New data-driven rating systems for association football
Acknowledgements

First and foremost, I sincerely thank my supervisor Marek Gągolewski for encouraging my research, being a tremendous mentor, constantly providing me with new ideas, giving me motivation and always be willing to help.

I greatly thank all my co-authors with whom I have had a pleasure to work: Sandjai Bhulai, Ulf Brefeld, Sebastian Mair and Zoltán Szlávik. I especially thank Zoltán Szlávik for introducing me to the fascinating world of sport analytics several years ago.

I also thank Hendrik Weber and Deutsche Fußball Liga (DFL) and Sportcast GmbH for providing positional data used in the parts of this thesis.

Finally, I acknowledge the support by the European Union from resources of the European Social Fund, Project PO KL Information technologies: Research and their interdisciplinary applications, agreement UDA-POKL.04.01.01-00-051/10-00 via the Interdisciplinary PhD Studies Program. Participating in this program was an excellent opportunity for me to learn and develop my scientific interests.
Abstract

Title: New data-driven rating systems for association football
Author: Jan Lasek, M.Sc.
Supervisor: Marek Gągolewski, Ph.D., D.Sc.
Keywords: association football, league formats, movement models, player ratings, prediction models, rankings, rating systems, team ratings, team strength modelling

In this thesis we discuss, analyse, develop and evaluate the methods for building accurate team ratings in sports with a particular focus on association football. We present several well-founded baseline approaches and how they can be optimised to yield even better results in terms of match outcome prediction accuracy. Further, we also present a bottom-up approach which is based on deriving team ratings via individual player ratings. We demonstrate that this approach constitutes an accurate rating system, provided that the player ratings are of good quality. We also present the theory underlying the prominent Elo model. This serves as an inspiration for developing new, accurate as well as interpretable rating systems. We propose several such schemes in which the ratings are updated after consecutive matches using transparent update rules. Finally, different models are compared on a quantitative basis via extensive simulation experiments.

As a further development of the bottom-up approach toward accurately measuring player skills, we propose a new model for player movements. The model is estimated using positional data that describe exact player positions during a match at a high frequency. It serves as a basis for match situation analysis, for which deriving zones of control is one of the most important applications. We show that the model possesses intuitive properties and evaluate it against standard approaches based on physical models of movement. In turn, it can be used to devise player and, in the next step, team ratings.

As for applications, we discuss how team rating models can be used to evaluate different league formats. This is an important issue in tournament design as domestic league formats vary significantly from country to country and can change from year to year. Using team rating systems as the true measurement of team strengths we compute objective metrics for tournament efficacy based on the agreement between the theoretical ranking and final league table. This study may help decision makers in sports to choose the optimal design that produces the most accurate team rankings.
Streszczenie

**Tytuł:** Nowe systemy ratingowe w piłce nożnej  
**Autor:** mgr Jan Lasek  
**Promotor:** dr hab. inż. Marek Gągolewski, prof. PW  
**Słowa kluczowe:** format ligi, modelowanie predykcyjne, modelowanie ruchu zawodników, modelowanie siły drużyn, piłka nożna, rankingi, ratingi drużyn, ratingi piłkarzy, systemy ratingowe


W ramach dalszego rozwoju podejścia oddolnego dla pomiaru umiejętności graczy proponujemy nowy sposób modelowania ruchów zawodników. Model ten jest oszacowany przy użyciu danych opisujących dokładne pozycje zawodników podczas meczu (o wysokiej częstotliwości próbkowania). Modele takie są podstawą analizy sytuacji meczowych, a w szczególności, jako jedne z ważniejszych zastosowań, przy podziale boiska na obszary kontrolowane przez danego zawodnika oraz drużynę. Prezentujemy intuicyjne własności modelu oraz porównujemy go ze standardowymi podejściami opartymi na fizycznych, uproszczonych modelach ruchu zawodników.

W zakresie zastosowań omawiamy sposób użycia systemów ratingowych dla oceny porównawczej różnych form rozgrywania ligi. Jest to istotne zagadnienie w zakresie planowania turniejów, gdyż w różnych krajach stosowane są zróżnicowane formy roz-
grywek, zmieniające się rokrocznie. Korzystając z systemów ratingowych jako narzędzia do pomiaru prawdziwej siły drużyn, wyznaczamy symulacyjnie obiektywne miary dokładności danego formatu rozgrywek na podstawie analizy zgodności między jego ostatecznym wynikiem a teoretyczną miarą siły drużyn. Badanie to może pomóc w procesie podejmowania decyzji w zakresie wyboru optymalnego systemu rozgrywek pod względem jego dokładności w odtwarzaniu prawdziwego, acz nieobserwowalnego, rankingu drużyn.
Contents

1 Introduction ........................................... 13
   1.1. Motivation and thesis scope ......................... 13
   1.2. Research hypothesis ............................ 15
   1.3. Original results .................................. 16
   1.4. Results already published ......................... 16
   1.5. Datasets used .................................... 17
   1.6. Thesis outline .................................... 18

2 Team rating systems for match outcome prediction ........ 19
   2.1. Preliminaries ...................................... 19
   2.1.1. Association football – background .............. 19
   2.1.2. Data representation ............................. 21
   2.1.3. Evaluating the predictive power of a model .... 21
   2.2. Team rating systems ............................... 26
   2.2.1. Elo rating system ................................ 26
   2.2.2. Ordinal logistic regression model ............... 29
   2.2.3. Basic Poisson regression model ................. 31
   2.2.4. Correlated Poisson regression model ............ 33
   2.2.5. Empirical evaluation of the correlated Poisson model 35
   2.2.6. Single parameter Poisson regression model .... 37
   2.3. Team rating systems built on the player level .... 38
   2.3.1. Related work .................................... 38
   2.3.2. EA Sports FIFA video game player ratings ....... 40
   2.3.3. Prediction models based on player ratings ....... 44
   2.3.4. Ordinal logistic regression model ............... 44
   2.3.5. Poisson regression model ....................... 46
   2.3.6. Team-specific adjustments based on performance 47
4.3.2. Overall analysis ................................................. 114
4.3.3. Influence of particular factors ................................ 116
4.3.4. Influence of the number of matches .......................... 117
4.3.5. Enhancing the 3RR + (1RR/1RR) format ...................... 117
4.4. Discussion .......................................................... 118
4.5. Supplement: Tournament metrics for selected parameter combinations .......................... 120

5 Summary and conclusions ................................. 125
5.1. Verification of the research goals and hypotheses .............. 125
5.2. Future research directions ....................................... 126

A Notation and abbreviations .................................... 129

Bibliography .......................................................... 131
Introduction

1.1 Motivation and thesis scope

This thesis looks at the design, analysis and evaluation of the rating systems for association football. The problem of designing accurate team rating models has long history and until very recently was mainly considered in applied statistics (see, e.g., Cattelan et al. 2013; Crowder et al. 2002; Dixon and Coles 1997; Elo 1978; Goddard 2005; Groll et al. 2015; Koning 2000; Maher 1982). It is worth noting that there are numerous applications of rating models in decision making in sports. These include constructing prediction models, providing team seedings for tournaments and qualifying rounds, scheduling tournaments, creating interesting match-ups, or even granting players work permit at the international level. The recent growth of e-sports has made rating systems even more important. Moreover, such settings are more computationally demanding as typically they involve a large number of contestants (Herbrich et al. 2006; Moulton 2014). The increasing availability of high-quality data makes this problem of interest in data mining and machine learning as well.

One of the most prominent applications of team rating models is their use in match outcome prediction. Many approaches to this problem have been put forward in applied statistics, data mining, and machine learning (see, e.g., Cattelan et al. 2013; Constantinou 2018; Constantinou et al. 2012; Crowder et al. 2002; Dixon and Coles 1997; Goddard 2005; Groll et al. 2015; Hubáček et al. 2018; Joseph et al. 2006; Lasek 2016; Maher 1982; Peeters 2018; Shin and Gaspanyan 2014; Sismanis 2010). Given this, the problem of match outcome prediction became a widely accepted method for evaluating and comparing different team rating models. The models with higher predictive performance are considered better rating systems (Barrow et al. 2013; Lasek et al. 2013; Ley et al. 2019). Moreover, it is often the case that a simple predictive model based on team ratings is competitive in terms of the accuracy of predictions as compared to more involved approaches (Ley et al. 2019). This additionally justifies all the efforts to design accurate rating models.
One of the major limitations of the most popular approaches is that they do not use low-level information on players to improve model performance. In this thesis, we develop models that employ such low-level data. We also propose methods for deriving accurate player ratings. Only some of the approaches for team rating and match prediction try to reduce (perhaps not directly) these issues to the lower-level task of rating individual players (Kharrat, 2016; Peeters, 2018; Shin and Gasparyan, 2014). Given player ratings, higher level team ratings may be obtained by, for example, averaging the individual players’ ratings. In this approach, it is important to design accurate methods for these micro-level ratings. Increasing the availability of different data sources opens up a wide array of possibilities. These include positional data in the form of a complete record of players’ moves during a match (Gudmundsson and Horton, 2017; Link, 2018).

At the very beginning it is important to distinguish the concepts of rating and ranking. A rating is a number used to quantify a given quality, for example, team strength. The ordering of teams based on these ratings provide a ranking. Such ranking systems typically employ some rating method as an intermediate step in providing rankings as a list of ordered objects with respect to their internal rating (although many other approaches also exist, e.g., Burges et al., 2005; Herbrich et al., 2000). In this way, the ranking is just a linear order of the set of teams. The rating can be considered a more general concept as the differences in the quality being evaluated typically matter, as opposed to rankings.

Rating systems are also important because they serve as a basis for answering higher level research questions. For example, their use is prevalent in operational research, particularly in the context of scheduling and tournament design (McGarry and Schutz, 1997; Ryvkin, 2010; Scarf et al., 2009). In fact, a league format can be viewed as a simple ranking method in which three points are awarded for a win, one for a draw and zero for a loss. Such league formats conform to a variety of structures and the number of points awarded for the result is an important parameter itself that needs to be considered. Designing leagues in a fair manner is an interesting topic not only from a theoretical point of view but also a practical one. Since many of the domestic championship leagues in Europe have undergone transformations in recent years, comparing the efficacy of different league formats in ranking teams is a worthwhile research issue (Lasek and Gagolewski, 2015a,b, 2018). Here, efficacy is defined as the ability of a given league format to accurately produce the true ranking of contestants that are based – this should come as no surprise – on team rating models.

From a different point of view, there is the problem of exploiting rating systems in order to optimise a team’s position in a given ranking (Lasek et al., 2016). There is a strong incentive to be ranked high in official rankings – it typically results in seeded positions for tournaments and qualifying rounds. As a result, a high-ranking team has a higher proba-
bility of being paired off against a potentially weaker opponent. It is therefore important to understand the theoretical underpinnings of a given rating system in order to optimise a team’s position. In this way, the team increases its probability of achieving its long-term objective.

Throughout the thesis, we focus on association football (or soccer, in American English), one of the most popular sports in the world. We study player and team ratings in this domain. In particular, the models proposed are discussed and evaluated in this context. We use datasets on football statistics in the experimental parts of the thesis. The considerations apply well to other sports – at least to a certain extent – but football is used as the sport to explain the ideas and provide the background.

Finally, sport analytics is becoming increasingly popular today. The definition of sport analytics is rather vague and it should be considered an umbrella term for sport-related data analysis. The recent growth in the number of dedicated journals (including special issues), workshops, conferences, or blog posts speaks out how much this field has expanded. The research presented in this thesis falls within this trend. Sport analytics is also increasingly available to a broader audience, with many interesting and interdisciplinary popular science books published in recent years (Kuper and Szymanski, 2012; Langville and Meyer, 2012; Sumpter, 2016).

1.2 Research hypothesis

In this dissertation as the main research objective and working hypothesis we propose that data-driven rating systems yield better predictive performance for predicting match outcome. They are more accurate and robust. We propose several changes to the existing methods as well as new approaches for deriving team and player ratings. The latter is particularly enabled by the increasing availability of data. We also study applications of team rating models in tournament design. In particular, we formulate the following auxiliary research hypotheses to be validated:

1) By exploratory data analysis, it is possible to improve parametric models that were proposed in the literature.
2) Accurate team rating systems can be designed on the basis of crowdsourcing data on players and teams rather than solely based on match outcome statistics.
3) It is possible to construct player ratings at a lower level that in turn can be used for team ratings.
4) It is possible to evaluate different league formats with respect to their efficacy in ranking teams. Efficacy is understood as the ability of a ranking system to produce the true (unobserved) ranking of teams.
As stated in the beginning, as a way of validating the hypotheses, we propose that a given model be chosen over others based on its predictive performance (Barrow et al., 2013; Breiman, 2001; Lasek et al., 2013; Ley et al., 2019). Given this, the issue of evaluating rating systems will be thoroughly discussed as a particularly important aspect of designing accurate team rating systems.

1.3 Original results

In the process of designing more accurate rating systems, we will reach a set of major and minor research goals. The results listed below are the main original achievements of this work:

- providing insights into the problem of evaluating experimental results by studying different evaluation functions and the relations between them (Section 2.1.3),
- designing a new model for rating teams based on the correlation structure for attacking and defensive team ratings (Section 2.2.4),
- designing a bottom-up approach for rating teams based on player ratings from video game data (Section 2.3),
- presenting the theoretical underpinnings of the Elo rating system as a special case of the gradient descent algorithm and, based on this, developing several other iterative rating models (Section 2.4),
- extending a method for modelling three-way outcomes by parametrising the method in order to improve its predictive accuracy (Section 2.4.4),
- developing a player movement model using positional data for improving various dimensions of individual player ratings (Section 3.2),
- comparing and discussing improvements in the accuracy of different league formats used for, e.g., top-tier football competitions in Europe (Chapter 4).

1.4 Results already published

Some of the results presented in the thesis have been published as journal articles or included in conference proceedings. Below is a complete list of the work I have been an author or co-author of and which is included in this thesis. However, many of the results have been extended.


Two other articles are related to the topic of the thesis. These contributions are referenced in the text, but are not a part of it:


1.5 Datasets used

In the computational experiments we used various data sources. First, in Chapters 2 and 4 we used data available at http://www.football-data.co.uk for match results and betting odds across a number of leagues. Second, we used two datasets available on Kaggle platform: https://www.kaggle.com/hugomathien/soccer and https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global. The former dataset conveys a rich source of information that includes player attributes from EA Sports FIFA video game and real-world match data for several seasons. It is used as the main dataset in the experiments in Section 2.3. The latter dataset is an auxiliary data source for inspecting player ratings in the same section. In Chapter 3 we used positional data on five matches. These are the exact player positions during a game recorded at the frequency of 25 Hz. In Chapter 4 we used supplementary data on the historical odds for the outright league winner obtained at https://www.sts.pl https://www.efortuna.pl and http://www.oddschecker.com. Finally, match attendance statistics for the Polish league were obtained at http://www.90minut.pl

When a particular dataset is used, we briefly recall its source for completeness.
1.6 Thesis outline

The thesis is structured as follows. In Chapter 2 we discuss various approaches for rating teams, their improvements, and new rating methods. We also discuss how different models are evaluated. In Chapter 3 we propose a spatial data-based movement model and discuss methods for assessing individual player qualities in order to improve overall team rating models. In Chapter 4 we discuss the application of rating systems as team strength models in order to evaluate different league formats according to the accuracy of those formats in ranking teams at the end of the season. Chapter 5 summarises the thesis and highlights some topics for future work in the area of sport rating systems.
Team rating systems for match outcome prediction

In this chapter, we discuss several prominent team rating systems widely used in sports. The classical ones are based on the ordered logit (or probit) and Poisson regression models \cite{AitchisonSilvey1957, BradleyTerry1952, Maher1982, McCullagh1980}. These two base methods have a long legacy in the literature and have inspired a variety of rating models. They are presented in Section 2.2 along with our own extension of the classic Poisson regression-based approach. The model proposed accounts for the correlation between teams' attack and defensive strengths. In Section 2.3, we then introduce team rating systems built upon player ratings from a video game. These approaches employ low-level information on players in order to compile a higher-level team rating rather than estimate it solely from match results. Finally, in Section 2.4 we present how to use a simple optimisation algorithm – gradient descent – as a workhorse for implementing new, accurate and interpretable rating systems that are, like the famous Elo rating system \cite{Elo1961}, updated iteratively after each match.

2.1 Preliminaries

We start with preliminary remarks on association football, the notation for the data used to estimate models and evaluation methods.

2.1.1 Association football – background

The data we are going to work with describe association football (football for short). In football, in a match lasting 90 minutes (plus a half time 15-minute break), two teams consisting of 11 players compete in an attempt to score more goals. Sometimes extra time is needed to determine the winner. Each of the players has a uniquely specialised role (position) on the pitch, for example, a goalkeeper, a midfielder, or a striker. For a more detailed description of football, its rules and history, see, e.g., \cite{Goldblatt2008} and \cite{IFAB2019}.
From a data analysis perspective, the important fact about football is that a match can end in one of three possible outcomes: one of the two teams wins, or they draw. Usually, the number of goals scored is not as important as simply having scored more goals than the opponent by the final whistle. There are some notable exceptions to this rule as the goals scored are often used to resolve ties. That is, when two or more teams are even on points in a tournament. In this regard, it is also interesting to analyse whether using the exact number of goals scored or only the final match result conveys more information for building rating systems or prediction models. As this is one of the basic distinctions between different modelling approaches, we will discuss this issue later in greater detail.

To provide insight into the average number of goals scored in a match and the frequency of particular results, Table 2.1 presents selected descriptive statistics for several European leagues for the 2017/18 season.

Table 2.1. Statistics for different leagues in the 2017/18 season: the number of matches, the average of goals scored by home (HG) and away team (AG), and result frequencies – home team wins, draws and away team wins (H, D and A, respectively).

<table>
<thead>
<tr>
<th>League</th>
<th>Matches</th>
<th>Avg. HG</th>
<th>Avg. AG</th>
<th>H</th>
<th>D</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belgium</td>
<td>240</td>
<td>1.5583</td>
<td>1.3042</td>
<td>0.4250</td>
<td>0.2750</td>
<td>0.3000</td>
</tr>
<tr>
<td>England</td>
<td>380</td>
<td>1.5316</td>
<td>1.1474</td>
<td>0.4553</td>
<td>0.2605</td>
<td>0.2842</td>
</tr>
<tr>
<td>France</td>
<td>380</td>
<td>1.5289</td>
<td>1.1895</td>
<td>0.4553</td>
<td>0.2526</td>
<td>0.2921</td>
</tr>
<tr>
<td>Germany</td>
<td>306</td>
<td>1.6013</td>
<td>1.1928</td>
<td>0.4542</td>
<td>0.2712</td>
<td>0.2745</td>
</tr>
<tr>
<td>Italy</td>
<td>380</td>
<td>1.4553</td>
<td>1.2211</td>
<td>0.4316</td>
<td>0.2184</td>
<td>0.3500</td>
</tr>
<tr>
<td>Netherlands</td>
<td>306</td>
<td>1.7222</td>
<td>1.4150</td>
<td>0.4477</td>
<td>0.2386</td>
<td>0.3137</td>
</tr>
<tr>
<td>Poland</td>
<td>296</td>
<td>1.4797</td>
<td>1.1588</td>
<td>0.4392</td>
<td>0.2838</td>
<td>0.2770</td>
</tr>
<tr>
<td>Portugal</td>
<td>306</td>
<td>1.5686</td>
<td>1.1307</td>
<td>0.5163</td>
<td>0.1993</td>
<td>0.2843</td>
</tr>
<tr>
<td>Scotland</td>
<td>228</td>
<td>1.3991</td>
<td>1.2061</td>
<td>0.4079</td>
<td>0.2456</td>
<td>0.3465</td>
</tr>
<tr>
<td>Spain</td>
<td>380</td>
<td>1.5474</td>
<td>1.1474</td>
<td>0.4711</td>
<td>0.2263</td>
<td>0.3026</td>
</tr>
<tr>
<td>Switzerland</td>
<td>180</td>
<td>1.6000</td>
<td>1.4111</td>
<td>0.4167</td>
<td>0.2167</td>
<td>0.3667</td>
</tr>
<tr>
<td>Overall</td>
<td>3382</td>
<td>1.5435</td>
<td>1.2173</td>
<td>0.4503</td>
<td>0.2442</td>
<td>0.3054</td>
</tr>
</tbody>
</table>

In an ordered pair of teams playing a match, the first team will typically be assumed to play at its home ground (stadium). This is important because playing at home gives an edge over the visiting team (see, e.g., Boyko et al. 2007; Neave and Wolfson 2003; Swartz and Arce 2014 for an analysis of this effect). As Table 2.1 shows, on average home teams tend to score more goals and win more matches than their guests. The data we analyse are taken from club competitions in which the home team advantage usually exists.

Data source: [http://www.football-data.co.uk](http://www.football-data.co.uk), last accessed on 29 January 2019.
On the other hand, in the case of the national teams or international club competitions, the matches are often played at a neutral ground. Table 2.1 also shows that a draw is the least likely match outcome.

Another quantitative aspect of football is that typically three points are awarded for a win, one for a draw and zero for a loss. This applies to domestic league championships and different tournaments involving group stages or their qualifying rounds. Naturally, in a knock-out, which is a tournament format where subsequent winners of previous rounds are paired off, the points are not awarded and it is only the final match outcome what matters. In Chapter 4 we will discuss in greater detail certain issues related to tournament design, including the number of points awarded for a given result.

2.1.2 Data representation

A dataset of match results will be denoted by $\mathcal{M}$. More precisely, it is a set of tuples representing individual matches,

$$\mathcal{M} = \{ (i^{(k)}, j^{(k)}, t^{(k)}, g_i^{(k)}, g_j^{(k)}, o_{ij}^{(k)}) \mid k = 1, 2, \ldots, m \}.$$ 

We shall symbolically write $k \in \mathcal{M}$ to denote match $k$ in the dataset. A match $k$ is represented as an ordered pair of teams $i^{(k)}$ and $j^{(k)}$ (playing at home and away, respectively), timestamp $t^{(k)}$, goals scored by the teams $g_i^{(k)}$, $g_j^{(k)}$ and the full-time match outcome $o_{ij}^{(k)} \in \{1, 2, 3\}$. The convention is that 1 stands for a home team win, 2 – a draw and 3 – an away team win (that is, $o_{ij}^{(k)} = 1 \iff g_i^{(k)} > g_j^{(k)}$, etc.). These events will be also sometimes denoted by the letters $H_{ij}^{(k)} = \{ o_{ij}^{(k)} = 1 \}$, $D_{ij}^{(k)} = \{ o_{ij}^{(k)} = 2 \}$ and $A_{ij}^{(k)} = \{ o_{ij}^{(k)} = 3 \}$.

In this text, where it is clear from the context, we omit subscripts or superscripts denoting a particular match or teams.

In the case of the Elo model discussed in Section 2.2.1 we will slightly abuse the notation and denote match outcome as $o_{ij}^{(k)} \in \{0, 0.5, 1\}$, where 1 stands for a home team win, 0.5 a draw and 0 for an away team win. The change in the convention applied will be pointed out and should be clear from the context.

2.1.3 Evaluating the predictive power of a model

In order to compare between the different models, metrics for evaluating the accuracy of predictions are needed. To introduce such metrics, more notation is required. Let $p = (p_1, p_2, p_3)$ denote three-way match outcome probabilities obtained from a given model with $p_1 + p_2 + p_3 = 1$, $p_i \geq 0$. Further, $q = (q_1, q_2, q_3)$ denotes the vector indicating the true outcome of a match with $q_1 + q_2 + q_3 = 1$, $q_i \in \{0, 1\}$. For example, if the home team wins a match, $q_1 = 1$ and $q_2 = q_3 = 0$. Below, we present definitions of several metrics commonly employed for evaluating football match result predictions (Constantinou and Fenton, 2012; Goddard, 2005; Hvattum and Arntzen, 2010; Pieters et al., 2012).
To simplify the notation, the metrics are described for a single match and a more general case of \( k \) possible outcomes is used for clarity (\( k = 3 \) in our application). Typically these metrics are computed and averaged to obtain aggregate performance over a dataset of matches.

**Accuracy.** This metric is defined as the proportion of correctly predicted results. In the case of a single match it equals

\[
1 \left( \arg \max(p) = \arg \max(q) \right). \tag{2.1}
\]

Here, the \( \arg \max \) function for a vector returns the index with the highest value among all coordinates of the vector.

**Logarithmic loss.** Logarithmic loss, or *logloss* for short, in the case of \( k \) possible outcomes is computed as

\[
- \sum_{i=1}^{k} q_i \cdot \log(p_i). \tag{2.2}
\]

We note that, mathematically, this metric is equivalent to two functions: the likelihood of results observed and information loss (Witten et al., 2011). These functions can be obtained from logloss by applying a strictly monotonic transformation to it. Thus, in the maximum likelihood setting, logloss is the criterion directly optimised by many prediction models, including the logistic regression model or the gradient boosting algorithm.

**Brier score.** Brier score, or *quadratic loss*, is computed by the following formula

\[
\frac{1}{k} \sum_{i=1}^{k} (p_i - q_i)^2. \tag{2.3}
\]

The normalising factor of \( \frac{1}{k} \) accounts for the number of possible outcomes.

**Ranked probability score.** The ranked probability score, or *RPS* for short, is computed as

\[
\frac{1}{k-1} \sum_{i=1}^{k-1} \left( \sum_{j=1}^{i} p_j - \sum_{j=1}^{i} q_j \right)^2. \tag{2.4}
\]

Again, the normalisation accounts for the number of outcomes (minus one). This metric accounts for the ordinal nature of the results. More precisely, according to RPS, an example prediction of \((0.5, 0.2, 0.3)\) is evaluated as a more accurate one than \((0.5, 0.3, 0.2)\) in the case of a home team win. The three other metrics presented, on the other hand, evaluate both predictions as equally accurate as they take into account only the single probability of the actual match result. We also note that RPS can be rewritten as
the quadratic distance between the empirical distribution functions for predictions $p$ and the actual match result observed $q$:

$$\frac{1}{k-1} \sum_{i=1}^{k-1} (F_p(i) - F_q(i))^2,$$

(2.5)

where $F_x(i) = \sum_{j=1}^{i} x_j$ for a vector $x$.

