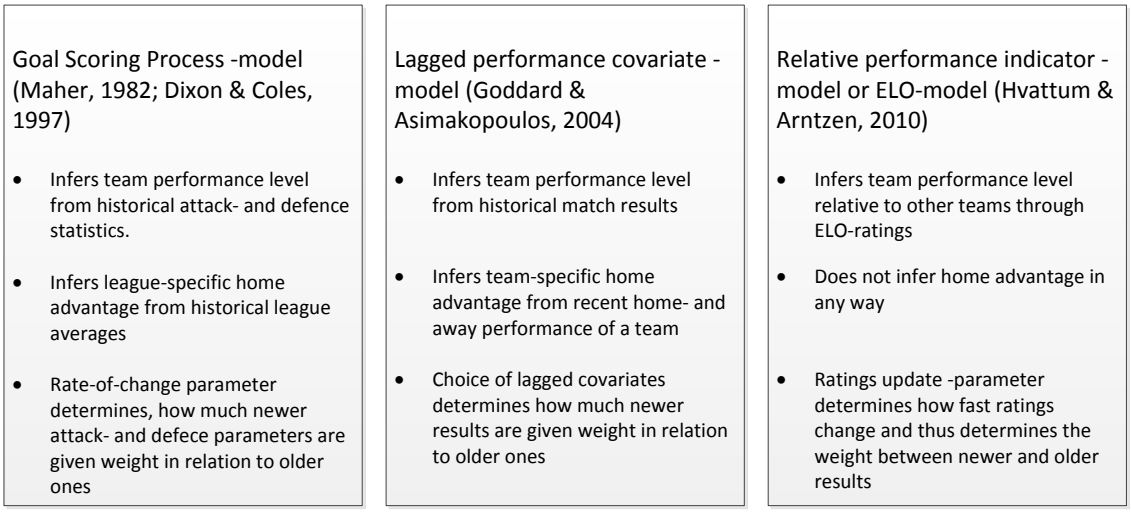


**Figure 2.3 – Summary of different methods for incorporating historical match data into a football forecast model**

As mentioned in Figure 2.3, all three modeling approaches also require the modeler to evaluate how older match data is weighted in relation to newer match data. In case of the lagged performance variable model, this is done by comparing the model fits of different lagged variable structures. In case of the other two models, a rate of change parameter is chosen on the basis of forecast accuracy in a training sample. Figure 2.3 also illustrates the differences in the modeling of home advantage: Out of the three approaches, only the lagged performance variable model accounts for team-specific home advantage. Based on the discussion done in 2.2.1 about the topic, this could be seen as a drawback to the other two models as there is some evidence that team-specific home advantage should be taken into account in forecasting.

While not shown explicitly in Figure 2.3, the three models differ in terms of forecast accuracy and parsimony. Based on the forecast accuracy measurements made by Goddard (2005) and Hvattum & Arntzen (2010), it seems that the lagged performance variable model is as good as the goal scoring process model, and that the ELO model is slightly better than the lagged performance variable model (at least with small sample sizes). And as for parsimony, out of the three models the ELO model is the most simple to specify, as it uses only a single variable

– the expected score – to forecast a match result. Hence, when judged by forecast accuracy and parsimony, the ELO model is superior to the other two approaches.

## 2.4 Conclusions of the literature review

We have now discussed football forecasting from three different perspectives. First, we discussed the efficiency of fixed-odds betting markets, as research on it effectively justifies research on football forecasting. Then we discussed what factors should be considered when building a football forecasting model. And finally, we discussed different methods of making inferences from historical match data.

Previous research on betting market efficiency indicates that fixed-odds betting markets for football are at least weak-form efficient and perhaps also semi-strong efficient. Previous studies have shown that although there are systematic biases and errors in odds, they cannot be profitably exploited by inferring betting opportunities from historical outcome distributions. Instead, more sophisticated forecasting methods are required, suggesting that fixed-odds betting markets for football are at least weak-form efficient. However, even studies in which information-rich statistical models have been used in forecasting, abnormal returns have not been achieved consistently during the past decade. While some authors have been able to use statistical forecasting models to exploit the biased odds in certain limited circumstances (like at the end of a league season), none of the studies that we reviewed were able to "beat the odds" consistently over a longer time period. Studies by, for example, Forrest et al. (2005) and Vlastakis et al. (2009) hypothesize that this increase in efficiency is due to the increased competition enabled by the internet.

Six different factors have been identified to be relevant in football results forecasting: team performance, team-specific home advantage, pairwise home advantage, relative match importance and effects of player injuries. The first two of the six – team performance and team-specific home advantage – have been identified as the most relevant to statistical forecasting, and many different ways to model these factors have emerged. According to our analysis, the ELO model by Hvattum & Arntzen (2010) is the best way to incorporate

historical match data into a statistical forecasting model, although the specification introduced by Hvattum & Arntzen (2010) does not model team-specific home advantage in any way.

Now that we have established what is relevant in football results forecasting and how forecasting models should be developed, we can move on to describe our empirical study on the topic. In the following chapter, we describe our test setup and thus return to the research questions originally presented in the introduction.

# 3 DATA AND RESEARCH METHODS

In this chapter, we describe how our forecast model is constructed and how its performance is assessed. This is done in four parts: First, we describe the data we use in our research. Second, we introduce the regression model we use. Third, we describe the variables used in the model. And finally fourth, we describe how the model is used in forecasting and how its forecasting performance is evaluated.

## 3.1 Data

Our primary dataset consists of the match results from the English Premier League (EPL) from seasons 1993-1994 to 2011-2012, totaling up to 7384 matches. From these matches the following information is collected: the names of the home and away teams, the game date and the end result of the match (home win, draw or away win). This dataset is provided by Footballdata.co.uk (2012a). This dataset is of our primary interest as the most important components of our model are built from it, and as parts of it are used in our model's performance assessment.

Several secondary sets of match results were also collected: historical FA Cup results provided by the FA (2012b), Europa and Champions League results provided by EuroCupsHistory.com (2012), and Football League Cup results provided by Capital One Cup (2012). All of these datasets contain match results of English football clubs from seasons 1995-1996 to 2011-2012. As with our primary dataset, the names of the home and away teams, the game date and the end result of the match are collected. From the Europa and Champions League datasets we also recorded which team proceeded to the next round in the knockout stages. These datasets are used to derive several additional variables for our forecast model.

In addition to the match results data, we have collected the coordinates of football stadiums in England. Coordinates were collected for those teams that played in the EPL during seasons 1995-1996 to 2011-2012. This information is provided by Google Maps (2013). Pairwise

distances between the stadiums are estimated from the coordinates to produce a distance-variable for our model.

Finally, we have also collected the average and maximum fixed odds for the EPL matches from seasons 2005-2006 to 2011-2012. These were provided by Footballdata.co.uk (2012a).[9] The odds are used to assess the performance of our forecast model, as they allow us to draw comparisons between the performance of our model and the forecasting abilities of the bookmakers.

The use of these datasets is described more specifically in later parts of this chapter. Before we get into those parts, we discuss the estimation procedure of our forecast model.

## 3.2 Using an ordered logit regression model for football results forecasting

Our choice of the model estimation method is determined by our forecasting need. To explain the rationale behind choosing an ordered logit model, we first briefly discuss several other alternatives.

As the purpose of this study is to estimate probabilities for future football match results, the chosen estimation method must be able to produce probability estimates for events with discrete outcomes. This requirement is evident from the fact that the result of a football match is always either an away win (0), a draw (0.5) or a home win (1). A common approach to model the probabilities of such discrete outcomes as a function of dependent variables is binary logistic regression. In this model, the logistic function is utilized in modelling the odds of two discrete outcomes as a linear combination of the independent variables (Rouhani-Kalleh, 2006). However, as the name of the model hints, binary logistic regression can only

[9] The average and maximum fixed odds are calculated by Footballdata.co.uk from data collected from Betbrain and Betbase. The odds are collected on Friday afternoons for the weekend games and on Tuesday afternoons for the midweek games (Footballdata.co.uk, 2012b).

account for an event with two discrete outcomes. As our problem has three, binary logistic regression is not suitable for our problem.

In order to account for more than two outcomes, the binomial model can be generalized into a multinomial version – a multinomial logit model. This model uses a combination of independent binary logistic regressions to produce the probability estimates for an event with more than two discrete outcomes. More specifically, in multinomial logit for $K$ discrete outcomes, $K-1$ independent binary logistic regressions are constructed (Greene, 2008, 720). In this setting, one outcome is chosen as the "baseline" outcome, and then the rest of the outcomes are regressed against this baseline (Greene, 2008, 720). It is notable that this approach requires the modeller to estimate more parameters than the binary logistic regression does. More specifically, $K-1$ more parameters are required when compared with the binary logistic regression.

While the multinomial logit seems like a suitable candidate model for our problem, it does not fulfil all of our requirements. More specifically, it does not account for the ordinality of the football results. The football results are ordinal by nature, as the distance between an away win and a draw is shorter than the distance between an away win and a home win (Constantinou, 2012b). To illustrate this, consider a simple example. Suppose that the home team has a lead of one goal. In this position, the away team is required to score one goal in order to convert a loss into a draw, and two goals in order to convert a loss into a victory. As the multinomial logit can only account for nominal outcomes, it fails to account for the ordinality between the different outcomes. Thus, even the multinomial logit is not sufficient to address our modelling requirements.[10]

To address the requirement of ordinality as well as the requirement of discreteness, we use an ordered logit model in our study. The ordered logit model fulfils both of our requirements as it

[10] This statement is a crude generalization done for the sake of keeping this introduction brief: While the multinomial logit does not model ordinality, there might still be cases where a multinomial logit model should be preferred over ordered the logit or some other ordinal model. We address this question more elaborately in part 3.3.2 of this chapter.

is suitable for producing probability estimates for outcomes that are discrete and ordinal (Davidson & MacKinnon, 2004, 446). As a further case in favour of the ordered logit, models similar to it have been widely used in the football results forecasting literature[11]. Given our requirements as well as the backing of previous authors, we specify our forecast model as an ordered logit model. In the next part of this chapter, we describe how this model is estimated and how the forecasts are produced with it.

### *3.2.1* Description of the general ordered logit model

The central idea of the ordered logit model is as follows. Suppose that there is an observable discrete variable and an unobservable continuous variable. Then, suppose that the observable discrete variable is governed by the continuous variable. More specifically, suppose that each value of the continuous variable corresponds to a set of probabilities, which states how likely it is to observe a particular value of the discrete variable. Finally, suppose that the continuous variable is a function of some independent observable variables plus a disturbance term. (Jackman, 2000)

Thus, the ordered logit model is essentially about establishing a connection between the independent variables and a discrete dependent variable with the help of an artificial (and thus unobservable) continuous dependent variable. This connection is done by assuming that the continuous variable is a function of the observable independent variables and a disturbance governed by a distribution whose properties are known. A more detailed illustration of the model is presented in the following paragraphs. We start this illustration by describing the assumptions of the model's stochastic part – the randomness governed by a standardized logistic distribution.

In the ordered logit model, it is assumed that the randomness – or the disturbance – is governed by a standardized logistic distribution. A logistic distribution is a symmetric

[11] See for example Goddard & Asimakopoulos (2004) or Forrest et al. (2005)

distribution which in many ways resembles the normal distribution[12]. Generally, a logistic distribution has a probability density function (PDF) of form

$$\frac{e^{-\frac{x-\mu}{s}}}{s\left(1+e^{-\frac{x-\mu}{s}}\right)^2},$$

a cumulative distribution function (CDF) of form

$$\frac{1}{1+e^{-\frac{x-\mu}{s}}}$$

and a variance of form

$$\frac{s^2\pi^2}{3}.$$

In this formulation, $s$ is a scaling parameter and $\mu$ is the mean value. For the standardized logistic distribution, $\mu = 0$ and $s$ is chosen so that the variance equals 1. Hence, the standardized logistic distribution has the following PDF

$$\frac{e^{-\eta}}{s(1+e^{-\eta})^2}$$

and the following CDF

$$\frac{1}{1+e^{-\eta}},$$

where $\eta = \frac{\sqrt{3}}{\pi}x$. In the model, we denote a disturbance originating from the standardized logistic distribution with the term $\varepsilon$.

[12] We discuss the similarities more elaborately later in this section.

As for the observable independent variables in the ordered logit model, we assume that there is a linear relationship between them and the unobservable continuous variable. More specifically, we assume that one part of the continuous variable is a product of two vectors: a vector of independent variables and an equally long vector of constants. To give an analytical expression for this relationship, we use the following notation:

$$y^* = unobservable\ continuous\ variable,$$
$$\boldsymbol{X} = a\ vector\ of\ n\ independent\ variables,$$
$$\boldsymbol{\beta^T} = a\ vector\ of\ n\ constants.$$

We recall from two paragraphs earlier, that $y^*$ is a function of the independent variables and a disturbance. Using the notation presented above and by recalling that the stochastic part is denoted with $\varepsilon$, we come up with the following expression for $y^*$ (Davidson & MacKinnon, 2004, 446):

$$y^* = \boldsymbol{\beta^T X} + \varepsilon.$$

Essentially, this expression states that the unobservable continuous variable is a sum of $n$ observable variables multiplied by $n$ coefficients (the dot product $\boldsymbol{\beta^T X}$) and a disturbance term governed by the standardized logistic distribution.

As we now have an expression for the unobservable continuous variable, we can proceed to explain how it relates to the observable discrete variable. In the ordered logit model, it is assumed that $y^*$ has some threshold points and that the value of the observable discrete variable depends on whether the value of $y^*$ has crossed a particular threshold. As a clarifying simplification, one can think that if $y^*$ is between certain threshold values, it implies that the observable discrete variable is likely to obtain a certain value and less likely to obtain other values. To give a more exact expression for this, we use the following notation:

$$y = the\ observable\ discrete\ variable\ (y = 1, 2, \dots, J),$$
$$j = index\ for\ jth\ discrete\ state\ y\ can\ get,$$
$$J = the\ total\ number\ of\ different\ states\ j,$$

$k = index\ for\ kth\ threshold\ value\ of\ y^*\ (k = 1, 2, \dots, J-1),$

$J - 1 = the\ number\ of\ thresholds\ required$

$to\ express\ y\ as\ a\ function\ of\ y^*,$

$\kappa_k = the\ kth\ threshold\ value.$

By using this notation and by recalling that the disturbance of the model is expressed by the standardized logistic distribution, we can find an analytical expression for the relationship between $y^*$ and $y$. We recall from a few paragraphs ago that the CDF of the standardized logistic distribution is $\frac{1}{1+e^{-\eta}}$. In the ordered logit model, we assume that the probability of observing state *j* of variable $y$ depends on $\boldsymbol{\beta^T X}$ and on the threshold values $\kappa_k$. To illustrate this, let us denote an individual pair of $y$ and $\boldsymbol{X}$ observations with the notation $y_i$ and $\boldsymbol{X_i}$. Given this, we can say that for each $y_i$ the probability of observing a particular state *j* is dependent on $\boldsymbol{X_i}$. Let us denote this dependency with the expression $P(y_i = j|\boldsymbol{X_i})$. Given this, the probabilities for all *J* states of $y_i$ are expressed in the ordered logit model with the following formulas (Davidson & MacKinnon, 2004, 446):

$$P(y_i = 1|\boldsymbol{X_i}) = \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_1}}$$

$$P(y_i = 2|\boldsymbol{X_i}) = \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_2}} - \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_1}}$$

…

$$P(y_i = j|\boldsymbol{X_i}) = \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_j}} - \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_{j-1}}}$$

…

$$P(y_i = J|\boldsymbol{X_i}) = 1 - \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_{J-1}}}.$$

As we can see, this expression implies that the probability of observing a particular value of $y$ depends on the corresponding independent variable vector $\boldsymbol{X}$, the constant vector $\boldsymbol{\beta^T}$, and the constant threshold parameters $\kappa_k$ $(k = 1, 2, \dots, J-1)$. As the expression is built around the

CDF of the standardized logistic distribution, it implies that the probability of observing a particular value of $y$ is also governed by the standardized logistic distribution. We can also see that this expression implies that the states $j$ of variable $y$ are ordinal. This is evident from the fact that the probability of observing states with larger indices $j$ grows as a function of the dot product $\boldsymbol{\beta^T X}$, which in turn grows as a function of the values in the variable vector $\boldsymbol{X}$. Hence, larger values of **X** imply larger probabilities for the larger state indices of $y$, and thus in this model the relationship between **X** and $y$ is ordinal.

It is worthwhile to note that in a model like this, the underlying distribution could be replaced with any other distribution whose PDF can be described analytically. For example, the most common alternative for the ordered logit model is an ordered probit model, which replaces the standardized logistic distribution with the standard normal distribution (Davidson & Mackinnon, 2004, 446). In terms of model selection purposes, the difference between the two is essentially the fact that the standardized logistic distribution has slightly "fatter tails", and thus it places more probability on outliers. Greene (2008, 832), however, argues that the differences between the two are likely to be trivial in practical applications, and thus the choice between the two distributions is more or less arbitrary. We chose the standardized logistic distribution for this thesis because its CDF has an easily understandable and interpretable analytical expression. Also, as previous authors have used either the probit or the logit in their forecast models,[13] it is not necessary to examine the differences between the two symmetric distributions in the context of this thesis.[14]

We have now established the general form for the model we use for producing forecasts in this thesis. Given this, we can proceed to part 3.2.2, where we discuss how this model is estimated on the basis of a data set.

---

[13] See for example Goddard & Asimakopoulos (2004) or Forrest et al. (2005).

[14] It should be noted that while the choice between the standard normal and the standardized logistic distribution is more or less arbitrary, the choice of the underlying distribution in general is not. While the previous literature has more or less settled on using symmetric distributions in these types of models, one could also argue that a skewed distribution might be suitable as well. We touch on this subject in the discussion part of this thesis.

### *3.2.2* Estimating the ordered logit model using maximum likelihood estimation

As a review, we state that the model presented in 3.2.1 consists of four parts: $y$ – the dependent discrete variable, $\boldsymbol{X}$ – the vector of independent variables, $\boldsymbol{\beta^T}$ – the vector of constants specific to $\boldsymbol{X}$, and $\kappa_k$ $(k = 1, 2, \dots, J-1)$ – the set of constant threshold parameters. Out of these four elements, two – $y$ and $\boldsymbol{X}$ – are observable and thus they are useful as they are as the inputs of the model. Meanwhile, $\boldsymbol{\beta^T}$ and $\kappa_k$ are something that is not directly observable from anywhere (at least in our application). Hence they have to be estimated in order to construct the ordered logit model. This estimation is done using maximum likelihood estimation.

The idea behind estimating the ordered logit model with maximum likelihood estimation is as follows. Find estimators for constants $\boldsymbol{\beta^T}$ and $\kappa_k$ that maximize the value of a likelihood function given a set of $(\boldsymbol{X}, y)$ pairs. By selecting the estimators for $\boldsymbol{\beta^T}$ and $\kappa_k$ this way, one ends up with constants that on average describe the relationship between $y$ and $\boldsymbol{X}$ most accurately, given the set of $(\boldsymbol{X}, y)$ pairs as inputs. (Davidson & MacKinnon, 2004, 399-444)

For a more specific description, let us define the following objects:

$$\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} = an\ estimate\ of\ \boldsymbol{\beta^T}$$
$$\hat{\kappa}_k = a\ set\ of\ estimates\ for\ \kappa_k\ (k = 1, 2, \dots, J-1)$$
$$\Lambda\left(\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{X_i} - \hat{\kappa}_k\right) = the\ value\ of\ standardized\ logistic\ distribution's\ CDF,\ given\ \left(\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{X_i} - \hat{\kappa}_k\right).$$

Given the notation above, the likelihood function whose minimum value has the "optimal" $\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}$ and $\hat{\kappa}_k$ $(k = 1, 2, \dots, J-1)$ for a given set of $(\boldsymbol{X}, y)$ pairs can be written as follows:

$$\ln L = \sum_{i, y_i = 1} \ln\left[\Lambda\left(\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{X_i} - \hat{\kappa}_1\right)\right] + \sum_{i, y_i = 2} \ln\left[\Lambda\left(\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{X_i} - \hat{\kappa}_2\right) - \Lambda\left(\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{X_i} - \hat{\kappa}_1\right)\right] + \dots$$

$$+ \sum_{i, y_i = j} \ln\left[\Lambda\left(\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{X_i} - \hat{\kappa}_j\right) - \Lambda\left(\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{X_i} - \hat{\kappa}_{j-1}\right)\right]$$

...

$$+ \sum_{i, y_i = J} \ln\left[1 - \Lambda\left(\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{X_i} - \hat{\kappa}_{J-1}\right)\right].$$

In this notation, each sum is taken over those $i$ that satisfy the given condition $y_i = j$. Maximum likelihood estimation is then used to find the minimum value of $\ln L$. Details of the computational procedures used in the maximum likelihood estimation and the proof behind the usefulness of this method to our problem are beyond the scope of this thesis. For a more detailed description of the maximum likelihood estimation and the computational estimation procedures used in producing it, see for example Davidson & MacKinnon (2004, 399-444). For the purpose of this thesis, it is enough to know that the estimation procedure starts by taking a dataset of $(\boldsymbol{X}, y)$ pairs as input. Arbitrary starting values are then given for the constants in vector $\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}$. After that, a computation is performed to find a set of threshold parameters $\hat{\kappa}_k$ which minimize the value of $\ln L$ in the equation presented above. This procedure is repeated several times with different starting values of $\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}$ to ensure that a global minimum for $\ln L$ is found. Once the values for estimators $\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}$ and $\hat{\kappa}_k$ are found this way, the estimation procedure constructs an ordered logit model that can generate probability estimates for $y$, given a vector of variables $\boldsymbol{X}$.

It is possible to do the estimation described in the previous paragraph properly only if the variance of the disturbance term $\varepsilon$ can be assumed to be constant across the sample of $(\boldsymbol{X}, y)$ pairs. If the variance cannot be assumed to be constant, vector $\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}$ cannot be assumed to be constant across the sample, and consequently $\hat{\kappa}_k$ cannot be assumed to be constant either (Long, 2001). This assumption of the constant disturbance term variance is called the proportional odds assumption. If this assumption does not hold, the estimated model is likely to be inconsistent and biased (Williams, 2009). The fact that values of the disturbance term $\varepsilon$ cannot be observed or estimated makes detecting the violations of this assumption

challenging. There are, however, some indirect techniques for testing if the proportional odds assumption is violated. These techniques are described in the next part of this thesis.

### *3.2.3* Testing the proportional odds assumption

One method to test the violation of the proportional odds assumption is to estimate a multinomial logit model and then compare its likelihood function value to the corresponding value produced by the estimation of the ordered logit model. As touched upon earlier in this chapter, in a multinomial logit model $K-1$ binary models are estimated to estimate the probabilities between $K$ outcomes. Thus, the number of estimated coefficients the multinomial logit model requires is $K-1$ times larger than the number of coefficients required for an ordered logit model. Hence the ordered logit is more parsimonious than the multinomial logit. The idea in testing the proportional odds assumption is to compare if the additional coefficients of the multinomial logit provide a significantly better fit over the more parsimonious ordered logit estimation. (Borooah, 2002)

The comparison is done by calculating the likelihood ratio statistic with the following formula:

$$LR = -2(\ln L_0 - \ln L_1),$$

where $\ln L_0$ is the log-likelihood of the estimated ordered logit model and $\ln L_1$ is the log-likelihood of the multinomial logit model estimated with the same data. The test statistic calculated this way follows a $\chi^2$-distribution with degrees of freedom equal to the difference in the number of parameters between the two models. If the test statistic is very large, it provides evidence of violating the proportional odds assumption, as it implies that estimating different sets of coefficients for different threshold values clearly provided a better fit for the data. A very low test statistic does not, however, indicate that the proportional odds assumption is not violated. This is because the test assumes that the ordered logit model is a constrained version of the multinomial logit model, while it is in fact a completely different specification. Hence, the result of the test can never confirm whether the proportional odds assumption is violated: it can only tell whether relaxing the assumption, and thus resorting to

the multinomial logit model over the ordered logit model, would provide a better fit for the data. (Borooah, 2002)

Another way to test for the violation of the proportional odds assumption is the Brant test (Williams, 2009, 551). While the likelihood ratio test described above tests for the entire model jointly, the Brant test tests for each variable of the model individually in the same manner (Long & Freese, 2004, 199-200). Thus it can identify whether there are specific variables in the model that cause the proportional odds assumption to be violated (Long & Freese, 2004, 199-200). Given that the Brant test is similar in the construct as the joint test, the same restrictions apply: the Brant test can only identify if the proportional odds assumption is violated and which variable is the likely cause of the violation. For the mathematical details of the Brant test, see Brant (1990).

### *3.2.4* General model specification of our thesis

As we have now introduced the general framework of our model as well as the estimation procedure behind it, we apply it to our case of football results forecasting. To express the general form of our model, let $y$ indicate the end result of a football match with $J = 3$ possible states:

$$y = 0, \quad an\ away\ win$$
$$y = 0.5, \quad a\ draw$$
$$y = 1, \quad a\ home\ win.$$

Now, the probability for $y_i$ obtaining a particular value is dependent on the values of the corresponding $\boldsymbol{X_i}$ in the following manner:

$$(y_i = 0|\boldsymbol{X_i}) = \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_1}}$$

$$P(y_i = 0.5|\boldsymbol{X_i}) = \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_2}} - \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_1}}$$

$$P(y_i = 1|\boldsymbol{X_i}) = 1 - \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_2}}.$$

With this model, we can express the relationship between $y$ and any $\boldsymbol{X}$. By using the method of maximum likelihood to estimate $\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}$ and $\hat{\kappa}_k$ $(k = 1, 2)$, we can construct a model which forecasts the probability of a particular match result as a function of $\boldsymbol{X}$.

We have now described our dependent variable and the general form of our model. The next step is to determine the independent variables that form the vector $\boldsymbol{X}$. The selection of independent variables of our model is described and justified in the next part of this chapter.

## 3.3 Independent variables used in our model

The description of the independent variables is divided into two parts. The first part describes how ELO ratings are used to derive an independent variable from the match history. The second part describes what other independent variables we derive for our model.

### *3.3.1* ELO rating based match expected score as an independent variable

As pointed out in the literature review, the ELO rating method seems like the best way to encode team performance from league match history. For this reason, Hvattum & Arntzen's formulation (2010) described in the literature view is the basis of our model. However, as also pointed out in the literature review, the authors' formulation does not model team-specific home advantage in any way, even though evidence from other authors would justify including it into the model. When testing with our sample data, we also found evidence of a need to account for team-specific home advantage when using the ELO rating method. These findings are illustrated in Figure 3.1.

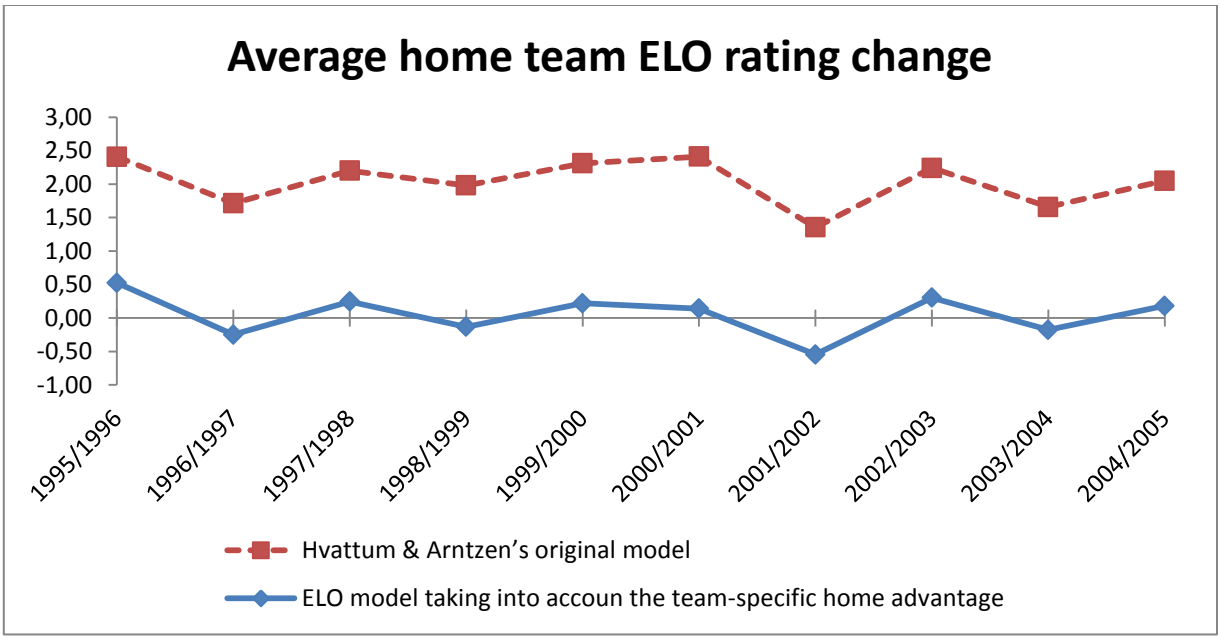


**Figure 3.1 – Comparison of the average home team rating change between two ELO rating systems from season 1995/1996 to season 2004/2005**

In Figure 3.1, the average home team ELO rating change across all matches of each season is plotted from season 1995/1996 to season 2004/2005 for two slightly different ELO rating methods. The dashed red plot is the average home team ELO rating change with Hvattum & Arntzen's (2010) rating assignment procedure that utilizes a single ELO rating for each team that describes the overall performance level of the team. The solid blue plot is the average home team ELO rating change with our enhanced model, where the home and away ELO ratings of a team are calculated separately. In other words, the Hvattum & Arntzen's method uses the same rating to account for both the home and away performance levels of a team – thus assuming that there is no team-specific element in home advantage, whereas our model assumes that there is such team-specific element. From Figure 3.1 we can see that the average home team rating change in the Hvattum & Arntzen's method is consistently above zero by approximately 1.5 to 2.5 rating points. This indicates that the ratings produced with Hvattum & Arntzen's system produce biased estimates of team performance: if the home team ratings are – on average – continuously adjusted upwards throughout the seasons, it means that the system producing the ratings consistently underestimates the performance level of the home

teams (by producing too high expected ELO scores for them) and overestimates the performance level of the away teams (by producing too low expected ELO scores for them). Moreover, as most teams accomplish – on average – better results from their home games than from the away games and because the match schedule for each team usually consists of home and away games alternating one after another, Hvattum & Arntzen's system leads to each team having – on average – a bit too high rating going into an away game (because the rating was last adjusted after a home game) and a bit too low rating going into a home game (because the rating was last adjusted after an away game).

Meanwhile, the average home team ELO rating change with our enhanced model (the solid blue line in Figure 3.1) is roughly zero across the whole sample. This effectively means that the ratings generated by our enhanced model do not seem to suffer from the same bias. As our enhanced model factors team-specific home advantage into the ratings, it seems that introducing this team-specific home advantage into the rating system rectifies a structural bias that is inherent in Hvattum & Arntzen's rating assignment procedure.

As already stated in one of our research questions, we are interested to see if introducing team-specific home advantage would improve the forecasts calculated with Hvattum & Arntzen's (2010) rating system. Based on the findings from previous literature and from our own empirical tests, it seems that introducing team-specific home advantage could lead to improvement in forecast accuracy. To verify this and thus answer our first research question comprehensively, we modify Hvattum & Arntzen's (2010) approach to cater for team-specific home advantage. More specifically, we record the home and away ratings for each team separately and thus assume that the expected (ELO) score of a match is dependent on the difference between the home rating of the home team and the away rating of the away team. With this formulation, we effectively account for two underlying factors (see Table 2.1 for reference) with one variable: by recording the home and away ratings separately, we produce a variable which assumes that team-specific home advantage is an intrinsic part of team's historical performance. Thus, our approach makes slightly different assumptions about the historical performance: on one hand our approach considers home advantage to be more important, but on the other hand we neglect the effect historical home matches might have on

future away games (and vice versa). The technical details of this modification are presented in the next part.

#### 3.3.1.1 Extending Hvattum & Arntzen's ELO model to consider team-specific home advantage

As in the original formulation introduced in part 2.3.2 of this thesis, let $\alpha^H$ and $\alpha^A$ represent the outcome scores of the home and away teams respectively so that

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

and $\alpha^A = 1 - \alpha^H$. Then, as a contrast to the original model, we define the respective expected scores of the home and away teams, $\gamma^H$ and $\gamma^A$, as functions of the teams' previous home and away ratings. Thus the expected scores for both teams participating in a game are defined as

$$\gamma^H = \frac{1}{1 + c^{\frac{\iota_{0H}^H - \iota_{0A}^A}{d}}} \quad \text{and}$$

$$\gamma^A = 1 - \gamma^H = \frac{1}{1 + c^{\frac{\iota_{0A}^A - \iota_{0A}^H}{d}}},$$

where $\iota_{0H}^H$ is the home team's pre-game rating derived from the home team's previous home matches, and $\iota_{0A}^H$ is the away team's pre-game rating derived from the away team's previous away matches. As in the original formulation, $c$ and $d$ are scaling parameters. Given that the home and away ratings are recorded separately, the post-match rating adjustment is made with the following formulas:

$$\iota_{1H}^H = \iota_{0H}^H + k(\alpha^H - \gamma^H) \text{ and}$$

$$\iota_{1A}^A = \iota_{0A}^A + k(\alpha^A - \gamma^A),$$

where $\iota_{1H}^H$ and $\iota_{1A}^A$ are the adjusted post-match (home and away) ratings of the home and away teams respectively and $k$ is a parameter which defines the rate of change in the ratings. As in

the original formulation, the home team's score expectation $\gamma^H$ is used as a variable in the forecast model.

As for the model's scaling parameters, we set $c = 10$ and $d = 400$ in accordance with Hvattum & Arntzen's model (2010). These two parameters serve only to set an appropriate scale for the ratings and alternative values would give identical rating systems. However, the value for the parameter $k$ that determines the rate of change of the ratings must be chosen carefully and the suitable value depends on the chosen scaling parameters $c$ and $d$. The selection of an appropriate value for $k$ is described in detail in the next part.

#### 3.3.1.2 Determining the appropriate value for the parameter k

As mentioned in Section 2.3 of the literature review, all forecast models that use historical data as proxy for the future performance face the question of how much the recent history should be weighted in comparison to the older data. In the ELO model the question boils down to the choice of the parameter $k$, as it defines how much the ratings change after each match. On one hand, if $k$ is too low, the model reacts too slowly to changes in the relative performance levels between the teams. On the other hand, if $k$ is too high, the model becomes too erratic as it "overreacts" to individual match outcomes by adjusting the ratings too much. To strike a balance between reactivity and stability, an empirical estimation is required.

Hvattum & Arntzen (2010, 465-466) used results from English football matches from seasons 1995/1996 to 1999/2000 to calibrate $k$. In their calibration, the authors recorded the average quadratic information losses of the forecasts over the dataset for different values of $k$, and thus established what value of $k$ minimized the quadratic information loss in the calibration sample (Hvattum & Arntzen, 2010, 465). Their testing indicated that the appropriate value for English football is $k = 20$.

As the calibration done by Hvattum & Arntzen (2010) is quite comprehensive, we build our model using the same value for $k$. While it would be interesting to do further testing with samples from – for example – different countries, the measurement done by the previous

authors suits the purposes of our research questions well enough. Hence, we do not perform further empirical estimations of $k$ for this thesis and settle for $k = 20$.

#### 3.3.1.3 Initial development of ratings and initial ratings of promoted teams

All ELO rating computations require that each team is provided with some pre-game rating. Thus, each team should have a rating even before the first match of our dataset. It is difficult to determine how these initial ratings should be assigned to each team and we have solved this by setting the initial home and away ratings of all teams to be equal. As a result, the ratings cannot be considered as reliable indicators of team strength until a sufficient amount of games have been played. We address this issue by using the first two seasons, 1993/1994 and 1994/1995, solely for ratings development. In other words, the match results from these seasons are only used for developing the ELO ratings of teams to an appropriate level. Two full seasons translate into 38 home games and 38 away games for each team and this should be enough to establish baseline ratings for individual teams.