There is a theoretical argument that for all three of the continuous metrics – logloss, quadratic loss and RPS – the optimal prediction is the one that yields the underlying true probabilities. This is a desirable feature of an evaluation method and such metrics are called proper [Gneiting and Raftery, 2007]. This justifies using either of the three metrics for evaluating football match predictions. In [Witten et al., 2011], this fact was demonstrated for logloss and Brier score. We maintain, and present an analogous argument below, that RPS is also a proper metric. Denoting with $p^* = (p^*_1, p^*_2, \ldots, p^*_k)^\top$ the true underlying distribution of the outcomes – that is, $E[q_i] = p^*_i$ – and taking the expected value of RPS (skipping scaling constant), we obtain

$$E\left[\sum_{i=1}^{k-1} (F_p(i) - F_q(i))^2\right] = \sum_{i=1}^{k-1} E[F^2_p(i)] - 2 \cdot E[F_p(i)F_q(i)] + E[F_q(i)] =$$

$$= \sum_{i=1}^{k-1} (F_p(i) - F_{p^*}(i))^2 + F_{p^*}(i) (1 - F_{p^*}(i)).$$

This is because $E[F_q(i)] = F_{p^*}(i)$ and $E^2_q(i) = F_q(i)$ as it equals either zero or one. Moreover, $E[F_p(i)] = F_p(i)$ and $E[F^2_p(i)] = F^2_p(i)$ as $F_p(i)$ and $F^2_p(i)$ are both constant with respect to the true underlying distribution of match outcome. The component $F_{p^*}(i)(1 - F_{p^*}(i))$ is a constant term related to the variance of the distribution $p^*$. This shows that, as required, RPS is minimised by distribution $p = p^*$.

**Metrics correlation.** To analyse the relations between the different metrics, we computed Kendall’s $\tau$ correlation between their values for the example predictions based on bookmaker odds for a sample of 7553 matches used in the experiments later in this Chapter (we detail how to derive the predictions from betting odds by the end of this section). Table 2.2 presents the results.

First, we observe that Brier score and logloss results are highly correlated ($\tau \approx 0.98$). That is, these metrics provide almost the same ordering of examples based on their values. On the other hand, their correlation with RPS is considerably lower, although it is still quite high. This observation leads to the conclusion that RPS indeed evaluates the results from a slightly different perspective than other metrics. Finally, the correlation of accuracy with the three other metrics is negative because, for this metric, a higher score means better results while for the other metrics a lower score is better (as such, they are often referred to as loss functions).
Table 2.2. Kendall \( \tau \) correlation for predictions based on bookmaker odds.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Brier</th>
<th>Logloss</th>
<th>RPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>1.0000</td>
<td>-0.7065</td>
<td>-0.7065</td>
<td>-0.3839</td>
</tr>
<tr>
<td>Brier</td>
<td>-0.7065</td>
<td>1.0000</td>
<td>0.9792</td>
<td>0.6356</td>
</tr>
<tr>
<td>Logloss</td>
<td>-0.7065</td>
<td>0.9792</td>
<td>1.0000</td>
<td>0.6363</td>
</tr>
<tr>
<td>RPS</td>
<td>-0.3839</td>
<td>0.6356</td>
<td>0.6363</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

**Metric choice.** There is some dispute as to which metric is most appropriate for evaluating of match outcome predictions (Constantinou and Fenton, 2012). The observation is that the three-way outcome should be considered on an ordinal scale. Hence a home team win is “closer” to a draw rather than an away team win. In other words, a forecast predicting a draw if the home team wins should be considered more accurate than the one predicting an away team win. Such an ordinal nature of the outcome can be addressed by metrics like the ranked probability score, which have been suggested by Constantinou and Fenton (2012). Another idea could be to employ cost-sensitive classification (Witten et al., 2011). On the other hand, Hubáček et al. (2018) analysed a set of predictions and concluded the probability of a draw is the lowest when two teams of similar strength face off (that is, when either of the teams is about equally likely to win the match). Further, in certain cases, a draw can be considered a loss, especially when a strong team faces a relatively weaker opponent. In this case, the strong team may take more risks to win the game and hence create scoring opportunities for its opponent. This could arguably result in the match ending in either of the teams winning rather than in a draw. Moreover, the draw can be considered the least stable outcome as a single goal changes the result from a tie to a lead by the side that scored the goal. These considerations undermine the assumption that the ordinality of possible match results always holds. Shifting the probability mass allocated for an underdog to a draw may not necessarily yield a more accurate forecast. This is especially visible in the case of high-scoring sports like handball or basketball. In these sports, draws are exceptionally rare, and in fact predicting a long-odds win may be a more accurate forecast than predicting a draw. However, this effect in football is less pronounced due to the far more frequent occurrence of draws.

**Predictions based on betting odds.** Another approach to evaluating forecast accuracy in football is to use a model’s predictions as the basis for constructing investment strategies on the betting market (Boshnakov et al., 2017; Constantinou and Fenton, 2013; Dixon and Coles, 1997; Hvattum and Arntzen, 2010). However, in addition to an accurate prediction model, this approach requires an investment strategy, which is beyond
the scope of this work. Nevertheless, we use probabilities derived from betting odds as benchmark predictions when evaluating different models. These predictions are derived as follows. If \( \mathbf{b} = (b_1, b_2, b_3)^\top \) represents betting odds in the decimal format (Table 2.3 presents an example), then the underlying probabilities equal

\[
\mathbf{p}^{(b)} = (p_1, p_2, p_3)^\top = \left( b_1^{-1}, b_2^{-1}, b_3^{-1} \right)^\top / \left( \sum_{i=1}^{3} b_i^{-1} \right). \tag{2.6}
\]

The normalising factor accounts for the fact that the inverted betting odds \( b_i^{-1} \) contain the bookmaker’s profit margin (typically several percent). For the example match in Table 2.3 we obtain \( 1/2.00 + 1/3.70 + 1/3.65 \approx 1.04 \). This normalisation assumes that the margin is equally distributed among the three possible outcomes. However, it may not necessarily hold true (Cain et al., 2000; Snowberg and Wolfers, 2010). There are methods that have been proposed to account for this (Strumbelj, 2014). For the purposes of our analysis, we consider the normalisation given in Equation (2.6) sufficient.

Table 2.3. Betting odds in decimal format for an example match. Odds of 2.00 for FC Barcelona means that a bet of 1 PLN placed on Barcelona earns 2 PLN if it actually wins the match. This implies the bookmaker probability of about 0.5 – in fact, slightly less due to inbuilt profit margin – for FC Barcelona to beat Real Madrid.

<table>
<thead>
<tr>
<th>Date</th>
<th>Match</th>
<th>Home</th>
<th>Draw</th>
<th>Away</th>
</tr>
</thead>
<tbody>
<tr>
<td>28/10/2018</td>
<td>FC Barcelona - Real Madrid</td>
<td>2.00</td>
<td>3.70</td>
<td>3.65</td>
</tr>
</tbody>
</table>

Sliding window evaluation. We should also discuss how model predictions are generated for a set of matches. This is done in a sliding window manner and proceeds as follows. Let consecutive matches be played in rounds. For domestic championships these may be the weekends when matchdays take place. If we are to generate model predictions for round \( k + 1 \), then we use all the preceding rounds up to round \( k \) as the training set to estimate the model. The procedure is repeated for round \( k + 2 \) with all the previous rounds up to \( k + 1 \) used as the training set. A model’s parameters are optimised by tuning its performance on a validation set consisting of rounds \( k + 1, k + 2, \ldots, k' \). Finally, the model’s performance is measured against a separate test set which includes future rounds \( k' + 1, k' + 2, \ldots, k'' \). Figure 2.1 presents an overview of this validation procedure.

Cross validation (Witten et al., 2011) is another approach to evaluation. However, when making validation splits, it happens that one uses future match outcomes for training to predict (and evaluate) a model with the past matches. Due to the temporal nature of data, we advocate using a sliding window approach for generating predictions and evaluating them so that future match outcomes are never used to predict past ones.
2.2 Team rating systems

In this section we discuss several rating models used in sports. As outlined in the introductory chapter, rating systems serve a variety of purposes in sports analytics. One of the most popular models is the Elo rating system (Elo, 1961, 1978) which is discussed first. Next, two basic team rating models are presented: a game outcome model based on ordinal logistic (logit) regression (McCullagh, 1980) and a model in which a Poisson distribution is assumed for the number of goals scored – the Poisson model for short (Maher, 1982). We also propose our extension of the Poisson model in Section 2.2.4.

2.2.1 Elo rating system

The Elo model was proposed by Arpad Elo for the purpose of rating chess players (Elo, 1961, 1978). Many extensions of this classic model have been proposed (Glickman, 1999; Herbrich et al., 2006) and it has been widely adapted for many sports (Stefani, 2011), among other things. It has been employed to assess areas as diverse as educational systems (see, e.g., Pelánek 2016 and the references therein), vulnerabilities in information security (Pieters et al., 2012) and dominance hierarchies within animal colonies (Pörschmann et al., 2010). There exist a few implementations of this system for football that propose some discipline-specific tweaks that we will discuss in greater details below. We begin, however, by describing the Elo model in the most basic setting.

Let $r^{(k)} = (r_1^{(k)}, r_2^{(k)}, \ldots, r_n^{(k)})^T$ denote the ratings for a set of $n$ teams. The superscript denotes match $k$ (or round $k$) and we assume that these ratings are given after an equal number of $k$ matches (rounds) have been played by each team (in principle they may have played a different number of matches, say $k_i$, but, to simplify notation, an equal number of

Figure 2.1. Splitting data into training, validation and test sets. The predictions for consecutive rounds are generated in a sliding window manner.
them is assumed). The ratings are initialised by some default value, typically \( r_i^{(0)} = 1500 \) for each team \( i \), and iteratively updated following each match.

The Elo rating system provides a simple, heuristic update rule\(^2\). First, the model generates its expectation for the possible outcome of the next match, say \( k \), based on the current team ratings updated after match \( k-1 \). This expectation is formulated in terms of the probability of the first team winning the match and computed using the logistic function

\[
p_{ij}^{(k)} = \frac{1}{1 + \exp\left(-\Delta_{ij}^{(k-1)}\right)},
\]

where \( \Delta_{ij}^{(k-1)} \) is the difference between team ratings before match \( k \)

\[
\Delta_{ij}^{(k-1)} = r_i^{(k-1)} - r_j^{(k-1)}.
\]

After observing the result of the match, the ratings are revised according to the rules

\[
\begin{align*}
r_i^{(k)} &= r_i^{(k-1)} + K \cdot (o_{ij}^{(k)} - p_{ij}^{(k)}), \\
r_j^{(k)} &= r_j^{(k-1)} - K \cdot (o_{ij}^{(k)} - p_{ij}^{(k)}),
\end{align*}
\]

where \( K \) is a scaling constant (referred to as \( K \)-factor) and – slightly abusing the notation here – \( o_{ij}^{(k)} \in \{0, 0.5, 1\} \) is the actual result of a match with the convention that 1 denotes a win for team \( i \), 0.5 – a draw and 0 – a victory for team \( j \). Finally, the magnitude of the updates is the same for both teams in absolute terms.

The Elo system offers an intuitive and plausible interpretation of the update rules described above. Namely, if the model expectation \( p_{ij}^{(k)} \) is lower than the actual result of game \( o_{ij}^{(k)} \) for team \( i \), then the ratings are adjusted upward. The team performed better than expected, so its rating should be increased. Moreover, the higher the discrepancy between the actual result of the game and the expected result, the greater the rating update will be. An analogous effect is observed when team \( i \) falls short of expectations. If the actual result observed is lower than the one predicted, then their rating is decreased accordingly. This works analogously for team \( j \).

**Implementing Elo for football.** There are several implementations of the Elo model for football. To illustrate one, we introduce a model proposed by Hvattum and Arntzen (2010). The update equations remain as above\(^3\), though with some changes to the \( K \)-factor governing the magnitude of updates. Typically, a domain adaptation of the Elo model involves modifying this parameter. There are two versions of the model, which

\(^2\)As we will see in Section 2.4, the update rule can be viewed as a single step in the gradient descent algorithm for minimizing the negative log-likelihood function of results observed.

\(^3\)In fact, Hvattum and Arntzen (2010) used exponentiation with base 10 but it is just a matter of scaling the ratings.
are denoted as the basic Elo\textsubscript{b} and the goals-based Elo\textsubscript{g}. The K-factor for Elo\textsubscript{b} is kept constant and equals, say, \( K_0 \). In the other formulation, the K-factor is additionally driven by the difference in goals scored

\[
K = K_0 \cdot (1 + \delta)^{\lambda_g},
\]

(2.9)

where \( \delta \) is the absolute goal difference in a match \( k \) – that is, \( \delta = |g^{(k)}_i - g^{(k)}_j| \) and \( \lambda_g > 0 \) is an additional parameter.

It is also noteworthy that the Elo rating system does not provide the probability of a draw (we will come back to this issue in Section 2.4.4). It is only the two-way outcome that is modelled by the logistic function in Equation (2.7) while the values of about 0.5 can be interpreted as a likely draw. To account for the three-way outcome in (Hvattum and Arntzen, 2010), the authors proposed to use the differences in ratings as covariates (features) to an ordinal logistic regression model, which is used a second-level model once the ratings differences are computed. The ordinal logistic regression model itself is presented in the context of rating teams in the next section and in a more general setting, making it possible to include extra covariates, in Section 2.3.4.

Other implementations of the Elo model for football include, e.g., EloRatings.net (2015) or both FIFA\textsuperscript{4} Men and Women World Rankings (FIFA 2015, 2019a). In particular, the Elo model was adopted in July 2018 as the official ranking for men’s football.

The previous procedure FIFA used for its Men’s World Rankings came in for criticism both in the media and among researchers that it does not reflect team strength accurately (Lasek et al., 2013, 2016; McHale and Davies, 2007). This may have been one of the reasons FIFA changed its methodology.

Last but not least, some implementations (including EloRatings.net, 2015 or FIFA 2019a) suggest the need to modify the difference in ratings \( \Delta^{(k)}_{ij} \) with a parameter \( h > 0 \) to account for the home advantage:

\[
\Delta^{(k)}_{ij} = h + r^{(k)}_i - r^{(k)}_j.
\]

The exact value of \( h \) depends on the ratings scale. Such enhancements are commonly applied in sports to account for the home team enjoying home-field advantage over the visiting team. There is a natural correspondence with this phenomenon in chess when a player making the first move (the white pieces) has some inherent edge over his or her opponent.

We proceed to describe the second model below.

\footnote{FIFA – Fédération Internationale de Football Association, international governing body of football, futsal and beach soccer.}
2.2.2 Ordinal logistic regression model

Let \( r = (r_1, r_2, \ldots, r_n)^T \) denote the ratings for a set of \( n \) teams. Here the superscript standing for the match index is dropped, unlike in the Elo model. This is to emphasise that while in the Elo model the ratings are updated iteratively after each match, in the ordinal logistic regression model (OLR for short), they are typically estimated using a whole set of games. The derivation procedure is as follows. Again, denote with \( \Delta_{ij} = r_i - r_j + h \) the difference in team ratings for two teams, \( i \) and \( j \), corrected for home-team advantage \( h > 0 \). In a general setting we assume that there is a latent, unobservable variable \( \Delta_{ij}^* = \Delta_{ij} + \epsilon_{ij} \), with \( \epsilon_{ij} \) being a random variable that follows the logistic distribution with mean zero and scale one. The idea behind the model is that we observe whether \( \Delta_{ij}^* \) falls into a specified interval. The possible intervals correspond to match outcomes in the following way

\[
o_{ij} = \begin{cases} 
1 & \text{for } \Delta_{ij}^* > c, \\
2 & \text{for } |\Delta_{ij}^*| \leq c, \\
3 & \text{for } \Delta_{ij}^* < -c,
\end{cases}
\]  

(2.10)

where \( c > 0 \) is an intercept governing the draw margin. Accordingly, the probabilities of the home and away teams winning and drawing are given as

\[
P(H_{ij}) = 1 - F(c - \Delta_{ij}) = 1 - \frac{1}{1 + \exp(-c + \Delta_{ij})};
\]
\[
P(D_{ij}) = F(c - \Delta_{ij}) - F(-c - \Delta_{ij}) = \frac{1}{1 + \exp(-c + \Delta_{ij})} - \frac{1}{1 + \exp(c + \Delta_{ij})};
\]
\[
P(A_{ij}) = F(-c - \Delta_{ij}) = \frac{1}{1 + \exp(c + \Delta_{ij})},
\]

(2.11)

where \( F \) is the cumulative distribution function of the random variable \( \epsilon_{ij} \),

\[
F(x) = \frac{1}{1 + \exp(-x)}.
\]

The parameter estimation procedure is based on the maximum likelihood principle. Given the outcome model, we can construct a loss function \( L(r, h, c|\mathcal{M}) \), defined as the negative penalised log-likelihood for the match results observed (Schauberger et al., 2018; Tutz and Gertheiss, 2016). The penalty introduced is of the form of an \( L_p \)-regularisation on the team rating parameters:

\[
L(r, h, c|\mathcal{M}) = - \sum_{k \in \mathcal{M}} \log P\left( a_{ij}^{(k)}|r, h, c \right) + \frac{\lambda}{p} ||r||_p^p,
\]

(2.12)

where \( P\left( a_{ij}^{(k)} \right) \) denotes the probability attributed by the model to the actual result of a match. Because the uncertainty factor is relatively large, the use of a regularisation term in the domain of match outcome prediction is advised. This usually helps to obtain more
accurate predictions (Groll et al., 2015; Lasek and Gagolewski, 2015b). The parameters of the model are found by minimising the above function or, equivalently maximising the penalised log-likelihood of results with respect to parameters $r$, $h$ and $c$. The choice of the regularisation parameter $\lambda$ is discussed later. In addition to preventing overfitting, regularisation ensures the parameters are identifiable. According to the model formulation given by Equation (2.11), any shift in the rating parameters by a constant yields the same probabilities (as their differences remain equal).

**Relation to other models and their extensions.** How this model is connected to the Elo rating system is worth noting. If $c = 0$ in Equation (2.11), the probability of a draw becomes zero and we arrive at the binary logistic regression model employed in this rating system for computing the expected outcome of a game.

Different specifications can be set on the distribution of the random component $\epsilon_{ij}$ to arrive at a different form of the prediction (or link) function in Equation (2.11). For example, assuming that it follows a normal distribution, another specification can be derived using the cumulative distribution function of the normal distribution. This leads to the ordinal probit regression model. Several works studied the differences between using either of the link functions. In terms of the models’ performance, little or no difference between them was identified (Chan, 2011; Stern, 1992). Chan (2011) found that the normal distribution yields slightly better results, albeit in an experiment with simulated data. That said, as well as due to analytical tractability and that facilitates interpretation, we opt to use the logistic distribution-based version of the model.

In the context of the rating systems, the connection of the two models – Elo and OLR – to the Bradley–Terry model should be recognised (Bradley and Terry, 1952). This model was proposed earlier for analysis of paired comparison data. In the case of the Elo and Bradley–Terry models, an analogous prediction function is used in both approaches. As such, they do not allow explicitly for modelling draws. On the other hand, OLR may be viewed as their extension for multiple-way outcomes. We also note that the Elo rating system is self-contained as it offers a complete and simple update equation. On the other hand, the team ratings derived by the Bradley–Terry or OLR models are typically found by using a numerical optimisation technique for the (penalised) likelihood function. Finally, as stated above, the Elo and OLR models are specified using the logistic distribution function. Using a normal distribution leads to the model proposed by Thurstone (1927). This model is often referred to as Thurstone–Mosteller model thanks to the influential work of the latter author in this area (Mosteller, 1951a,b,c).

Other extensions of the OLR model include, for example, a dynamic rating system based on exponentially weighted moving average processes by Cattelan et al. (2013).
2.2.3 Basic Poisson regression model

The Poisson model is based on the assumption that the number of goals scored by the two teams in a match are random variables that follow a Poisson distribution. Maher (1982) suggests modelling scores by the two teams competing in a match as independent Poisson variables. This was one of the first approaches specifically crafted for modelling association football scores and it serves as a basis for the more involved models discussed below. The following introduces the Maher model in greater detail.

Let $G_i$ and $G_j$ be random variables that express the goals scored in a game by home team $i$ and away team $j$. We assume that these random variables are independent and follow the Poisson distributions with means $\mu_i$ and $\mu_j$, respectively:

$$P(G_i = x, G_j = y | \mu_i, \mu_j) = \frac{\mu_i^x}{x!} \exp(-\mu_i) \cdot \frac{\mu_j^y}{y!} \exp(-\mu_j).$$

When a log-linear model for the goal scoring rates is assumed, $\log(\mu_i) = c + h + a_i - d_j$ and $\log(\mu_j) = c + a_j - d_i$, where $c$ is an intercept, and $a_i$, $a_j$ and $d_i$, $d_j$ stand for attack and defence ratings of teams $i$ and $j$, respectively. As usual, parameter $h$ is introduced to capture home-field advantage.

Again, the model parameters are estimated by the maximum likelihood principle. We also impose the parameter regularisation (Hoerl and Kennard, 1970; Schauberger et al., 2018). Let $r = (a, d) = (a_1, a_2, \ldots, a_n, d_1, d_2, \ldots, d_n)^T$ be team rating parameters. Let us denote by $L(r, h, c | \mathcal{M})$ the loss function as the negative penalised log-likelihood of the results observed in the dataset $\mathcal{M}$:

$$L(r, h, c | \mathcal{M}) = -\sum_{k \in \mathcal{M}} \log P\left(g_i^{(k)} | r, h, c\right) + \log P\left(g_j^{(k)} | r, h, c\right) + \frac{\lambda}{2} ||r||_2^2. \quad (2.13)$$

We note that regularisation also enables identification of the parameters. Finally, since this approach takes into account the exact number of goals scored by the teams rather than only the full-time three-way outcome, it can be expressed by (omitting the subscripts):

$$P(o_{ij} = 1) = P(Z_{ij} > 0) = 1 - F_{Z_{ij}}(0),$$

$$P(o_{ij} = 2) = P(Z_{ij} = 0) = F_{Z_{ij}}(0) - F_{Z_{ij}}(-1),$$

$$P(o_{ij} = 3) = P(Z_{ij} < 0) = F_{Z_{ij}}(-1) \quad (2.14)$$

for the random variable $Z_{ij} = G_i - G_j$ that follows a Skellam distribution with parameters $(\mu_i, \mu_j)$ and $F_{Z_{ij}}$ as its cumulative distribution function. The Skellam distribution is defined as the distribution of the difference of two Poisson variables (Skellam, 1946). While its distribution function does not lead to as simple an analytical form as that of the Elo or the ordinal logistic regression models, it can be easily evaluated numerically using any of the popular scientific computing environments (Jones et al., 2001; Lewis et al., 2016).
Extensions of the basic model. The use of the basic Poisson model and its modifications have been extensive in the sports modelling literature (Boshnakov et al., 2017; Crowder et al., 2002; Dixon and Coles, 1997; Groll and Abedieh, 2013; Karlis and Ntzoufras, 2003; Kharrat, 2016; Koopman and Lit, 2015; Owen, 2011; Rue and Salvesen, 2000; Scarf et al., 2009). To account for the dependence of goals scored between the teams, some bivariate Poisson models (Dixon and Coles, 1997; Karlis and Ntzoufras, 2003) or copula-based (Nelsen, 2006) approaches have been proposed (Kharrat, 2016; McHale and Scarf, 2007, 2011). In order to account for changing team strength, several time-varying parameter models have also been put forward (Crowder et al., 2002; Koopman and Lit, 2015; Owen, 2011; Rue and Salvesen, 2000), or likelihood weighting using time (Dixon and Coles, 1997). Yet another approach was recently proposed in (Kharrat, 2016) and (Boshnakov et al., 2017) as a more general model for the scoring rates. More precisely, an implicit assumption behind modelling goals scored by a team using Poisson distribution is that goal inter-arrival times are exponentially distributed. The authors proposed a Weibull model for these arrival times. The Poisson model is a special case of this approach. Moreover, the authors also addressed the dependence of the scores between the teams using a copula function for modelling the bivariate distribution of goals scored by both teams. These modifications were shown to provide an improved fit to the data. This comes at the cost of losing analytical tractability of the basic version of the model based on Poisson distribution. We henceforth focus on further improvements in the basic version of the model that maintain its interpretability (which in particular will be our goal in Section 2.4).

On the other hand, while much effort has been put into handling the dependence between goals scored by the two teams, relatively little attention has been devoted to analysing a team’s attack and defence ratings themselves and the relationship between them. As an alternative approach, we propose extending the Poisson model with a correlation component for attack and defence ratings. The approach resembles the model proposed by Stenerud (2015) and is based on the observation that a team’s attack and defence strengths are correlated. To the best of our knowledge, the first attempt toward exploiting the correlated structure of team strength parameters in the Poisson regression for modelling sport results. The differences stem from implementation details. Stenerud (2015) sketched the idea in a fully Bayesian approach to estimate parameters. Here, we propose including the correlation structure in the regularisation component for model parameters and employing the maximum likelihood for parameter estimation. Moreover, we use a Bayesian interpretation of the regularisation term to aid in model analysis and we evaluate the predictive power of the approach against the basic version of the Poisson model. Our model is presented in the next section.
2.2.4 Correlated Poisson regression model

We recently proposed in [Lasek and Gagolewski 2018] the following extension to the model given in Equation (2.13). As demonstrated below, the estimated parameter pairs for teams \((a_i, d_i)\) exhibit a positive correlation. We suggest extending the regularisation operator by a correlation component and minimising the following objective function:

\[
L(r, h, c | \mathcal{M}) = -\sum_{k \in \mathcal{M}} \log P(g_i^{(k)} | r, h, c) + \log P(g_j^{(k)} | r, h, c) + \lambda \left( \frac{\|r\|_2^2}{2} - \rho \langle a, d \rangle \right),
\]

(2.15)

where \(\rho \in [-1,1]\) is a correlation parameter. Thus, highly positively correlated attack and defence team parameters reduce the penalty component. Henceforth we refer to this model as the correlated Poisson model, and its counterpart given by Equation (2.13) as the basic Poisson model.

In order to examine this model in greater detail, we investigate whether the optimisation problem given by minimising the negative penalised log-likelihood function in Equation (2.15) is well-defined. Moreover, to aid in interpretation, we discuss the regularisation component in a Bayesian setting. For simplicity, we will focus on the attack and defence ratings \((a, d)\) for a single team (we omit the subscripts in order not to clutter the notation). By exponentiating the penalty term, we obtain

\[
\exp \left( -\lambda \cdot \left( \frac{1}{2}a^2 + \frac{1}{2}d^2 - \rho \cdot \langle a, d \rangle \right) \right) = \exp \left( -\frac{1}{1 - \rho^2} \cdot \frac{a^2 + d^2 - 2\rho ad}{2\sigma^2} \right),
\]

(2.16)

with \(\sigma^2 = (\lambda(1 - \rho^2))^{-1}\). This can be recognised as the (not normalised) bivariate Gaussian density with mean 0, variance \(\sigma^2\) in both dimensions and correlation \(\rho\) between them. In general, for all teams, the regularisation component for vector \(r = (a, d)\) can be viewed as a 2n-dimensional Gaussian distribution with mean 0 and correlation matrix \(\Sigma = [\Sigma_{ij}] \in \mathbb{R}^{2n \times 2n}\), where

\[
\Sigma_{ij} = \begin{cases} 
\sigma^2 & \text{for } i = j, \\
\rho \sigma^2 & \text{for } |i - j| = n, \\
0 & \text{otherwise.}
\end{cases}
\]

(2.17)

Let us now consider the penalty as a function of the model’s parameters. For \(|\rho| \neq 1\) the inverse \(\Sigma^{-1}\) exists and the penalty term can be rewritten as

\[
F_\lambda(a, d) = \lambda \cdot \left( \frac{1}{2}||r||_2^2 - \rho \cdot \langle a, d \rangle \right) = \frac{1}{2} \cdot r \Sigma^{-1} r^\top.
\]

For the optimisation problem given in Equation (2.15) to be well-posed, this function needs to be bounded from below. This means that matrix \(\Sigma\) needs to be positive semidefinite. This is the case if and only if it gives a proper nondegenerate Gaussian distribution.
In the case described here this is satisfied when $\rho^2 < 1$. In the special case $|\rho| = 1$, the optimisation problem is also well-posed. However, in practice $|\rho| \approx 1$ results in numerical stability issues. Moreover, such cases heavily restrict parameter search space as the attack and defence ratings are then strongly correlated.