Another issue to be addressed is how to determine a proper rating for those teams that have just been promoted to the Premier League. These teams either have not played at the Premier league level before, in which case we do not have any match history data – and therefore no rating – for them, or – if they have played in the Premier League before – the existing rating could be outdated as it has not been updated since the team was relegated. Hvattum & Arntzen (2010), who examined multiple levels of English football leagues, made the assumption that ratings are directly transferred from one league to another without any transformations. As we do not have data from other English leagues, we cannot do similar transfers. Instead, we approximate the ratings of the promoted teams with the ratings of those teams that were relegated at the end of the previous season. More specifically, we calculate the arithmetic mean of the ratings of the relegated teams from the previous season, and assign this mean rating to all promoted teams as their initial rating for the current season.

### *3.3.2* Other independent variables

Now, as we have described an independent variable that accounts for the historical performance and team-specific home advantage, we can move on to modelling the rest of the

factors discussed in Section 2.1. Table 3.1 presents a summary of the variables we construct for our model.

| Factor | Description | Corresponding variables(s) in our model |
|---|---|---|
| Team performance level | How well the team has recently performed | $\gamma^H$ - ELO rating based expected score of home team |
| Team-specific home advantage | How well the team has recently performed at home in comparison to away performance | $\gamma^H$ - ELO rating based expected score of home team |
| Pairwise home advantage | How home advantage manifests in specific fixtures | $DIST$ - Natural logarithm of the distance between teams' stadiums |
| Cup effects | How participation in external cups affects league match outcomes | $PreviousExtMatchResult$ – Home and away team's latest match results in external competitions |
| | | $FutureExtMatch$ – Indicates, if participating teams have external matches in near future |
| | | $PastGames$ – Indicates differences in historical schedule congestion for participating teams |
| Player injuries | How injuries of key players affect match outcomes | Not modelled |
| Relative match importance | How differences in match importance affect match outcome | Not modelled |

**Table 3.1 – Summary of the factors affecting match outcomes and their corresponding variables in our model**

As shown in Table 3.1, in addition to using the $\gamma^H$ variable, we construct several other variables to model the factors that previous authors have shown to have an effect on football match outcomes. To expand on the summary, we next describe these additional variables more explicitly.

### 3.3.2.1 Estimating pairwise home advantage with pairwise distance

As discussed in Section 2.2.1 of our literature review, pairwise home advantage has been shown to be a significant determinant of match outcomes. Moreover, as previous authors have shown that the pairwise distance between the stadiums can be used as proxy for it, the factor is relatively easy to model. Because of these reasons, we include this variable, $DIST$, in our model as well. For this variable, approximations of pairwise distances are calculated from the stadium coordinates using the great circle formula (Vincenty, 1975). Goddard & Asimakopoulos (2004, 54) suggest that the natural logarithm of the distance provides a better fit for the model than the distance by itself. This makes sense when we consider that the distance is proxy for the audience distribution: in case of longer distances, additional kilometres do not necessarily change the audience distribution significantly, and thus it makes sense to "squeeze" the distances from the ends with a logarithmic transformation. Figure 3.2 illustrates how the transformation affects the variable's values.

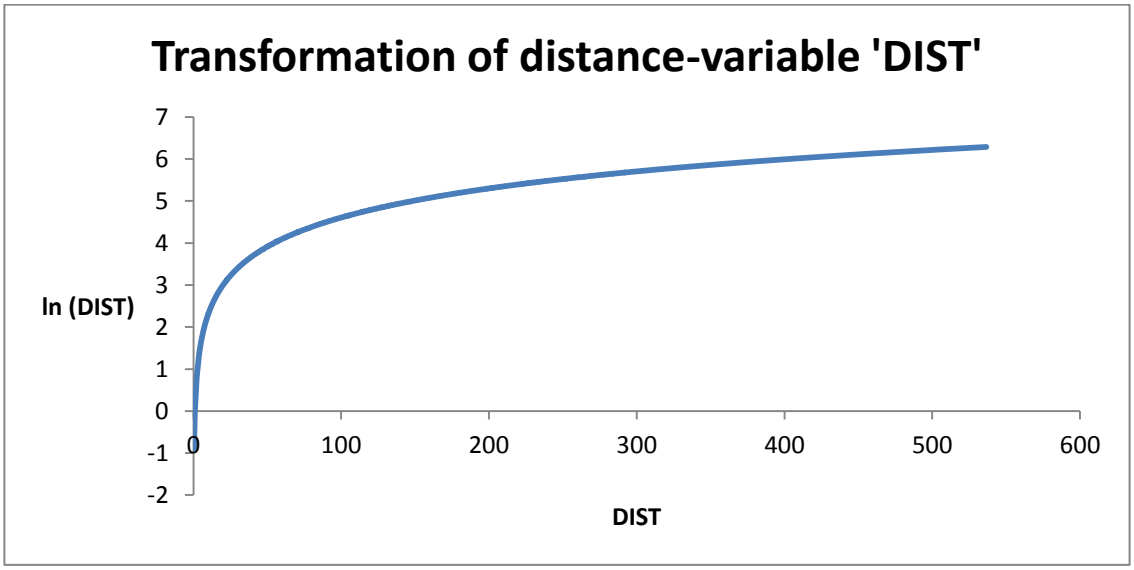


**Figure 3.2 – Relationship between logarithmic distance and distance in kilometres**

### 3.3.2.2 Modeling cup effects

As for the cup effects, our literature review concluded that external competitions – which in the case of the English Premier League are the League Cup, the FA Cup, the Europa League and the Champions League – can affect the league match results. As already mentioned,

Goddard & Asimakopoulos (2004) discovered that participation in the FA Cup had a positive effect on the league performance across professional English Football Leagues, and Constantinou et al. (2012) modeled for fixture-congestion induced fatigue in their model design. In the spirit of these findings, we construct the following variables into our model:

$$PreviousExtMatchResult,$$
$$FutureExtMatch_{LC},$$
$$FutureExtMatch_{FA},$$
$$FutureExtMatch_{CL},$$
$$FutureExtMatch_{EL},$$
$$PastGames.$$

In this notation, the subscripts of $FutureExtMatch$ are as follows: LC corresponds to the League Cup, FA corresponds to the FA Cup, CL corresponds to the Champions League and EL corresponds to the Europa league (and hence all external cup competitions are tracked with separate variables).

The first variable, $PreviousExtMatchResult$, indicates the difference between the results the home and away teams got from their previous external matches if the external matches were played within ten days of the game whose result we want to forecast[15]. The encoding of the variable is represented in Table 3.2.

[15] The variable also considers each external match only once. For example, if an external match is followed by two league games within the ten-day period, the external match only affects the first one.

| Variable value | External match result of home team | External match result of away team |
|---|---|---|
| -2 | LOST | WON |
| -1 | DRAW or No recent external match | WON |
| 0 | DRAW or No recent external match | DRAW or No recent external match |
| 0 | WON | WON |
| 0 | LOST | LOST |
| 1 | WON | DRAW or No recent external match |
| 2 | WON | LOST |

**Table 3.2 – Encoding of variable $PreviousExtMatchResult$**

As shown in Table 3.2, the idea of this variable is to measure the short-term effect recent success or failure in an external cup match could have on the result of a future league match. This kind of 'momentum' is not captured in the ELO rating based expected scores and thus it needs to be measured separately. As Goddard & Asimakopoulos (2004) reported that staying in the FA Cup increased the future league performance of a team, we expect this variable to behave in the same way. Thus, our null hypothesis is that the coefficient of this variable is positive, as then the home team's success (failure) in an external cup competition would imply a higher chance of a home win (away win) and vice versa for the away team.

The $FutureExtMatch$-variables (all variations of it), consider whether the result of a league match is affected by the fact that either (or both) of the teams has an external match upcoming in the near future. More specifically, for each variation of this variable the value of is encoded according to Table 3.3.[16]

[16] As with the variable $PreviousExtMatchResult$, these variables also takes each external match into account only once.

| Variable value | External match coming for the home team in the next 5 days | External match coming for the away team in the next 5 days |
|---|---|---|
| -1 | NO | YES |
| 0 | YES | YES |
| 0 | NO | NO |
| 1 | YES | NO |

**Table 3.3 – Encoding of "$FutureExtMatch$"-variables**

As shown in Table 3.3, the idea of these variables is to measure if teams place more emphasis on the external competition than on the league games (e.g. by resting key players in the league match). Thus, if negative coefficients are estimated for these variables, then it is likely that the teams emphasize the given external competitions over their league efforts on the short term. Conversely, if positive coefficients are estimated, it is likely that an upcoming external match could have a morale-increasing effect on the performance similar to the one Goddard & Asimakopoulos (2004) reportedly observed in the case of the FA Cup and English Leagues.

The final variable concerning the cup effects, $PastGames$, measures the effect a congested match schedule (induced partly by external cup competitions) has on the league performance of a team. The variable is encoded according to the rules presented in Table 3.4.

| Variable value | Home team had a match three days before forecasted match | Away team had a match three days before forecasted match |
|---|---|---|
| -1 | YES | NO |
| 0 | YES | YES |
| 0 | NO | NO |
| 1 | NO | YES |

**Table 3.4 – Encoding of variable $PastGames$**

As shown in Table 3.4, the variable measures what effect relative differences in short-term schedule congestion has on future league performance. As a null hypothesis, we expect this variable to have a positive coefficient, as it is likely that the team that has had more rest has an edge in the upcoming match.

As already evident in the specifications, all cup effect variables are encoded to measure the differences between the teams. In other words, these variables assume that their effects are symmetric for the home and away teams and are thus independent of home advantage. This is a plausible assumption, as team-specific and pairwise home advantage is already accounted for in variables $\gamma^H$ and $DIST$.

#### 3.3.2.3 Factors omitted from our model

As shown in Table 3.1, we have chosen not to include player injuries or the relative match importance in our model. These factors are left out because we consider that modelling them is relatively complex in comparison to the benefits they could bring.

As mentioned in the literature review, variables that model player injuries are often omitted from forecast models due to the fact that constructing them requires accurate and timely data about the health statuses of different players. As there are no easily accessible records of what players were injured before a historical match, collecting this kind of information for historical match results is quite difficult. In addition, using the information in a statistical model would be quite challenging. On one hand, it would not be plausible to model every single player, and on the other hand, it would be difficult to objectively say which players are important enough to justify individual modelling. Selecting the important players would be difficult also because the relative importance of an individual player within a team can shift even within a season. For these reasons, we have chosen to omit any player injury related variables from our model. Even though their inclusion could provide improvements to the forecasts, they are too difficult to use for the purposes of this study.

As also mentioned in the literature review, Goddard & Asimakopoulos (2004) were able to identify differences in the relative match importance between the opposing teams. More specifically, the authors developed an algorithm that identified fixtures that had uneven performance incentives between the opposing teams (see Section 2.2.2 of this thesis for details). While the variables produced by the algorithm were proven to be significant in predicting match results (Goddard & Asimakopoulos, 2004, 56) in the top four English football leagues, we omit them from our study because we expect them to be less efficient

when only English Premier League (EPL) matches are analysed. This is because the incentives are likely to be different in the EPL than what they are in English football in general. In the EPL, there are relatively high financial incentives for finishing as high as possible in the final standings, and thus every match should be important for every team. As the prize structure in the leagues below the EPL does not have similar incentives, promotion and relegation carry relatively higher weight in them, and thus variables modelling relative match importance are likely to be efficient predictors only in lower leagues. As the algorithm that produces the variable is relatively complicated to construct and as we expect it to be relatively inefficient, we choose to omit it from this study.

### *3.3.3* Summary of independent variables used in our study

As explained in the paragraphs above, we have constructed a group of candidate variables for our model. Table 3.5 below presents a summary of these variables.

| Independent variable | Description |
|---|---|
| $\gamma^H$ | Measures the expected score for the home team of the forecasted match based on historical home and away performances of home and away teams. |
| $DIST$ | Measures logarithmic distance between the home and away teams' home stadiums (a measure of pairwise home advantage). |
| $PreviousExtMatchResult$ | Measures differences in recent performance in matches of external competitions |
| $FutureExtMatch - variables$ | Measure differences in upcoming external cup competition schedules |
| $PastGames$ | Measures differences in short-term schedule congestion |

**Table 3.5 – Summary of model's candidate variables for forecasting future match results**

As shown in Table 3.5, we have constructed eight possible variables for our model. In the next part of this chapter, we describe in more detail how these variables are used in our model to derive forecasts.

## 3.4 Forecasting procedure and measurement of model's forecasting efficiency

As already established in the literature review, the efficiency of a football results forecast model can be measured by comparing its output with bookmakers' fixed-odds and their implied forecasts. In this part of the chapter, we explain how the forecasting with our model is done in this study, and what metrics are used to evaluate our model.

### *3.4.1* Description of our forecasting procedure

In our forecasting procedure, we split our match data sample – the EPL results from season 1993/1994 to season 2011/2012 – into three parts: a preliminary sample for computing the starting ELO ratings for all teams, an initial training sample and an evaluation sample. The preliminary sample consists of the first two seasons – 1993/1994 and 1994/1995 – and is used only for developing the (initially arbitrary) ELO ratings to a reliable level as described in 3.3.1.3. After this, the initial training sample is used for fitting the forecasting model while the

evaluation sample is used for measuring the efficiency of the model. The initial training sample has match data from season 1995/1996 to season 2004/2005 and the evaluation sample has match data from season 2005/2006 to season 2011/2012. A more detailed description of the use of the training and evaluation samples is presented next.

The initial training sample is used to fit the model. Once this fitting is completed, forecasts for the first matchday of the evaluation sample are produced. For the purposes of this study, a matchday is defined as a date on which one or more matches are played. Thus, a football season consists of several dozens of matchdays. Once the forecasts for the first matchday are produced, they are evaluated against the actual outcomes of the matches on that day and against the odds offered by bookmakers for the same matches.[17] After the evaluation is done, the information from the first matchday is added to the training sample. Then a new model is fitted using the – now several matches larger – training sample, and this new model is used to produce forecasts for the second matchday. Similar evaluation comparisons between the actual results and the odds are made for the second matchday. After these comparisons, the training sample is again extended with the data from the second matchday and a new model is again fitted to produce forecasts for the third matchday. This procedure is continued until we have produced forecasts for all the matchdays in seasons 2005/2006-2011/2012. The procedure is carried out with a program written with statistical software package R and its cumulative link model extension (Christensen, 2012). Figure 3.3 summarizes how this forecasting procedure is carried out in our study.

[17] The comparison methods are explained in more detail later in this chapter.

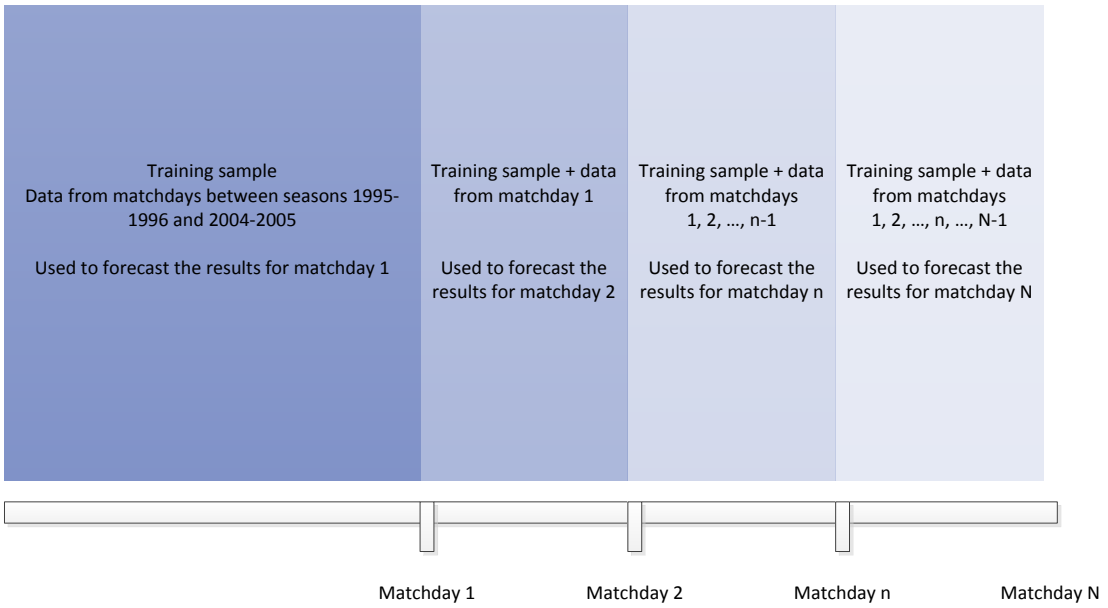


**Figure 3.3 – Summary of our forecasting procedure**

As shown in Figure 3.3, the forecasts generated with this procedure will use all information that would have been available before any given matchday. Hence it makes efficient use of history, but is not gullible for any hindsight bias that would make efficiency measurements unreliable. As also evident in Figure 3.3, the model is fitted $N$ times in the entire procedure, where $N$ is the total number of matchdays in our evaluation sample. When compared with similar studies, the number of model fittings done in this study is quite large. Goddard & Asimakopoulos (2004) and Hvattum & Arntzen (2010), for example, fitted their models only at the start of each season. However, fitting the model that seldom can result in the loss of model efficiency, as maximum likelihood models generally benefit from the increased sample size in terms of increased consistency and efficiency. This can be relevant especially during the last matchdays of each season, as by then there are already several hundred matches that could improve the model fit. Forrest et al. (2005, 560) also recommend fitting the model as often as possible, as doing it would be plausible in the case of real betting as well and, as it could result in improved forecast accuracy. Hence, the only reasons for not fitting the model as often as possible are related to restricted computer processing time. As we did not experience abnormally long processing times when using R for this procedure, we have no reason to restrict our model fitting on the basis of performance restrictions.

The independent variables for the model are generated in the following manner. The ELO ratings are updated after each matchday, and thus the expected scores, $\gamma^H$'s, for each upcoming match are influenced by what happened in all previous matchdays for the teams participating in any given match. The other variables (see Table 3.5) are generated by combining the information of the secondary datasets (e.g. the FA Cup fixtures) with the information of the primary dataset. For example, $FutureExtMatch_{CL}$ for a match is derived by determining whether either of the participating teams has a Champions league match scheduled within the maximum of five days after the EPL match of interest.[18]

While we expect $\gamma^H$ to be significant throughout the whole training sample, it is possible that our relatively small sample size can cause problems in the case of other variables. Sample size can become an issue especially in the cases of those variables that indicate the participation in external competitions, as the number of teams participating in an individual external competition is relatively small throughout the entire sample. Hence it is possible that some variables start to improve the model fit only after the training sample has grown to be substantially larger than what it is at the start of the season 2005/2006. We address this issue by fitting the model with R's backward/forward stepwise-algorithm. The algorithm tries many combinations of the variables, and then fits the model that minimizes the information loss, estimated with Akaike Information Criterion (Akaike, 1974). With this procedure, forecasts for each matchday are produced with those variables that – based on historical results – minimize the information loss given the possible variables listed in Table 3.5

#### 3.4.1.1 Sensitivity analysis by running multiple permutations of our model

In addition to doing a stepwise model selection across the eight available variables, we run two alternative versions of the model as well. This is done to address some of our research questions more explicitly.

[18] The algorithms used to derive all the variables are available from the authors upon request.

The first two alternative versions relate to our research question about improving Hvattum & Arntzen's model. As stated in the introduction, we are interested to see whether our specification of the ELO model is significantly better than the specification proposed by Hvattum & Arntzen (2010). To do this, we generate variable $\gamma^{H*}$ with Hvattum & Arntzen's specification introduced in Section 2.3.2 of this thesis, and then fit the model for each matchday with $\gamma^{H*}$ as the only independent variable (and thus without the need for stepwise selection). To compare whether $\gamma^{H*}$ is a better predictor than $\gamma^{H}$, we run the whole forecasting procedure described in the previous paragraphs also with $\gamma^{H*}$. If there are significant differences in the results produced by these two specifications, we are able to conclude which way of deriving ELO ratings is better for forecasting. For generating values for $\gamma^{H*}$ we use the same parameters as we used with $\gamma^{H}$.

By running these two alternative versions, we are also able to compare whether the version with multiple variables (which employs secondary data sources) is superior to the simple specification. Therefore we also get an answer to the research question that considers the role of secondary data sources in football results forecasting.

#### 3.4.1.2 Start of the season adjustments

One issue that also needs to be addressed in forecasting is the fact that at the start of the season the ELO ratings inherited from the end of the previous season do not necessarily reflect the true performance levels of the teams. This is because player transfers and managerial changes during the summer can change the structure of teams considerably between the seasons. Moreover, as discussed in 3.1.1.3, the ELO ratings for the promoted teams are approximated with the ratings of those teams that were relegated at the end of the previous season, and this approximation is not necessarily realistic. Given these circumstances, it is possible that our estimations for the first few matchdays of a season can be, if not biased, at least somewhat inaccurate.

We address this issue by not including the first two home and away games of each team in each season in our model fitting or forecasting procedures. By doing this, we let the ELO ratings converge to a more accurate level for each team before using them in model fitting,

and thus we do not decrease the efficiency of the future forecasts by including matches, whose variables would most likely not reflect the teams' true performance. As we do not forecast these games either, we effectively omit the start of the season from our study. While this results in inability to say anything about how well our model fares in the first few matches of each season, the omitted number of matches is only about 10% of the entire evaluation sample. Thus, this treatment does not diminish the reliability of our study when the long-run performance of a forecast model is considered.

### *3.4.2* Measuring our model's forecasting efficiency

As mentioned earlier, match outcomes and bookmakers' historical odds are the basis for assessing the performance of our model. More specifically, we use two types of metrics to assess the efficiency of our model: forecast accuracy metrics that assess the forecasting accuracy directly, and economic metrics that assess the forecasting accuracy indirectly via betting profits. Metrics related to these estimations are described below. After introducing the metrics, we also discuss statistical tests that are used to make inferences about the differences between the different forecast models.

#### 3.4.2.1 Forecast accuracy metrics

Hvattum & Arntzen (2010) recommend using the Brier score as a measure of forecast accuracy. The Brier score is a tool for measuring the accuracy of probabilistic forecasts that assign probabilities for mutually exclusive and collectively exhaustive outcomes. This measurement is done by measuring the quadratic distance of each estimated probability from the actual outcome of the event (Brier, 1950). Hence, the smaller the total quadratic distance is, the better the forecast is. The average Brier score for a set of forecast-event pairs is defined as follows (Brier, 1950):

$$BS = \frac{1}{M}\sum_{t=1}^{M}\sum_{i=1}^{R}(f_{ti} - o_{ti})^2 ,$$

where $M$ is the number of forecast-event pairs, $R$ is the total number of possible outcomes for an event, $f_{ti}$ is the estimated probability for event $t$'s outcome $i$ to occur, and $o_{ti}$ is a binary

indicator for each outcome possibility of match $t$ (1 if the outcome occurred, 0 otherwise). For an individual forecast-event pair, the Brier Score is

$$BS_{indiv} = \sum_{i=1}^{R} (f_{ti} - o_{ti})^2,$$

with similar notation as in the formulation for the average Brier score. In this thesis, the individual Brier scores are needed in order to test whether two different forecast methods produce significantly different average Brier scores. The method for testing this is covered later in this chapter.

Constantinou & Fenton (2012) note that the Brier score does not address the ordinality of football match results properly when scoring the forecasts. Constantinou & Fenton (2012) argue that the Rank Probability Score (RPS) is a more appropriate measure of forecast accuracy, as it accounts for this ordinality by acknowledging that the distance between a home win and an away win is smaller than the distance between a home win and a draw or an away win and a draw. The RPS was first introduced by Epstein (1969). In principle the RPS is quite similar to the Brier Score. The major difference is that the score given to a forecast with the RPS is weighted with the number of possible outcomes. The average RPS is defined as follows (Epstein, 1969):

$$RPS = \frac{1}{M} \sum_{t=1}^{M} \frac{1}{R-1} \sum_{i=1}^{R-1} (f_{ti} - o_{ti})^2,$$

where $M$ is the number of forecast-event pairs, $R$ is the total number of possible outcomes for an event, $f_{ti}$ is the estimated probability for event $t$'s outcome $i$ to occur, and $o_{ti}$ is a binary indicator for each outcome possibility of match $t$ (1 if the outcome occurred, 0 otherwise). As with the Brier score, a low average RPS indicates high forecast accuracy, and thus a forecast model that produces a lower average RPS is superior to one that produces a higher average RPS. Similarly, as with the Brier score, the individual Rank Probability Scores of forecast-

event pairs are required for statistical testing. For this purpose, the RPS of an individual forecast-event pair is

$$RPS_{indiv} = \frac{1}{R-1}\sum_{i=1}^{R-1}(f_{ti} - o_{ti})^2,$$

with similar notation to the average RPS presented above.

Constantinou & Fenton (2012) also propose that Epstein's formulation of the RPS should be modified to measure the absolute distance instead of the squared distance. This transformation is necessary as it takes into account the fact that football matches are drawn often because the inferior team plays very defensively (Constantinou & Fenton, 2012). Hence, draws are often a result of the inferior team not even trying to win. Therefore, in cases where the dominant team draws with an inferior one, the result can often be due to pure chance, as the dominant team is likely to have had more opportunities to score (and thus win the match). Thus in the case of a draw, the scoring algorithm should not penalize the forecast which placed more emphasis on the team that was a priori estimated to be more dominant in the match (Constantinou & Fenton, 2012). Because of this reasoning, we use the absolute RPS in addition to using the Brier score and the RPS to measure forecast accuracy. The average absolute RPS is defined as follows (Constantinou & Fenton, 2012):

$$RPS_A = \frac{1}{M}\sum_{t=1}^{M}\frac{1}{R-1}\sum_{i=1}^{R-1}|f_{ti} - o_{ti}|,$$

where $M$ is the number of forecast-event pairs, $R$ is the total number of possible outcomes for an event, $f_{ti}$ is the estimated probability for event $t$'s outcome $i$ to occur, and $o_{ti}$ is a binary indicator for each outcome possibility of match $t$ (1 if the outcome occurred, 0 otherwise). As with the other two metrics, the individual absolute Rank Probability Scores of forecast-event pairs have to be tracked for testing purposes. The absolute RPS for an individual event is

$$RPS_{A\ indiv} = \frac{1}{R-1}\sum_{i=1}^{R-1}|f_{ti} - o_{ti}|,$$

where similar notation as in the case of $RPS_A$ applies.

The three measures described above are used for two purposes. First, they are used for making comparisons between the different model specifications. This is done to infer which independent variables are the most relevant in forecasting. Second, these measures are used for making comparisons between our best model and the forecasts implied by bookmakers' average odds. As explained in 2.1.1, bookmakers' forecasts can be estimated from the odds they offer. Thus by estimating forecasts from the average odds, we get estimates for "the consensus forecasts" bookmakers have made for each game. When the performance of our forecasts is compared with the performance of these consensus forecasts, we can infer how well our model fares against the market.

#### 3.4.2.2 Economic metrics

In addition to the direct accuracy metrics, we evaluate our model also indirectly by measuring how profitable betting with the model could have been based on historical odds. In order to measure this, we use three different betting strategies that place hypothetical bets on the best (maximum) odds that were found for the matches in our sample. In all strategies (unless noted otherwise) a bet is placed on an odd, if the expected value of the odd is larger than one, given the forecast produced by the underlying forecast model. This way we are able to simulate a situation, where a bettor first searches for the best possible odds for a given match, and then – based on her forecast and a predetermined betting rule – decides which odds she would bet on and by how much. This measurement procedure was originally introduced by Hvattum & Arntzen (2010).

The first betting strategy that we use in our evaluation is the unit bet strategy (UNIT). In the UNIT strategy, a bet of one unit is placed on an odd if, given the forecast, the expected value of the wager is estimated to be larger than one (Hvattum & Arntzen, 2010, 464). The profits of the bets placed this way are then averaged across the entire sample to see how profitable our model would have been in conjunction with UNIT. The drawback of UNIT is its naivety: as UNIT makes no difference between large and small expected values, it is not able to vary bet sizes according to them. This can result in situations, where bets are placed too excessively on

outcomes with low probabilities (high odds). Because of this drawback, two more sophisticated betting strategies are used as well.

The second, a more advanced, betting strategy used in our study is the Kelly bet size strategy (KELLY). In KELLY, the bets on odds are placed according to the Kelly criterion (Kelly, 1956). The Kelly criterion is defined as follows (Kelly, 1956):

$$f^* = \frac{bp - q}{b} = \frac{p(b+1) - 1}{b},$$

where $f^*$ is the size of the bet recommended for the outcome expressed as a fraction of the available bankroll, $b$ is the net gain available from placing a bet on the odd, $p$ is the forecasted probability for the outcome and $q = 1 - p$ (Kelly, 1956). In this study, we apply the Kelly criterion by using it as a rule to determine the fraction of the bankroll that should be placed on each wager. If the probabilities $p$ are the true probabilities of match outcomes, betting this way should maximize returns for the long series of bets.[19] As KELLY scales bets in relation to the probability and net odds, it is more sophisticated strategy than UNIT. The drawback of KELLY is, however, the fact that most likely we are not able to estimate the true probabilities of different match outcomes. Hence, it is not likely that the returns are maximized by using KELLY in conjunction with any forecast model. Even if we were able to estimate the true probabilities, the Kelly criterion is very prone to placing very large fractions of the bankroll on high-yield odds. If this high-yield bet fails, then KELLY effectively wastes a large portion of the bankroll with a single bet, and thus it has a high risk of ruining the entire bankroll on short term. Given also the fact that the probability estimates generated by our model are likely to be inaccurate, this problem can potentially be even more severe.

The third betting strategy used in our economic evaluation is the UNIT WIN strategy. In UNIT WIN, the bets are placed on matches so that the expected gross win from the bet is equal to one unit (Hvattum & Arntzen, 2010, 464). Like KELLY, this approach scales bets

[19] For proof of this, see Kelly (1956).

according to the probabilities and odds, but it is less likely to overbet on high-yield odds. Hence when compared with KELLY, this approach should be more conservative.

The results produced by these betting strategies are used for assessing what would happen in the long run, if bets were placed on the odds according to the forecasts produced by our model. This kind of assessment reveals if our model is able to produce returns that are abnormal or even profitable.

#### 3.4.2.3 Statistical tests for observing differences between the different forecasting methods

In order to make conclusions about the results of the metrics described above, three types of t-tests for means are used. These tests are described briefly in the following paragraphs.[20]

As our model produces forecasts for all matches in the evaluation sample, we can use the pairwise t-test to test for the differences in forecast accuracy between different models. This test is applicable to all three forecast accuracy metrics presented in 3.4.2.1. The test statistic for this test is calculated as follows:

$$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{M}}},$$

where $\bar{d}$ is the mean difference of the forecast metric values over a sample of forecast pairs for an event, $s^2$ is the sample variance of these differences, $M$ is the number of events that forecasts are produced for and $t$ is the value of the Student's t-distribution the observed sample represents. With this test, we can make inferences about the mean forecast accuracies the different models produce.

As an example of how the test works, suppose that the forecast model A generates the individual Brier scores of 0.2, 0.3 and 0.4 for events 1, 2 and 3 respectively. Then, suppose

[20] For a thorough description of the principles of t-tests see, for example, Greene (2008).

that the forecast model B generates the individual Brier scores of 0.25, 0.25 and 0.35 for the same events. In this case

$$\bar{d} = \frac{(0.2 - 0.25) + (0.3 - 0.25) + (0.4 - 0.35)}{3} = 0.0167,$$

$$s^2 = 0.333,$$

$$M = 3 \text{ and}$$

$$t = 0.5.$$

With the sample size of three, the test has two degrees of freedom, and thus the corresponding p-value of $t = 0.5$ is $0.67$. Thus based on the test, it is not possible to say which of the forecast models is better.

In the case of the economic metrics, we cannot make pairwise comparisons as we are measuring long-run profitability when betting (selectively) on market odds. In order to infer if the returns of a forecast model/betting strategy combination are different from the average return of the maximum odds (and hence abnormal), a two-sample t-test for unequal sample sizes and unequal variances is done. The test statistic for this test is calculated as follows:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{m_1} + \frac{s_2^2}{m_2}}},$$

where $\overline{X_1}$ and $\overline{X_2}$ are the mean returns of the compared models, $s_1^2$ and $s_2^2$ are the sample variances of the returns, $m_1$ and $m_2$ are the sample sizes and $t$ is again the value of the Student's t-distribution the observed sample pair represents. For our purposes, the sample sizes $m_1$ and $m_2$ are the number of bets each betting simulation places.

In addition to knowing if our returns are abnormal, we are also interested to see if any of the forecast model/betting strategy combinations are profitable. This is tested with one-sample t-test, which tests whether the average return observed with a forecast model/betting strategy combination is different from 0. The test statistic for this test is defined as follows:

$$t = \frac{\bar{X} - 0}{\frac{s}{\sqrt{m}}},$$

where $\bar{X}$ is the mean return of the examined model across the sample, $s$ is the standard deviation of the returns, $m$ is the number of bets and $t$ is again the value of the Student's t-distribution.

It is worthwhile to notice that all three tests used in our study assume that the observed means (or the differences between the means in the case of the pairwise test) are normally distributed. Even though none of the forecast accuracy metrics or betting strategies produce normally distributed data by themselves, the assumption of the normality at the mean is fulfilled with our data. This is because both the forecast accuracy measures and the betting returns have a well-defined mean and well-defined variance, and thus by the central limit theorem, their means follow the normal distribution given that the sample size from where the mean is observed is large enough. Since our sample size consists of over 2000 observations for all measures, we can safely make inferences with t-tests.

## 3.5 Concluding remarks on research methods and data

To summarize, the purpose of this study is to assess how good ELO rating based statistical forecast models are in predicting league football match results. The performance of the models is assessed with two types of metrics: forecast accuracy metrics that compare the forecasts with the outcomes and betting return metrics that compare the returns different forecast models deliver when used in conjunction with a betting strategy. To answer the research questions presented in the introduction, different models are compared with each other and with the market odds.

All models are estimated as ordered logit models using the maximum likelihood estimation. Hence, the models tested in this study differ only in terms of the variables used by the different models. The models estimated in this manner generate probabilistic forecasts for mutually exclusive and collectively exhaustive outcomes; a home win, a draw and an away

win. This kind of forecasting technique is necessary because we want to compare our predictions with the predictions made by the bookmakers which represent the market for betting odds.

The data used in this analysis consists of league matches played in the EPL between the years 1993-2012. The data is divided into two chronologically ordered parts: the training period and the evaluation period. The training period is used for the estimation of the models while the evaluation period is used for forecasting and betting simulations. Hence, the models are evaluated on the basis of the forecasts they generate for the evaluation period.

The forecasting procedure used in this study estimates a model for each matchday based on all match history preceding this matchday. This is done by moving match records from the evaluation sample to the training sample once forecasts have been produced for them. Hence the forecasting procedure simulates a situation, where the forecaster uses all previous information about the examined league to produce forecasts for the upcoming matches. While this procedure emulates the behaviour of a rational bettor, it also allows several weaker variables to "jump" into the model once they become significant for forecasting: while the ELO rating based variables are likely to be significant predictors across both samples, it is possible that several other variables used in our study become relevant only after enough seasons have passed. In each model estimation, AIC is used as a criterion for determining whether a variable is used in forecasting the results of the upcoming matchday.

Now that we have covered our research methods and data, we can move on to presenting the results of our study. This is done in the next chapter.

# 4 RESULTS

In this part of our thesis, we report the results of our testing. The results are organized as follows. First, we report our observations about the initial model specification. Then, we report the observed differences between the ELO rating based variables $\gamma^{H*}$ and $\gamma^{H}$. After these observations, we present the results of the forecast accuracy measurements and betting simulations.

## 4.1 Results of initial model specification

The following part of this chapter describes the results of testing done with the initial training sample. In other words, these results relate to the model specification that is done before any forecasts are calculated for the evaluation sample.