We note that any form of a positive semi-definite matrix $\Sigma$ could be used here. For example, it can convey cases where the ratings of two different teams are correlated as these teams are competing for relegation or championship and the form of one team may influence the other team. It can also be used when a separate set of the home and away ratings are proposed, as in (Maher 1982), (Constantinou and Fenton 2013) or (Ley et al. 2019). While a team at home may perform differently when it is away, intuition tells us that there is a positive correlation between these performances. This can be accounted for by considering a special correlation structure, as exemplified here.

The model presented here was discussed in the framework of generalised linear models with parameter regularisation (e.g., Hastie et al. 2009). We introduced the objective function based on the likelihood of results observed and the penalty for the parameters. The penalty term was interpreted as a prior distribution for the parameters in the Bayesian setting. Another but equivalent perspective is to look at it as a generalised linear mixed model (Bates and DebRoy 2004; Robinson 1991). In this setting, the intercept and the home team advantage parameters are considered fixed effects and the attack and defence strengths are considered random effects. We also provided a detailed correlation structure for the random effects that depends on two parameters $\sigma$ and $\rho$. These parameters are considered known and will be set by optimising the predictive performance as discussed in the next two sections on model fitting.

The approach presented here employs the empirical observation and intuition that good teams tend to have both strong attack and solid defence (the converse is true for weak teams) and incorporates this in the regularisation term. Along those lines, an interesting model for rating chess players was proposed by Sismanis (2010). The author defined the regularisation component of the model in such a way that player ratings are of similar magnitude to their opponents’ ratings. This stems from the observation that players tend to compete with those of similar strength.

The question remains as to whether the extended version constitutes a better model than the basic one. This will be discussed in the context of a comparative study of the two models’ predictive performance.
2.2.5 Empirical evaluation of the correlated Poisson model

To evaluate the model we use data from 24 seasons (1993/94–2016/17) for five major European leagues – English, French, German, Italian and Spanish\footnote{Data source: http://www.football-data.co.uk}. First, we observe that the parameter pairs \((a_i, d_i)\) exhibit a positive correlation. Figure 2.2 illustrates the estimated pairs of coefficients for each team during each of the seasons (a pair of rating per season per team) for the basic model given in Equation 2.13 with a relatively small penalty parameter \(\lambda = 0.001\) to ensure model identification. The observation that the attack and defence ratings are positively correlated will also be exploited when simulating team ratings throughout the season in a study of league format efficacy in Chapter 4.

![Figure 2.2](image-url)

**Figure 2.2.** Attack and defence ratings for a group of teams with a linear trend line. Linear correlation between the two ratings is ca. 0.467.

In order to verify the usefulness of the model for prediction, we propose the following evaluation procedure. In case of the basic Poisson model, for a given season, we use first 40% of data as the training set, we generate predictions (as detailed in Section 2.1.3) and choose the optimal parameter \(\lambda\) that minimises logloss on a grid of values from 0 to 75 with a step size of 0.5, which appears to be sufficiently small. In the case of the correlated Poisson model, for a given value of parameter \(\rho\) from range \(0.01, 0.05, 0.1, 0.15, \ldots, 0.9, 0.95, 0.99\) we choose optimal \(\lambda\) in the same manner (values 0.01 and 0.99 are examined instead of 0.0 and 1.0, respectively, to avoid numerical stability problems). Next, we compute the fraction of the total number out of 120 trials – for 5 leagues and 24 seasons – in which the correlated model produced better results than its basic version in terms of logloss. This is depicted in Figure 2.3 (as the present season results) for the given range of the correlation values.
The results may be overly optimistic since they are determined using the same sample of data to build and evaluate the model, so we also validate the optimal parameter choice \((\lambda, \rho)\) using next season’s data. That is, for a given parameter pair, we compute the fraction of the total number of 115 trials – for 5 leagues and 23 seasons (23 because for the first season the results are not available) – in which logloss was lower for the correlated model. This procedure assures that the parameters are optimised and validated on an out-of-sample basis. There is an implicit assumption that the optimal value of \(\lambda\) exhibits some persistence as we used the model optimised for a given season and then used the next season to evaluate it. Figure 2.3 presents this fraction as the next season’s results. Additionally, Table 2.4 presents the success fraction for a subset of parameter values.

![Figure 2.3](image)

**Figure 2.3.** The fraction of seasons in which the correlated Poisson model achieves lower error rate than the basic model.

Looking at the present season’s evaluation, the highest success fraction is observed for \(\rho \in [0.01, 0.75]\). On the other hand, results for the next season’s evaluation show that there has been no improvement for \(\rho = 0.01\). In this case, \(\rho \in [0.1, 0.6]\) produces the best results. Based on proportion tests for any \(\rho\) in this range we conclude that the fraction of superior results is significantly different from 0.5 (\(p\)-values < 0.01), which would be expected under no effect (as in the case of \(\rho = 0.01\)). The best results are observed for \(\rho\) equal 0.1 or 0.3 and are equal 0.67. We also observe that increasing parameter \(\rho\) decreases the overall success rate, that is, the fraction of seasons in which the logloss improved. In fact, the correlation values greater than 0.95 produce inferior results to the basic model. Finally, the median and average improvement in logloss over the 115 tests are equal 0.0005 and 0.0003, respectively. This is a modest improvement. However, based on the analysis above we conclude that the correlated Poisson model provides a significant improvement over its basic version for correlation values in the range \(\rho \in [0.1, 0.6]\) and a greater than 0.6 success rate is observed for those values.
Table 2.4. The success fraction for a given correlation parameter in both evaluation settings. Values in bold represent results significantly greater than 0.5 (at the 0.01 level).

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>0.01</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present season</td>
<td><strong>0.65</strong></td>
<td><strong>0.65</strong></td>
<td><strong>0.64</strong></td>
<td><strong>0.63</strong></td>
<td><strong>0.63</strong></td>
<td><strong>0.64</strong></td>
<td><strong>0.62</strong></td>
<td><strong>0.62</strong></td>
<td>0.59</td>
<td>0.55</td>
<td>0.53</td>
</tr>
<tr>
<td>Next season</td>
<td>0.50</td>
<td><strong>0.67</strong></td>
<td><strong>0.65</strong></td>
<td><strong>0.67</strong></td>
<td><strong>0.63</strong></td>
<td><strong>0.63</strong></td>
<td><strong>0.59</strong></td>
<td>0.55</td>
<td>0.53</td>
<td>0.50</td>
<td>0.47</td>
</tr>
</tbody>
</table>

2.2.6 Single parameter Poisson regression model

In this section, we discuss the Poisson model in which attack and defence strengths are reduced to a single parameter. Since the Maher’s (1982) seminal work most Poisson regression-based approaches in football have focused on two parameter models (Crowder et al., 2002; Dixon and Coles, 1997; Groll et al., 2015; Karlis and Ntzoufras, 2003; Koopman and Lit, 2015; Rue and Salvesen, 2000). In fact, in his work Maher considered a general model in which each team is described by two parameters: attack and defence strength, which become four parameters when one accounts for the differences home and away matches introduce. The author considered several simpler models nested within the general model and concluded that the model described in Section 2.2.3 is the most appropriate for modelling purposes. However, he did not explicitly consider a model in which each team is described by a single parameter indicating its overall rating. More recently, such a model was studied, e.g., by Ley et al. (2019).

The parameters of the corresponding Poisson variables in that model are assumed to have the following form

\[
\log(\mu_i) = c + h + r_i - r_j, \\
\log(\mu_j) = c + r_j - r_i, 
\]

(2.18)

Here the attack and defence strengths of a team are equal, \( a_i = d_i \). The objective function is defined analogously as in previous models (Equations 2.13 and 2.15). The prediction function remains the same as described in Equation (2.14).

We note that this model is, in a way, a marginal case of the correlated Poisson regression model presented in Section 2.2.4 for highly correlated attack and defence strengths. Rewriting the penalty term in Equation (2.15) with \( \rho = 1 \) yields

\[
\lambda \left( \frac{||r||^2}{2} - \rho \langle a, d \rangle \right) = \frac{\lambda}{2} \sum_{i=1}^{n} (a_i - d_i)^2.
\]

If the regularisation parameter \( \lambda \) is set to some high enough value, then setting \( a_i = d_i \) cancels the penalty out and the single parameter model is obtained. However, this is rather only a theoretical argument. In practice, setting a large \( \lambda \) and \( \rho = 1 \) results in convergence problems due to numerical stability issues.
2.3 Team rating systems built on the player level

So far, we presented the rating systems that infer team strengths solely based on previous match results. In this section, we discuss the idea of compiling team rating based on its player ratings. To this end, such player ratings are needed. In our application, we will use the EA Sports FIFA video game player ratings. This may be viewed as a bottom-up approach for rating teams. First, however, we discuss several studies for background.

2.3.1 Related work

Several authors have used video games for the analysis of player performance and real-world match results. Thanks to the growing availability of data, such approaches are increasingly popular. For example, Prasetio and Harlili (2016) considered overall team ratings from the EA Sports FIFA video game as a strength indicator. The authors proposed a model for predicting matches using logistic regression based on teams’ attack and defence ratings. There are three shortcomings to their analysis. First, the model was not compared to a benchmark approach. Moreover, the prediction task was reduced to a binary classification by ignoring draws, which most do not do. Finally, the model used only team-level statistics as input, without referring to individual player qualities. In a related study, Shin and Gasparyan (2014) used a data-intensive approach to building a prediction model based on various player attributes from EA Sports FIFA video game. The authors proposed a set of features derived from the game statistics and also built a prediction model for match outcomes. They also considered a simplified version of the prediction task by looking at three different binary prediction problems in a one-vs-all way: predicting home team wins, away team wins, or draws against the two other possible results. In both these studies, the accuracy was used the an evaluation metric. Unfortunately, this metric provides little insight into the quality of probabilities obtained from a model. Moreover, due to the class imbalance in football caused by home advantage and the relatively low frequency of draws (see also Table 2.1), accuracy should not be the evaluation metric of first choice.

Kharrat’s (2016) work is the most relevant to our study. The author proposed a predictive model based on overall player ratings from the EA Sports FIFA video game. The model is based on a bivariate Weibull count model with the attack and defence team strengths obtained from individual player ratings as input. We propose another approach utilising a similar idea but focusing on a model based on a single covariate composed from the average player ratings. This line of research is relatively new in rating systems and match outcome prediction. The results described by these authors are encouraging further experiments in this unexplored area of sport analytics.
As for exploratory analysis of player ratings, Cotta et al. (2016) used an analogous dataset from EA Sports FIFA video game to study the differences between the Brazilian and German national teams, focusing on the 2014 World Cup semi-final between two teams. The study discusses the changing differences between these teams prior to the aforementioned game. The authors also analysed FC Barcelona’s midfield for the period 2008–12. Another example of exploratory analysis of player ratings is a clustering of players based on their attributes (Soto-Valero 2017). This study shows how different attributes discriminate between player roles (e.g., defender or striker) and which attributes are the most important doing so.

Video game data are one example of a detailed player dataset that can be used to conduct interesting analyses and build prediction models. Another example comes via Transfermarkt, a website that keeps player statistics. In particular, the website provides market evaluations that are collected by crowdsourcing among the website’s users. In (Peeters 2018), the author built a model that used the average player valuations to describe a team. The model turned out to be more accurate in forecasting football match results than those based on the Elo model or FIFA rankings. As the player valuations are based on the opinions of the website’s users, the results support the wisdom of crowds principle (Surowiecki 2005). In a different context, Kudenko and Zheng (2010) used positional data on players from another video game – Championship Manager – and used it for automatic commentary generation. Example commentaries are text messages such as a dangerous attack for a given team, the events of passing and receiving the ball, or dribbling. This study is yet another example of how a video game can be used on a large scale to build (train) a system for a specified task.

Besides using player ratings from an auxiliary source of data, previous research has also focused on devising player ratings from different events in a match and its result. For example, McHale et al. (2012) proposed using the Poisson model and different match events (shots on goal, assists, passes, tackles, clean sheets, etc.) as covariates to the model to attribute points obtained by a team to its players. The model was officially employed as a player rating system in the English Premier League. Kharrat et al. (2017) adapted the Plus-Minus methodology for rating players previously proposed for basketball or hockey (see, e.g., Macdonald 2012; Sill 2010). The idea is to infer the ratings using a regression model in which players are encoded by dummy variables and the target variable is the difference in a given quantity between the teams: goals scored, expected goals (Lucey et al. 2014) or expected points. Attempts have also been made to construct player ratings from positional data. We postpone discussing these approaches until the next chapter.

Previous studies have proposed some interesting ideas on using auxiliary data to construct prediction models. Building upon them, we propose to derive team ratings by averaging player ratings. The utility of such an approach will be evaluated for predicting match outcomes. The average player rating will be used as input to the two models presented in the previous section – the ordinal logistic regression and the single parameter Poisson model. First, however, we discuss the data used in the analysis in greater detail.

2.3.2 EA Sports FIFA video game player ratings

In the following experiments we use a dataset available on Kaggle platform. This database constitutes player ratings for a set of various attributes for eleven European leagues spanning eight seasons, from 2008/09 to 2015/16. The result of each match and information on the lineups for both teams are also provided. This makes it possible to bring together player-level ratings and match results in a rating system. The typical top-down approach is to build a model based on match results rather than use player data for this purpose.

In the EA Sports video game there are in total 31 ratings describing various player qualities. A complete list of them is presented in Table 2.5. The ratings are divided into seven different categories. There is also an overall player rating that we will use for modelling later in this section. Each attribute is assigned a value in the interval [0, 100] for evaluating a particular quality. Exactly how the ratings are compiled is not a matter of public information. EA Sports sheds some light on this by stating that the ratings are evaluated by experts. A single expert typically tracks a single or several teams in a given league.

Both the scope of the ratings and the frequency of their updates vary across the dataset. For more recent seasons, the updates are more frequent. Currently, the ratings are updated weekly and can be obtained at SoFIFA website. Let us focus on the overall player ratings – an aggregate of different skills into a single number. We inspected the relation between player attributes and the overall rating with the use of a related dataset, which is also available on Kaggle platform. An application of linear regression for the overall rating as a dependent variable and the player qualities as covariates reveals that for a given position (striker, central defender, etc.) the overall rating is compiled as a convex combination of different attributes (about 15) considered as particularly important for this position (with a nearly perfect fit of $R^2 > 0.99$). For

[https://sofifa.com](https://sofifa.com), last accessed on 29 January 2019.
[https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global](https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global), last accessed on 29 January 2019. In addition to player ratings, this dataset includes player positions (roles) which enables to discover the relation between the overall rating and different player features per position.
Table 2.5. EA Sports FIFA video game player qualities.

<table>
<thead>
<tr>
<th>Overall</th>
<th>Attacking</th>
<th>Skill</th>
<th>Movement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall rating</td>
<td>Crossing</td>
<td>Dribbling</td>
<td>Acceleration</td>
</tr>
<tr>
<td>Potential</td>
<td>Finishing</td>
<td>Curve</td>
<td>Sprint speed</td>
</tr>
<tr>
<td></td>
<td>Heading accuracy</td>
<td>Free kick accuracy</td>
<td>Agility</td>
</tr>
<tr>
<td></td>
<td>Short passing</td>
<td>Long passing</td>
<td>Reactions</td>
</tr>
<tr>
<td></td>
<td>Volleys</td>
<td>Ball control</td>
<td>Balance</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Power</th>
<th>Mentality</th>
<th>Defending</th>
<th>Goalkeeping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shot power</td>
<td>Aggression</td>
<td>Marking</td>
<td>GK diving</td>
</tr>
<tr>
<td>Jumping</td>
<td>Interceptions</td>
<td>Standing tackle</td>
<td>GK handling</td>
</tr>
<tr>
<td>Stamina</td>
<td>Positioning</td>
<td>Sliding tackle</td>
<td>GK kicking</td>
</tr>
<tr>
<td>Strength</td>
<td>Vision</td>
<td></td>
<td>GK positioning</td>
</tr>
<tr>
<td>Long shots</td>
<td>Penalties</td>
<td></td>
<td>GK reflexes</td>
</tr>
</tbody>
</table>

example, in the case of strikers, the top five attributes are: finishing (about 18%), positioning (13%), shot power, ball control and heading accuracy (about 10% each). Most likely, there are also some minor upward adjustments in the ratings of the most recognisable players. Finally, along with the overall rating, the data provide a player’s potential rating. This can be considered an upper bound of a player’s overall rating throughout his career. The overall and potential ratings for a couple of example players and teams are presented below.

Individual player ratings. Figure 2.4 presents the ratings for two example players: Marco Reus and Daley Blind within the seasons available in the dataset. The vertical lines indicate a player’s current club. The dots in the figures represent the time in which a player rating is revised.

We observe that when a player is transferred to a presumably better club, for example when Marco Reus was about to move, at the beginning of the 2009/10 season, from Rot Weiss Ahlen to Borussia Mönchengladbach, his rating increased substantially. This was a transfer from the German second league to a team competing in the top-tier league – the Bundesliga. Another player, Daley Blind, was an even more interesting case. His rating decreased following a transfer from AFC Ajax to FC Groningen. His rating then

---

10See also https://www.kaggle.com/dpwynne/goalkeeper-overall-ratings for a related analysis (last accessed on 29 January 2019).
rose again on the heels of a successful 2014 World Cup performance, which was perhaps decisive in earning him a move to Manchester United. These observations suggest that the ratings do not always reflect a player’s performances or evaluation by scouts. Rather, once the player is transferred to a higher league or a stronger club, both his overall and potential rating are adjusted accordingly. On the other hand, when playing in a stronger club, the player has a greater chance to develop his potential. The ratings are therefore adjusted around the dates when a player is about to change clubs.

**Team ratings.** Analogously, we can look at team ratings compiled as the average ratings of the players in the teams’ starting lineups. Figure 2.5 presents the overall and potential ratings for two teams – Leicester City and Aston Villa – in the 2015/16 English Premiership season. At the end of that season, these teams occupied the extreme ends of the league table – Leicester was crowned the champions and Aston Villa, finishing at
the bottom of the table, was relegated to a lower division. Throughout the season we observed a steady increase and decrease in the teams’ ratings. Presumably, the player’s rating is largely dependent on his current team performance. At the same time, his potential rating is also being adjusted. While this metric is intended as a player’s (and, in turn, his current team’s) potential, it appears to be heavily influenced by the team’s form. During the 2015/16 season, other teams’ ratings remained relatively stable, which may indicate that the teams performed as expected with respect to their players’ evaluations.

![Figure 2.5](image)

**Figure 2.5.** The overall and potential ratings for Leicester City (top) and Aston Villa in consecutive rounds in the 2015/16 season.

While a player’s rating may depend on both his current club and the club’s performance and the ratings stem from a subjective expert opinion, we conjecture that they carry useful information that can be used for modelling. We will also discuss the problem of setting accurate player ratings for selected qualities in Chapter 3.
2.3.3 Prediction models based on player ratings

We now proceed to describe the models used to validate the efficacy of using overall player ratings to construct a team’s rating. First, to build a prediction model, we propose the rating of each team be compiled as the average rating of players

\[ r_i^{(k)} = \frac{1}{L} \sum_{j=1}^{L} r_{i,j}^{(k)}, \tag{2.19} \]

where \( r_{i,j}^{(k)} \) denotes the overall rating of player \( j \) from team \( i \) in match \( k \) and \( L = 11 \) is the lineup size. For a given match, we use player ratings that are closest to the match date. This is necessary as the ratings evolve in time. This rating is the team rating and it will be used as input for a prediction model.

We note that for constructing club ratings, we may use the average of players within the team’s current squad, which is typically fixed and should be reported in advance to the competition’s organiser. For rating national teams, however, we propose to use average player ratings from several of the given team’s most recent matches. This is because national team squads frequently change due to new caps, left out or resigning players. Averaging should perhaps be done using only competitive matches (e.g., major tournament matches or their qualifiers) and avoid including friendlies since often in these matches the lineups are experimental. Hence, the exact construction of a team rating at the international level involves certain issues that need to be considered. Here, we focus on the predictive power of such a rating system for club teams, considering the average player ratings in the starting eleven as the team rating.

We now proceed to describe how the aggregate player ratings are included in the prediction models.

2.3.4 Ordinal logistic regression model

The average player rating will be used as a single feature (covariate) in ordinal logistic regression. We shall now discuss this model in a more general setting than in Section 2.2.2, one that allows us to include features. In particular, it allows to estimate team ratings with an appropriate data encoding in a design matrix. The exact forms of these matrices for the models studied will be presented for an example dataset of five matches, following description of the model below.

In its general form, the ordinal logistic regression model is described with a sequence of \( p \) coefficients, \( \beta = (\beta_1, \beta_2, \ldots, \beta_p)^\top \). For a single match \( k \) (the match index is omitted for brevity), the input data is a sequence of \( p \) numbers, \( x = (x_1, x_2, \ldots, x_p)^\top \), describing the two competing teams. As discussed earlier, there are also two extra parameters \((c, h)\) to be estimated. Parameter \( c > 0 \) is an intercept governing the draw margin and \( h \).
is the parameter introduced to account for home advantage. In the model, for a given match, the probabilities of a home team win, a draw and an away team win are given as

\[
P(o_{ij} = 1) = 1 - \frac{1}{1 + \exp(-c + h + \beta^T x)},
\]

\[
P(o_{ij} = 2) = \frac{1}{1 + \exp(-c + h + \beta^T x)} - \frac{1}{1 + \exp(c + h + \beta^T x)},
\]

\[
P(o_{ij} = 3) = \frac{1}{1 + \exp(c + h + \beta^T x)}.
\] (2.20)

In our setting, the model utilises a single feature, which is the difference in team ratings defined by Equation (2.19); that is, \(x_k = r_i^{(k)} - r_j^{(k)}\) in match \(k\) between teams \(i\) and \(j\). In turn, vector \(\beta\) reduces to a single parameter that needs to be estimated. Additionally, as a preprocessing step, the rating differences were scaled to the interval \([-1, 1]\) by dividing them by 100. The description of the model estimation procedure follows a more general setting with a parameter vector \(\beta\) and a feature vector \(x\) of arbitrary length.

The model is estimated using maximum likelihood with a penalty on the model parameters. The penalty is \(L_p\) regularisation on the model coefficients for \(p = 1\) or \(p = 2\). The couple of constants \((c, h)\) is not subject to regularisation. The exact form of the objective function is

\[
L(\beta, h, c | M) = -\sum_{k \in M} w^{(k)} \cdot \log P \left( o_{ij}^{(k)} | \beta, h, c \right) + \frac{\lambda}{p} \| \beta \|_p^p,
\] (2.21)

where \(P \left( o_{ij}^{(k)} | \beta, h, c \right)\) denotes the probability attributed by the model to the actual result of a match according to Equation (2.20) and \(w^{(k)}\) is a time-dependent weighting of the match. The weighting is introduced to account for how long ago the match was played (see, e.g., Dixon and Coles 1997; Lasek 2016; Ley et al. 2019). More specifically, we employ exponential weighting. For match \(k\) that was played \(t\) days prior to the prediction time, the weight is set to \(w^{(k)} = \exp(-at)\), where \(a > 0\) is a decay parameter to be specified. In principal, the weighting scheme may not be necessary here as the ratings are themselves dynamic. However, it is included for completeness. Henceforth, this model is referred to as OLR\(_1\) model.

**Example.** We demonstrate how the model is estimated in practice using matrix representation with a small illustrative example with \(n = 4\) teams and \(m = 5\) matches given in Table 2.6. In this example, the first match is played between teams 1 and 2.

We use this example to present data encoding, or a design matrix, employed in the computations. The models also involve constants \((c, h)\), which can be included as appropriate vectors in the design matrices. For brevity, as \(c\) is uniform across all matches and \(h\) indicates the home team, the corresponding columns for these parameters are skipped in the presentation.
Table 2.6. An example schedule of five matches played by four teams.

<table>
<thead>
<tr>
<th>No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match</td>
<td>1 – 2</td>
<td>2 – 3</td>
<td>2 – 4</td>
<td>4 – 2</td>
<td>4 – 1</td>
</tr>
</tbody>
</table>

For OLR, the feature vector is given by the differences in team ratings. That is, for match \( k \), \( x_k = (x_k) = (r_i^{(k)} - r_j^{(k)}) \) and the model has a single parameter \( \beta \) to estimate. Hence \( X \in \mathbb{R}^{m \times 1} \). In our small example, the design matrix is of the form

\[
X = \begin{bmatrix}
    r_1^{(1)} - r_2^{(1)} & r_2^{(2)} - r_3^{(2)} & r_3^{(3)} - r_4^{(3)} & r_4^{(4)} - r_2^{(4)} & r_4^{(5)} - r_1^{(5)}
\end{bmatrix}^T
\]

For the sake of completeness we also present an analogous representation for the ordinal logistic regression ratings model (Section 2.2.2). In the case of this model, the feature vector for a single match, \( x_k = (x_{k1}, x_{k2}, \ldots, x_{kn})^T \), with team \( i \) playing against team \( j \), has the following form: \( x_{ki} = 1, x_{kj} = -1 \) and \( x_{kl} = 0 \) for \( l \notin \{i, j\} \). In other words, the features across all matches constitute a sparse encoding of the match schedule. With this representation, the model parameters \( \beta = (\beta_1, \beta_2, \ldots, \beta_n)^T \) represent team ratings. Hence, the design matrix is \( X \in \mathbb{R}^{m \times n} \). In the small example given in the table above, it is of the form

\[
X = \begin{bmatrix}
    1 & -1 & 0 & 0 \\
    0 & 1 & -1 & 0 \\
    0 & 1 & 0 & -1 \\
    0 & -1 & 0 & 1 \\
    -1 & 0 & 0 & 1
\end{bmatrix}
\]

### 2.3.5 Poisson regression model

Analogously to the OLR model above, in the single parameter Poisson model described in Section 2.2.6, the goal scoring rates may depend on some features \( x \). More precisely, given the same notation and data representation as those used in OLR, the rates become

\[
\log(\mu_i) = c + h + \beta^T x, \\
\log(\mu_j) = c + \beta^T x.
\]  

(2.22)

Accordingly, the variables for scores modelling are modified and the objective function becomes

\[
L(r, h, c|\mathcal{M}) = - \sum_{k \in \mathcal{M}} w^{(k)} \cdot \left( \log \mathbb{P} \left( g_i^{(k)} | \beta, h, c \right) + \log \mathbb{P} \left( g_j^{(k)} | \beta, h, c \right) \right) + \frac{\lambda}{p} \| \beta \|_p^p. \quad (2.23)
\]

This model will be referred to as PR.
A more refined approach would be to extend the Poisson model (described in Section 2.2.3) based on the attack and defence player ratings. This approach was pursued by Kharrat (2016), who included player ratings grouped by player role (i.e., defending or attacking player) as covariates in a regression model for match results. The model used was the Weibull count model which extends the Poisson regression. The author also reported very accurate results for predicting matches. They were so accurate, in fact, that they were even shown to outperform the predictions based on bookmaker odds. However, here we focus on deriving a simple rating procedure using the player ratings and validate its usefulness rather than pushing the prediction accuracy of the models to the limits.