### *4.1.1* Findings about individual independent variables

Test runs with the initial training sample revealed the following information about the variables initially specified for our model:

1. Both ELO rating based variables $\gamma^{H}$ and $\gamma^{H*}$ were significant in all of the test runs, and they had larger coefficients than any other tested variable.
2. The variable measuring the pairwise home advantage, $DIST$, had a significant coefficient, but in absolute terms the coefficient is quite small. Thus the effect of $DIST$ on the forecasts is also small.
3. The variables that measure the effect of the upcoming cup matches and the results of the past cup matches have relatively few observations with non-zero values.
4. $PastGames$ variable, which models the match congestion, does not have a statistically significant coefficient. Whether this is due to the small sample size or the quality of the variable, cannot be determined based on the initial training sample.
5. $PreviousExtMatchResult$ variable, which indicates how opposing teams fared in their recent non-league games relative to each other, is not significant. Whether this is

due to the small sample size or the quality of the variable, cannot be determined based on the initial training sample.

6. The variable group $FutureExtMatch$, which carries information about external cup matches in the near future, contributed mixed results. None of the initially specified variables was statistically significant by itself, but a joint variable that accounted for both European Cups (the Champions League and the Europa League) had a statistically significant positive coefficient. Meanwhile, the variables that carried information about the future FA or League Cup matches were not significant. Based on the results, it seems that participation in the European cup competitions indicates a league success whereas participation in the domestic cups does not.

Based on the observations listed above, we modify the variable group $FutureExtMatch$. More specifically, those variables that consider the Champions League and the Europa League are combined into a single variable, $FutureExtMatch_{Euro}$, while the rest of the variables in the group are left as they are. This allows our model to use information about the European Cup matches in a manner that seems to provide the most significant information based on the training sample. Remaining variables are used as already described in Table 3.5 of Section 3.3.3.

### *4.1.2* Findings about model estimation

The initial model – that is the model used for forecasting results of first matchday – is estimated with matches from the seasons 1995/1996 – 2004/2005. When two first home and away games for each team are excluded from each season, a total of 3384 observations were used in this estimation. The ordered logit model produced with this sample using AIC based stepwise variable selection method is presented in Table 4.1.

| Variable | Estimated Coefficient | Std Error | z-value of Wald test | P >\|z\| |
|---|---|---|---|---|
| $\gamma^H$ | 4.234 | 0.268 | 15.802 | <0.001 |
| $DIST$ | 0.044 | 0.029 | 1.501 | 0.133 |
| $FutureExtMatch_{Euro}$ | 0.222 | 0.115 | 1.937 | 0.053 |
| **Threshold Coefficients:** | | | | |
| $\hat{\kappa}_1$ | 1.63 | 0.212 | 7.68 | <0.001 |
| $\hat{\kappa}_2$ | 2.879 | 0.216 | 13.31 | <0.001 |

**Table 4.1 – Ordered logit model estimated for forecasting results of 1st matchday**

As shown in Table 4.1, the stepwise procedure based on AIC chose only three variables for the initial model: $\gamma^H$, $DIST$, and $FutureExtMatch_{Euro}$. Table 4.1 further illustrates that the coefficient of $\gamma^H$ is significantly larger than the coefficients of other variables.

To see if there is evidence of breaking the proportional odds assumption, the likelihood ratio and Brant tests (see 3.2.3) are done for the model. The results of these tests are presented in Table 4.2.

| Variable | $\chi^2$ | $p > \|\chi^2\|$ | df |
|---|---|---|---|
| Entire model | 4.87 | 0.432 | 3 |
| $\gamma^H$ | 0.19 | 0.296 | 1 |
| $DIST$ | 1.09 | 0.243 | 1 |
| $FutureExtMatch_{Euro}$ | 0.02 | 0.88 | 1 |

**Table 4.2 – Results of likelihood ratio- and Brant tests for model presented in table 4.1**

As the p-values in Table 4.2 show, under the null hypothesis of no evidence of the assumption violation, the probability of observing absolute $\chi^2$ values as large as shown in the table is relatively common. This applies for both the entire model jointly (the Brant test) and for the individual variables (the likelihood ratio test). Therefore, we do not reject the null hypothesis of not observing evidence of breaking the proportional odds assumption. Hence, we find no reason to use some other model than the ordered logit for our forecasts. In addition, as we have a large initial sample of over 3000 observations, we also assume that the variables

selected for this initial model behave in a similar fashion in later models. This is a plausible assumption, since the variables are likely to become only more stable as the sample size increases. Therefore, we do not test for the proportional odds assumption in models that are used to predict the results for the matchday two and onwards. However, if new variables appear into the model in the stepwise selection during the course of the estimations, then the entire model is tested again with the Brant and likelihood ratio tests.

## 4.2 Results of forecast accuracy measurements

In the following paragraphs, we present the results of the forecast accuracy measurements. The results are presented as pairwise comparisons between the different models. This is done to extract the results that are relevant to our research questions. The evaluation sample, which runs from season 2005/2006 to season 2011/2012 and consists of 2364 matches, is used for all models, and hence the results observed for different models are comparable with each other.

The rest of this section is organized as follows. First, we compare the results produced by the univariate models that use $\gamma^{H*}$ and $\gamma^{H}$ respectively as the model's sole independent variable. Then, we compare the $\gamma^{H}$ univariate model with the stepwise-estimated model, which may (based on AIC) choose its variables for each matchday freely from the list of candidate variables we have specified in Section 3.3.3 and adjusted in Section 4.1.2. After this, we compare a model that uses both $\gamma^{H*}$ and $\gamma^{H}$ with the univariate model that uses only $\gamma^{H*}$. Then, based on our observations, we select the best model and compare its forecast accuracy with the forecast accuracy of the average odds.

### *4.2.1* Comparison between univariate $\gamma^{H*}$, and $\gamma^{H}$ models

As stated in the introduction, one of our research questions is about determining whether the model formulation proposed by Hvattum & Arntzen (2010) could be improved when team-specific home advantage is accounted for in the ELO ratings. We test this by comparing two univariate models that use $\gamma^{H*}$ and $\gamma^{H}$ respectively as the sole variables to predict match results. With this comparison, we are able to determine which method of encoding match history is superior (and thus should serve as the basis for models that use other variables as

well). The comparison is done by using the forecast procedure presented in Section 3.4 for both models, with the exception that stepwise model selection is not done at all as both model specifications have only one independent variable. The results of this comparison and relevant test statistics are presented in Table 4.3.

| Model | Mean Brier Score ($BS$) | Mean rank probability Score ($RPS$) | Mean absolute rank probability score ($RPS_A$) |
|---|---|---|---|
| $\gamma^{H*}$ | 0.1922 | 0.1936 | 0.3898 |
| $\gamma^{H}$ | 0.1931 | 0.195 | 0.3901 |
| *P-value of pairwise t-test for means* | *0.1475* | *0.1203* | *0.7265* |

**Table 4.3 – Results of forecast accuracy comparison between models $\gamma^{H*}$ and $\gamma^{H}$**

As Table 4.3 shows, the model which uses Hvattum & Arntzen's (2010) original specification, $\gamma^{H*}$, produces forecasts with slightly better accuracy averages as all scoring rules give lower scores to $\gamma^{H*}$. However, when the averages are compared with the pairwise t-tests (the bottom row), we observe that the differences between the models are not statistically significant. Hence based on our test, it seems that $\gamma^{H*}$ and $\gamma^{H}$ are equally good predictors when used by themselves.

### *4.2.2* Comparison between $\gamma^{H}$ and stepwise-estimated model with multiple variables

Another research question presented in the introduction is about determining if modelling indirect effects with auxiliary datasets would improve the forecast efficiency over the scenario where only league match history based variables are used in calculating the forecasts. This question is answered by comparing forecast accuracy means between the univariate $\gamma^{H}$ model and the stepwise-estimated model. For the model with multiple variables, the variables are selected for each matchday based on AIC during the forecasting procedure.

Before presenting the results of this comparison, we present the specifics of the stepwise-selected model that is used to predict the results for the final matchday of our evaluation sample. In other words, the model presented next is estimated with a sample of ~5300 matches

(the training sample + the evaluation sample – the matches on the final matchday of the evaluation sample). This model is presented to illustrate how the model with larger sample size differs from the initial model presented in Section 4.1.2. The model for the final matchday is presented in Table 4.4.

| Independent Variable | Estimated Coefficient | Std Error | z-value of wald test | P (>\|z\|) |
|---|---|---|---|---|
| $\gamma^H$ | 4.294 | 0.185 | 23.267 | <0.001 |
| $DIST$ | 0.053 | 0.022 | 2.386 | 0.017 |
| $FutureExtMatch_{Euro}$ | 0.228 | 0.084 | 2.727 | 0.006 |
| **Threshold Coefficients:** | | | | |
| $\hat{\kappa}_1$ | 1.6994 | 0.152 | 11.16 | <0.001 |
| $\hat{\kappa}_2$ | 2.9566 | 0.156 | 18.98 | <0.001 |

**Table 4.4 – Ordered logit model estimated for forecasting results of the final matchday**

As shown in Table 4.4, the stepwise-procedure that estimated the model for the final matchday chose exactly the same variables as it chose for the first matchday (see Table 4.1). This result implies that the $PreviousExtMatchResult$ variable, the domestic cup variables or the $PastGames$ variable do not improve the model fit enough to warrant their inclusion even with dataset spanning over 17 league seasons and over 5000 observations. Meanwhile, the variables $\gamma^H$ , $DIST$ and $FutureExtMatch_{Euro}$ are significant in both samples. This observation highlights the fact that the variables which were robust with a sample of 3000 were robust enough to be useful more universally as well.

Given the results shown in Table 4.4, the univariate $\gamma^H$ model is compared with the $\gamma^H + DIST + FutureExtMatch_{Euro}$ model to see whether the extra information provided by the additional variables produces significantly better forecasting accuracy. This comparison is done by comparing the values of the forecast accuracy metrics for two forecast runs on the same sample: the first run produces forecasts to all matchdays using only the $\gamma^H$ variable. The second run uses the variables $\gamma^H$, $DIST$ and $FutureExtMatch_{Euro}$. It should be noted that during the course of the forecast run, both models increase their sample size after each

matchday, as with the passing of each matchday the amount of past information (that is potentially useful for the model) grows. The results of this comparison are presented below in Table 4.5. It should also be noted that the pairwise t-test is done to the sample of differences in the forecast accuracy metrics/forecasted match.[21]

| Model | Mean Brier Score ($BS$) | Mean rank probability score ($RPS$) | Mean absolute rank probability score ($RPS_A$) |
|---|---|---|---|
| $\gamma^H$ | 0.1931 | 0.1950 | 0.3901 |
| $\gamma^H + DIST + FutureExtMatch_{Euro}$ | 0.1928 | 0.1946 | 0.3897 |
| *P-value of pairwise t-test for means* | *0.1985* | *0.1918* | *0.1547* |

**Table 4.5 – Results of forecast accuracy comparison between models $\gamma^H$ and $\gamma^H + DIST + FutureExtMatch_{Euro}$**

As shown in Table 4.5, the inclusion of the additional variables seems to improve the forecast accuracy with all the three forecast metrics. However, as the p-value of the pairwise t-tests indicates, none of the metrics shows that this improvement is statistically significant. Based on this, we therefore say that complementing the ELO rating based model with variables that carry auxiliary information does not improve forecast accuracy substantially.

### *4.2.3* Comparison between "$\gamma^H + \gamma^{H*}$" -model and "$\gamma^{H*}$"-model

Based on the comparisons done in 4.2.1, it seems that $\gamma^{H*}$ and $\gamma^H$ do not differ substantially from each other in terms of forecast accuracy. This is most likely because both variables have different kinds of structural flaws. As illustrated in 3.3.1, $\gamma^{H*}$ is biased when compared to $\gamma^H$. On the other hand, $\gamma^H$is flawed as well since it does not take into account the fact that past home games can explain the results of future away games and vice versa.

[21] For example, for game one of matchday one the first model produces a Brier Score of 0.2500 and the second model produces a Brier Score of 0.2600. The difference is then 0.2500-0.2600 = -0.100, or slightly in favor of first model. The t-test then tells if the differences are on average significantly different from zero.

Given this situation, it is possible that forecasting with a model that utilizes both $\gamma^{H*}$ and $\gamma^{H}$ could result in improved forecast accuracy. Theoretically, this model could outperform the previously introduced models as it has a component that tracks the past performance irrespective of home advantage ($\gamma^{H*}$) in addition to having a component that takes the team-specific home advantage into account ($\gamma^{H}$). To see whether this combined model is better, we run the model estimation/forecasting procedure in which stepwise-algorithm can also choose $\gamma^{H*}$ in addition to choosing the other variables. While the results presented in 4.2.2 indicated that the indirect-effect variables did not improve the forecast accuracy significantly, they did improve the averages of all metrics. Hence for the sake of making the most of our model components, it is reasonable to consider the auxiliary variables in this estimation as well. The model estimated in this way for the final matchday is presented in Table 4.6.

| Independent Variable | Estimated Coefficient | Std Error | z-value of Wald test | P (>\|z\|) |
|---|---|---|---|---|
| $\gamma^{H*}$ | 3.157 | 0.438 | 7.207 | <0.001 |
| $\gamma^{H}$ | 1.214 | 0.463 | 2.624 | 0.009 |
| $DIST$ | 0.050 | 0.022 | 2.312 | 0.021 |
| $FutureExtMatch_{Euro}$ | 0.172 | 0.084 | 2.056 | 0.039 |
| **Threshold Coefficients:** | | | | |
| $\hat{\kappa}_1$ | 1.4086 | 0.174 | 8.952 | <0.001 |
| $\hat{\kappa}_2$ | 2.674 | 0.16 | 16.672 | <0.001 |

**Table 4.6 – Ordered logit model estimated for forecasting results of the final matchday**

As seen in Table 4.6, both ELO rating based variables are highly significant and due to their large coefficients they affect forecasts more than the other variables in the model. Out of the two variables, $\gamma^{H*}$ seems to be the more relevant to forecasting due to its coefficient being almost three times larger than $\gamma^{H}$'s coefficient. It is also worthwhile to notice that the coefficient of the $FutureExtMatch_{Euro}$ variable has shrunk about 25% in this model when compared with the model presented in Table 4.4. Based on this estimation it seems that this specification could be superior to the other builds we have experimented with.

The forecast accuracy metrics calculated from this comparison are presented in Table 4.7. The model's results are again compared with the ones produced by the univariate $\gamma^{H*}$ model to see whether the improvements are significant when compared with Hvattum & Arntzen's (2010) original specification. The comparison is done with the procedure described already in Section 4.2.2.

| Model | Mean Brier Score ($BS$) | Mean rank probability score ($RPS$) | Mean absolute rank probability score ($RPS_A$) |
|---|---|---|---|
| $\gamma^{H*}$ | 0.1922 | 0.1936 | 0.3898 |
| $\gamma^{H*} + \gamma^{H} + DIST + FutureExtMatch_{Euro}$ | 0.1919 | 0.1933 | 0.3882 |
| P-value of pairwise t-test for means | 0.3686 | 0.3869 | <0.0001 |

**Table 4.7 – Results of forecast accuracy comparison between models "$\gamma^{H*}$" and "$\gamma^{H*} + \gamma^{H} + DIST + FutureExtMatch_{Euro}$"**

As Table 4.7 indicates, the model build with four variables improves the average forecast accuracy across all three metrics when compared with Hvattum & Arntzen's (2010) build. This improvement is, however, statistically significant only in the case of metric $RPS_A$ as shown by the pairwise t-test result reported in the bottom row of the table.

### *4.2.4* Comparison to forecasts implied by average odds

Based on the forecast accuracy results presented above, it seems that only the model "$\gamma^{H*} + \gamma^{H} + DIST + FutureExtMatch_{Euro}$" provides any improvement in forecast accuracy over Hvattum & Arntzen's (2010) original model build. To see how this improvement gauges against the betting market, we compare the results already presented in Table 4.7 with the forecast accuracy implied by the average odds. The results of this comparison are presented in Table 4.8.

| Model | Mean Brier Score ($BS$) | Mean rank probability score ($RPS$) | Mean absolute rank probability score ($RPS_A$) |
|---|---|---|---|
| Average odds | 0.1907 | 0.1916 | 0.3901 |
| $\gamma^{H*} + \gamma^{H} + DIST + FutureExtMatch_{Euro}$ | 0.1919 | 0.1933 | 0.3882 |
| $\gamma^{H*}$ | 0.1922 | 0.1936 | 0.3898 |

**Table 4.8 – Results of forecast accuracy comparison between average odds, $\gamma^{H*} + \gamma^{H} + DIST + FutureExtMatch_{Euro}$ model, and $\gamma^{H*}$ model**

As shown in Table 4.8, based on the forecast accuracy metric means, the forecasts implied by the average odds seem to be superior to both ELO rating based models in terms of $BS$ and $RPS$, but inferior in terms of $RPS_A$. To see behind these means, we present the p-values of the pairwise t-tests, which compare the forecast accuracies of both ELO based models with the forecast accuracy implied by the average odds. The pairwise t-tests are run against the average odds for both models separately and for each forecast accuracy metric. These results are presented in Table 4.9. The upper row shows the p-values of the comparison between the $\gamma^{H*} + \gamma^{H} + DIST + FutureExtMatch_{Euro}$ model and the average odds. The bottom row shows the p-values of the comparison between the $\gamma^{H*}$ model and the average odds. As mentioned, these p-values are based on the samples whose averages are reported in Table 4.8 above.

| Model | P-value of pairwise t-test with $BS$ | P-value of pairwise t-test with $RPS$ | P-value of pairwise t-test with ($RPS_A$) |
|---|---|---|---|
| $\gamma^{H*} + \gamma^{H} + DIST + FutureExtMatch_{Euro}$ | 0.0639 | 0.0708 | 0.0491 |
| $\gamma^{H*}$ | 0.0212 | 0.0248 | 0.7290 |

**Table 4.9– Results of pairwise t-tests between the ELO rating based forecast models and the average odds**

The results presented in Table 4.9 indicate that the accuracy of the 4-variable model is – depending on the forecast accuracy metric – either equally good or better when compared with the average odds. As shown on the top row, no statistically significant difference in the means is observed when the averages of $BS$ or $RPS$ are compared, and in the case of $RPS_A$ a statistically significant difference in favour of the 4-variable model is observed on a 0.05 significance level. As shown on the bottom row, Hvattum & Arntzen's $\gamma^{H*}$ model is inferior to the average odds in terms of $BS$ and $RPS$ on a 0.05 significance level, and equally good in terms of $RPS_A$.

Based on this comparison, we can say that the $\gamma^{H*} + \gamma^{H} + DIST + FutureExtMatch_{Euro}$ model is equally good or superior when compared with the average odds: based on BS and RPS, there is no significant difference and based on $RPS_A$ there is (in favour of our model). We can also say that Hvattum & Arntzen's (2010) model, the $\gamma^{H*}$ model, is inferior or equally good when compared with the average odds: based on the BS and RPS, there is significant difference (in favour of the average odds) and based on $RPS_A$ there is no difference.

## 4.3 Results of betting simulations

As the results of the forecast accuracy measurements done in 4.2 indicate, the model that uses four variables – the $\gamma^{H*} + \gamma^{H} + DIST + FutureExtMatch_{Euro}$ model – provides the best forecast accuracy out of the models we experimented with. Therefore, it is relevant to present the results of the betting simulations only for this model.

All betting simulations are done with the evaluation sample which runs from season 2005/2006 to season 2011/2012 and consists of 2364 matches. For each possible outcome of each match, we have identified the maximum odds out of the odds that were available in our dataset. All betting strategies bet on these maximum odds, and hence they emulate a decision rule where the bettor first assesses the probability of the different outcomes, and then compares the estimated probabilities with the best odds available on the market. Unless otherwise noted, all betting strategies used in the simulations place a bet on the odd if the

expected value of the bet is larger than zero, given the probability assessment provided by the underlying forecast model.[22] It is also worthwhile to note that all betting strategies can find more than one value bet per match. Therefore the number of bets the betting strategies place can exceed the number of matches in the sample.

The results of these simulations are divided into three parts. First, we look at the results on an aggregate level across the evaluation sample. Then, we present the results of a sensitivity analysis where the results are grouped by the ex-ante expected value of the bet (indicated by the forecast model). After this, we present the results of the bias-analysis where we evaluate our model's ability to exploit documented biases in the odds.

### *4.3.1* Aggregate results over the evaluation sample

Table 4.10 presents the average betting returns per match for each betting strategy where the $\gamma^{H*} + \gamma^{H} + DIST + FutureExtMatch_{Euro}$ model is used as the underlying forecast model of the betting decision. With this procedure, a total of 3243 bets were placed on 2292 matches. The mean return of the maximum odds over all the matches (2364) is also presented to compare the performance of the model with the scenario where bets would be placed on maximum odds at random.

[22] For example, if the forecast model predicts that the probability of a draw is 0.25 and if an odd of 5 is offered by a bookmaker for the draw, then the expected value of betting on a draw is greater than 0. Thus, a bet should be placed on a draw in that game ($0.25 * 5 - 0.75 * 1 = 0.5 > 0$).

| | Average return | Standard deviation |
|---|---|---|
| UNIT BET | -0.01 | 2.081 |
| KELLY | 0.001 | 0.092 |
| UNIT WIN | 0.005 | 0.454 |
| Maximum odds | -0.011 | 0.021 |

**Table 4.10 – Average returns and standard deviations of different betting strategies with $\gamma^{H*} + \gamma^{H} + DIST + FutureExtMatch_{Euro}$ as the underlying forecast model**

The results presented in Table 4.10 indicate that our forecast model produces returns circulating around zero regardless of the betting strategy. The results also indicate that the standard deviations of returns differ greatly across all betting strategies. It should also be noted that betting on the maximum odds at random seems to – on average – yield a loss of 1.1% of for each bet.

To see behind these averages and standard deviations, a series of two sample t-tests for means assuming uneven samples with uneven variances is done. The test is done for three pairs of samples: one comparing the returns of UNIT BET and maximum odds, another comparing the returns of KELLY and maximum odds and the third one comparing the returns of UNIT WIN and maximum odds. With these comparisons, we are able to see if any of the betting strategies outmatch betting on maximum odds at random. The results of the tests are presented below in Table 4.11.

| | P-value of t-test for means (comparison to average returns of maximum odds) |
|---|---|
| UNIT BET | 0.988 |
| KELLY | < 0.001 |
| UNIT WIN | 0.103 |

**Table 4.11 – P-values of two-sample t-tests for means: returns of betting strategies compared against the average return of maximum odds**

The results shown in Table 4.11 show that the betting returns deviate significantly from the average return of the maximum odds only in the case of the KELLY strategy. The p-value of $< 0.000$ indicates that the returns produced by KELLY are higher than the returns of the maximum odds by a very high level of certainty. UNIT BET produces returns that are, on average, identical to betting on the maximum odds at random. UNIT WIN produces the highest average returns out of the three strategies, but due to the high standard deviation of the returns of this strategy, this could be due to chance as indicated by the p-value of 0.103. It is worthwhile to note that when the same tests are run with $\gamma^{H*}$ as the underlying forecast model, the conclusions are the same as they are in Table 4.11.

It is also relevant to know if any of the betting strategies produce returns that are significantly different from 0. Table 4.12 represents the results of one-sample t-tests that test for this difference.

| | **P-value of one sample t-test (against 0)** |
|---|---|
| UNIT BET | 0.835 |
| KELLY | 0.636 |
| UNIT WIN | 0.652 |
| Maximum odds | < 0.001 |

**Table 4.12 – P-values of one-sample t-tests for difference from average return of 0**

As the results of Table 4.12 show, none of the betting strategies produces returns that are significantly different from zero, whereas the average return of the maximum odds is significantly lower than 0. On one hand, this indicates that the underlying forecasting model does not seem to be able to produce forecasts which would lead to profitable (or unprofitable) betting in the long run. On the other hand, this also confirms that betting on the maximum

odds at random is unprofitable. As with the previous comparison, when $\gamma^{H*}$ is used as the underlying model, the conclusions of the tests are almost identical.

### *4.3.2* Effect of expected value on betting returns

In order to understand how our model behaves when its probability estimates differ greatly from the market odds, we conduct a sensitivity analysis where the average returns are divided into different brackets based on the ex-ante perceived expected value of a bet. The results of this analysis are presented below in Table 4.13.

| | | Average return | | |
|---|---|---|---|---|
| Required EV | N of Bets | UNIT BET | KELLY | UNIT WIN |
| 1.0 | 3243 | -0.007 | 0.001 | 0.003 |
| 1.1 | 1546 | -0.016 | 0.001 | 0.004 |
| 1.2 | 809 | 0.004 | 0.003 | 0.011 |
| 1.3 | 455 | -0.092 | -0.002 | -0.006 |
| 1.4 | 270 | -0.062 | -0.001 | 0.001 |
| 1.5 | 166 | -0.233 | -0.009 | -0.019 |
| 1.6 | 101 | -0.125 | -0.004 | -0.010 |

**Table 4.13 – Average returns of different betting strategies as a function of required expected value for bets**

Table 4.13 shows the average returns of different betting strategies as a function of the required expected value. In other words, it tells how the returns of the different strategies behave when the deviations between our model's forecast and the forecast implied by the maximum odds grow larger. Based on the averages themselves, it is hard to say anything specific about the relationship. To see if there is a downward or upward trend, we present (t-distribution based) confidence intervals of these returns. These intervals for each strategy are shown in Table 4.14.

| | | 95 % Confidence Intervals on average return (t-test) | | | | | |
|---|---|---|---|---|---|---|---|
| Required EV | N of Bets | UNIT BET | | KELLY | | UNIT WIN | |
| 1.0 | 3243 | -0.073 | 0.059 | -0.002 | 0.003 | -0.114 | 0.018 |
| 1.1 | 1546 | -0.126 | 0.095 | -0.004 | 0.007 | -0.016 | 0.025 |
| 1.2 | 809 | -0.178 | 0.185 | -0.007 | 0.013 | -0.016 | 0.037 |
| 1.3 | 455 | -0.348 | 0.164 | -0.018 | 0.013 | -0.038 | 0.025 |
| 1.4 | 270 | -0.443 | 0.318 | -0.024 | 0.022 | -0.037 | 0.040 |
| 1.5 | 166 | -0.748 | 0.281 | -0.041 | 0.022 | -0.060 | 0.022 |
| 1.6 | 101 | -0.900 | 0.649 | -0.050 | 0.052 | -0.062 | 0.043 |

**Table 4.14 – Confidence intervals of average returns per betting strategy as a function of required expected value**

The intervals presented in Table 4.14 further illustrate what happens with all betting strategies as the required expected value for placing a bet is increased: while the deviation of the returns increases, the returns also gradually shift below zero and thus indicate that betting with our model becomes increasingly unprofitable when large deviations between our forecasts and the forecasts implied by the maximum odds are observed. It is worthwhile to notice that this gradual shift to sub-zero returns is slower with KELLY and UNIT WIN than with UNIT BET: the confidence intervals of the first two strategies start to lean significantly towards negative values only after a relatively high required edge of 1.5, whereas clear shifts are observed in UNIT BET already at the required expected value of 1.3.

### *4.3.3* Biases in odds

As discussed in Section 2.1.3.1 of this thesis, several authors have observed that fixed odds have built-in biases. Further on, some authors have argued that these biases could be exploited in betting.

To see if these claims apply to our forecast model, we present the results of several bias-exploitation tests. First, we test whether our model is able to identify and exploit the favourite-longshot bias discussed by, for example Cain et al. (2000). After this, we test how well our model is able to exploit biases related to home advantage (see part 2.1.3.1 of this thesis).

To test for the favourite-longshot bias, we have broken our betting returns to subsets that differ in respect to how likely our model has identified a home or an away win to be. In other words, we examine what kind of betting returns are observed on average in matches where our model has identified either of the teams to be a strong favourite to win. The average returns per betting strategy yielded by this analysis are presented in Table 4.15.

| | | Average return | | |
|---|---|---|---|---|
| **Probability of home- or away win estimated by our forecast model** | **N of matches** | **UNIT BET** | **KELLY** | **UNIT WIN** |
| P(favourite) > 0 (all matches) | 2292 | -0.010 | 0.001 | 0.005 |
| P(favourite) > 0.5 | 1322 | -0.023 | 0.001 | 0.007 |
| P(favourite) > 0.6 | 712 | -0.026 | 0.003 | 0.017 |
| P(favourite) > 0.7 | 337 | -0.064 | -0.002 | 0.015 |

**Table 4.15 – Average returns of bets placed on matches where a favourite is identified**

As the figures in Table 4.15 illustrate, there is some variation in the average returns when bets are placed on favourites. It is interesting to notice that different betting strategies seem to yield quite different returns: The average returns of KELLY and UNIT WIN are larger in cases of moderate favourites $P(\text{favourite}) > 0.6$, whereas the returns on UNIT BET decrease as a function of the forecasted winning probability of a favourite. It is also worthwhile to note that the average returns of all strategies drop in in the case of $P(\text{favourite}) > 0.7$ – the matches where our model identifies a very strong favourite.

To get a better idea how the average returns develop, we present the estimated confidence intervals of these averages. The intervals are reported in Table 4.16 below.

| | | 95 % confidence interval on average returns (t-test) | | | | | |
|---|---|---|---|---|---|---|---|
| **Probability of home- or away win estimated by our forecast model** | **N of matches** | **UNIT BET** | | **KELLY** | | **UNIT WIN** | |
| P(favourite) > 0 (all matches) | 2292 | -0.095 | 0.075 | -0.003 | 0.005 | -0.014 | 0.023 |
| P(favourite) > 0.5 | 1322 | -0.151 | 0.105 | -0.005 | 0.006 | -0.017 | 0.031 |
| P(favourite) > 0.6 | 712 | -0.233 | 0.181 | -0.005 | 0.011 | -0.013 | 0.046 |
| P(favourite) > 0.7 | 337 | -0.418 | 0.290 | -0.012 | 0.007 | -0.024 | 0.055 |

**Table 4.16 – 95% confidence intervals of average returns of bets placed on matches, where a favourite is identified**

The intervals presented in Table 4.16 further highlight what the averages in Table 4.15 suggested: the average returns of the more intelligent betting strategies (KELLY, UNIT WIN) stay the same or grow in the case of moderate favourites, but decline heavily in the case of very strong favourites. These results suggest that there is some degree of favourite-longshot bias present in the odds. However, as all confidence intervals have their lower bound below zero, the results also indicate that our forecast model is not able to exploit the bias profitably in any of the tested circumstances.

To investigate how home advantage is related to the favourite-longshot bias, we present additional subsets of our simulation results: the bets on home-favourites, home-longshots and away-longshots are examined separately to see, if the observations made by Vlastakis et al. (2009) about home advantage related biases could be exploited with our model. Because our model identified less than a hundred cases of away-favourites with expected value larger than 1, we omitted them from our results as we cannot infer anything reliably from such a small sample. The results of the other three cases are presented below, starting with the results of bets placed on home favourites presented in Table 4.17. As the conclusions of this analysis did not differ significantly between the different betting strategies, the results are presented only for the KELLY strategy.

| Home-favourite | | | | |
|---|---|---|---|---|
| **Probability of home win** | **N of Bets** | **Average return** | **95 % confidence interval on average returns (t-test)** | |
| P(Home) > 0.5 | 667 | 0.003 | -0.004 | 0.010 |
| P(Home) > 0.6 | 333 | 0.005 | -0.005 | 0.016 |
| P(Home) > 0.7 | 147 | -0.002 | -0.010 | 0.006 |

**Table 4.17 – Average returns and confidence intervals of bets placed on home-favourites with the KELLY strategy**

The averages and their confidence intervals presented in Table 4.17 indicate modest increases in returns, with the exception of betting only on very strong home-favourites ($P(\text{Home}) > 0.7$). While this shows that our model is again unable to properly detect strong favourites, the improvement in the more moderate cases ($P(\text{Home}) > 0.5$ and $P(\text{Home}) > 0.6$) could be interpreted as evidence of the home-favourite bias. To further shed light on these results, similar results for home-longshots are presented in Table 4.18.

| Home-longshot | | | | |
|---|---|---|---|---|
| **Probability of home win** | **N of Bets** | **Average return** | **95 % confidence interval on average returns (t-test)** | |
| P(Home) < 0.5 | 723 | 0.005 | -0.003 | 0.012 |
| P(Home) < 0.4 | 445 | 0.005 | -0.006 | 0.017 |
| P(Home) < 0.3 | 308 | 0.008 | -0.006 | 0.022 |

**Table 4.18 – Average returns and confidence intervals of bets placed on home-longshots with the KELLY strategy**

The results in Table 4.18 show a modest increase in returns even in the category of very unlikely home-longshots ($P(\text{Home}) < 0.3$). The confidence intervals also show that the bets were not very far from being profitable in any of the probability categories. These results together with the results on home-favourites (Table 4.16) could be interpreted as evidence of underestimation of home advantage in the odds.

Next, we present the returns of bets placed on away-longshots. The results of this analysis are presented below in Table 4.19.

| Away-longshot | | | | |
|---|---|---|---|---|
| **Probability of away win** | **N of Bets** | **Average return** | **95 % confidence interval on average returns (t-test)** | |
| P(Away) < 0.5 | 861 | -0.005 | -0.010 | 0.000 |
| P(Away) < 0.4 | 768 | -0.004 | -0.010 | 0.001 |
| P(Away) < 0.3 | 606 | -0.007 | -0.013 | -0.001 |

**Table 4.19 – Average returns and confidence intervals of bets placed on away-longshots with the KELLY strategy**

As shown in Table 4.19, the bets placed on away-longshots fare significantly worse than the bets in any other examined category (or our bets in general for that matter). This indicates that our model quite systematically overestimates the winning probabilities of the away longshots. Therefore it is unable to detect any irregularities that might exist within the odds for these matches.

# 5 DISCUSSION AND CONCLUSIONS

This part of the thesis is structured to explicitly answer the research questions presented in the introduction. As a refresher, the three research questions are presented again below:

- Question 1: Is it possible to improve Hvattum & Arntzen's (2010) ELO rating based variable by introducing team-specific home advantage into the variable?
- Question 2: Should other information than the direct league match history be used in forecasting?
- Question 3: Is it possible to build a model that would outperform the fixed odds betting market?

In the following sections, we reflect upon our findings to answer these research questions. After this, we proceed to chapter six where we present several suggestions for further research.

## 5.1 Research question 1

To summarize, our results indicate that introducing team-specific home advantage into the model introduced by Hvattum & Arntzen (2010) does remove the systematic bias persistent in the authors' model. However, as this bias correction is done by tracking the home and away performance separately, the estimated variable neglects the effect past away games have on the future home games, and thus some information about recent performance is lost. Therefore as a result, the bias correction does not translate into remarkably large improvements in the forecast accuracy. Based on further tests, it seems that using the biased and non-biased ELO rating variable together yields results that outperform Hvattum & Arntzen's model by some criteria. Thus, we can say that modest improvements to the original build were accomplished in this thesis. The details of this conclusion are presented in the following paragraphs.

As our observations presented in Section 3.3.1 show, Hvattum & Arntzen's method of deriving ELO rating based variable values produces systematically biased expected scores for the match outcomes. As discussed in the same section, this bias is due to the method's insensitivity to team-specific home advantage. As shown in part 3.3.1 the method that this

thesis introduces for deriving the variable no longer exhibits this bias, and hence based on these observations it would seem that our method would improve the model build. However, as shown in part 4.2.1, this bias correction does not translate into improvements in the average forecast accuracy. On the contrary, our measurements indicate that the forecasts produced by the bias-correcting build are worse on average, although the differences between the averages are not large enough to be statistically significant. This similarity in the observed forecast accuracy is most likely because of the fact that while our build corrects for the bias, it does so at the expense of tracking the impact a team's away performance has on its home performance and vice versa. Hence, the bias-correcting variable does not use all the information about recent performance that is available to it. This negligence results in more inaccurate estimates of the differences in teams' performance levels, and thus some forecast accuracy is lost. Based on our forecast accuracy measurements, the loss of accuracy caused by this negligence is large enough to outdo any accuracy improvements created by the bias correction. Hence, neither of the variables is superior to the other when used as the only independent variable.