In the next section, we discuss an extension of the two basic approaches presented so that they include both player ratings and match results to estimate ratings.

2.3.6 Team-specific adjustments based on performance

The extension of the models presented above is intended to allow for modelling team-specific performance effects. This may be viewed as a method to aggregate player-level and result-based modelling. Intuitively, the extension makes it possible to adjust the base models presented above according to a team’s performance.

Let $\beta_*$ be a global parameter to account for the difference in team ratings on results and $\beta'_i$ be a team-specific adjustment. Hence, there are in total $n + 1$ parameters to be estimated. Given an appropriate representation of the parameter and feature vectors, $\beta$ and $x$, respectively, which is detailed in the example below, computing the probabilities for a given match is analogous to the models previously discussed. However, the product of model parameters and team ratings produces

$$\beta^T x = (\beta_* + \beta'_i)r_i - (\beta_* + \beta'_j)r_j.$$ 

For $\beta'_i > 0$, this can be viewed as an extra contribution to the component attributed to team $i$, which equals $(\beta'_i + \beta_*)r_i$, which can be viewed as a synergy effect between its players. That is, there is an extra contribution to the overall coefficient $\beta_*$. On the other hand, $\beta'_i < 0$ can be interpreted as a team performance inferior to the implied rating. Such a model can potentially help with biased estimates of team strength by the average player ratings $r^{(k)}_i$. This rating is verified against team results and the parameters $\beta'_i$ serve as correcting factors for the teams that perform significantly better (or worse).

In the case of ordinal logistic regression, the estimation procedure based on the negative penalised log-likelihood function, is intended to minimise the following expression

$$L(\beta, h, c | M) = - \sum_{k \in M} w^{(k)} \cdot \log P \left( o^{(k)}_{ij} | \beta, h, c \right) + \frac{\lambda_1}{p} |\beta_*|^p + \frac{\lambda_2}{p} \sum_{i=1}^n |\beta'_i|^p. \quad (2.24)$$
This model will be referred to as OLR timezone. For the Poisson regression, the objective becomes

\[ L(r, h, c | M) = - \sum_{k \in M} w^{(k)} \cdot \left( \log P \left( g_i^{(k)} | \beta, h, c \right) + \log P \left( g_j^{(k)} | \beta, h, c \right) \right) + \frac{\lambda_1}{p} |\beta_s|^p + \frac{\lambda_2}{p} \sum_{i=1}^n |\beta'_i|^p \]  

and the model will be referred to as PR timezone. In this way, we allow different regularisation for the global parameter and the team specific adjustments.

Example. Both models discussed here conform to the same data representation. For match \( k \), the feature vector equals \( x_k = (x_{k1}, x_{k2}, \ldots, x_{k(n+1)})^\top \) with \( x_1 = r_i^{(k)} - r_j^{(k)} \), \( x_{k(i+1)} = r_i^{(k)} \), \( x_{k(j+1)} = -r_j^{(k)} \) and \( x_{kl} = 0 \) otherwise. The parameter vector becomes \( \beta = (\beta_s, \beta'_1, \beta'_2, \ldots, \beta'_n)^\top \). In turn, \( X \in \mathbb{R}^{m \times (n+1)} \). In our small example from Table 2.6, we obtain

\[
X = \begin{bmatrix}
    r_1^{(1)} - r_2^{(1)} & r_1^{(1)} & -r_2^{(1)} & 0 & 0 \\
    r_2^{(2)} - r_3^{(2)} & 0 & r_2^{(2)} & -r_3^{(2)} & 0 \\
    r_2^{(3)} - r_4^{(3)} & 0 & r_2^{(3)} & 0 & -r_4^{(3)} \\
    r_4^{(4)} - r_2^{(4)} & 0 & -r_4^{(4)} & 0 & r_4^{(4)} \\
    r_4^{(5)} - r_1^{(5)} & -r_4^{(5)} & 0 & 0 & r_4^{(5)}
\end{bmatrix}
\]

The models with team-specific performance adjustments can be viewed as a combination of the basic model utilising differences in team ratings as features and the rating systems presented in Section 2.2.2 and Section 2.2.6 for the OLR and basic Poisson model, respectively. The matrix representation above provides some intuition that the model can be viewed as a mixture of the two approaches. That is, the core part of the model represented by parameter \( \beta_s \) is the same as in the model specified previously. The team-specific parameters \( \beta'_i > 0 \) are driven by team results as in the basic ordinal logistic regression ratings. In a special case, by setting all \( \beta'_i = 0 \), this model reduces to the OLR timezone model presented in the previous section. This can also be achieved by setting the team-specific regularisation parameter \( \lambda_2 \) to a high value, which forces all \( \beta'_i \) to tend to zero.

As in the case of the correlated Poisson model presented in Section 2.2.4, the model presented here is referred to in one of three ways in the context of the statistical modelling literature (Pinheiro and Bates, 2000): as a generalised linear model with random effects, a multilevel model or a hierarchical model. In machine learning, our model can also be formulated in a regularised multi-task learning framework (Evgeniou and Pontil, 2004). We assume that there is an overall parameter governing the influence of the rating difference on the results, \( \beta_s \), which is considered a fixed effect, and that \( \beta' \) are random effects that describe the individual differences (random deviations) for the teams from
the global fixed effect $\beta_*$. The assumption is that the individual random effects are normally distributed around $\beta_*$. That is, $\beta_i^r \sim \mathcal{N}(\beta_*, \sigma^2)$. Analogously, the model can also be formulated as a hierarchical Bayesian model assuming this distribution as a prior for parameters $\beta_i^r$. Finally, in a typical random effects framework we assume that the variances and the correlation structure are unknown. Here, we consider their equivalents – regularisation strengths $\lambda_1$ and $\lambda_2$ – as extra parameters and choose them by optimising the predictive performance of the model on a validation set. This will be discussed later as a part of the experiment setup. First, however, we will introduce baseline models in the next section.

2.3.7 Baseline approaches

We focus here on the models that use a single parameter per team as its rating. The proposed models are compared with the following baseline approaches. The baselines are the two versions of the Elo rating system described in Section 2.2.1 – abbreviated as Elo_b and Elo_g for the basic and goal-based version of the model, respectively – and ordinal logistic regression and Poisson regression-based ratings presented in Section 2.2.2 and Section 2.2.6 (referred to as OLR_0 and PR_0, respectively). These two models are also extended by including temporal weighting of matches as outlined in Equations (2.21) and (2.23). As for the Elo model versions, we set their parameters to the values obtained in the original formulation (Hvattum and Arntzen, 2010).

Finally, the results for bookmaker-based forecasts (see Section 2.1.3) are provided as well. As a bottom-line approach, probabilities obtained from the frequency of particular results in the previous seasons are used. These baselines are often considered upper and lower bounds for prediction accuracy.

2.3.8 Experiment setup

Below we describe the steps for running the computational experiment to compare different methods. In particular, we discuss how the parameters are optimised, the predictions are generated and how to compare the accuracy of predictions between different models.

**Setting parameters.** As for the parameter choice, the two regression-based models involve the regularisation operator ($L_1$ or $L_2$), regularisation parameters ($\lambda_1, \lambda_2$) and time decay parameter $a > 0$. The regularisation parameter was chosen on an approximately geometric scale from values $0, 10^k, 2.5 \cdot 10^k, 5 \cdot 10^k$ for $k = -2, -1, 0, 1, 2$. The time decay values range was set from 0.0 to 0.01 with a step of 0.001. Including regularisation operator choice, these parameters span a large space of possible values, especially for OLR_2.
and PR₂. To mitigate the computational burden, we employ random search (Bergstra and Bengio [2012]) by sampling without replacement from all the possible combinations of 250 parameter settings to compute the predictions for the validation set (discussed below) and choose the optimal setup. We conclude that it is a sufficient number of trials by inspecting the results. Finally, the models’ objective functions are optimised using the BFGS algorithm (Nocedal and Wright [2006]).

**Generating predictions.** The procedure for generating predictions for consecutive rounds follows the description presented in Section 2.1.3. The models were trained using data from 2008/09 to 2013/14, where the 2008/09–2010/11 seasons were used only for training and the 2011/12 and 2012/13 seasons together as a validation set for measuring model performance for different parameters. The models were then evaluated using three seasons 2013/14–2015/16 using all the previous data for training. In total, this produced 7553 matches in the final test set (data on one match are missing).

In the experiments, eight leagues were used (presented in Table 2.8) and the models were fitted on a per league basis. From the eleven leagues available in the dataset, the Belgian, Dutch and Polish leagues were dropped from the analysis due to incomplete or presumably faulty data describing their player ratings. In these cases, the models built on the player level did not result in sensible predictions even on the training set.

**Significance testing.** The significance of differences is determined using pairwise application of a series of $t$-tests (Seltman [2018]). The assumption behind the $t$-test is that the compared values are normally distributed. Although normality tests are rejected for the metric values distributions for the three continuous methods – logloss, Brier score and RPS – the $t$-test was found to be relatively robust to the normality assumption for large sample sizes (Fagerland [2012]), as in our experiment. Moreover, anomalous observations may influence the result of a $t$-test. However, no major outliers were identified by inspecting the histograms of metric values (this is partially due to the fact that both Brier score and RPS are bounded metrics by two and one, respectively). This confirms the validity of the approach. Finally, significance testing by $t$-test was also applied in previous studies on match outcome prediction (e.g., Constantinou et al. [2012] Hvattum and Arntzen [2010] Peeters [2018]).

### 2.3.9 Results

Model performance is reported in two cases. Table 2.7 presents aggregate results over all the leagues while Table 2.8 presents the results for the individual leagues. The bottom two rows of Table 2.7 represent the results for two baseline methods based on results
frequency and bookmaker odds. The values in bold represent the best performance according to a given criterion in the first table and for a given subset of predictions (league) in the second, ignoring the bookmaker odds. These predictions are unanimously the most accurate ones. Unsurprisingly, the class frequency baseline is the least accurate method.

Table 2.7. Aggregate prediction statistics for the three test seasons.

<table>
<thead>
<tr>
<th>Model</th>
<th>Logloss</th>
<th>Accuracy</th>
<th>RPS</th>
<th>Brier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elo_b</td>
<td>0.9884</td>
<td>0.5171</td>
<td>0.2024</td>
<td>0.1963</td>
</tr>
<tr>
<td>Elo_g</td>
<td>0.9860</td>
<td>0.5218</td>
<td>0.2017</td>
<td>0.1958</td>
</tr>
<tr>
<td>OLR_0</td>
<td>0.9904</td>
<td>0.5146</td>
<td>0.2031</td>
<td>0.1969</td>
</tr>
<tr>
<td>OLR_1</td>
<td><strong>0.9801</strong></td>
<td><strong>0.5301</strong></td>
<td><strong>0.1996</strong></td>
<td><strong>0.1946</strong></td>
</tr>
<tr>
<td>OLR_2</td>
<td>0.9812</td>
<td>0.5226</td>
<td>0.2001</td>
<td>0.1949</td>
</tr>
<tr>
<td>PR_0</td>
<td>0.9917</td>
<td>0.5170</td>
<td>0.2035</td>
<td>0.1972</td>
</tr>
<tr>
<td>PR_1</td>
<td>0.9850</td>
<td>0.5297</td>
<td>0.2010</td>
<td>0.1956</td>
</tr>
<tr>
<td>PR_2</td>
<td>0.9850</td>
<td>0.5242</td>
<td>0.2011</td>
<td>0.1956</td>
</tr>
<tr>
<td>Class frequency</td>
<td>1.0683</td>
<td>0.4496</td>
<td>0.2295</td>
<td>0.2153</td>
</tr>
<tr>
<td>Bookmaker odds</td>
<td>0.9695</td>
<td>0.5326</td>
<td>0.1966</td>
<td>0.1922</td>
</tr>
</tbody>
</table>

To differentiate between the models, a series of t-tests was applied. Tables 2.13–2.15 at the end of this chapter, present the results of these tests. Below we focus on the main findings of the comparative study. First, we conclude that among the models examined, the two proposed versions of the OLR model – OLR_1 and OLR_2 – yield more accurate predictions than the baseline approaches – Elo_g, OLR_0 and PR_0 models. The performance of Poisson-regression based models – PR_1 and PR_2 is on a par with Elo_g. We also note that the goals-based version of the Elo model performs better than the basic version. In general, we consider the baselines of Elo_b, OLR_0 and PR_0 performing at a similar level of accuracy. At the same time, they produce inferior results to other methods studied. On the other hand, the model pairs OLR_1/OLR_2 and PR_1/PR_2 provide equally accurate predictions. The differences between these models are investigated more closely below.

Finally, looking at the aggregate statistics in Table 2.7, we observe that the evaluation according to the different metrics yields nearly identical results insofar as the relative ordering of the methods is the same. Hence, there is fair agreement between the metrics considered.

Table 2.8 decomposes the aggregate statistics for individual leagues. Skipping the bookmaker odds, which of course yield the most accurate predictions, the bold values in the table show the best-performing model for a given league. In most cases, the two versions...
Table 2.8. Performance of the models according to logloss for individual leagues for the three test seasons.

<table>
<thead>
<tr>
<th>League</th>
<th>Elo_b</th>
<th>Elo_g</th>
<th>OLR_0</th>
<th>OLR_1</th>
<th>OLR_2</th>
<th>PR_0</th>
<th>PR_1</th>
<th>PR_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>England</td>
<td>0.9965</td>
<td>0.9914</td>
<td>0.9977</td>
<td>1.0011</td>
<td>0.9945</td>
<td>0.9960</td>
<td>1.0020</td>
<td><strong>0.9913</strong></td>
</tr>
<tr>
<td>France</td>
<td>1.0099</td>
<td>1.0066</td>
<td>1.0087</td>
<td><strong>0.9967</strong></td>
<td>0.9968</td>
<td>1.0118</td>
<td>1.0029</td>
<td>1.0005</td>
</tr>
<tr>
<td>Germany</td>
<td>0.9884</td>
<td>0.9852</td>
<td>0.9926</td>
<td>0.9960</td>
<td>0.9911</td>
<td>0.9986</td>
<td>0.9982</td>
<td>0.9983</td>
</tr>
<tr>
<td>Italy</td>
<td>0.9958</td>
<td>0.9945</td>
<td>1.0031</td>
<td><strong>0.9826</strong></td>
<td>0.9863</td>
<td>0.9997</td>
<td>0.9901</td>
<td>0.9899</td>
</tr>
<tr>
<td>Portugal</td>
<td>0.9547</td>
<td>0.9529</td>
<td>0.9521</td>
<td><strong>0.9384</strong></td>
<td>0.9505</td>
<td>0.9590</td>
<td>0.9507</td>
<td>0.9587</td>
</tr>
<tr>
<td>Scotland</td>
<td>1.0019</td>
<td>0.9994</td>
<td>1.0087</td>
<td>0.9867</td>
<td><strong>0.9860</strong></td>
<td>1.0076</td>
<td>0.9909</td>
<td>0.9942</td>
</tr>
<tr>
<td>Spain</td>
<td>0.9536</td>
<td>0.9530</td>
<td>0.9507</td>
<td><strong>0.9388</strong></td>
<td>0.9448</td>
<td>0.9526</td>
<td>0.9423</td>
<td>0.9478</td>
</tr>
<tr>
<td>Switzerland</td>
<td>1.0198</td>
<td>1.0197</td>
<td>1.0267</td>
<td>1.0129</td>
<td><strong>1.0125</strong></td>
<td>1.0254</td>
<td>1.0153</td>
<td>1.0142</td>
</tr>
</tbody>
</table>

of the OLR-based model turn out to be the most accurate. This confirms the superiority of these models.

Summing up, we conclude that OLR₁, based on player ratings, provides reasonably accurate predictions. In particular, it outperforms the Elo rating system and the basic rating methods based on ordinal logistic or Poisson regression models. The conclusion to be drawn here is that *video game statistics can be successfully used to construct real-world team ratings*.

Analysis of team-specific adjustments. As discussed in Section 2.3.7, in the extended version of the model, each team is described by an extra parameter $\beta'_i$, which allows for modelling, e.g., the synergy effects or underperformance of teams with respect to the average team rating. First, we present the evaluation results in Table 2.9 for the 2015/16 season. As noted in Section 2.3.2, this was a particularly surprising season as Leicester City – a team with relatively lower rated players – won the title. The accuracy of predictions is significantly lower for both OLR₁ and PR₁ methods as compared to the baseline OLR₀ and PR₀ and Elo-type models. However, the extended versions of the models using team-specific adjustments, OLR₀ and PR₀ described in Section 2.3.6 perform considerably better. The extended models are able to adjust to the overrated and underrated ratings based on team performance.

Table 2.9. Performance of the models according to logloss for the Premiership 2015/16 season.

<table>
<thead>
<tr>
<th>Elo_b</th>
<th>Elo_g</th>
<th>OLR₀</th>
<th>OLR₁</th>
<th>OLR₂</th>
<th>PR₀</th>
<th>PR₁</th>
<th>PR₂</th>
<th>Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0448</td>
<td><strong>1.0379</strong></td>
<td>1.0508</td>
<td>1.0724</td>
<td>1.0515</td>
<td><strong>1.0403</strong></td>
<td>1.0628</td>
<td><strong>1.0398</strong></td>
<td><strong>1.0319</strong></td>
</tr>
</tbody>
</table>
We also present the results for estimating OLR for the 2015/16 season. Regularisation parameters \((\lambda_1, \lambda_2)\) were set to their optimal estimates, determined on the validation set to be \((0.025, 10.0)\) for \(L_2\) penalty operator. The estimated regularisation coefficient is higher for the team-specific coefficients. This is a typical observation across different leagues studied – better results are obtained by setting a higher penalty value for the team adjustments. Figure 2.6 presents the estimated coefficients for the teams in that season. The estimate of the global parameter for this season is \(\beta_* = 3.3\).

![Coefficient values for teams](image)

**Figure 2.6.** Team-specific coefficients for the Premiership 2015/16 season.

We observe that the coefficient for Leicester City is the highest, indicating that the team outperformed relative to the average ratings for its players. On the other hand, Aston Villa, Newcastle United and Norwich City have the lowest coefficient values, so the opposite effect held true for them. Indeed, they were all relegated to the lower division that season.

### 2.3.10 Discussion

In this section, we present a new method for compiling team ratings based on individual player statistics from the EA Sports FIFA video game. The proposed rating is used as an input variable for the ordinal logistic and Poisson regression models and compared to other rating systems on the basis of prediction accuracy. It turns out that a simple rating system based on averaged player ratings provides quite accurate predictions as compared to several baseline approaches.

The proposed model offers many advantages. By using the lineup for deriving team ratings, it accounts for temporal changes such as resting key players or player injuries. This
makes the model inherently dynamic. Moreover, it is constructed to solve possible problems that arise with the performance-based approaches with handling newly promoted teams for which little or no data are available. In fact, the more accurate predictions for newly promoted teams that player rating-based models produce give them an edge over other methods. Moreover, it is robust to result uncertainty early in a season start that makes result-based approaches struggle.

Player rating-based models may be used to predict results in which a team has not been used in the estimation procedure. The only requirement is that player ratings be up to date. For OLR\textsubscript{2} and PR\textsubscript{2} we simply use parameter $\beta_*$ and assume that the team specific adjustment equals zero. For a score-based model like the Elo rating system or OLR\textsubscript{0}, the initial team rating should be specified, which often results in setting an uninformed average rating. This can be considered another advantage of the proposed player ratings-based approach. In fact, once the parameters for the OLR\textsubscript{1}/PR\textsubscript{1} or OLR\textsubscript{2}/PR\textsubscript{2} models are obtained, they can be used to predict the result of any match in any league, provided that the player statistics are available.

On the flip side, this rating system all comes with certain limitations. First, the model has demanding maintenance requirements as we need to store up-to-date player ratings before the match as well as team lineups to generate team ratings, or prediction for a given encounter. On the other hand, in the case of the Elo model and other methods estimated from match results, only the final score is needed. This is considerably less data. Moreover, if the player ratings data are faulty or biased, then the models based on them will perform poorly. This was perhaps observed for the surprising Premiership 2015/16 season. In Chapter 3, we will come back to the issue of designing accurate player ratings from positional data.

### 2.4 Iterative team rating systems

In this section, we discuss rating systems that are based on iterative updates after each match. Such rating systems are valuable in that they are transparent, easily updated when new data arrive and, therefore, are convenient to use. Thanks to their iterative design they are also scalable which is an advantage for large-scale scenarios in e-sports. The Elo rating system is a prominent example of such a rating scheme. Its elegant formulation, accuracy and interpretability contributed to its popularity for various applications. To design new rating systems, we discuss the details of how the likelihood functions are optimised. Typically, the inference in the models provided in Equations (2.12), (2.13), or (2.24) is performed by numerical methods like the BFGS algorithm. The algorithm itself is not the focus of the rating system. However, there is a particular optimisation method
– stochastic gradient or steepest descent (e.g., Bottou 2010; Robbins and Monro 1951) – which can be viewed as a rating system with dynamic updates. Recently, such a reasoning was proposed in work by Koopman and Lit (2017) and referred to as score-driven models in the domain of econometric time series modelling (Creal et al., 2013). These models define a rating update equation as an autoregressive process with an extension to account for the derivative of predictive likelihood with respect to the modelled ratings (see the authors’ original work for the exact details). While these ideas are similar, we propose to directly use an optimisation routine, like gradient descent, to optimise a predefined loss function. We also formulate the Elo model as a part of this framework.

In a different context, Moulton (2014) used a version of the gradient descent algorithm (Adagrad – an extension of the basic gradient descent by Duchi et al., 2011) to build a rating system for an online multiplayer video game. This can also be viewed as an approach that fits the general framework of using the gradient descent algorithm for optimising the log-likelihood function. However, the exact problem setting is slightly different as the goal was to rate individual players based on their team’s score, analogously as the TrueSkill rating system does (Herbrich et al., 2006).

The goal of this section is to develop new accurate, iterative and interpretable rating systems. First, we discuss the theory underlying the Elo model in the next section. Then, based on these considerations, we design several iterative rating systems based on ordinal logistic regression and the Poisson model. We discuss the details on how the Elo model handles ties in Section 2.4.4, then turn to a more advanced optimisation technique in Section 2.4.5. Finally, we compare different approaches and discuss the results.

2.4.1 The Elo rating system as a stochastic gradient descent algorithm

One of the most appealing features of the Elo rating system (outlined in Section 2.2.1) is its simplicity. The update equations have a plausible and intuitive interpretation. However, Pelánek (2016) noted that this rating model is in fact a special case of the stochastic gradient descent algorithm. Typically, the stochastic gradient descent operates on a sample of examples (referred to as batches). Moreover, the reason the term stochastic is used is that the function is iteratively optimised by randomly choosing batches in one scan over the whole dataset (epoch). With the Elo rating system, the batch is composed of a single match. Moreover, the examples are ordered by the time a match takes place. Considering the whole history of matches as a dataset, we may say that the algorithm uses a single epoch to derive the ratings with the matches order given by their timestamps.

We formalise this discussion with exact specification of the algorithm below. Let us
define the following function \( L \), which is the negative log-likelihood of the observed results:

\[
L(r) = - \sum_{k \in M} o_{ij}^{(k)} \log \left( p_{ij}^{(k)} \right) + \left(1 - o_{ij}^{(k)} \right) \log \left(1 - p_{ij}^{(k)} \right),
\]

(2.26)

where \( o_{ij}^{(k)} \in \{0, 0.5, 1\} \) is the actual result of the match and \( p_{ij}^{(k)} \) is a prediction function defined as follows. The function \( p_{ij}^{(k)} \) yields the probability that team \( i \) wins over team \( j \) in match \( k \) which is modelled using the logistic function

\[
p_{ij}^{(k)} = \frac{1}{1 + \exp(-r_i + r_j - h)},
\]

(2.27)

where as usual \( r_i \) and \( r_j \) denote the ratings and \( h \) is the home team advantage parameter (we shall assume parameter \( h \) as given and not optimised directly). This means that we simply model match results using the logistic regression model where the team ratings \( r = (r_1, r_2, \ldots, r_n) \) are the parameters to estimate. As such, this is a convex optimization problem ([Boyd and Vandenberghe 2004]) and we may effectively use gradient descent to minimise the objective function to find the ratings.

To use gradient descent to optimise the specified loss, the derivatives with respect to rating parameters are needed. First, the objective function can be decomposed as the sum of losses for individual matches, \( L(r) = \sum_{k \in M} l^{(k)}(r) \). Taking the derivatives of a single component for match \( k \) with respect to ratings, we obtain

\[
\frac{\partial l^{(k)}}{\partial r_i} = p_{ij}^{(k)} - o_{ij}^{(k)}
\]

(2.28)

and analogously for team \( j \). The gradient descent algorithm operates iteratively to find a local minimum of a function. In general, one step gradient descent update is to perform a coordinate-wise move in the counter-gradient direction (steepest descent). Let us assume that current strength estimate from the previous iteration of the algorithm is \( r_i^{(k-1)} \). Then the update rule for the rating parameters is

\[
r_i^{(k)} = r_i^{(k-1)} - \gamma \cdot (p_{ij}^{(k)} - o_{ij}^{(k)}),
\]

(2.29)

where \( \gamma > 0 \) is the learning rate in the gradient descent algorithm and the prediction function \( p_{ij}^{(k)} \) in Equation (2.27) is evaluated using the previous rating values \( r_i^{(k-1)} \) and \( r_j^{(k-1)} \). This corresponds to the rating updates in the Elo rating system given in Equation (2.8) with \( K = \gamma \). Thus, the Elo model has a theoretical background as a single scan over the dataset of matches in the order given by the dates they appear.

There are some differences to the way the \( K \)-factor, or learning rate, is formulated. For example, in [Hvattum and Arntzen 2010], \( K \)-factor is amplified by the difference in the goals scored by the teams, as shown in Equation (2.9). The larger the difference, the greater the update to the current team ratings. Finally, setting the initial values for the ratings \( r_i^{(0)} \) corresponds to choosing prior ratings in the Elo model.
2.4.2 Iterative ordinal logistic regression model

The Elo rating system formulation outlined above provides us with the inspiration to optimise other models and their likelihood function formulations for the results observed. Now our job will be to design other rating systems in this way. Below, we propose to adopt these ideas for the OLR model presented in Section 2.2.2.