When the bias-correcting variable $\gamma^H$ was used jointly with the biased but more reactive variable $\gamma^{H*}$, an improvement over Hvattum & Arntzen's model build (2010) was observed. As shown in the measurements presented in 4.2.3, the joint model outperformed the original build on average in case of all three forecast accuracy metrics, and this improvement was statistically significant in the case of the Absolute Rank Probablity Score metric. These results indicate that accounting for team-specific home advantage in forecasting has some merit. It is, however, debatable how much merit it has, as we observed statistically significant differences between averages with only single forecast accuracy metric. Given this discrepancy between the conclusions the different metrics yield, the question of whether we made improvements or not boils down to the interpretation of the different forecast metrics.

Constantinou & Fenton (2012) argue that the RPS is more suitable for measuring the efficiency of football results forecasts than the Brier score, as football results are ordinal in nature. Given the sound theoretical evidence the authors present in favor of the RPS over the Brier score, it is interesting to notice that the conclusions made based on the RPS and the Brier score do not differ from each other in any of the scenarios we created. While this

observation could be attributed to similarities between the models in case of comparisons between the ELO rating based models, it is interesting to see that both metrics gave similar conclusions also in those comparisons that contrasted ELO rating based models to forecasts implied by the average odds (see 4.2.4). Hence, based on our empirical tests, it seems that the Brier score is not necessarily as inadequate in practice as Constantinou & Fenton (2012) claim it to be. On theoretical grounds, the RPS should be more suitable in football results forecasting than the Brier Score. However, based on our results, it seems that for practical purposes choosing one metric over the other does not result in different conclusions.

As mentioned in the paragraph above, our best model is superior to Hvattum & Arntzen's (2010) model build if the absolute RPS was used as the criterion. As the other two forecast accuracy metrics find no substantial difference between our build and the original, this raises interesting questions about the absolute RPS as a forecast accuracy metric. As discussed in part 3.4.2.1, Constantinou & Fenton (2012) argue that the absolute RPS does not punish forecast models too heavily on occasions, where the model places large probability mass to a win (on either side) and when subsequently a draw occurs. Given these arguments and our results, we can thus conclude that our joint model differs from the original build in the way it identifies and handles favorites: apparently, our model identifies strong favorites more often than the original build does. Moreover, in those cases where a strong favorite draws with a clear underdog, our model is inferior in terms of all other metrics than the absolute RPS. This is the most likely reason for the fact that the absolute RPS gives a different conclusion about the forecast accuracy from that of the other two metrics. Based on literature, it is hard to say which interpretation, the one offered by the RPS and the Brier score or the one offered by the absolute RPS, is more correct. While the arguments Constantinou & Fenton (2012) give in favor of the absolute RPS are solid in terms of common sense, we did not find any formal theoretical justifications for favoring the absolute RPS over the RPS.

In spite of these unresolved questions regarding the forecast accuracy metrics, the results of the comparisons to average odds speak more in favor of our joint model. As the results presented in Tables 4.7 and 4.8 show, our best model is either equally good or better than the forecasts implied by the average odds in the market in terms of forecast accuracy. As also

shown in Tables 4.7 and 4.8, similar comparison between the Hvattum & Arntzen's model (2010) and the forecasts implied by the average odds yielded a conclusion, by which the original build is either inferior or equally good when compared with market forecasts. Thus, it seems that our model should be preferred over the original build introduced by Hvattum & Arntzen (2010) when forecasting football match results.

To conclude, we can say with confidence that modeling team-specific home advantage does improve Hvattum & Arntzen's (2010) ELO rating based build, if the efficiency issues related to cross tracking the home and away performance can be accounted for. The magnitude of this improvement is, however, quite small, as our forecast accuracy metrics could not uniformly agree on whether our model was equally good or superior. Therefore our results are, to an extent, in line with the results Clarke & Norman (1995, 514) observed: there is evidence on the existence of team-specific home advantage, but statistically the evidence is relatively weak. Based on our results, it seems that modelling for team-specific home advantage is important in forecasting, but not nearly as important as modelling for performance across home and away games is in general.

## 5.2 Research question 2

To summarize, our results indicate that using auxiliary information sources such as external cup match histories and pairwise stadium distances do not substantially improve forecast accuracy over the case of using only ELO rating based variables. Our results also indicate that most of the variables derived with the help of the auxiliary information sources only serve as proxies for inadequate use of direct match history data. The details of these conclusions are presented in the paragraphs below.

As model the specifications represented in Tables 4.1 and 4.4 show, the only auxiliary variables that provided substantial improvement to the model fit over our evaluation sample were $DIST$ and $FutureExtMatch_{Euro}$. While both of them were deemed useful for forecasting in terms of AIC, they behave slightly differently and thus reveal interesting

conclusions about the use of the auxiliary variables alongside the historical match results variables.

As shown in Tables 4.1, 4.4 and 4.6, out of the two variables $DIST$ was the only one whose coefficient stayed relatively stable regardless of the other variables present in the model. The sign of the coefficient is also positive in all model estimations done for this thesis. These observations suggest that $DIST$ most likely measures exactly what it is supposed to measure: as hypothesized in Chapters 2 and 3, it works as proxy for pairwise home advantage by raising the probability of home team victory as a function of geographical distance between the home stadiums of two teams. However, in spite of this consistency, the effect $DIST$ has on the forecast accuracy is relatively small. When the model with multiple variables was compared with the model with only a single ELO based variable, no significant improvements were observed in the forecast accuracy. This suggests that even though $DIST$ is a relatively stable variable with firm theoretical backing, its impact on the forecast accuracy is almost negligible. It is possible that a more pronounced improvement in forecast accuracy could be observed when the measurements were done on larger samples than ours. This is actually quite plausible as in our model estimations, the variable's standard error decreased as a function of the sample size (see Tables 4.1, 4.4 and 4.6). Hence our empirical results confirm the postulations Clarke & Norman (1995, 515-516) and Goddard & Asimakopoulos (2004, 56) have made earlier: the geographical distance can be used as proxy for measuring pairwise home advantage. Even though using $DIST$ might not yield large improvements in forecast accuracy even in larger datasets, the robustness of the variable, the firm theoretical backing and the fact that it is easy to derive, justify its use in any models that forecast league football results.

As for $FutureExtMatch_{Euro}$, almost similar conclusions apply. Similarly as with $DIST$, the variable was found to be relevant to forecasting in terms of AIC. However, the contribution of the variable to the forecast accuracy was also shown to be relatively small as we can observe from its estimated coefficients in Tables 4.1 and 4.4 and from the forecast accuracy results presented in Table 4.5. However, the comparison between the models presented in Tables 4.4 and 4.6 reveals interesting facts about the variable itself. When variables $\gamma^{H*}$ and $\gamma^{H}$ are used

together in a model, the estimated coefficient of $FutureExtMatch_{Euro}$ is roughly 24% smaller than it is when only $\gamma^H$ is used. In other words, the contribution of $FutureExtMatch_{Euro}$ to the forecasts decreases significantly when match history is used more efficiently! These results suggest that the variable – at least partly – compensates for the inadequate information about the performance level of a team. If a team has an upcoming match in the Champions League or the Europa League, it is likely that the team's performance is good to begin with. Hence, the upcoming match only serves as an indicator of the participating team's good long-term performance instead of revealing anything about the team's motivation in that particular situation. The fact that similar variables for the FA Cup and the League Cup did not yield any effect speaks in favor of this interpretation: because teams from the lower league levels participate in those cups, it is very likely that a mediocre EPL team can enjoy lasting success in the FA or League Cup without faring too well in the EPL. In addition, as the variables tracking these cups were not significant in any of the model fittings, it is safe to say that $FutureExtMatch_{Euro}$ at least partially models a good performance level instead of changes in team's motivation regarding cup-competitions.

This conclusion has interesting implications for Goddard & Asimakopoulos's (2004) FA Cup variable which is formulated in a similar fashion as our $FutureExtMatch$ variables. As mentioned in part 2.2.2 of this thesis, the authors found that staying in the FA Cup had a favorable effect on the team's performance on the league side. As the authors made this conclusion based on data from the top four levels of English football, they most likely ended up using the FA Cup participation as proxy for team's performance in similar fashion as we ended up using the $FutureExtMatch_{Euro}$ variable. If a team from 2nd, 3rd or 4th division stays in the FA Cup for long, it must be able to beat teams from the higher league levels after the first few qualification rounds. Moreover, if the team is able to do this, it is likely that it performs well in its own league as well. Thus, the variable most likely measures good long term performance instead of measuring how the success in the FA Cup affects the success in league games. In addition, the fact that the FA Cup variables had relatively large coefficients in Goddard & Asimakopoulos's (2004) model thus suggests that the authors' model captures information about recent league performance inefficiently. Hvattum & Arntzen's (2010)

measurements, which compared Goddard & Asimakopoulos's match history variables with ELO rating based variable, also suggest this as their results indicated that the ELO rating based variable was superior in forecasting and thus made better use of match history.

The fact that our remaining auxiliary variables remained insignificant for forecasting in all model fittings can be interpreted in two ways. On one hand, one could argue that our variable formulations were poor and hence unable to measure the effect they were supposed to measure. On the other hand, one could also argue that modeling external effects can – even at its best – generate only a marginal contribution, if the variables that measure long-term league performance are correctly specified. In the light of our results, the latter interpretation seems more plausible. Based on our results, even the statistically significant auxiliary variables either contributed very little or merely compensated for the inadequate use of match history. Hence, it is very likely that our remaining variables could not be improved enough to make them relevant to football results forecasting.

To conclude, we say that using auxiliary information is, in general, less important than the proper use of match history. While our experimentation with the $DIST$ variable showed that – if properly modelled – the external factors can contribute to forecasting, the contribution these variables bring is relatively small when compared with the contribution the teams' match history brings to the forecasts. Thus, our results indicate that the ELO based forecasting methods encode information about the team performance level so well, that the contribution of any external information sources becomes marginal when they are used alongside the ELO based variable(s).

## 5.3 Research question 3

To summarize, our results indicate that while our model produces better returns than placing bets at random, it does not produce returns that are greater than zero when bets are placed on the maximum odds (i.e. the best available odds). The results also indicate that the model is incapable of exploiting systematic biases which – according to other authors – exist in these odds. However, the forecasting accuracy of our model is not significantly different from the

forecast accuracy implied by the average market odds. Together these results suggest that the failure to produce profits is due to two factors. First, the maximum odds do not – on average – deviate enough from the average odds to enable easy profitable betting opportunities. Second, our model is inferior to bookmakers in several clearly defined scenarios. The details of these conclusions are presented in the following paragraphs.

As the results shown in Table 4.10 in Section 4.3.1 indicate, our model's returns are larger than the average return of maximum odds when the Kelly criterion is used in determining the bet size. These results are in line with the forecast accuracy results presented in 4.2.4 which showed that the forecast accuracy of our model is not significantly different from the accuracy of the forecasts implied by the average odds. By Kyuper's (2000) definition of betting market efficiency (see part 2.1.3), these results indicate that our model was able to outperform the market, and thus show that fixed odds for the English Premier League have been inefficient during our evaluation period of seasons 2005/2006 to 2011/2012. However, the results presented in Table 4.12 in Section 4.3.1 show that our model did not produce returns that were significantly different from zero. This result indicates that the aforementioned inefficiency is not large enough to generate profits with our model. This result is consistent with the observation made by Forrest et al. (2005): the fact that bets are offered online for global audiences has forced bookmakers to be more efficient, and thus opportunities for long-term profits in fixed odds betting have started to disappear.

As our sample consisted only of odds for the English Premier League, which is the most followed professional football league in the world, one could argue that the odds on the results of the less popular football leagues could be more inefficient. The results of simulations made by Hvattum & Arntzen (2010), however, argue against this claim. In their work, the authors examined the top four divisions of English football with a model specification which, according to our tests (see part 4.2.3), does not differ too much from our best model in terms of forecast accuracy. Their betting simulations with this model and with identical betting strategies yielded returns that were similar to returns observed with our best model. As the betting results of these two experiments are so similar in spite of the differences in the sample,

it might be that betting on the fixed odds of lower (English) football leagues does not improve the profitability of betting over the case of betting on the most followed league.

As discussed in part 2.1.3.1, the inefficiencies in fixed odds are largely due to the structural biases that are built into the odds by the bookmakers on purpose. To see if our model is useful in exploiting any of the bias types detected by Cain et al. (2000) and Vlastakis et al. (2009), we divided our betting simulation results into subsamples that focus on specific bias types established in the literature. By examining the returns our model produced within these subsamples, we were able to see how acknowledging the biases affects the model's returns. The results of this analysis presented in part 4.3.3 showed that while our model was able to detect some underestimation of home advantage, it was not able to exploit the biases in a manner that would make the betting profitable in any of the settings postulated by Cain et al. (2000) and Vlastakis et al. (2009).

As for our model specification, the bias-analysis also revealed that our model itself is slightly biased. As the results of Table 4.18 show, the returns of the bets placed on away-longshots yielded average returns whose 95% confidence interval's upper limit was either zero or negative. This result has two possible interpretations: if the odds on these results are efficient, then our model produces estimates that are slightly biased upwards when it comes to the winning probabilities of the away-longshots. On the other hand, if there is a favorite-longshot bias as proposed by Cain et al. (2000) and hinted by the results of Hvattum & Arntzen (2010), then our estimates are very heavily biased upwards, as they produce losses even on biased odds! Based on our results, we cannot determine which one of these two interpretations prevails. Neither of the interpretations is, however, too flattering towards our model. Based on our study, it is difficult to say why our model overestimates the winning probability of the away-longshots. However, several suggestions for further research on this topic are presented in the part six of this thesis.

Another concern regarding the estimates of our model is related to the forecasts that differ very much from the forecasts of bookmakers. As shown in Tables 4.13 and 4.14 in part 4.3.2, the returns of our model decrease as a function of the estimated ex-ante expected value of a

bet. Similar observation was made by Hvattum & Arntzen (2010) when they conducted a similar analysis on their ELO rating based model. As observing a large expected value on a bet indicates that the distance between our forecast and the bookmakers' forecast is large, our results mean that in the cases where our estimate deviates substantially from bookmakers' estimate, our estimate is usually wrong. The fact that the bets placed on very strong (home) favorites result in decreased returns (see Tables 4.15 and 4.17) is most likely related to this problem as well, although we cannot verify this based on our results.

The most likely explanation for this phenomenon is the fact that the bookmakers use more information to produce their forecasts than what our model does. First, as mentioned in 2.1.3.2 of literature review, it is highly likely that the bookmakers incorporate player-specific information into their forecasts. In addition, as mentioned in part 3.3.2, we chose to omit player-specific information due to the difficulties of incorporating it into a statistical model. Second, as also mentioned in 2.1.3.2, the odds are often at least partially formed as a composite opinion by the panels of experts. Hence, they can contain subjective views about different fixtures – something that our model is incapable of doing. As the forecast models of the betting agencies are carefully guarded business secrets, we – or any other authors – cannot verify whether these hypotheses are true or not. However, these two factors are widely accepted by numerous authors as the primary sources of bookmaker's competitive edge against the bettors[23] and thus they are the most likely reasons for the observed results.

To conclude the answer to this research question, we say that our model is able to forecast as well as the bookmakers forecast, but this does not translate into profitable betting. The reason for this failure to make profit is twofold. On one hand, the odds are very efficient due to the pressure exerted on the bookmakers by the global availability of odds. On the other hand, our model has several clearly defined problems that hinder its performance.

[23] See Forrest et al. (2005) for more elaborate discussion on this topic.

# 6 SUGGESTIONS FOR FUTURE RESEARCH

Our results open several avenues for new research on the topic of football forecasting. These topics are discussed in this chapter.

As mentioned in part 5.3, our model has difficulties in two areas. First, as seen in the results presented in 4.3.2, a significant portion of our bets have a very high ex-ante expected value. As discussed in 4.3.2 and 5.3, the expected values are high because in many cases the forecasts produced by our model differ significantly from the forecasts produced by the bookmakers. Second, it seems that our model fares poorly when bets are placed on away-longshots. Even though we attributed both of these problems to the inaccuracy of our model, a solution to them might not be related to tweaking the forecast model at all. Instead, more emphasis should perhaps be placed on the bet selection criterion (Milliner et al., 2009). Further on, Milliner et al. (2009, 93) propose that the bet selection criterion should be calibrated based on the historical profits different types of bets yield. Empirical tests on odds showed that the profits are maximized when the bet selection criterion rejects bets on longshots as well as bets for which the probability estimates deviate very much from the market estimates (Milliner et al., 2009, 89). These results are consistent with our observation of the poor performance with away-longshots and high expected values, and thus they suggest that the betting rules used in this research might be a bit too naive. Given this, it would be interesting to see how our results would change if the historical odds were used to calibrate a bet selection criterion dynamically in a similar fashion as our forecast models are calibrated.

While it could be possible to compensate the inaccurate forecasting with more sophisticated bet selection criteria, it would also be worthwhile to examine why our forecast model is inaccurate in the circumstances discussed earlier in this thesis. In addition to the hypotheses of incomplete information and the lack of subjective knowledge discussed in part 5.3, it could also be worthwhile to examine if the inaccuracy could be due to some structural problems of the model. While we found no evidence of breaking the proportional odds assumption (see part 4.1.2), it does not automatically indicate that our approach of using the ordered logit estimation is necessarily the best one. Generally, in sports forecasting literature the idea of

forecasting match results with the ordered logit or probit models has been accepted without much criticism, and a critical examination of the estimation technique itself could yield improvements to forecasting accuracy. One possibility for improvement could be to change the underlying distribution of the model from a symmetric one to an asymmetric one. One example of this could be the Weibull-distribution that is sometimes referred to as the complementary-log-log (Christensen, 2012, 9). As this distribution is skewed, it might be better at accommodating the distribution of football match results since they are heavily skewed towards the home victory. As the estimations done with the Weibull-distribution would most likely place less probability mass on the away wins, this improvement could possibly reduce the amount of bets our model places on the away-longshots and thus alleviate some of the problems we encountered. Another possibility for improvement could be to use an ordered response model where the thresholds between the discrete outcomes can vary as a function of a set of variables. Sanders (2000) proposed a model like this, and it could have applications in football results forecasting. If it turns out that, for example, the ratio of the draw and away win probability varies as a function of the home win probability, it could be justified to allow the thresholds to vary as a function of data-specific variables as well.