To obtain updates in the gradient descent, derivatives with respect to the model parameters are needed. We shall consider the parameters $c$, $h$ and $\lambda$ as fixed, focusing only on optimising the ratings. To provide iterative updates for the likelihood function given in Equation (2.12), it is useful first to obtain the derivatives of the logarithm of probability functions (Equations 2.11)

$$\frac{\partial \log \mathbb{P}(H_{ij})}{\partial r_i} = 1 - \mathbb{P}(H_{ij}),$$

$$\frac{\partial \log \mathbb{P}(D_{ij})}{\partial r_i} = \mathbb{P}(A_{ij}) - \mathbb{P}(H_{ij}),$$

$$\frac{\partial \log \mathbb{P}(A_{ij})}{\partial r_i} = \mathbb{P}(A_{ij}) - 1,$$

and symmetrically for the derivatives for the other team $j$. These derivatives have a simple closed form, unlike the related model in the case of the probit link function in [Koopman and Lit, 2017]. Now, we introduce the update equations for the model. For simplicity, we use $L_2$ regularisation operator for the ratings. As in the presentation of the Elo model above, the step in gradient descent algorithm is performed for a single match. Here additionally the regularisation component is added to each update. That is, it is included in the rating update formula. Hence, assuming rating $r_i^{(k-1)}$ from the previous iteration as given, the update equations become (this works symmetrically for team $j$)

$$r_i^{(k)} = \begin{cases} r_i^{(k-1)} - \gamma \cdot \left( \mathbb{P}\left(H_{ij}^{(k)}\right) - 1 + \lambda r_i^{(k-1)} \right) & \text{for } o_{ij}^{(k)} = 1, \\ r_i^{(k-1)} - \gamma \cdot \left( \mathbb{P}\left(H_{ij}^{(k)}\right) - \mathbb{P}\left(A_{ij}^{(k)}\right) + \lambda r_i^{(k-1)} \right) & \text{for } o_{ij}^{(k)} = 2, \\ r_i^{(k-1)} - \gamma \cdot \left( 1 - \mathbb{P}\left(A_{ij}^{(k)}\right) + \lambda r_i^{(k-1)} \right) & \text{for } o_{ij}^{(k)} = 3. \end{cases}$$

We note that the result probabilities of match $k$, denoted with $\mathbb{P}\left(H_{ij}^{(k)}\right)$ and $\mathbb{P}\left(A_{ij}^{(k)}\right)$ above, are evaluated using the previous rating estimates $r_i^{(k-1)}$. The update equations can be interpreted intuitively as follows. For the sake of it, let $\lambda = 0$. First, if a home team win is observed, then the home team rating update is proportional to the confidence in this event measured by the difference $1 - \mathbb{P}\left(H_{ij}^{(k)}\right)$. The more unexpected this result was, the greater the decrease team $i$’s rating. The opposite holds if the away team wins. On the other hand, in the case of a draw, both the home and away team ratings are changed so that the rating of the team expected to win based on the expression $\mathbb{P}\left(H_{ij}^{(k)}\right) - \mathbb{P}\left(A_{ij}^{(k)}\right)$ is decreased and the rating of the other team is increased. Finally, the regularisation
component equal to $\lambda r_{i}^{(k-1)}$ acts as the rating decay by shrinking it, regardless of the actual match result. We will denote this rating model with $\text{OLR}_0^I$ with the superscript “$I$” standing for the iterative version of the model.

### 2.4.3 Iterative Poisson rating systems

For the Poisson model given in Section 2.2.4, we can adopt analogous ideas to arrive at the iterative version of the model that constitutes an independent rating system. To obtain updates in the gradient descent, derivatives with respect to the model parameters are needed. We consider the parameters $c$, $h$ and $\lambda$ as fixed, focusing on the attack and defence ratings. For single match $k$ with teams $i$ and $j$, ending in score $g_i^{(k)}$ to $g_j^{(k)}$, taking derivatives with respect to the current rating values $a_i$ and $d_i$ and including regularisation we obtain

$$
\frac{\partial l^{(k)}}{\partial a_i} = \mu_i - g_i^{(k)} + \lambda \cdot (a_i - \rho d_i),
$$

$$
\frac{\partial l^{(k)}}{\partial d_i} = g_j^{(k)} - \mu_j + \lambda \cdot (d_i - \rho a_i),
$$

(2.32)

where $\log (\mu_i) = c + h + a_i - d_j$ and $\log (\mu_j) = c + a_j - d_i$. Let us assume that the current rating estimates are $a_i^{(k-1)}$ and $d_i^{(k-1)}$ (and analogously for team $j$). Then, in one step of the gradient descent algorithm for optimising the loss function $l$, we arrive at the following update equations

$$
a_i^{(k)} = a_i^{(k-1)} - \gamma \cdot \left[(\mu_i^{(k)} - g_i^{(k)}) + \lambda \cdot (a_i^{(k-1)} - \rho d_i^{(k-1)})\right],
$$

$$
d_i^{(k)} = d_i^{(k-1)} - \gamma \cdot \left[(g_j^{(k)} - \mu_j^{(k)}) + \lambda \cdot (d_i^{(k-1)} - \rho a_i^{(k-1)})\right].
$$

(2.33)

Here the rating update equations also have a plausible interpretation. Recall that $\mu_i^{(k)}$ and $\mu_j^{(k)}$ are the pre-match average goal scoring rates – the means of the corresponding Poisson variables computed for the rating estimates in the previous iteration $k - 1$. Now, if team $i$ has scored more goals than would be expected based on parameter $\mu_i^{(k)}$ – which also depends on the opponent’s defensive rating – then its attack rating increases accordingly. Analogously, if this team loses by fewer goals than expected from the overall goal scoring rate of its opponent, $\mu_j^{(k)}$, then its defensive rating increases. Moreover, the higher the differences, the larger the updates. The regularisation component in the update equations push the attack and defence ratings in the direction of zero and the correlation $\rho$ forces the ratings to remain closer to each other. We will refer to this rating system as $\text{PR}_{(a,d)}^{\rho,I}(a,d)$, where the subscript refers to the version of the model based on the decomposition of the rating into attacking and defensive skills.

Below, we also work out the updates for the single parameter Poisson model presented in Section 2.2.6. The derivative of likelihood function for a single match $l^{(k)}$ plus
regularisation yields
\[ \frac{\partial l^{(k)}}{\partial r_i} = (\mu_i - \mu_j) - (g_i^{(k)} - g_j^{(k)}) + \lambda r_i. \] (2.34)
Hence, the rating update equation becomes
\[ r_i^{(k)} = r_i^{(k-1)} - \gamma \cdot \left[ (\mu_i^{(k)} - \mu_j^{(k)}) - (g_i^{(k)} - g_j^{(k)}) + \lambda r_i^{(k-1)} \right], \] (2.35)
and analogously for team \( j \). This model can also be interpreted intuitively, based on the expected (pre-match) and observed win margin, \( \mu_i^{(k)} - \mu_j^{(k)} \) and \( g_i^{(k)} - g_j^{(k)} \), respectively. Assuming \( \lambda = 0 \), if the expected margin exceeds the observed margin, team \( i \)'s rating is adjusted downwards. We note that analogous margin-based models inspired by the Elo rating system can be found in the literature, though with scant theoretical foundations behind them (see, e.g., Carbone et al., 2016). This formulation is again a natural consequence of applying a gradient descent algorithm to estimate a model’s parameters. We shall refer to this version of the model as PR\( I_0 \).

In the next section, we discuss how to handle three-way outcomes in the Elo model.

### 2.4.4 Handling three-way outcomes

Several extensions were proposed to account for the restriction of the Elo rating system for binary outcomes in both the model estimation and prediction (Davidson, 1970; Glenn and David, 1960; Rao and Kupper, 1967). Here we will focus on a simple yet effective heuristic approach proposed more recently by Schrader (2012). Given the probability of a team \( i \) win in Equation (2.27), denoted with \( p \), the idea is to allocate the probability mass of \( \frac{4}{3}p(1-p) \) to a draw by subtracting half of it from each of the teams’ win probability. The intuition is that a draw is most likely when the prediction is of the highest uncertainty (variance); that is, \( p = \frac{1}{2} \). Then the probability of a draw attains a maximal value of \( \frac{1}{3} \).

We would suggest extending this heuristic by the following parametrisation. We introduce parameter \( \xi \), which governs the probability mass allocated to the draw. More precisely, the respective probabilities for \( (H, D, A) \) events become
\[ \left( p - \frac{\xi}{2}p(1-p), \xi p(1-p), 1 - p - \frac{\xi}{2}p(1-p) \right), \] (2.36)
where \( \xi \in [0, 2] \). In the original formulation, \( \xi = \frac{4}{3} \). In the experimental part below, we are going to show that setting \( \xi < \frac{4}{3} \) helps to obtain more accurate predictions. Accordingly, when the original prediction is ca. \( \frac{1}{2} \), the probability of a draw becomes lower than \( \frac{1}{3} \) which is in line with the result frequencies observed (see also Table 2.1). The Elo method that includes this heuristic in generating draw probability will be denoted as Elo\( \xi \).

Presenting the Elo rating system as the gradient descent algorithm also helps one recognise more clearly how this model handles ties. Namely, from the log-likelihood function optimised by the Elo model given in Equation (2.26), we conclude that a draw is
considered a “half-win, half-loss”. More precisely, we assume that a draw enters the likelihood function for the results observed as the geometric mean of win and loss probabilities, \( \sqrt{p_{ij} \cdot (1 - p_{ij})} \). This convention was later proposed by Glickman (1999). As demonstrated above, it is an implicit assumption behind the Elo rating system.

As for the Poisson-based models, a plausible feature is that they offer a relatively simple way of generating predictions for three-way outcomes. This is done via the set of Equations (2.14) using the cumulative distribution function of a Skellam random variable.

### 2.4.5 Momentum updates

In the above model formulations we considered a basic version of the gradient descent algorithm. There are many developments and extensions of this method including momentum updates (see, e.g., Goh 2017; Rumelhart et al. 1986), or more recently proposed variations including Adagrad (Duchi et al., 2011) or Adam (Kingma and Ba, 2014).

We will focus here on the extension of the algorithm involving momentum which has plausible interpretation and was studied extensively in sports (see discussion below). The method introduces one extra parameter \( \eta \in [0, 1) \) to account for weighting consecutive gradients in the update step. We will now describe the modified updates in the case of the Elo model. The first step is to accumulate the gradient

\[
v_i^{(k)} = \eta \cdot v_i^{(k-1)} + \left( p_{ij}^{(k)} - o_{ij}^{(k)} \right),
\]

where the initial value of accumulation parameter (before the first match takes place for a given team) is \( v_i^{(-1)} = 0 \). The update equation then becomes

\[
r_i^{(k)} = r_i^{(k-1)} - \gamma \cdot v_i^{(k)}.
\]

The modified update equations for other models are analogous. For \( \eta = 0 \), the model reduces to the standard gradient descent updates.

Intuitively, momentum results in higher rating updates for the teams in consecutive runs of wins and, conversely, larger decreases for the teams with a record of consecutive losses. This has a natural interpretation in sports. The question of whether the runs of results are revealing anything of substance about the next game has been studied, for example, by Dobson and Goddard (2003) and Goddard (2006) who all concluded that there is zero or even negative momentum effect (referred to as persistence effect).

That is, in a simulation study, the authors showed that for a sequence of matches of consecutive wins (losses), the probability of winning another match does not increase (decrease). On the other hand, Heuer and Rubner (2009) found that teams on a losing streak are more likely to lose the next match. As for a winning streak, the authors did not
observed any effect, or even a slight decrease the probability of winning the next match. Their conclusions were drawn using empirical data of football match results.

Using momentum in the gradient descent formulation is an alternative way of validating the hypothesis on the existence of persistence in sports. If the update equations with some parameter value \( \eta > 0 \) lead to a better performing model, we shall consider the persistence effect to be in force. In the forthcoming, for each model we are going to investigate its two variations with and without momentum updates. This will be denoted with additional superscript “\( M \)”.  

In the next section we discuss a comparative experiment between the Elo and the iterative versions of the ordinal logit and Poisson models presented.

2.4.6 Experiment setup

In this experiment we follow the same validation procedure and data we used in the previous section. This allows us to compare different methods more extensively. The seasons 2008/09–2010/11 are used as the training “burn-in” data for the rating systems. Using the two next seasons as a validation set – 2011/12 and 2012/13 – we chose the models’ optimal parameter values. Finally, the last three seasons are used as the test set to measure the models’ performance – 2013/14, 2014/15 and 2015/16.

The Elo model parameters are again set to their default values from the original formulation (Hvattum and Arntzen, 2010). For the proposed extension for the three-way outcome probabilities, parameter \( \xi \) needs to be specified. As we opted not to use the second-level OLR model to generate predictions, parameter \( h \) accounting for home advantage needs to be set as well. These two parameters are optimised on a grid of values from 1 to 1.4 and from 40 to 120 with a step of 0.1 and 10, respectively.

The iterative Poisson model requires the parameters \( c, \lambda \) and, in its attack-defence version, correlation parameter \( \rho \). Moreover, all the models require the learning rate and momentum parameters \((\gamma, \eta)\) to be set. We define the possible values for each parameter and then the combination of values to produce a large grid of parameter settings. Again, to limit the computations, we employ random search by examining only 1000 parameter settings sampled without replacement from all the possible values. Unlike in the previous experiment in Section 2.3, these parameters are set globally for all the leagues at the same time. Finally, the rating parameters need to be assigned initial values. Here we employ a simple strategy: assign the initial value of the rating parameters across all teams to zero.

The results, presented in the next section, are evaluated according to logloss, accuracy, Brier score, and RPS.
2.4.7 Results

Table 2.10 presents aggregate statistics for the test season predictions according to all the metrics considered. For completeness, we also recall the results of the baseline methods from the previous section: Elo$g$, OLR$0$, PR$0$, class frequency, and bookmaker odds. The significance of differences between methods is determined using the $t$-test. The corresponding $p$-values are provided at the end of this chapter in Tables 2.13–2.15. First, we observe that momentum versions of the models do not yield any significant improvement. In fact, these versions turn out to be less accurate in predicting match results than their base versions. The discussion will therefore henceforth focus on the base models. Second, the Elo$\xi$ model, which includes the parameterised heuristic to account for draws, performs significantly worse than its previous version (see Table 2.7) and other methods presented here. We conclude that although the method described in Section 2.4.4 is appealing in its simplicity, it is less accurate than the approach to modelling draws proposed by Hvattum and Arntzen (2010).

Interestingly, the iterative versions of the regression-based rating models – OLR$I_0$ and PR$I_0$ – perform better than their counterparts – OLR$0$ and PR$0$ estimated using BFGS algorithm. However, only in the case of the Poisson regression-based approach we do observe a significant improvement. Finally, the iterative Poisson regression-based models turn out to be more accurate than both the Elo model versions and the ordinal logistic regression-based approaches. We conclude that goal-based modelling provides better predictions than that based solely on the three-way match outcome. This confirms the results from previous studies on this issue (see, e.g., Goddard, 2005; Hvattum and Arntzen, 2010; Koopman and Lit, 2017; Ley et al., 2019; Barrow et al., 2013 also analysed this difference for other sports than football).

The predictions for individual leagues are decomposed in Table 2.11. We focus here on logloss, dropping other metrics. Moreover, we skip the momentum versions of the models that were found to produce less accurate predictions. The predictions based on the two versions of the iterative Poisson model are more accurate for all the leagues. This additionally confirms their superior performance.

2.4.8 Discussion

We now focus on the details of optimal parameter setting found by optimising the predictions based on logloss. The detailed parameter specification is presented in Table 2.12. As for the Elo model, we identified $h = 70$ and $\xi = 1.2$ to be the best home team advantage parameter and the draw weight in the discussed tie-handling procedure. In particular, we note that $\xi = 1.2 < 1.3$ as in the original formulation (Schrader, 2012). This means that
Table 2.10. Aggregate prediction statistics for the test seasons.

<table>
<thead>
<tr>
<th>Model</th>
<th>Logloss</th>
<th>Accuracy</th>
<th>RPS</th>
<th>Brier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elo&lt;sub&gt;g&lt;/sub&gt;</td>
<td>0.9860</td>
<td>0.5218</td>
<td>0.2017</td>
<td>0.1958</td>
</tr>
<tr>
<td>Elo&lt;sub&gt;ξ&lt;/sub&gt;&lt;sup&gt;g&lt;/sup&gt;</td>
<td>0.9979</td>
<td>0.5191</td>
<td>0.2046</td>
<td>0.1979</td>
</tr>
<tr>
<td>Elo&lt;sub&gt;ξ&lt;/sub&gt;+&lt;sup&gt;M&lt;/sup&gt;&lt;sup&gt;g&lt;/sup&gt;</td>
<td>0.9979</td>
<td>0.5185</td>
<td>0.2047</td>
<td>0.1980</td>
</tr>
<tr>
<td>OLR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>0.9904</td>
<td>0.5146</td>
<td>0.2031</td>
<td>0.1969</td>
</tr>
<tr>
<td>OLR&lt;sub&gt;0&lt;/sub&gt;&lt;sup&gt;I&lt;/sup&gt;</td>
<td>0.9879</td>
<td>0.5198</td>
<td>0.2020</td>
<td>0.1962</td>
</tr>
<tr>
<td>OLR&lt;sub&gt;0&lt;/sub&gt;&lt;sup&gt;I+M&lt;/sup&gt;</td>
<td>0.9887</td>
<td>0.5162</td>
<td>0.2023</td>
<td>0.1964</td>
</tr>
<tr>
<td>PR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>0.9917</td>
<td>0.5170</td>
<td>0.2035</td>
<td>0.1972</td>
</tr>
<tr>
<td>PR&lt;sub&gt;0&lt;/sub&gt;&lt;sup&gt;I&lt;/sup&gt;</td>
<td><strong>0.9825</strong></td>
<td>0.5235</td>
<td>0.2007</td>
<td><strong>0.1951</strong></td>
</tr>
<tr>
<td>PR&lt;sub&gt;0&lt;/sub&gt;&lt;sup&gt;I+M&lt;/sup&gt;</td>
<td>0.9829</td>
<td>0.5220</td>
<td>0.2008</td>
<td>0.1953</td>
</tr>
<tr>
<td>PR&lt;sub&gt;ρ,I&lt;/sub&gt;&lt;sup&gt;(a,d)&lt;/sup&gt;</td>
<td>0.9826</td>
<td><strong>0.5246</strong></td>
<td><strong>0.2006</strong></td>
<td>0.1952</td>
</tr>
<tr>
<td>PR&lt;sub&gt;ρ,I&lt;/sub&gt;&lt;sup&gt;(a,d)+M&lt;/sup&gt;</td>
<td>0.9830</td>
<td>0.5239</td>
<td>0.2007</td>
<td>0.1953</td>
</tr>
<tr>
<td>Class frequency</td>
<td>1.0683</td>
<td>0.4496</td>
<td>0.2295</td>
<td>0.2153</td>
</tr>
<tr>
<td>Bookmaker odds</td>
<td><strong>0.9695</strong></td>
<td><strong>0.5326</strong></td>
<td><strong>0.1966</strong></td>
<td><strong>0.1922</strong></td>
</tr>
</tbody>
</table>

Table 2.11. Performance of the models according to logloss for individual leagues for the test seasons.

<table>
<thead>
<tr>
<th>League</th>
<th>Elo&lt;sub&gt;g&lt;/sub&gt;</th>
<th>Elo&lt;sub&gt;ξ&lt;/sub&gt;&lt;sup&gt;g&lt;/sup&gt;</th>
<th>OLR&lt;sub&gt;0&lt;/sub&gt;</th>
<th>OLR&lt;sub&gt;0&lt;/sub&gt;&lt;sup&gt;I&lt;/sup&gt;</th>
<th>PR&lt;sub&gt;0&lt;/sub&gt;</th>
<th>PR&lt;sub&gt;0&lt;/sub&gt;&lt;sup&gt;I&lt;/sup&gt;</th>
<th>PR&lt;sub&gt;ρ,I&lt;/sub&gt;&lt;sup&gt;(a,d)&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>England</td>
<td>0.9914</td>
<td>1.0045</td>
<td>0.9977</td>
<td>0.9928</td>
<td>0.9960</td>
<td><strong>0.9846</strong></td>
<td>0.9859</td>
</tr>
<tr>
<td>France</td>
<td>1.0066</td>
<td>1.0130</td>
<td>1.0087</td>
<td>1.0099</td>
<td>1.0118</td>
<td><strong>1.0062</strong></td>
<td>1.0067</td>
</tr>
<tr>
<td>Germany</td>
<td>0.9852</td>
<td>1.0001</td>
<td>0.9926</td>
<td>0.9913</td>
<td>0.9986</td>
<td><strong>0.9837</strong></td>
<td>0.9838</td>
</tr>
<tr>
<td>Italy</td>
<td>0.9945</td>
<td>1.0040</td>
<td>1.0031</td>
<td>0.9938</td>
<td>0.9997</td>
<td><strong>0.9907</strong></td>
<td>0.9924</td>
</tr>
<tr>
<td>Portugal</td>
<td>0.9529</td>
<td>0.9602</td>
<td>0.9521</td>
<td>0.9480</td>
<td>0.9590</td>
<td>0.9470</td>
<td><strong>0.9462</strong></td>
</tr>
<tr>
<td>Scotland</td>
<td>0.9994</td>
<td>1.0189</td>
<td>1.0087</td>
<td>1.0127</td>
<td>1.0076</td>
<td>1.0004</td>
<td><strong>0.9999</strong></td>
</tr>
<tr>
<td>Spain</td>
<td>0.9530</td>
<td>0.9649</td>
<td>0.9507</td>
<td>0.9506</td>
<td>0.9526</td>
<td>0.9465</td>
<td><strong>0.9442</strong></td>
</tr>
<tr>
<td>Switzerland</td>
<td>1.0197</td>
<td>1.0380</td>
<td>1.0267</td>
<td>1.0227</td>
<td>1.0254</td>
<td><strong>1.0176</strong></td>
<td>1.0186</td>
</tr>
</tbody>
</table>

for equally rated teams the predictions yields ca. \((0.44, 0.29, 0.27)\) and \((0.35, 0.3, 0.35)\) for \(h = 70.0\) and \(h = 0\), respectively. These results are close to the overall frequency of results in a league with and without home team advantage (see Table 2.1). In the case of the Poisson-based models, parameters \(c = 0.001\) and \(h = 0.3\) mean that for equally rated teams the prediction function yields ca. \((0.45, 0.27, 0.28)\) and \((0.34, 0.31, 0.35)\) for \(h = 0\). Again, this is in line with the result frequencies observed. As for the correlation parameter in the model regularisation component \(\rho\), it equals 0.75 for the version with
Table 2.12. Optimised parameter values for different versions of the iterative rating models.

<table>
<thead>
<tr>
<th>Model</th>
<th>c</th>
<th>h</th>
<th>γ</th>
<th>λ</th>
<th>η</th>
<th>ρ</th>
<th>ξ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elo(ξ)</td>
<td>–</td>
<td>70.0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1.2</td>
</tr>
<tr>
<td>Elo(ξ+M)</td>
<td>–</td>
<td>70.0</td>
<td>–</td>
<td>–</td>
<td>0.1</td>
<td>–</td>
<td>1.2</td>
</tr>
<tr>
<td>OLR(I)₀</td>
<td>0.65</td>
<td>0.4</td>
<td>0.1</td>
<td>0.001</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>OLR(I+M)₀</td>
<td>0.65</td>
<td>0.4</td>
<td>0.1</td>
<td>0.02</td>
<td>0.25</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>PR(ρ,I)₀(a,d)</td>
<td>0.001</td>
<td>0.3</td>
<td>0.02</td>
<td>0.02</td>
<td>–</td>
<td>1.0</td>
<td>–</td>
</tr>
<tr>
<td>PR(ρ,I+M)₀(a,d)</td>
<td>0.001</td>
<td>0.3</td>
<td>0.02</td>
<td>0.02</td>
<td>0.1</td>
<td>0.75</td>
<td>–</td>
</tr>
<tr>
<td>PR(I)₀</td>
<td>0.001</td>
<td>0.3</td>
<td>0.01</td>
<td>0.002</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>PR(I+M)₀</td>
<td>0.001</td>
<td>0.3</td>
<td>0.01</td>
<td>0.02</td>
<td>0.15</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

momentum and 1.0 for the version without. This is a relatively large value which partially explains why the model performance is close to the single parameter model as the attack and defence ratings are highly correlated.

Finally, we observe that the optimal values for the momentum parameter $\eta$ are relatively low. This explains why the model’s performance is similar to the model versions without momentum updates – similar, or, in fact, even slightly worse. Based on these findings we conclude that the momentum is not really present in the sequences of consecutive results.

To sum up, in this experiment iterative rating systems based on the OLR and Poisson models proved highly accurate. In particular, the two versions of the Poisson models turned out to be more accurate than the Elo rating system, and hence they rate teams more effectively. The proposed framework based on the (stochastic) gradient descent algorithm for maximising log-likelihood function is a simple and effective method for devising team rating systems in an online fashion. The Poisson model directly uses information in the goals scored rather than, as in the Elo model (subject to some modifications included in the $K$-factor), only the final match result. In our application, the goals-based Poisson model produced better results than the variations of the Elo rating system presented here.

The rating systems based on the gradient descent algorithm can be computed quickly. In terms of a dataset of $m$ matches played by $n$ teams, the models presented in this section are of $O(m)$ time complexity. They also have minimal memory requirements of $O(n)$ as only the team ratings need to be stored. By construction, it is easy to revise the ratings using a new set of matches. On the other hand, the complexity of estimating the ratings using, e.g., the BFGS or L-BFGS-B algorithms (or another quasi-Newton method) is, in general, of higher order. Moreover, updating the ratings when new data arrive requires
such an algorithm to operate in an on-line manner. The benefits of building a rating system using the gradient descent algorithm are evident, particularly in large-scale scenarios (for example, in e-sports). Yet another advantage is that this method provides transparent and interpretable update rules.

In this section, the plausible features of the Elo model as a fast, analytically tractable rating system have been successfully translated to other rating systems. By recognising it as a special case of the gradient descent algorithm, we discussed its theoretical underpinnings. The proposed ordinal logistic and Poisson regression-based rating systems are derived in an analogous way and also enjoy the appealing features that made the Elo model popular in various application scenarios in sport and beyond.

In this chapter, we also studied team ratings based on the EA Sports FIFA video game. In particular, in Section 2.3.6 we proposed a hybrid method that utilises player-level information adjusted on a per team basis, driven its recent performance. This model may also be formulated in an iterative way. We leave extending this model as a part of future work in this area.

In the next chapter, we further investigate micro-level data on players and analyse player movements recorded over the course of a match. The main goal is to explore the possibilities for building accurate movement models and, in turn, evaluate player qualities using such a model and positional data.

### 2.5 Supplement: Testing the differences in the predictive power

Tables 2.13–2.15 present the \( p \)-values for \( t \)-tests applied to determine the significance of the differences in the results between the methods discussed in Sections 2.3.9 and 2.4.7.

To be concise, some methods are dropped in this report:

- the version of the Elo model that includes the heuristic for modelling draw probability discussed in Section 2.4.4,
- momentum versions of the models described in Section 2.4.5,
- class frequency and predictions based on bookmaker odds.

The first two methods proved less accurate in the accuracy of predictions than their base versions and hence are dropped. As for class frequency and bookmaker-based predictions, these methods turned out to be the least and the most accurate in each comparison, respectively.

When discussing the differences, we assumed the significance level of 0.05. We conclude that one approach is more accurate than another if at least two out of three test results indicate significantly better predictions.
When comparing different methods we decide to forego testing the accuracy of predictions. The other three metrics provide a better picture of how accurate the predictions are as they take into account the exact probabilities that are derived by a given model. Moreover, due to class imbalance, accuracy often fails to provide an objective measure of goodness of fit. Nevertheless, it is included in the reports for intuition and completeness.
Table 2.13. Results (p-values) of the tests for significance of the differences in logloss for the models studied.

<table>
<thead>
<tr>
<th></th>
<th>Elo_b</th>
<th>Elo_g</th>
<th>OLR_0</th>
<th>OLR^I_0</th>
<th>OLR_1</th>
<th>OLR_2</th>
<th>PR_0</th>
<th>PR^I_0</th>
<th>PR_1</th>
<th>PR_2</th>
<th>PR^p,I_{(a,d)}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elo_b</td>
<td>–</td>
<td>&lt; 0.01</td>
<td>0.36</td>
<td>0.63</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>0.22</td>
<td>&lt; 0.01</td>
<td>0.25</td>
<td>0.2</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Elo_g</td>
<td>&lt; 0.01</td>
<td>–</td>
<td>&lt; 0.05</td>
<td>0.15</td>
<td>&lt; 0.05</td>
<td>&lt; 0.05</td>
<td>&lt; 0.01</td>
<td>0.73</td>
<td>0.69</td>
<td>&lt; 0.05</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>OLR_0</td>
<td>0.36</td>
<td>&lt; 0.05</td>
<td>–</td>
<td>0.24</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>0.41</td>
<td>&lt; 0.01</td>
<td>0.05</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>OLR^I_0</td>
<td>0.63</td>
<td>0.15</td>
<td>0.24</td>
<td>–</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>0.13</td>
<td>&lt; 0.01</td>
<td>0.31</td>
<td>0.26</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>OLR_1</td>
<td>&lt; 0.01</td>
<td>&lt; 0.05</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td></td>
<td>0.45</td>
<td>&lt; 0.01</td>
<td>0.33</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>OLR_2</td>
<td>&lt; 0.01</td>
<td>&lt; 0.05</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>0.45</td>
<td>–</td>
<td>&lt; 0.01</td>
<td>0.57</td>
<td>&lt; 0.05</td>
<td>&lt; 0.05</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>PR_0</td>
<td>0.22</td>
<td>&lt; 0.05</td>
<td>0.41</td>
<td>0.13</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>–</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>PR^I_0</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>0.33</td>
<td>0.57</td>
<td>&lt; 0.01</td>
<td>–</td>
<td>0.29</td>
<td>0.24</td>
<td>0.86</td>
</tr>
<tr>
<td>PR_1</td>
<td>0.25</td>
<td>0.73</td>
<td>&lt; 0.05</td>
<td>0.31</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>0.29</td>
<td>–</td>
<td>0.96</td>
<td>0.31</td>
<td>–</td>
</tr>
<tr>
<td>PR_2</td>
<td>0.2</td>
<td>0.69</td>
<td>&lt; 0.01</td>
<td>0.26</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>0.24</td>
<td>0.96</td>
<td>–</td>
<td>0.26</td>
<td>–</td>
</tr>
<tr>
<td>PR^I_{(a,d)}</td>
<td>&lt; 0.01</td>
<td>&lt; 0.05</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>0.32</td>
<td>0.54</td>
<td>&lt; 0.01</td>
<td>0.86</td>
<td>0.31</td>
<td>0.26</td>
<td>–</td>
</tr>
</tbody>
</table>
Table 2.14. Results (p-values) of the tests for significance of the differences in Brier score for the models studied.

<table>
<thead>
<tr>
<th></th>
<th>Elo_b</th>
<th>Elo_g</th>
<th>OLR_0</th>
<th>OLR_0</th>
<th>OLR_1</th>
<th>OLR_2</th>
<th>PR_0</th>
<th>PR_0</th>
<th>PR_1</th>
<th>PR_2</th>
<th>PR_{\rho,I}^{(a,d)}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elo_b</td>
<td>–</td>
<td>&lt; 0.01</td>
<td>0.23</td>
<td>0.46</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>0.12</td>
<td>&lt; 0.01</td>
<td>0.24</td>
<td>0.22</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Elo_g</td>
<td>&lt; 0.01</td>
<td>–</td>
<td>&lt; 0.05</td>
<td>0.24</td>
<td>&lt; 0.05</td>
<td>0.06</td>
<td>&lt; 0.05</td>
<td>&lt; 0.05</td>
<td>0.69</td>
<td>0.71</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>OLR_0</td>
<td>0.23</td>
<td>&lt; 0.05</td>
<td>–</td>
<td>0.12</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>0.36</td>
<td>&lt; 0.01</td>
<td>&lt; 0.05</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>OLR_0'</td>
<td>0.46</td>
<td>0.24</td>
<td>0.12</td>
<td>–</td>
<td>&lt; 0.01</td>
<td>&lt; 0.05</td>
<td>0.05</td>
<td>&lt; 0.01</td>
<td>0.35</td>
<td>0.33</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>OLR_1</td>
<td>&lt; 0.01</td>
<td>&lt; 0.05</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>–</td>
<td>0.44</td>
<td>&lt; 0.01</td>
<td>0.34</td>
<td>&lt; 0.01</td>
<td>&lt; 0.05</td>
<td>0.33</td>
</tr>
<tr>
<td>OLR_2</td>
<td>&lt; 0.01</td>
<td>0.06</td>
<td>&lt; 0.01</td>
<td>&lt; 0.05</td>
<td>0.44</td>
<td>–</td>
<td>&lt; 0.01</td>
<td>0.59</td>
<td>0.07</td>
<td>&lt; 0.05</td>
<td>0.57</td>
</tr>
<tr>
<td>PR_0</td>
<td>0.12</td>
<td>&lt; 0.05</td>
<td>0.36</td>
<td>0.05</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>–</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>PR_0'</td>
<td>&lt; 0.01</td>
<td>&lt; 0.05</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>0.34</td>
<td>0.59</td>
<td>&lt; 0.01</td>
<td>–</td>
<td>0.41</td>
<td>0.31</td>
<td>0.86</td>
</tr>
<tr>
<td>PR_1</td>
<td>0.24</td>
<td>0.69</td>
<td>&lt; 0.05</td>
<td>0.35</td>
<td>&lt; 0.01</td>
<td>0.07</td>
<td>&lt; 0.01</td>
<td>0.41</td>
<td>–</td>
<td>0.91</td>
<td>0.44</td>
</tr>
<tr>
<td>PR_2</td>
<td>0.22</td>
<td>0.71</td>
<td>&lt; 0.01</td>
<td>0.33</td>
<td>&lt; 0.05</td>
<td>&lt; 0.05</td>
<td>&lt; 0.01</td>
<td>0.31</td>
<td>0.91</td>
<td>–</td>
<td>0.34</td>
</tr>
<tr>
<td>PR_{I}</td>
<td>&lt; 0.01</td>
<td>&lt; 0.05</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>0.33</td>
<td>0.57</td>
<td>&lt; 0.01</td>
<td>0.86</td>
<td>0.44</td>
<td>0.34</td>
<td>–</td>
</tr>
</tbody>
</table>
Table 2.15. Results (p-values) of the tests for significance of the differences in RPS for the models studied.

<table>
<thead>
<tr>
<th></th>
<th>$\text{Elo}_b$</th>
<th>$\text{Elo}_g$</th>
<th>$\text{OLR}_0$</th>
<th>$\text{OLR}_0^I$</th>
<th>$\text{OLR}_1$</th>
<th>$\text{OLR}_2$</th>
<th>$\text{PR}_0$</th>
<th>$\text{PR}_0^I$</th>
<th>$\text{PR}_1$</th>
<th>$\text{PR}_2$</th>
<th>$\text{PR}_{(a,d)}^{\rho,I}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Elo}_b$</td>
<td>$-$</td>
<td>$&lt; 0.01$</td>
<td>0.32</td>
<td>0.16</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>0.17</td>
<td>$&lt; 0.01$</td>
<td>0.11</td>
<td>0.1</td>
<td>$&lt; 0.01$</td>
</tr>
<tr>
<td>$\text{Elo}_g$</td>
<td>$&lt; 0.01$</td>
<td>$-$</td>
<td>$&lt; 0.05$</td>
<td>0.28</td>
<td>$&lt; 0.05$</td>
<td>$&lt; 0.05$</td>
<td>$&lt; 0.01$</td>
<td>0.44</td>
<td>0.48</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
</tr>
<tr>
<td>$\text{OLR}_0$</td>
<td>0.32</td>
<td>$&lt; 0.05$</td>
<td>$-$</td>
<td>0.12</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>0.33</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
</tr>
<tr>
<td>$\text{OLR}_0^I$</td>
<td>0.16</td>
<td>0.28</td>
<td>0.12</td>
<td>$-$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.05$</td>
<td>$&lt; 0.01$</td>
<td>0.21</td>
<td>0.22</td>
<td>$&lt; 0.01$</td>
</tr>
<tr>
<td>$\text{OLR}_1$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.05$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>$-$</td>
<td>0.34</td>
<td>$&lt; 0.01$</td>
<td>0.16</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
</tr>
<tr>
<td>$\text{OLR}_2$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.05$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>0.34</td>
<td>$-$</td>
<td>$&lt; 0.01$</td>
<td>0.34</td>
<td>0.07</td>
<td>$&lt; 0.05$</td>
<td>0.39</td>
</tr>
<tr>
<td>$\text{PR}_0$</td>
<td>0.17</td>
<td>$&lt; 0.05$</td>
<td>0.33</td>
<td>$&lt; 0.05$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>$-$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
</tr>
<tr>
<td>$\text{PR}_0^I$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>0.16</td>
<td>0.34</td>
<td>$&lt; 0.01$</td>
<td>$-$</td>
<td>0.67</td>
<td>0.51</td>
<td>0.64</td>
</tr>
<tr>
<td>$\text{PR}_1$</td>
<td>0.11</td>
<td>0.44</td>
<td>$&lt; 0.01$</td>
<td>0.21</td>
<td>$&lt; 0.01$</td>
<td>0.07</td>
<td>$&lt; 0.01$</td>
<td>0.67</td>
<td>$-$</td>
<td>0.8</td>
<td>0.61</td>
</tr>
<tr>
<td>$\text{PR}_2$</td>
<td>0.1</td>
<td>0.48</td>
<td>$&lt; 0.01$</td>
<td>0.22</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.05$</td>
<td>$&lt; 0.01$</td>
<td>0.51</td>
<td>0.8</td>
<td>$-$</td>
<td>0.46</td>
</tr>
<tr>
<td>$\text{PR}_{(a,d)}^{\rho,I}$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>0.18</td>
<td>0.39</td>
<td>$&lt; 0.01$</td>
<td>0.64</td>
<td>0.61</td>
<td>0.46</td>
<td>$-$</td>
</tr>
</tbody>
</table>
Individual movement models for player ratings

In the previous chapter, we proposed a rating system that includes player-level information from a video game. In recent years, the availability of new data sources in sports enabled a wide variety of innovative approaches and new solutions to existing problems. In this chapter, we propose a new method for estimating player movement models using positional (or trajectory) data. Such a model lays the groundwork for a number of analyses. In particular, positional data and the movement models can be used for quantitative assessment of different player qualities discussed in Section 2.3.

This chapter is organised as follows. First, in the next section, we discuss the work related to our study. Section 3.2 presents the method for estimating the individual movement model using positional data. In Section 3.3, we then discuss how to compute zones of control using the proposed model. Finally, we discuss the results in the light of quantitative approaches for devising player ratings and some issues related to estimating the model using positional data.

The results presented in this chapter were published as a part of the Special Issue on Machine Learning for Soccer in Machine Learning Journal (Brefeld, Lasek and Mair, 2018). This chapter also includes several new results.

3.1 Preliminaries

Arguably, the most important aspect in team sports is player coordination. In football, collective movements are crucial in positional play, counter-attack, or pressing. Models that quantify the probability of reaching a given location for a player in a given time are therefore of ultimate interest. Such models are called movement models.

Traditional approaches ground on elementary physical models for a player’s ability to move and are typically based on simplified assumptions (Fujimura and Sugihara, 2005; Taki and Hasegawa, 2000; Taki et al., 1996). Moreover, it is often the case that these movement models treat every player the same by assuming that a single movement model
serves all players equally well and hence, ignore individual differences between the players.

Movement models are important as they lay the groundwork for many applications. One of them is computation of zones of control, also often referred to as dominant regions. The zone of control for a single player is defined by the area that the player can attain before any other player. These zones for individual players together form a team’s dominant region. These two concepts serve as a basis for deriving different player ratings.

In this chapter, we propose to estimate individual player movement models from positional data and demonstrate how they can be turned into zones of control. We pursue a probabilistic approach that leverages players’ positions and velocities and results in a distribution of all reachable positions in a given time. When compared to traditional methods, our approach leads to realistic zones of control. On the quantitative basis, we compare different models for the problem of predicting pass outcomes. Below, we start with reviewing the work related to the study presented here. In Section 3.4 we will specifically discuss related studies in the area of devising player ratings.

Trajectory analyses are often carried out via wearable devices like smartphones or smartwatches that are equipped with accelerometers and gyroscopes (Mazimpaka and Timpl, 2016; Zheng, 2015). Often, the trajectories serve only as proxies for a higher level research question such as road defects identification (see, e.g., Byrne et al., 2013; Mohan et al., 2008), price discrimination by insurance companies (Paefgen et al., 2011), or activity recognition (Avci et al., 2010; Lasek and Gagolewski, 2015c; Zdravevski et al., 2017).

Similarly, trajectory data in sports is used to identify movement patterns. At an individual level, Zhao et al. (2016) use Gaussian Processes to model velocity (flow) of athletes in ski races. Laube et al. (2005) proposed to analyse relative motions and different temporal patterns across many subjects. As an exemplary application, the authors analysed positional data to retrieve patterns from coordinated team motions. The problem of pattern identification in groups of moving objects was also studied by Gottfried (2008, 2011). The author proposed qualitative descriptions of movement patterns using a set of atomic motions as building blocks to analyse and describe more complex behaviours. Sprado and Gottfried (2009) applied this method to RoboCup and real-world football matches. Knauf et al. (2016) proposed spatio-temporal convolution kernels as a similarity measure over time and space and identify game initiations and attacking patterns using a clustering approach. Similarly, Janetzko et al. (2014) grouped attacking patterns of strikers. In general, frequent patterns in trajectory data can be retrieved by applying episode mining algorithms (Haase and Brefeld, 2014).

Zhang et al. (2016) proposed to a method to visualise time interval data to analyse player and team performance. The analysis includes a variety of features including player velocities or ball possession as a measure of team dominance. Other methods include, for
example, estimating the probability of a shot being made or a goal being scored (Harmon et al., 2016; Link et al., 2016; Lucey et al., 2014). Zheng et al. (2016) and Le et al. (2017) proposed to model player trajectories with recurrent neural networks for improving player positioning in basketball and football. Similarly, convolutional neural networks were used by Harmon et al. (2016) to estimate the probability of scoring opportunities. Generally, applying neural networks to player trajectories, represented either as sequences or images, renders the need for engineering hand-crafted features unnecessary and may thus be beneficial in certain situations. Finally, Memmert et al. (2016) and Gudmundsson and Horton (2017) provided a general overview of analyses based on positional data in team sports. Other interesting applications include pass quality evaluation (Brooks et al., 2016; Chawla et al., 2017; Horton et al., 2015) or injury prediction (Rossi et al., 2017).

We proceed to discuss the previous research focusing specifically on estimating player movement models, which are most relevant for our analysis. Taki and Hasegawa (2000) proposed a movement model which is based on a player’s velocity and an acceleration profile along different directions. The authors discussed the dependency of acceleration on velocity and its direction and emphasised that the acceleration decreases with increasing speed. However, for a practical application, the authors ignored physical details and focused on a basic version of their model in which a player is able to move with the same acceleration regardless of the direction he is facing. Moreover, the acceleration is assumed to be constant and, in particular, independent of the player’s current speed. This results in the possibility of unbounded speed. Fujimura and Sugihara (2005) extended this approach by introducing a resistive force to prevent velocities growing infinitely. With this in mind, the two approaches simplify physical laws to a large extent to model player movements. Moreover, both approaches are designed in one-serves-all way as these models are not tailored to account for differences between players. More recently, Gudmundsson and Wolle (2014) sketched how such an individual movement model can be estimated from positional data. They suggest approximating a player’s reachable region in time $t$ by constructing a convex polygon for all the points he reached within this time given his actual position. However, neither technical nor algorithmic details of their approach were presented.

Once a movement model is established, it serves as a foundation for various applications in match analysis. Perhaps one of the most important is the computation of zones of control, also referred to as dominant regions. This concept was introduced by Taki and Hasegawa (2000) and defined as the part of the pitch that can be attained by a player before all others. Consequently, zones of control are useful for a variety of analyses. These include methods to evaluate pass quality (Gudmundsson and Wolle, 2014; Horton et al., 2013; Nakanishi et al., 2009; Taki and Hasegawa, 2000), pressing (Taki and Hasegawa, 2000), analysis of team behaviour and interaction (Fonseca et al., 2012), or organization
and positioning in both attack and defence (Ueda et al., 2014). A player movement model and the implied zones of control are therefore fundamental concepts in understanding the game dynamics and, in turn, compiling player ratings in a data-driven way.

### 3.2 Estimating individual movement models

Before we proceed to describe different approaches to modelling player movements we introduce some notation. Let $p^{(k)}(t) = (x^{(k)}(t), y^{(k)}(t))_{t \in \mathbb{R}_{>0}}$ be the trajectory of player $k$ describing his positions in area $F \subset \mathbb{R}^2$ over time and let $v^{(k)}_t \in \mathbb{R}^2$ be his velocity vector at time $t \in \mathbb{R}_{>0}$ with its magnitude (speed) equal $v^{(k)}_t = \|v^{(k)}_t\|$. Time index $t$ is typically discrete as samples are generated at equidistant timestamps $t_1, t_2, \ldots, t_n$, where $t_{i+1} - t_i = \tau > 0$ is fixed. The trajectories and the associated velocities form dataset $D = \{(p^{(k)}_i, v^{(k)}_i)\}_{i=1}^n$. The goal is to generate a probabilistic model of the player’s whereabouts in time horizon $t_\Delta > 0$ given his current position $p^{(k)}_t$ and velocity $v^{(k)}_t$:

$$P_{t_\Delta}\left(((x, y) \mid p^{(k)}_t, v^{(k)}_t)\right).$$

To not clutter the notation unnecessarily, we discard the player index $k$ whenever possible and focus on data for a single player.

#### 3.2.1 Existing approaches

We start with briefly reviewing existing approaches for modelling player movements. The simplest model assumes that all players are able to move in all directions equally fast at a constant speed. Thus, there is no acceleration or direction of movement and the resulting zones of control are equal to Voronoi tessellations (Voronoi, 1908) of the pitch using the players as centre points. This model is referred to as Voronoi. Taki and Hasegawa (2000) improve on this by incorporating the notion of velocity and acceleration. Their model is based on the assumption that every player is able to accelerate in each direction equally fast with magnitude of $a_{\text{max}} > 0$. Thus, at time $t = 0$ the player begins to move with acceleration $a_{\text{max}}$ in a direction given by an angle $\phi \in [-\pi, \pi)$. Assuming that a player has been moving with speed $v$ in the direction of the $x$-axis, in time $t$ his position $p$ is given by

$$p = (x, y) \quad \text{with} \quad \begin{cases} x = \frac{1}{2}a_{\text{max}} \cdot \cos(\phi) \cdot t^2 + vt, \\ y = \frac{1}{2}a_{\text{max}} \cdot \sin(\phi) \cdot t^2. \end{cases} \quad (3.1)$$

\footnote{The velocity can be directly estimated from positional data in case it is not provided.}
In other words, the set of points that can be reached exactly in time \( t \) forms a circle centred at \( c \in \mathbb{R}^2 \) with radius \( r > 0 \), where
\[
c = (vt, 0) \quad \text{and} \quad r = \frac{1}{2} a_{\text{max}} t^2.
\]

The details of setting the model’s parameters are relegated to Section 3.2.3. We refer to this model as Taki & Hasegawa.

Fujimura and Sugihara (2005) introduced a resistive force proportional to a player’s current speed to render the movement model more realistic. The resistive force prevents the speed to grow infinitely and even clips it at maximal value \( v_{\text{max}} > 0 \). At time \( t = 0 \), a player accelerates in direction \( \phi \in [-\pi, \pi) \) with the underlying assumption that he can exert the maximum speed in any direction. The position \( p \) of the player at time \( t \) is given by
\[
p = (x, y) \quad \text{with} \quad \begin{cases} x = v_{\text{max}} \cdot \cos(\phi) \cdot \left( t - \frac{1 - \exp(-\alpha t)}{\alpha} \right) + v \cdot \frac{1 - \exp(-\alpha t)}{\alpha}, \\ y = v_{\text{max}} \cdot \sin(\phi) \cdot \left( t - \frac{1 - \exp(-\alpha t)}{\alpha} \right), \end{cases} \tag{3.2}
\]
where \( v \) is the initial velocity in the direction of the \( x \)-axis and the parameter \( \alpha > 0 \) accounts for for the resistive force. Hence, the set of points within reach of the player in time \( t \) forms a circle centred at \( c \in \mathbb{R}^2 \) with radius \( r > 0 \), where
\[
c = \left( v \cdot \frac{1 - \exp(-\alpha t)}{\alpha}, 0 \right) \quad \text{and} \quad r = v_{\text{max}} \cdot \left( t - \frac{1 - \exp(-\alpha t)}{\alpha} \right).
\]

The model is referred to as Fujimura & Sugihara.

Figure 3.1 presents movement profiles for the three discussed approaches – Voronoi, Taki & Hasegawa, and Fujimura & Sugihara-based models (in columns) – for two different values of player initial speed moving in the direction of \( x \)-axis (in rows). For a slowly moving player, all the three models result in similar circular-shaped boundaries of reachable region in a given time. The differences between the models become more pronounced when the player moves at a higher speed. Taki & Hasegawa-based reachable regions in a given time become drop-shaped. Moreover, this approach is largely focused on the part of the area in front of the player. The approach by Fujimura & Sugihara leads to nested circles centred at a short distance in front of the player. On the other hand, the Voronoi-based approach does not take into account the velocity at all. These are the consequences of using simplified physical models. One would expect elliptical curves for the reachable region. In the next section, we propose to estimate the movement model from positional data. We will demonstrate that this approach indeed leads on elliptically shaped movement profiles. Moreover, the model is tailored to on a per player basis.
Figure 3.1. Movement models visualisation for a player moving with speed 7 km/h (top row) and 24 km/h (bottom row) in direction of x-axis (Brefeld, Lasek and Mair, 2018).

3.2.2 Constructing movement models from positional data

We now describe how to compute probabilistic movement models from positional data. The computation is based on triplets \((p_s, p_t, p_u)\) with \(s < t < u\), \(t - s = t\delta\) and \(u - t = t\Delta\) that are drawn from a player’s trajectory. Coordinates \(p_s\) and \(p_t\) will be used to estimate the direction in which the player moves while \(p_u\) will be used to estimate his ability to move. First, the triplet is transformed by a translation so that point \(p_t\) is centred at \((0, 0)\) followed by a rotation so that the vector \(p_s p_t = (x_t - x_s, y_t - y_s)\) is aligned with the x-axis. This way, the transformed position \(p_u\) describes the point the player reaches assuming his current position is the origin, moving in direction of x-axis with velocity \(v\) at the given speed \(\|v\|_2 = \|v_t\|_2 = v_t\). Figure 3.2 provides an overview of this approach.

Formally, for a position triplet \((p_s, p_t, p_u)\) within a player’s trajectory, let \(\psi\) be a function that maps such a triplet to \((x, y)\) coordinates:

\[
\psi : \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2, \quad (p_s, p_t, p_u) \mapsto p = (x, y).
\]  

We obtain the destination point \(p = (x, y)\) using a representation in polar coordinates

\[
(x, y) = (r \cdot \cos(\theta), \ r \cdot \sin(\theta)),
\]

where \(\theta\) is a signed angle and \(r\) the distance. Angle \(\theta\) is computed via the following
Figure 3.2. Left: Illustration of the $\psi$ function. Centre and right: Example for a time horizon of $t_\Delta = 1$ second and a player velocity of $14–20$ km/h. Position triplets $(p_s, p_t, p_u)$ are used to obtain data samples which are then smoothed using a KDE with a Gaussian kernel to obtain the movement model.

The direct calculation

$$\theta = \angle(p_s p_t, p_t p_u) = \arctan2(y_t - y_s, x_t - x_s) - \arctan2(y_u - y_t, x_u - x_t)$$

(3.4)

for $p_s \neq p_t, p_t \neq p_u$, where $\arctan2(y, x)$ is a function that yields an angle between point $(x, y)$ and the positive $x$-axis. The distance is given by

$$r = \|p_t p_u\|_2.$$  

(3.5)

Figure 3.2 (left) illustrates how the mapping $\psi$ processes data triplets to derive the position $p$. Samples of the transformed positions and the associated speed values are collected within set $S_{t_\Delta}$. This approach is summarised in Algorithm 1.