Another question that needs to be answered is related to the interpretation of the different forecast accuracy metrics. As already discussed in part 5.1, in the light of our results, it is not completely clear which forecast accuracy metric best distinguishes between the different models. There is robust theoretical backing that explains the difference between the Brier score and the Rank Probability Score (Constantinou & Fenton 2012), but distinguishing between the latter and the Absolute Rank Probability Score is not as clear theoretically, although good arguments in the favor of the latter exist (Constantinou & Fenton 2012). One way to approach this issue would be to develop betting simulations so that they are better at distinguishing between the different models. Our results, as well as the results of previous authors like Hvattum & Arntzen (2010) suffer slightly from the fact that the models that produce different forecast accuracies do not produce significantly different betting returns. Based on the current knowledge of the subject, it is hard to say whether this is due to the model similarity or the crudeness of betting simulations. In this sense, the developments

proposed for the bet selection criteria (see two paragraphs above) could also help to verify empirically which forecast accuracy metrics best reflect the practical usefulness of the model (that is, profitability). Our observations about the returns with the different bet sizing strategies also relate to this problem. As all of our simulation results indicate that crude strategies like UNIT BET produce returns with high standard deviation, they are not necessarily good measures for comparing returns across models. Based on our results, it seems that more intelligent bet sizing strategies like KELLY and UNIT WIN generate more consistent returns, and are thus more capable of distinguishing between the different underlying forecast models. Hence, future research on sports forecasting models should also place more emphasis on the development of the bet sizing rules, as different sizing rules can yield very different interpretations about the practical usefulness of the forecast models.

The final development idea relates to using our forecast model as a basis for predicting in-game events. Pullein (2009) suggests that many in-game variables, such as the goal difference or the number of corners per game, can be forecasted with models that combine estimates about the match outcome probabilities to estimates about each team's propensity to cause the in-game events in question. As we can observe from the results in part 4.2.4, our model fares quite well in the forecast accuracy against the average odds. Hence, our best model could be used as a starting point for a model which would try to predict, for example, the number of corners per game. We found no research about the market efficiency of odds for in-game events, and therefore developing these kinds of models and testing them could reveal interesting insights about the odds markets that have not enjoyed widespread attention in the academic literature.

# REFERENCES


Ali, M. M. 1977, Probability and utility estimates for racetrack bettors, *Journal of Political Economy*, Vol. 85, pp. 803–815.

Akaike, H. 1974, A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, pp. 716–723.

Buraimo, B., Forrest, D., Simmons, R. 2010, The 12th man?: Refereeing bias in English and German football, *Journal of the Royal Statistical Society, Series A: Statistics in Society*, Vol. 173 No. 2, pp. 431-449.

Borooah, V. K. 2002, *Logit and Probit: Ordered and Multinomial Models*, SAGE Publications, Issue 138.

Boyko, R.H., Boyko, A.R., Boyko, M.G 2007, Referee bias contributes to home advantage in English Premiership football, *Journal of Sports Sciences*, Vol. 25 No. 11, pp. 1185-1194.

Brier, G.W 1950, Verification of forecasts expressed in terms of probability, *Monthly Weather Review*, Vol. 78, No. 1, pp. 1-3.

Brant, R. 1990, Assessing proportionality in the proportional odds model for ordinal logistic regression, *Biometrics*, Vol. 46, pp. 1171-1178.

Bwin.party 2011a, Online Sports Betting, *bwinparty.com*, available at: http://www.bwinparty.com/AboutUs/OurMarkets/OnlineSportsBetting.aspx, accessed 05 January 2013.

Bwin.party 2011b, Online Sports Betting, *bwinparty.com*, available at: http://www.bwinparty.com/AboutUs/OurMarkets/The%20online%20gaming%20market.aspx, accessed 05 January 2013.

Cain, M., Law, D., Peel, D. 2000, The favourite-longshot bias and market efficiency in UK football betting, *Scottish Journal of Political Economy*, Vol. 47, No. 1, pp. 25-36.

Cain, M., Law, D. and Peel, D. 2003, The favourite-longshot bias, bookmaker margins and insider trading in a variety of betting markets, *Bulletin of Economic Research*, Vol. 55, pp. 263-273.

Capital One Cup 2013, Fixtures and Results, *Capital One Cup*, available at: http://www.capitalonecup.co.uk/fixtures-results/fixtures-and-results/, accessed 15 September 2012.

Christensen, R. H. B. 2012, Analysis of ordinal data with cumulative link models – estimation with the R-package 'Ordinal', *cran.r-project.org*, available at: http://cran.r-project.org/web/packages/ordinal/vignettes/clm_intro.pdf, accessed 07 April 2013.

Clarke, S. R., & Norman, J. M. 1995, Home ground advantage of individual clubs in English football, *Statistician*, Vol. 44, pp. 509-521.

Constantinou, A. & Fenton N.E 2012, Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models, *Journal of Quantitative Analysis in Sports*, Vol. 8, No. 1.

Constantinou, A. 2012, *Bayesian networks for prediction, risk assessment and decision making in an inefficient association football gambling market*, Risk & Information Management (RIM) Research Group, Queen Mary University of London.

Courneya, K. S., & Carron, A. V. 1992, The home advantage in sport competitions: a literature review, Journal of Sport & Exercise Psychology, Vol. 14, pp. 18-27.

Davidson, R. & MacKinnon J.G. 2004, Econometric Theory and Methods, Oxford University Press, USA.

Deloitte 2012 Football Finance - 2012 Football Money League 2012, Deloitte UK, available at: http://www.deloitte.com, accessed 10 October 2012.

Dixon, M. J. & Coles, S. C. 1997, Modeling association football scores and inefficiencies in the football betting market, Applied Statistics, Vol. 46, pp. 265-280.

Dixon, M. J. & Robinson, M. E. 1998, A birth process model for association football matches, Statistician, Vol 47, pp. 523-538.

Drawer, S. & Fuller, C.W. 2002, An economic framework for assessing the impact of injuries in professional football, Safety Science, Vol. 40, No. 6, pp. 537-556.

Elo, A. E. 1978, The rating of chess players, past and present, New Arco Publishing, New York, USA.

Encyclopaedia Britannica 2012a, Gambling, Encyclopaedia Britannica, available at: http://www.britannica.com/EBchecked/topic/224836/gambling, accessed 16 January 2013.

Encyclopaedia Britannica 2012b, Bookmaking, Encyclopaedia Britannica, available at: http://www.britannica.com/EBchecked/topic/73591/bookmaking, accessed 16 January 2013.

Epstein, E. 1969, A Scoring System for Probability Forecasts of Ranked Categories, Journal of Applied Meteorology, Vol. 8, 985-987.

Eurocupshistory.com 2012, European club competitions, Eurocupshistory.com, available at: http://www.eurocupshistory.com/eurocups, accessed 27 September 2012.

Fama E. F. 1970, Efficient capital markets: a review of theory and empirical work, Journal of Finance, Vol. 33, pp. 383-423.

Football by the Numbers 2011, Home Field Advantage: What You See Depends On Where You Look (And What You Make Of Draws), Football by the Numbers, available at: http://www.footballbythenumbers.com/2011/03/home-field-advantage-what-you-see.html, accessed 23 February 2013.

The FA 2012a, The financial impact of the FA Cup, The FA, available at: http://www.thefa.com/TheFACup/News/2012/Jan/~/media/421387323A984557A3686BA18D383E6D.ashx, accessed 05 October 2012.

The FA 2012b, Past Results, The FA, available at: http://www.thefa.com/TheFACup/More/PastResults, accessed 14 November 2012.

Footballdata.co.uk 2012a, Data files England, Footballdata.co.uk, available at: http://www.football-data.co.uk/englandm.php, accessed 01 August 2012.

Footballdata.co.uk 2012b, Notes for Football Data, Footballdata.co.uk, available at: http://www.football-data.co.uk/notes.txt, accessed 08 May 2013.

Forrest, D., Goddard, J., Simmons R. 2005, Odds-setters as forecasters: The case of English football, International Journal of Forecasting, Vol. 21, No. 3, pp. 551-564.

Google Maps 2013, Google Maps, available at: https://maps.google.fi/, accessed 08 May 2013.

Goddard, J. & Asimakopoulos, I. 2004, Forecasting football results and the efficiency of fixed-odds betting, Journal of Forecasting, Vol. 23, pp. 51–66.

Goddard, J. 2005, Regression models for forecasting goals and match results in association football, International Journal of Forecasting Vol. 21 pp. 331-340.

Greene W. H. 2008, *Econometric analysis, 6th edition*, Prentice Hall.

Greer D.L. 1983, Spectator booing and the home advantage: a study of social influence in the basketball arena, *Social Psychology Quarterly*, Vol. 46 pp. 252-261.

Hvattum, L.M. & Arntzen, H. 2010, Using ELO ratings for match result prediction in association football, *International Journal of Forecasting*, Vol. 26, pp. 460-470.

Jackman, S. 2000, *Models for Ordered Outcomes*, Lecture material from course 'Political Science 200C', Standford University.

Jamieson, J. P. 2010, The home field advantage in athletics: a meta-analysis, *Journal of Applied Social Psychology*, Vol. 40, pp. 1819-1848.

Jennett, N. 1984, Attendances, uncertainty of outcome and policy in Scottish league football, Scottish Journal of Political Economy, Vol. 31, pp. 176-198.

Karlis, D., & Ntzoufras, I. 2003, Analysis of sports data by using bivariate Poisson models, *Statistician*, Vol. 52, pp. 381-393.

Kelly, J.L. JR 1956, A new interpretation of information rate, *The bell system technical journal*, pp. 917-926.

Kyupers, T., 2000, Information and efficiency: an empirical study of a fixed odds betting market, *Applied economics*, Vol. 32, pp. 1353-1363.

Long, J. S. 2001, *Regression Models for Categorical and Limited Dependent Variables Using STATA*, STATA Corporation, Texas.

Long, J. S. and Freese, J. 2004, *Regression Models for Categorical Dependent Variables Using Stata*, *Second Edition*, Stata Press.

Maher, M. J. 1982, Modeling association football scores, Statististica Neerlandica, Vol. 36, 109- 118.

Makropoulou, V. & Markellos, Raphael N. 2011, Optimal Price Setting in Fixed-Odds Betting Markets Under Information Uncertainty, *Scottish Journal of Political Economy*, Vol. 58, No. 4, pp. 519-536.

Milliner, I., White, P. and Webber, D. 2009, A statistical development of fixed odds betting rules in football, *Journal of Gambling, Business and Economics*, Vol. 3 No. 1 pp. 89-99.

Nevill, A. M. & Holder, R.L. 1999, Home Advantage in Sport: an Overview of Studies on the Advantage of Playing at Home, *Sports Medicine*, Vol. 28, pp. 221-236.

Nevill A. M., Balmer, N. & Williams, M. 1999, The influence of Crowd noise and experience upon refereeing decisions in football, *Journal of Personality and Social Psychology*, Vol. 96, pp. 135-154.

OLBG.com 2012, Football Bets, *OLBG.com*, available at: http://www.olbg.com/school/football_bets.php, accessed 06 May 2013.

O'Connor, N. 2012, Betting Market Lessons No 1: Understanding the Overround, *Bettingmarket.com*, available at: http://www.bettingmarket.com/overround.htm, accessed 21 January 2013.

Paton, D., Siegel, D. S. & Vaughan W. L. 2002, A policy response to the e-commerce revolution: the case of betting taxation in the UK, *Economic Journal*, Vol. 112, pp. 296-314.

Preston M.G. & Baratta, P. 1948, An experimental study of the auction-value of an uncertain outcome, *The American Journal of Psychology*, Vol. 61, No.2, pp. 183-193.

Pollard R. 1986, Home advantage in football: a retrospective analysis, Journal of Sport Sciences, Vol. 4, pp. 237-48.

Pullein, K. 2009, *Definitive guide to betting on football*, Raceform/Racing Post, UK.

Schwartz B. & Barsky S.F. 1977, The home advantage, *Social Forces*, Vol. 55, pp. 641-661.

Rue, H. & Salvesen, Ø. 2000, Prediction and retrospective analysis of football matches in a league, *The Statistician*, Vol. 49, pp. 399-418.

Rouhani-Kalleh, O. 2006, *Analysis, Theory and Design of Logistic Regression Classifiers Used For Very Large Scale Data Mining*, University of Illinois.

Sanders, M. S. 2000, Uncertainty and Turnout, *Political Analysis*, Vol. 9 pp. 45–57.

Sharpe, W.F. 1964, Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk, *The Journal of Finance*, Vol. 19, No. 3 pp. 425-442.

Shin, H. S. 1993, Measuring the incidence of insider trading in a market for state-contingent claims, *Economic Journal*, Vol. 103, pp. 1141-1153.

Smith M.A., Paton, D. &Vaughan W. L. 2006, Market efficiency in person-to-person betting, *Economica*, Vol. 73 pp. 673-689.

Vincenty, T. 1975, *Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations*, Directoreate of Overseas Surveys, Ministry of Overseas Development, Surrey.

Vlastakis, N., Dotsis, G. & Markellos R.N. 2009, How Efficient is the European Football Betting Market? Evidence from Arbitrage and Trading Strategies, *Journal of Forecasting*, Vol. 28, pp. 426-444.

Webby R. & O'Connor M. 1996, Judgemental and statistical time series forecasting: A review of the literature, *International Journal of Forecasting*, Vol. 12, pp. 91–118.

Williams R. 2009, Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients Across Groups, *Sociological Methods Research*, Vol. 37, No. 4, pp. 531-559.

# APPENDIX 1 – LIST OF SELECTED MATHEMATICAL EXPRESSIONS

Determining the after-match performance ratings of the home and away teams:

$$\iota_1^H = \iota_0^H + k(\alpha^H - \gamma^H),$$

$$\iota_1^A = \iota_0^A + k(\alpha^A - \gamma^A).$$

Determining the expected scores of a match based on the pre-match ratings and scaling parameters:

$$\gamma^H = \frac{1}{1 + c^{\frac{\iota_0^H - \iota_0^A}{d}}},$$

$$\gamma^A = 1 - \gamma^H = \frac{1}{1 + c^{\frac{\iota_0^A - \iota_0^H}{d}}}.$$

The probability density function, the cumulative density function and the variance of the logistic distribution:

$$\frac{e^{-\frac{x-\mu}{s}}}{s\left(1 + e^{-\frac{x-\mu}{s}}\right)^2},$$

$$\frac{1}{1 + e^{-\frac{x-\mu}{s}}},$$

$$\frac{s^2\pi^2}{3}.$$

Expressions for deriving the probabilities for the forecasted events from a (general) ordered logit model for an event with $J$ ordered, mutually exclusive and collectively exhaustive outcomes:

$$P(y_i = 1|\boldsymbol{X_i}) = \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_1}}$$

$$P(y_i = 2|\boldsymbol{X_i}) = \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_2}} - \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_1}}$$

$$\dots$$

$$P(y_i = j|\boldsymbol{X_i}) = \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_j}} - \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_{j-1}}}$$

$$\dots$$

$$P(y_i = J|\boldsymbol{X_i}) = 1 - \frac{1}{1 + e^{\boldsymbol{\beta^T X_i} - \kappa_{J-1}}}.$$

The likelihood function of the ordered logit model for an event with $J$ ordered, mutually exclusive and collectively exhaustive outcomes:

$$\begin{aligned}
\ln L = & \sum_{i, y_i = 1} \ln\left[\Lambda\left(\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{X_i} - \hat{\kappa}_1\right)\right] \\
& + \sum_{i, y_i = 2} \ln\left[\Lambda\left(\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{X_i} - \hat{\kappa}_2\right) - \Lambda\left(\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{X_i} - \hat{\kappa}_1\right)\right] \\
& \dots \\
& + \sum_{i, y_i = j} \ln\left[\Lambda\left(\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{X_i} - \hat{\kappa}_j\right) - \Lambda\left(\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{X_i} - \hat{\kappa}_{j-1}\right)\right] \\
& \dots \\
& + \sum_{i, y_i = J} \ln\left[1 - \Lambda\left(\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{X_i} - \hat{\kappa}_{J-1}\right)\right].
\end{aligned}$$

The likelihood-ratio statistic for testing the proportional odds assumption of an ordinal regression model:

$$LR = -2(\ln L_0 - \ln L_1),$$

Expressions for calculating the average Brier scores, the Rank Probability Scores and the absolute Rank Probability scores for a sample of forecast-event pairs:

$$BS = \frac{1}{M}\sum_{t=1}^{M}\sum_{i=1}^{R}(f_{ti} - o_{ti})^2\,,$$

$$RPS = \frac{1}{M}\sum_{t=1}^{M}\frac{1}{R-1}\sum_{i=1}^{R-1}(f_{ti} - o_{ti})^2,$$

$$RPS_A = \frac{1}{M}\sum_{t=1}^{M}\frac{1}{R-1}\sum_{i=1}^{R-1}|f_{ti} - o_{ti}\,|.$$

Expressions for calculating the Brier scores, the Rank Probability Scores and the absolute Rank Probability scores for individual forecast-event pairs (for the purposes of conducting t-tests):

$$BS_{indiv} = \sum_{i=1}^{R}(f_{ti} - o_{ti})^2,$$

$$RPS_{indiv} = \frac{1}{R-1}\sum_{i=1}^{R-1}(f_{ti} - o_{ti})^2,$$

$$RPS_{A\ indiv} = \frac{1}{R-1}\sum_{i=1}^{R-1}|f_{ti} - o_{ti}|.$$

The Kelly criterion for determining bet sizes on the basis of the forecast and the size of the odds:

$$f^* = \frac{bp - q}{b} = \frac{p(b+1) - 1}{b}.$$

The t-test statistics for the pairwise test, the two-sample t-test assuming uneven means and uneven variances and the one-sample test:

$$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{M}}},$$

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{m_1} + \frac{s_2^2}{m_2}}},$$

$$t = \frac{\bar{X} - 0}{\frac{s}{\sqrt{m}}}.$$