**Algorithm 1** Computation of movement samples.

**Input:** Dataset $D = \{(p_t, v_t)\}_{i=1}^n$, $t_\delta$, $t_\Delta$

**Output:** Set $S_{t_\Delta}$ of attained positions in time $t_\Delta$ including initial velocities

1. for $s < t < u$ s.t. $s = t - t_\delta, u = t + t_\Delta$ do
2. $p = \psi(p_s, p_t, p_u)$ $\triangleright$ transformed destination as outlined in Equation (3.3)
3. $S_{t_\Delta} = S_{t_\Delta} \cup \{(p, v_t)\}$ $\triangleright$ append the derived sample to set $S_{t_\Delta}$
4. end for

Having obtained set $S_{t_\Delta}$, it is possible to define a probability distribution over possible player whereabouts given his position and initial velocity. This can be done with a two-dimensional kernel density estimate (KDE). For a dataset of $n$ points, $(x_i, y_i)_{i=1}^n$, it is defined as

$$\text{KDE}(x, y) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i, y - y_i),$$

where $K_h$ is a kernel function parametrised by a bandwidth parameter $h$. Typical examples are uniform, triangle, Epanechnikov, or Gaussian kernels (see, e.g., Turlach, 1993).
Due to practical considerations, we suggest to discretise the speed range and include it in the model at a lower level of granularity. To this end, we define a subset of points

\[ S_{t, V} = \{ (p, v) \mid v \in V \} \subseteq S_t \]

for a range of velocity values in an interval \( V = [v_L, v_U] \) and compute the individual movement model using a KDE based on the samples from this set. We obtain several KDEs depending on different velocity ranges denoted as \( P_{t, V}^{KDE} \). To evaluate the likelihood of attaining a given position \( p \in \mathbb{R}^2 \), we use

\[
P_{t, v} (p \mid p_{t}, v_t) = P_{t, v} (p \mid p_{t}, p_{t-t_S}, v_t) = P_{t, V}^{KDE} \left( \psi(p_{t-t_S}, p_t, p) \right)
\]

for \( v_t \in V \). We introduce an extra conditioning on the previous player’s position \( p_{t-t_S} \) is used to estimate the direction (angle \( \theta \)) in which the player moves. Figure 3.2 (centre) presents a set of samples collected and Figure 3.2 (right) a corresponding movement model based on a Gaussian KDE. In this figure, the bandwidth parameter \( h \) – which in this case is the standard deviation of the underlying Gaussian density – is set arbitrarily to 0.7 (see, e.g., Turlach, 1993 for an overview of bandwidth selection methods).

The model relies on a particular discretisation of the speed range. Analogously, different models are obtained for different values of time horizon parameter \( t_\Delta \). In fact, we are interested in several values of this parameter for different time horizons (of about one second) in a given interval.

In some cases, the triplets of points used to estimate the model can contain outliers. These may come from an interruption during a match (e.g., due to a foul or a corner kick) or errors in the data collection process. Hence, triplets containing outliers should be discarded. Finally, given that a player’s ability to move should be symmetric with respect to the direction he is facing, the set of positions reached can be augmented with \( (\bar{p}, v) \), where \( \bar{p} = (x, -y) \) for each sample \( (p, v) \in S_{t, V} \).

### 3.2.3 Example estimation

There are two typical methods of collecting positional data in sports. The first is to attach sensors to players and ball to monitor their positions (Grünn et al., 2011; Mutschler et al., 2013; Pettersen et al., 2014). The second is to use computer vision algorithms for retrieving players’ and ball’s trajectories in consecutive frames (Barris and Button, 2008; D’Orazio and Leo, 2010). The positional data we use in the experiments stem from the latter and are recorded at 25 Hz. For a single match, this usually yields over \( 25 \cdot 60 \cdot 90 = 135,000 \) samples (due to possible extra time by the end of each half). The dimensions of a typical football field are 105 by 68 meters and the coordinates of positions in the trajectory data
are given relative to the origin of the field, which is set to \((0, 0)\). Hence, player coordinates \((x, y)\) are within \(F = [-52.5, +52.5] \times [-34.0, +34.0] \subset \mathbb{R}^2\).

Except for the Voronoi-based approach, the models discussed in Section 3.2 involve user-defined parameters that need to be specified. For Taki \& Hasegawa, the acceleration parameter \(a_{\text{max}}\) can be derived from the corresponding speed samples \(v_t\) using \(a_t = \frac{1}{h}(v_{t+h} - v_t)\). Here, these are computed for a time horizon of \(h = 1\) second using data from a single match. Based on this, we set \(a_{\text{max}} = 4.2\, \text{m/s}^2\), which is equal to the 0.999-quantile of the derived values. This quantile instead of the maximum acceleration observed is used to ignore outliers. The model by Fujimura \& Sugihara includes two parameters: \(\alpha\) and \(v_{\text{max}}\). We use \(\alpha = 1.3\), which is the value proposed in the original paper [Fujimura and Sugihara, 2005], and \(v_{\text{max}} = 8.0\, \text{m/s}\), where the latter corresponds to the 0.999-quantile of the observed speed values (analogously as in the case of \(a_{\text{max}}\) parameter in the previous model).

To compute the individual movement models presented in Section 3.2.2, we use \(t_\delta = 0.2\) and \(t_\Delta = 1\) second in Algorithm 1. We also use five different speed intervals shown in Table 3.1. The discretisation employed is a common way to bin velocities to account for sparseness in real data, as the number of samples per speed interval may vary significantly (Coutts et al., 2010; Gudmundsson and Wolle, 2014; Lago-Peñas et al., 2009). Table 3.1 also presents speed distributions for three different players: a goalkeeper, a defender, and an attacking midfielder. On average, the field players walk and jog and save their energy for only a few sprints.

Table 3.1. Distribution of speed classes for three different players. Differences between the defender and the midfielder are significant according to a \(\chi^2\)-test.

<table>
<thead>
<tr>
<th>Speed Range (km/h)</th>
<th>Goalkeeper</th>
<th>Defender</th>
<th>Midfielder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stand &lt; 1</td>
<td>0.2719</td>
<td>0.1125</td>
<td>0.1201</td>
</tr>
<tr>
<td>Walk 1–7</td>
<td>0.6668</td>
<td>0.5322</td>
<td>0.5090</td>
</tr>
<tr>
<td>Jog 7–14</td>
<td>0.0544</td>
<td>0.2757</td>
<td>0.2811</td>
</tr>
<tr>
<td>Run 14–20</td>
<td>0.0057</td>
<td>0.0597</td>
<td>0.0636</td>
</tr>
<tr>
<td>Sprint &gt; 20</td>
<td>0.0012</td>
<td>0.0200</td>
<td>0.0262</td>
</tr>
</tbody>
</table>

Movement estimates for these three players are presented in Figure 3.3 using a Gaussian KDE with the bandwidth set to one. Note that there are small but significant differences between the players’ ability to move. For example, the goalkeeper has a significantly lower probability to reach distant positions compared to the field players. The reason lies, however, not in his limited ability to move but in the lack of corresponding observations: goalkeepers hardly push forward and usually cover a smaller radius than field players.
The figure also reveals that the midfielder covers a wider area and is, on average, moving faster than his peers. The issue of few data samples collected for the goalkeeper can be balanced with an average model, see discussion in Section 3.4.

Figure 3.3. Individual movement models for three different players with initial speed in the range 14–20 km/h in the direction of the x-axis: goalkeeper (left), defender (centre), and midfielder (right).

3.3 Zones of control

Movement models can be used to compute zones of control (or dominant regions) for individual players and teams as a whole (Gudmundsson and Wollg, 2014; Horton et al., 2015; Taki and Hasegawa, 2000). Below we formally define dominant regions for the models presented in the previous section. To do so, it is beneficial to recall the definition in the case of the traditional movement models driven by physical laws. For the probabilistic model it is analogous and is discussed later.

Let function $\Gamma : \mathbb{R}^2 \rightarrow \mathbb{R}_{\geq 0}$ yield the time needed to reach position $p \in \mathbb{R}^2$ for a player $k$ at position $p^{(k)}_t$ moving with velocity $v^{(k)}_t$ in a given direction, i.e., $\Gamma(p | p^{(k)}_t, v^{(k)}_t)$. This function is specific to a given physical model governing player movements. In other words, for a given player, function $\Gamma$ yields the minimal time that satisfies Equations (3.1) and (3.2) for the Taki & Hasegawa and the Fujimura & Sugihara models, respectively. In (Taki and Hasegawa, 2000), the concept of a player’s zone of control is defined as the subset $D^{(k)}$ of the field $F$, where player $k$ can arrive before any other player $i \neq k$. Formally, this is to say that $D^{(k)} \subseteq F$ is defined such that $\forall p \in D^{(k)}$

$$k = \arg \min_{i \in \{1, 2, ..., K\}} \Gamma(p | p^{(k)}_t, v^{(k)}_t).$$

This provides a decomposition of the playing area into $K$ disjoint subsets, $F = \bigcup_{i=1}^{K} D^{(i)}$ with $D^{(i)} \cap D^{(j)} = \emptyset$ for $i \neq j$. It is also noteworthy that interdependencies between
players may be complex enough to produce a player’s zone of control that is not a single
connected region (Taki and Hasegawa, 2000).

The zone of control of a team is defined analogously. In our approach a different per-

spective is taken by considering probabilistic movement models for a given time horizon.
That is, the zones of control are derived on the basis of density functions of possible
players’ whereabouts. Therefore, we obtain probability distributions for locations of indi-

gual players over the playing area. The computation of those regions using probabilistic
movement models is detailed below.

3.3.1 Probabilistic formulation

Let $P^{(k)}_{t\Delta}(p | p_t, p_{t-t_k}, v_t)$ be the movement model of player $k$ introduced in Section 3.2.2.
The model quantifies the likelihood of player $k$ to reach position $p$ given his current and
previous positions, $p_t$ and $p_{t-t_k}$, respectively, speed $v_t$, and time horizon $t\Delta$. A position $p$
is controlled by player $k$ having the highest likelihood

$$k = \arg \max_{i \in \{1, 2, \ldots, K\}} P^{(i)}_{t\Delta}(p | p_t, p_{t-t_k}, v_t).$$

(3.7)

Now, to partition a playing area $F$ into zones controlled by players it is necessary to
iterate over each position $p \in F$ to obtain the player who controls it. In a practical
application, the playing area $F$ is typically discretised into a finite grid of relatively small-
sized squares (Franks et al., 2015; Lucey et al., 2012; Nakanishi et al., 2009; Narizuka
et al., 2014). The playing area is covered by $n_x \cdot n_y$ squares of side $\Delta$ aligned with sides
of the playing area. Naturally, the smaller $\Delta$, the better the approximation. The player
who controls the centre of a square as defined in Equation (3.7) is then assumed to
control the region of the field given by that small square. This procedure is employed for
visualisation purposes below.

3.3.2 Examples

Below we demonstrate the differences between the zones of control implied by different
movement models: Voronoi, Taki & Hasegawa, Fujimura & Sugihara, and the proposed
data-driven model for an example match situation. Figure 3.4 presents the resulting re-
gions. The arrows indicate a player’s direction and are proportional to his speed. The arrow
origins are fixed at a player position one and two seconds before the situation took place.

The top left part of the figure demonstrates Voronoi tessellation of the field. It is
achieved by assuming a simplistic movement model that every player is able to run in
each direction equally fast. As such, this approach ignores the players’ current velocities.
First, we observe that the black player that is trying to get the ball from the white
Figure 3.4. Zones of control implied by different movement models. The black team plays from top to down [Brefeld, Lasek and Mair 2018].
player controlling it in the midfield has an improbable zone of control in the back of him, especially given that he moves quickly. Other models account for his high speed and produce controlled region in front of him. Further issues with the Voronoi approach may be observed at the wings of the pitch. For example, the right winger (or fullback) of the white team is moving quickly toward opponents half, however, his controlled zone in that area is of modest size. This is because the left defender of the black team is closer to that area even though he is moving slowly toward the centre of the pitch. On the other hand, other models imply that the white player controls a large part of the wing, which is intuitive. Finally, the white player in the top left part of the picture is also moving toward the centre while his controlled zone goes deeply into the opponent team’s half.

The top right part of the figure demonstrates the zones of control implied by the Taki & Hasegawa model. First, we note that this model is heavily focused towards the regions in front of a player. More precisely, due to unlimited growth of speed combined with its high initial value, the model produces zones of control expanding in the areas in front of the player. Moreover, the model also implies somewhat limited ability to turn. This produces drop-like controlled zones (see also Figure 3.1) as demonstrated by the white player at the right wing. Yet another consequence of this model is that it implies a large part of the white team defensive zone (the lower right corner of the playing area) controlled by the black team despite the fact that there are several white players (including the goalkeeper) closer to it.

The bottom left figure presents the zones of control underlying the Fujimura & Sugihara model. This approach corrects some of the limitations of the Voronoi tessellation and the Taki & Hasegawa model. First, it produces more regular region boundaries than the latter model. Still, it squeezed the zone of control behind the uppermost white player to nearly empty while the closest black player is much further from that area. Moreover, the model produces boundaries that are straight lines, analogous to the Voronoi tessellation. In the case of the white player moving toward the bottom left corner of the pitch we would expect an expanding zone of control for this player given that he is moving quickly. This is not the case for the Fujimura & Sugihara model, which may be considered a limitation of this approach.

The bottom right part of Figure 3.4 depicts zones of control implied the probabilistic movement model. First, we observe that the zone of control for the white right winger discussed is realistically large as this player is running into the open space in the opponent team’s half. Second, it produces an intuitive shape of the zone of control for the uppermost white player. Finally, we observe an extending zone of control of the white player moving toward the bottom left corner of the field. Based on visual inspection of the zones of control implied by different models, we conclude that the proposed probabilistic approach
yields realistic, more regular, elliptically shaped controlled regions that are in line with the player velocities and the distances between them. Any of the alternatives, grounding on simplistic physical models of movement, results in different artefacts of the implied zones of control. For a more thorough insight into the zones of control underlying different models we present several more example situations in Figures 3.5–3.8 at the end of this chapter.

As a remark, we note that the zones of control for the three baseline approaches are identical when no player is moving. This can be seen by setting $v = 0$ in Equations (3.1) and (3.2) for Taki & Hasegawa and Fujimura & Sugihara, respectively, which then reduces the resulting zones of control to the Voronoi tessellation. The time needed to reach an arbitrary position is now a strictly increasing function of the distance to that position. This implies that the resulting zones of control are identical to the Voronoi tessellation of the field.

3.3.3 Results for pass outcome prediction

To evaluate different models on a quantitative basis we consider the problem of pass outcome prediction once it has been made by a given player. Different movement models and their implied zones of control will be used to predict pass outcomes. The motivation for considering such kind of a problem is twofold. First, a model’s performance in the pass outcome prediction problem may serve as a metric for optimising the model’s parameters, e.g., the bandwidth of the KDE in our model. Second, such a metric may be also used to evaluate the quality of zones of control induced by different movement models and, in turn, quantitatively compare the underlying movement models.

In previous research, the pass outcome prediction problem was studied by, e.g., Nakanishi et al. (2009) who considered this task for RoboCup Small Size League tournament. The authors found that the zones of control, based on the Taki & Hasegawa model, can effectively predict the outcome of a pass with the accuracy of 95%. However, the study covered a rather small sample of 65 passes to draw far-reaching conclusions based on it.

We detail now how the outcome of a pass is determined using a movement model below. First, a pass is described by a sequence of timestamps, $T = (t_1, t_2, \ldots, t_m)$, and the corresponding ball trajectory $\mathbf{p}_{\text{ball}}^{t_i} = (x_{t_i}, y_{t_i}), t_i \in T$. Moreover, at its starting and end time, $t_1$ and $t_m$, respectively, a flag for the ball possession is provided. The pass may also go outside playing area, but such cases are discarded. If the pass is blocked, we assume that the team which blocked the pass possesses the ball. The pass outcome is successfully predicted if the ball possession is correctly assigned by a given method. The prediction starts at timestamp $t_2$, which is the first moment in the pass trajectory following its making.
To predict pass outcome, the following procedure is employed. For simplicity, we describe the problem considering a physical movement model that yields the time needed to reach a given position. First, the positions of all players are fixed at time $t_1$ to perform the prediction. The analysis proceeds for consecutive timestamps until a stopping criterion is met. At a given timestamp $t_i$, the shortest time to reach the position of the ball, $p_{i}^{\text{ball}}$, is computed for every player. If for a player this time is shorter than $t_i - t_1$ then the player is assumed to control the ball. The ball is in the player’s zone of control restricted by time horizon $t_{\Delta_i} = t_i - t_1$. We refer to such an event as the interception below. Otherwise, if no player is able to reach the ball, the analysis proceeds to next timestamp $t_{i+1}$. If this timestamp is $t_m$, we reached the end of the pass trajectory. In this case, if the ball still is not assigned to any of the players, we assume that the ball is controlled by the player who is able to get the ball fastest. We refer to this case as the pass completion.

For a set of passes, we use the accuracy and the area under ROC curve (AUC) as evaluation metrics (Witten et al., 2011). In the case of probabilistic movement model proposed in this chapter, the time needed to reach the ball is not given but rather the likelihood of attaining it in a specified time horizon. Hence, there is a need for a decision rule that states whether a player is able to control the ball. More precisely, if for a player this likelihood at timestamp $t_i$ exceeds a certain threshold, say $\xi_i$, then the player is assumed to control the ball. For the decision rule proposed here this threshold is linearly dependent on time horizon $t_{\Delta_i}$ for every moment within the pass trajectory. Player $k$ is assumed to control the ball if the log-likelihood of this happening is greater than $\xi_i$:

$$\log P_{t_{\Delta_i}}^{(k)}(p_{i}^{\text{ball}} \mid p_{t_1}, p_{t_1} - t_1, v_{t_1}) \geq \xi_i = a \cdot t_{\Delta_i} + b.$$ \hspace{1cm} (3.8)

Note that we keep the positions of all the players fixed at the moment the pass is made; that is $t_1$. The ball position in the analysis is taken ex-post from trajectory data.

In our experiment we manually labelled 1194 passes from two selected matches (about 600 passes per match) by identifying their start and end timestamps along with possession flags. To optimise the parameters of the probabilistic model, we use 10-fold cross validation. More precisely, the model parameters are estimated using nine training folds and the tenth fold is used as a test set for which passes are classified using the model with its parameters determined on the training set. The procedure is repeated ten times to produce predictions for the entire dataset. These predictions are next evaluated using accuracy and AUC.

Parameters of the baselines are set to their default values from their original formulations (Fujimura and Sugihara, 2005; Taki et al., 1996). For our model we need to choose a kernel function and to set some parameters. Threshold $\xi_i$ in Equation (3.8) depends on $a$ and $b$, and for a kernel function $K_h$ bandwidth $h$ has to be set. In order to find the optimal parameter values we use the grid search in the intervals $[-0.3, 0]$ and $[-1.2, -0.5]$.
with a step size of 0.05 for \(a\) and \(b\), respectively. We also test several kernel functions: Epanechnikov, Gaussian, exponential, linear and uniform. As a search space for parameter \(h\) we define the interval \([0.5, 10]\) with a step of 0.5 for each of the kernel functions.

The results for all methods for the pass outcome prediction are presented in Table 3.2. In addition, 0.95-Clopper–Pearson confidence intervals for the observed success rates are provided (Clopper and Pearson, 1934). As for AUC, the confidence intervals were obtained for 10,000 bootstrap replications. For each kernel function, we report the corresponding bandwidth parameter value that provides the best results with respect to AUC and the 0.95-confidence intervals.

### Table 3.2. Pass classification results for different models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voronoi</td>
<td>0.509 (0.480, 0.538)</td>
<td>0.506 (0.477, 0.533)</td>
</tr>
<tr>
<td>Taki &amp; Hasegawa</td>
<td>0.832 (0.809, 0.852)</td>
<td>0.826 (0.796, 0.841)</td>
</tr>
<tr>
<td>Fujimura &amp; Sugihara</td>
<td><strong>0.915 (0.897, 0.93)</strong></td>
<td><strong>0.908 (0.891, 0.925)</strong></td>
</tr>
<tr>
<td>Probabilistic (Epanechnikov, (h = 4.0))</td>
<td><strong>0.916 (0.899, 0.931)</strong></td>
<td><strong>0.913 (0.891, 0.926)</strong></td>
</tr>
<tr>
<td>Probabilistic (Gaussian, (h = 0.5))</td>
<td><strong>0.915 (0.898, 0.931)</strong></td>
<td><strong>0.912 (0.891, 0.925)</strong></td>
</tr>
<tr>
<td>Probabilistic (exponential, (h = 0.5))</td>
<td><strong>0.914 (0.898, 0.931)</strong></td>
<td><strong>0.911 (0.888, 0.923)</strong></td>
</tr>
<tr>
<td>Probabilistic (linear, (h = 4.0))</td>
<td>0.913 (0.895, 0.928)</td>
<td>0.91 (0.887, 0.922)</td>
</tr>
<tr>
<td>Probabilistic (uniform, (h = 3.0))</td>
<td><strong>0.906 (0.888, 0.922)</strong></td>
<td><strong>0.907 (0.878, 0.914)</strong></td>
</tr>
</tbody>
</table>

We can see in Table 3.2 that the Voronoi model performs worst among all competitors. The Taki & Hasegawa model is the second best performing method while the two best methods are ex-aequo Fujimura & Sugihara and the proposed probabilistic model. As for different kernel functions, there is little difference in model performance. The best performing one is the Epanechnikov kernel with bandwidth parameter set to 4.

To provide context for the numbers, we report that the overall success rate of passes is 0.936. That is, when passing the ball, there is about 0.936 probability that it is successful, meaning that the ball is still possessed by the same team. In general, a player’s ability to win a one-on-one depends on many factors including his physical strength or ability to block or control the ball. The models considered do not take these factors into account. In turn, it appears that a passing player has relatively high accuracy in evaluating dominant regions as compared to each of the methods studied.

The results presented in Table 3.2 can be decomposed into two cases: interceptions and completions. In the case of an interception, a pass outcome is determined at the first timestamp \(t_i\) for which the stopping criterion is met. Pass completion can only be evaluated at the end timestamp \(t_m\) of the ball trajectory. In this case, the ball is assigned to
the player with the shortest time to reach it (or the highest likelihood of getting the ball). Table 3.3 presents this decomposition conditioned on the fact whether the ball is intercepted or not. In this analysis, we focus on the best performing probabilistic movement model obtained for the Epanechnikov kernel. We note that the fraction of intercepted passes varies strongly between models. For example, in the case of the Voronoi model, each pass is intercepted. This is because there always exists a timestamp in the ball trajectory when the ball is closer to some other player than to the pass maker.

Table 3.3. Model performance decomposition for interception and completion tasks.

<table>
<thead>
<tr>
<th>Model</th>
<th>Interception</th>
<th>Accuracy\text{_{interception}}</th>
<th>AUC\text{_{interception}}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voronoi</td>
<td>1.0</td>
<td>0.509 (0.48, 0.538)</td>
<td>0.506 (0.478, 0.533)</td>
</tr>
<tr>
<td>Taki &amp; Hasegawa</td>
<td>0.571</td>
<td>0.937 (0.916, 0.954)</td>
<td>0.932 (0.907, 0.949)</td>
</tr>
<tr>
<td>Fujimura &amp; Sugihara</td>
<td>0.8</td>
<td>0.945 (0.928, 0.958)</td>
<td>0.939 (0.923, 0.955)</td>
</tr>
<tr>
<td>Probabilistic</td>
<td>0.388</td>
<td>0.978 (0.961, 0.99)</td>
<td>0.978 (0.956, 0.989)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Completion</th>
<th>Accuracy\text{_{completion}}</th>
<th>AUC\text{_{completion}}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voronoi</td>
<td>0.0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Taki &amp; Hasegawa</td>
<td>0.429</td>
<td>0.691 (0.649, 0.731)</td>
<td>0.69 (0.645, 0.724)</td>
</tr>
<tr>
<td>Fujimura &amp; Sugihara</td>
<td>0.2</td>
<td>0.795 (0.738, 0.844)</td>
<td>0.793 (0.739, 0.843)</td>
</tr>
<tr>
<td>Probabilistic</td>
<td>0.612</td>
<td>0.877 (0.851, 0.9)</td>
<td>0.874 (0.843, 0.895)</td>
</tr>
</tbody>
</table>

The probabilistic model turns out to be more accurate in both tasks and outperforms its competitors significantly. The model is characterised by confident predictions in the two scenarios though its overall performance is on a par with the Fujimura & Sugihara model (Table 3.2).

3.4 Discussion

The movement models proposed along with the implied zones of control may be employed for a number of analyses. In particular, they can be used to devise different dimensions of player ratings. We sketch some ideas for quantifying player qualities in the next section. Next, we discuss some issues related to the proposed approach.

Using positional data for devising player attributes. When proposing the rating system based on player attributes in Section 2.3, we discussed a rich database of player ratings from the EA Sports FIFA video game. In particular, Table 2.5 lists the player qualities that are evaluated by experts in the game development process. We note that
these concepts are not exactly defined by the creators of the database. Certainly, there are some overlaps in their meaning. However, they can be understood intuitively. Below we consider some of the attributes are viable for the analysis using positional data.

First, we focus on the set of attributes pertaining to the movement group. These are:

- acceleration,
- sprint speed,
- agility,
- reactions,
- balance.

The acceleration or sprint speed can be computed directly from a player’s trajectory data. The players with higher values should be awarded a higher rating for these skills. Computing agility, reactions or balance may be slightly more involved. These concepts are definitely related with one another. Evaluating them in quantitative way may be done as follows. For example, for a set of position changes in a specified time window, we may compute the convex hull for player’s positions reached. The larger it is, the higher should be the rating of, for example, agility.

Another group of attributes related to the power:

- shot power,
- jumping,
- strength,
- long shots.

In this group, there are attributes like shot power or long shots. Provided that the event of a shot is labelled, shot power may be computed from the speed of the ball. Since the exact position of the player making the shot is given, evaluation of long shots taking and its efficacy is also viable. This may be done, for example, by studying the accuracy and effectiveness of shots made, say, outside the penalty area. As an additional prerequisite, this requires labels for different events in a match, in particular, shots on goal.

Rating features related to making and receiving passes, like short and long passing, positioning or vision may also be performed. Previous research demonstrated how a pass quality may be evaluated (Chawla et al., 2017; Horton et al., 2015). In particular, these analyses involved movement models for deriving features that were next used as input to a classifier for assessing pass quality. Along these lines are also works of Bransen (2018) and Bransen and Van Haaren (2018). The authors proposed methods for measuring the expected reward from a given pass as a contribution to goal scoring opportunity made by a team. In Bransen (2018) it was shown that the derived metrics exhibit high correlation with EA Sports FIFA video game player attributes related to passing. On the other hand, the defending team may intercept the pass or prevent an attacking player from getting it.
In turn, evaluating skills like interceptions or marking can also be performed using analogous tools. These methods can be used for computing data-driven player skills related to passes in an informed way.

The discussion so far focused on analysing particular dimensions of player skills. At a higher level, one wants to obtain an overall player rating. Decroos et al. (2017) proposed an approach to distribute the account for goals scored for the players which took part in a given sequence of actions. These actions include, for example, passes, dribbles, crosses, and shots. The approach proceeds by analysing similar sequences (including their spatial context) to evaluate how likely an action is to result in a goal being scored. The approach is more focused toward rating attack performance. While the primarily focus is to obtain overall player ratings, it may be employed well for rating player features related to these actions.

In general, an interesting approach toward deriving player ratings may ground on recent advancements in the algorithms for assessing the probability of a goal being scored. These methods take into account players’ and ball’s positions and estimate the likelihood of a shot or a goal being made. Such approaches may involve developing hand-crafted features as in (Lucey et al., 2014), (Link et al., 2016) or (Eastwood, 2017), or more generic approaches involving convolutional neural networks as presented in (Harmon et al., 2016) for basketball or in (Wagenaar et al., 2017) for football. For the former group of methods, the proposed movement model can be used for deriving extra features or revising the ones that ground on such models. In the latter approach, the models and the resulting zones of control may be used as additional input channels to a neural network. This is a method of enhancing the input representation which may lead to improved training convergence and a model’s performance (Eastwood, 2017).

Once an accurate model for evaluating goal scoring opportunities is given, it may serve as a basis for deriving player attributes. More precisely, by postprocessing the model output, it is possible to decompose the attribution of a player’s action for increasing or decreasing value of the action for a given team. Here, the action is understood as a single pass, a cross, or a move to open space or to attract defenders. The value of the action of a team may be viewed as the probability of a shot or a goal being made. Such an approach may well be used for defensive actions like interceptions, player marking or tackles. The greater the decrease of the action value for the opposing team, the more important a given defensive action is.

We shall also note that there are some limitations in using positional data for rating certain attributes. The reason is that the data are flat in the sense that they are recorded only in the 2D coordinate system. In particular, the records of the height at which the ball is, are not present in data. Because of this, for example, whether a player makes a standing
or a sliding tackle cannot be distinguished using the positional data alone (or perhaps with some uncertainty). The same argument applies to a pass or shot made by foot or a header. A possible remedy to this problem is to obtain the labels of particular events along with the positional data.

**Other issues with estimating movement models.** The empirical analysis in Section 3.3.2 reveals some shortcomings of the existing movement models. While the quantitative results for predicting pass outcomes yield similar results for the proposed approach and the Fujimura & Sugihara model, the qualitative comparison of the models exhibits the limitations of the models relying on oversimplified physics. The idea of this work is to develop a data-driven movement model that is based on the records of player trajectories observed during a match. In this way, the model avoids defining too simplistic physical model and is based on a player’s moves observed in the real-world.

However, using positional data for estimating movements of players all comes with certain limitations. Angle estimation from trajectory data via Equation (3.4), for instance, is based on the assumption a player always moves forward. In other words, the model assumes that the direction a player is facing is in line with his movement. This is not always the case as players often need to move so that they are facing the opponent or the ball and, in particular, sometimes move backwards. The model would thus provide a biased estimate of the time needed for turning around depending on the actual change of direction. A possible remedy would be to use a better approximation of angle $\theta$ rather than by Equation (3.4). This could be achieved by devising the angle from an auxiliary data source. Using positional data in the form presented and used here is, however, not sufficient to solve this matter.

Another issue of the proposed approach is that some players may seldom run with all their might. The goalkeeper is an example of such a player as it is evident from Table 3.1. Another issue is that for a new player no data is available to estimate his movement model. This is known as the cold-start problem, which often occurs in the domain of recommender systems (Son, 2016). To overcome this limitation, two separate models may be used: for an average player and the individual movement model. The idea is to blend the personalised component with the average component until the former is accurate enough to be used alone. More precisely, if we denote by $P_{nk}^{(k)}$ the movement model estimated using $n_k$ samples for a player $k$ and by $P$ the average model, then the model becomes

$$P^{(k)} = \alpha \cdot P_{nk}^{(k)} + (1 - \alpha) \cdot P,$$

with $\alpha = \min \left( \frac{n_k}{N}, 1 \right)$ and $N$ being parameter that informs how many samples are needed to obtain an accurate model. When $n_k = 0$, only the average component is used. With the growing number of samples, the role of individual component becomes increasingly
important and eventually when $N = n_k$ it is used alone. The required number of observations depends both on the domain and the player’s speed. For a field player, typically data from a single match suffices.

In this chapter, we propose a novel data-driven method for constructing individual movement models for players using positional data. It provides the estimate of probability of reaching a given position in a given time using a kernel density estimator. The model is compared both qualitatively and quantitatively to the previously proposed physical movement models. We also mentioned some limitations of the proposed approach. The proposed movement model and the underlying zones of control may serve as a basis for improving many existing analyses and methods for rating player qualities in a data-driven way. These in turn, at an aggregate level, may produce more accurate team ratings.
Figure 3.5. The probabilistic movement model identifies the possibility of a scoring chance for the white team. According to it, the white player in the midfield controls the area of the opponent’s half in front of the goal. Passing the ball there may create a one-on-one chance for the white team. While this is also considered an opportunity by the Taki & Hasegawa model, both the Voronoi and Fujimura & Sugihara approaches fail to detect this. The limitation of the analysis is that the possibility of blocking or committing an infamous tactical foul by the defenders against the white attacking player is not taken into account.
Figure 3.6. Implicit assumptions in baselines constrain possible shapes of zones. In the case of the Voronoi and Fujimura & Sugihara models they result in polygon-like shapes. For the Taki & Hasegawa approach, they are drop-shaped.
Figure 3.7. The Taki & Hasegawa model possibly overestimates the influence of high speeds. On the other hand, the probabilistic movement model implies that the leftmost white player (winger) in the opponent half controls a larger part of his team’s wing in comparison to other models. In this case, we consider it a less intuitive finding of the proposed approach.
Figure 3.8. The Taki & Hasegawa and Fujimura & Sugihara models imply zones of control that are sensitive to high speeds while the Voronoi approach completely ignores them.
In recent years there have been many changes as to how domestic football championships (leagues) are organized. Given a variety of league formats, some objective comparison between them as to their fairness is needed. In this chapter, we analyse the efficacy of different league formats with respect to their accuracy in producing a ranking of teams that conforms with their true, latent strengths. To this end, we use selected rating systems proposed in Chapter 2 as team strength indicators. Various performance metrics are estimated reflecting the agreement between latent teams’ strength parameters and their final rank in the league table. The tournament designs studied here are used in the majority of top-tier divisions in European countries. We compare between different league designs to select the most accurate ones.

The chapter is set out as follows. The next section provides background on different tournament designs in the UEFA\(^1\) countries. Section 4.2 details the team strength models used, their setup and the methodology proposed to evaluate different league formats. Finally, Section 4.3 we analyse the results and Section 4.4 discuss their practical side.

The results presented in this chapter are an extended version of the article published as a part of the Special Issue on Statistical Modelling for Sports Analytics in *Statistical Modelling* journal [Lasek and Gagolewski, 2018]. The code to reproduce the experiments is available online as a GitHub repository\(^2\).

### 4.1 Preliminaries

Revealing truthfully the latent abilities of competing agents is an important issue in many domains. For example, when choosing among job applicants, one wants to adapt a mechanism that helps us rank them according to their skills [Breaugh and Starkel, 2000].

---

\(^1\) *Union of European Football Associations*, the governing body for association football in Europe. The map of UEFA countries roughly corresponds to the geographical region of Europe with some exceptions (e.g., Israel and Kazakhstan).

\(^2\) [https://github.com/janekl/league-formats-efficacy](https://github.com/janekl/league-formats-efficacy)
information retrieval, search engines employ algorithms to select relevant items by ranking them (Langville and Meyer 2006, Li 2011). In the multi-armed bandit problem (Katherine and Veinott 1987) the goal is to maximise total payoffs from slot machines with unknown rewards distributions. In sports (including e-sports), the issue of designing an efficacious tournament in selecting the best participants among competing individuals or teams occurs naturally (Langville and Meyer 2012, Lasek et al. 2016, Stefani 1997). Thus, the efficacy of competition formats is among the most important criteria taken into account in tournament design.

In this application, we focus on league formats employed in the top-level association football divisions in countries belonging to UEFA. The main goal of this chapter is to study which tournament types rank teams according to their strengths best. To this end, we propose a detailed simulation methodology based on team strength (rating) models discussed in the previous chapter. The models are calibrated to real-world football data and analysed for a grid of parameter values.

It is important to emphasise the particular aspects of tournament design considered in this study. In general, the choice of tournament format is driven by a variety of factors (see, e.g., Goossens and Spieksma 2012, Szymanski 2003, Wright 2014), including, e.g., the number of teams taking part in the competition. On the one hand, the goal may be to produce accurate rankings with respect to the teams’ true abilities. In this way, designs that minimise the uncertainty of a tournament’s outcome are desired. We refer to such designs as being efficacious in the sense that they produce accurate team rankings. On the other hand, the uncertainty of a match’s outcome contributes to the fans’ excitement and the overall beauty of sport. What is more, the choice of a tournament format impacts multiple organisations involved in buying and selling the TV broadcast rights. Our study takes the perspective of tournament fairness based on its accuracy in ranking participants which is among the most important aspects of the tournament design problem.

4.1.1 Evaluation of different competition designs

Before presenting different league designs in greater detail, we shall discuss two particular tournament designs commonly applied in sports. They form a basis for an array of hybrid systems (e.g., McGarry and Schutz 1997). The first tournament structure is a k-round-robin (kRR) tournament, in which every team plays against each other k times. With n competing teams, such a tournament requires $k \cdot (n-1)$ rounds (assuming time-constrained scheduling) and $k \cdot \binom{n}{2}$ matches to be played. Single elimination (or knock-out, KO for short) and double elimination tournaments are also amongst the most popular tournament structures. In a knock-out tournament, the teams are paired off in successive rounds.
with a loss causing immediate elimination and the final match determines the winner. The double elimination tournament extends this structure by allowing first-time losers to be paired off with only the second loss resulting in elimination.

When investigating the efficacy of different tournament structures, a typical approach is to assume a theoretical model for participants’ strengths and generate results of pairwise comparisons (for example, match outcomes in sport) using this model. The results are aggregated according to the rules of a specific tournament and next the outcome is compared to the latent team strengths. Due to the complexity of the problem, the vast majority of the current studies address it by means of simulations.

Scarf et al. (2009) studied a range of tournament formats – KO, RR and multiple combinations thereof – and their ability to rank teams according to their latent strengths. For modelling teams’ strength, the authors used the basic Poisson model (Maher, 1982) described in Section 2.2.3. The conclusion was that 2RR is the most effective tournament design among the alternatives considered. Ryvkin (2010) considered a theoretical model of players’ strength, and also concluded that RR is a more efficacious tournament in comparison to KO and contests (a tournament in which each participant performs individually once and next all the participants are ranked according to their performance measured by a specified criterion). The efficacy of RR comes at a high cost because it requires relatively large number of matches to be played. The author also studied the dependency of the expected rank of a winner in relation to the number of participants, which turns out to exhibit non-monotonic dependency on the number of competing agents. McGarry and Schutz (1997), by considering various tournament designs involving eight teams, concluded that in general RR is the most efficacious format. However, enhanced versions of single and double elimination tournaments (by, e.g. seeding teams) are also competitive in terms of their accuracy. Notably, they may be preferred due to a smaller number of matches to be played. Mendonca and Raghavachari (2000) studied multiple RR tournaments and the methods of aggregating their results into a single ranking for all players. These rankings are then compared to latent teams abilities. The authors used two different team strength models. Different methods are found to perform better depending on the distribution of initial team strengths. The study provides guidelines for ranking participants based on many RR tournaments in which not all of them participate in each tournament.

In general, RR is considered to be the most efficient competition format that produces a ranking of teams that conforms with their latent strengths. This may justify its prevalence among different structures for domestic championships. Since RR-type tournaments require the number of matches played to be a quadratic function of the teams involved, they are considered to be costly. On the other hand, KO requires relatively few matches –
linear in the number of teams. However, due to this reason it produces less stable results with respect to the true team abilities. There is a trade-off between tournament efficacy and the number of required matches to complete it.

The studies discussed above analyse different combinations of $KO$ and $RR$ tournaments or team rankings based on the outcomes of multiple $RR$ tournaments. Lasek and Gagolewski (2018) was the first comprehensive study of real-world tournament formats that are commonly applied in domestic championships in European football. This is especially important as some countries recently employed quite non-standard scheduling schemes. In particular, Poland introduced a new format in the 2013/14 season: after an initial 2$RR$ tournament, the points are halved and there is an extra 1$RR$ tournament played in the top and bottom half of the league table separately. On the other hand, from the 2017/18 season the league format was maintained but halving points after the first stage was abandoned. Without a deeper investigation it is unclear what are the advantages (if any) of such schedule upgrades. Therefore, our main contribution is the analysis and comparison of different tournament designs employed for football leagues in the UEFA countries.

4.1.2 Overview of league formats

Most domestic championships in the UEFA countries operate as a $kRR$ tournament or one of its creative variations. In particular, a 2$RR$ tournament is prevalent among league designs with every pair of teams playing against one another twice – a home and an away game. However, there are a few noteworthy exceptions. For example, in the 2018/19 season the league formats in Finland or Hungary are 3$RR$, the leagues in Estonia or Switzerland are 4$RR$, and the domestic championship in Armenia is 6$RR$. Moreover, sometimes the competition runs in two stages. In an initial phase all teams compete against each other in a $kRR$ tournament. Next, the league table is split into two parts. This gives two sets of teams which compete further on in the championship and relegation groups. The competition lasts within each of the groups separately based on another $kRR$ tournament. We will refer to such designs as two-stage league formats. Such a league format has been applied in, among others, Belgium, Cyprus, Israel, Poland, Romania, Serbia and Ukraine. Notably, in certain leagues – including the Belgian, Romanian and Serbian – the points gained by the teams after the first stage are divided by two and halves are rounded up if necessary. We will refer to such structures as league formats with a points division.

All the league formats employed in UEFA countries in the 2018/19 season are detailed in Table 4.1. For brevity, $kRR_n$ denotes a $k$ round-robin tournament with $n$ teams involved. The “$+$” symbol denotes that a given entity employs a two-stage format and “$/”
Table 4.1. Overview of league formats in UEFA member countries in the 2018/19 season. The number of teams, the total number of rounds (possibly different in championship and relegation groups) and matches are denoted by $N$, $K$ and $M$, respectively.

<table>
<thead>
<tr>
<th>Country</th>
<th>Format</th>
<th>$N$</th>
<th>$K$</th>
<th>$M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albania</td>
<td>$4RR_{10}$</td>
<td>10</td>
<td>36</td>
<td>180</td>
</tr>
<tr>
<td>Andorra</td>
<td>$3RR_{8} + (2RR_{4}/2RR_{4})$</td>
<td>8</td>
<td>27</td>
<td>108</td>
</tr>
<tr>
<td>Armenia</td>
<td>$6RR_{6}$</td>
<td>6</td>
<td>30</td>
<td>90</td>
</tr>
<tr>
<td>Austria</td>
<td>$4RR_{10}$</td>
<td>10</td>
<td>36</td>
<td>180</td>
</tr>
<tr>
<td>Azerbaijan</td>
<td>$4RR_{8}$</td>
<td>8</td>
<td>28</td>
<td>112</td>
</tr>
<tr>
<td>Belarus</td>
<td>$2RR_{16}$</td>
<td>16</td>
<td>30</td>
<td>240</td>
</tr>
<tr>
<td>Belgium</td>
<td>$\frac{1}{2} \cdot 2RR_{16} + (2RR_{6}^{1-6}/2RR_{6}^{7-15}/2RR_{6}^{7-15})$</td>
<td>16</td>
<td>40</td>
<td>330</td>
</tr>
<tr>
<td>Bosnia and H.</td>
<td>$2RR_{12} + (2RR_{6}/2RR_{6})$</td>
<td>12</td>
<td>32</td>
<td>192</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>$2RR_{14} + (2RR_{6}^{1-6}/2RR_{4}^{7-14}/2RR_{4}^{7-14})$</td>
<td>14</td>
<td>36/32</td>
<td>236</td>
</tr>
<tr>
<td>Croatia</td>
<td>$4RR_{10}$</td>
<td>10</td>
<td>36</td>
<td>180</td>
</tr>
<tr>
<td>Cyprus</td>
<td>$2RR_{14} + (2RR_{6}^{1-6}/2RR_{6}^{7-12})$</td>
<td>14</td>
<td>36</td>
<td>242</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>$2RR_{16}$</td>
<td>16</td>
<td>30</td>
<td>240</td>
</tr>
<tr>
<td>Denmark</td>
<td>$2RR_{14} + (2RR_{6}^{1-6}/2RR_{4}^{7-14}/2RR_{4}^{7-14})$</td>
<td>14</td>
<td>36/32/32</td>
<td>236</td>
</tr>
<tr>
<td>England</td>
<td>$2RR_{20}$</td>
<td>20</td>
<td>38</td>
<td>380</td>
</tr>
<tr>
<td>Estonia</td>
<td>$4RR_{10}$</td>
<td>10</td>
<td>36</td>
<td>180</td>
</tr>
<tr>
<td>Faroe Islands</td>
<td>$3RR_{10}$</td>
<td>10</td>
<td>27</td>
<td>135</td>
</tr>
<tr>
<td>Finland</td>
<td>$3RR_{12}$</td>
<td>12</td>
<td>33</td>
<td>198</td>
</tr>
<tr>
<td>France</td>
<td>$2RR_{20}$</td>
<td>20</td>
<td>38</td>
<td>380</td>
</tr>
<tr>
<td>Macedonia</td>
<td>$4RR_{10}$</td>
<td>10</td>
<td>36</td>
<td>180</td>
</tr>
<tr>
<td>Georgia</td>
<td>$4RR_{10}$</td>
<td>10</td>
<td>36</td>
<td>180</td>
</tr>
<tr>
<td>Germany</td>
<td>$2RR_{18}$</td>
<td>18</td>
<td>34</td>
<td>306</td>
</tr>
<tr>
<td>Greece</td>
<td>$2RR_{16}$</td>
<td>16</td>
<td>30</td>
<td>240</td>
</tr>
<tr>
<td>Hungary</td>
<td>$3RR_{12}$</td>
<td>12</td>
<td>33</td>
<td>198</td>
</tr>
<tr>
<td>Iceland</td>
<td>$2RR_{12}$</td>
<td>12</td>
<td>22</td>
<td>132</td>
</tr>
<tr>
<td>Israel</td>
<td>$2RR_{14} + (2RR_{6}^{1-6}/2RR_{8}^{7-14})$</td>
<td>14</td>
<td>36/40</td>
<td>268</td>
</tr>
<tr>
<td>Italy</td>
<td>$2RR_{20}$</td>
<td>20</td>
<td>38</td>
<td>380</td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>$2RR_{12} + (2RR_{6}/2RR_{6})$</td>
<td>12</td>
<td>32</td>
<td>192</td>
</tr>
<tr>
<td>Latvia</td>
<td>$4RR_{8}$</td>
<td>8</td>
<td>28</td>
<td>112</td>
</tr>
<tr>
<td>Lithuania</td>
<td>$4RR_{8}$</td>
<td>8</td>
<td>28</td>
<td>112</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>$2RR_{14}$</td>
<td>14</td>
<td>26</td>
<td>182</td>
</tr>
<tr>
<td>Malta</td>
<td>$2RR_{14}$</td>
<td>14</td>
<td>26</td>
<td>182</td>
</tr>
</tbody>
</table>
(Table 4.1 continued from the previous page).

<table>
<thead>
<tr>
<th>Country</th>
<th>Format</th>
<th>N</th>
<th>K</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moldova</td>
<td>4RR_{10}</td>
<td>10</td>
<td>36</td>
<td>180</td>
</tr>
<tr>
<td>Montenegro</td>
<td>4RR_{10}</td>
<td>10</td>
<td>36</td>
<td>180</td>
</tr>
<tr>
<td>The Netherlands</td>
<td>2RR_{18}</td>
<td>18</td>
<td>34</td>
<td>306</td>
</tr>
<tr>
<td>Northern Ireland</td>
<td>3RR_{12} + (2RR_{6}/2RR_{6})</td>
<td>12</td>
<td>38</td>
<td>228</td>
</tr>
<tr>
<td>Norway</td>
<td>2RR_{16}</td>
<td>16</td>
<td>30</td>
<td>240</td>
</tr>
<tr>
<td>Poland</td>
<td>2RR_{16} + (RR_{8}/RR_{8})</td>
<td>16</td>
<td>37</td>
<td>296</td>
</tr>
<tr>
<td>Portugal</td>
<td>2RR_{18}</td>
<td>18</td>
<td>34</td>
<td>306</td>
</tr>
<tr>
<td>Republic of Ireland</td>
<td>4RR_{10}</td>
<td>10</td>
<td>36</td>
<td>180</td>
</tr>
<tr>
<td>Romania</td>
<td>$\frac{1}{2} \cdot 2RR_{14} + (2RR_{8}^{1-6}/2RR_{6}^{7-14})$</td>
<td>14</td>
<td>36/40</td>
<td>268</td>
</tr>
<tr>
<td>Russia</td>
<td>2RR_{16}</td>
<td>16</td>
<td>30</td>
<td>240</td>
</tr>
<tr>
<td>Scotland</td>
<td>3RR_{12} + (RRk_{6}/RR_{6})</td>
<td>12</td>
<td>38</td>
<td>228</td>
</tr>
<tr>
<td>Serbia</td>
<td>$\frac{1}{2} \cdot 2RR_{16} + (RR_{8}/RR_{8})$</td>
<td>16</td>
<td>37</td>
<td>296</td>
</tr>
<tr>
<td>Slovakia</td>
<td>2RR_{12} + (2RR_{6}/2RR_{6})</td>
<td>12</td>
<td>32</td>
<td>192</td>
</tr>
<tr>
<td>Slovakia</td>
<td>4RR_{10}</td>
<td>10</td>
<td>36</td>
<td>180</td>
</tr>
<tr>
<td>Spain</td>
<td>2RR_{20}</td>
<td>20</td>
<td>38</td>
<td>380</td>
</tr>
<tr>
<td>Sweden</td>
<td>2RR_{16}</td>
<td>16</td>
<td>30</td>
<td>240</td>
</tr>
<tr>
<td>Switzerland</td>
<td>4RR_{10}</td>
<td>10</td>
<td>36</td>
<td>180</td>
</tr>
<tr>
<td>Turkey</td>
<td>2RR_{18}</td>
<td>18</td>
<td>34</td>
<td>306</td>
</tr>
<tr>
<td>Ukraine</td>
<td>2RR_{12} + (2RR_{6}/2RR_{6})</td>
<td>12</td>
<td>32</td>
<td>192</td>
</tr>
<tr>
<td>Wales</td>
<td>2RR_{12} + (2RR_{6}/2RR_{6})</td>
<td>12</td>
<td>32</td>
<td>192</td>
</tr>
</tbody>
</table>

denotes round-robin tournaments played simultaneously in championships and relegation groups. If after the first stage the ranking table is not split equally, superscripts $i$ and $j$ in $kRR_{n}^{i-j}$ denotes that teams ranked at $i$th to $j$th position compete in a separate $kRR$ (with a total of $j - i + 1$ teams involved). Note that if this is the case, the number of rounds played by the teams may differ. Finally, prefix “$\frac{1}{2}$” denotes that the points after the first stage are divided by two and rounded if necessary.

For example, the Cypriot league format is denoted with $2RR_{14} + (2RR_{6}^{1-6}/2RR_{6}^{7-12})$. It means that in the first stage there are 14 teams competing in a double round-robin tournament. Next, the league table is split into two parts: the first consist of the teams ranked from 1 to 6 and the second one from 7 to 12 (the last two teams are directly relegated). In both subgroups, another $2RR$ tournament is played (with six teams each). Hence, there are 36 rounds and $2 \cdot \left( \binom{14}{2} \right) + 2 \cdot 2 \cdot \left( \binom{6}{2} \right) = 242$ matches are played in total.

There are two countries missing in the comparison. In Liechtenstein, no domestic
championship is played (some teams compete in Switzerland), only a domestic cup takes place. In San Marino, there are 15 teams and league starts with two groups (7 and 8 teams each) in which $3RR$ is played. Following that, a double elimination tournament is played to determine the champion.

Perhaps one of the most complicated league format is employed in Belgium. After the first stage the teams ranked 1–6 compete in the championship group. The teams ranked 7–15 are coupled with teams from the second division to compete for one extra spot in international cups. The lowest ranked team is directly relegated. In Bulgaria and Denmark, with some modifications, the league table is also split into three groups after the first stage of the competition.

In certain countries – including Belgium, Bulgaria, Denmark, Iceland and the Netherlands – after a regular season, an extra round of play (typically a single elimination tournament) is employed in order to determine the league champion or the teams participating in the international cups competition in the season to follow.

**Points division.** We now look at the leagues which operate in a two-stage manner with the points division. Most importantly, given a standard way of allocating three points for a win, one point for a draw and no points for a loss, the division of points changes payoffs for winning and drawing to be effectively 1.5 and 0.5, respectively. As a result, one may expect it to impact a team’s attitude and motivation in the initial part of the competition compared to its final stage when the wins are worth more points. However, it should be emphasised that possible changes in a team’s attitude may be also partially attributed to the final stage of the competition when many high-stake (and many irrelevant too) matches take place. During this part of the season there is no room for mistakes and the matches are played under a higher pressure. This is something to bear in mind when considering any possible influence of the differences in the number of points awarded for a match.

As far as the number of points for a particular result is concerned, prior to 1995, two points for a win and one point for a draw were officially awarded. This was then changed to allocating three points for winning a match. The change was introduced by FIFA to promote a more attacking style of play. The impact on the teams’ attitude and tactic after introducing the extra point for a win has been studied in the literature. For example, Sumpter (2016), Moschini (2010) and Dilger and Geyer (2009) concluded that both the fraction of draws decreased and the number of goals increased under the three-points-for-a-win system. Sumpter (2016) also analysed the change introduced in a game theoretic setting and concluded that more attacking style of play pays off with the new rule. There is some evidence that introducing the new point allocation system changed
the game process itself, however, the conclusions are mixed (Hon and Parinduri 2016). In case of league formats, dividing points by two may result in analogous effects.

To investigate the influence of points division, we decided to compute the average number of goals scored in a match and the fraction of draws for the Polish league before and after a two-stage league format was introduced. Namely, from the 2013/14 season the league operates as a two-stage tournament which replaced the standard $2RR$ structure. For four seasons in which the new format is in force – from 2013/14 to 2016/17 – we found that the average number of goals scored in a match during the first part of the season equals 2.668 and the fraction of draws equals 0.284. This compares to 2.328 and 0.264 for four seasons in the past – from 2009/10 to 2012/13 – in which the Polish league operated as a $2RR$ tournament. Thus, the difference in the average number of goals scored increased significantly (based on the Mann–Whitney–Wilcoxon test, $p$-value < 0.0001). Moreover, the number of draws slightly increased (but not significantly; test for equality of proportions, $p$-value = 0.331). We note that the effect observed for the number of goals scored is the opposite to what some of the authors observed after the change FIFA introduced. However, this may well be attributed to the overall increase in the level of play.

Another feature of a two-stage league format is that it has two major breakpoints – prior to the table split and at the end of the competition. Undoubtedly, the interest of supporters skyrockets during these parts of the season. For example, Figure 4.1 presents the match attendance for Polish league matches, averaged over the four seasons discussed in which it operates in a two-stage manner with the first stage finishing after 30 rounds of play. While the match attendance hits an all-season high at the end of the league, it is also significantly higher at the first stage ending (see also Pawlowski and Nalbantis 2015 for a study of other factors influencing match attendance).

![Figure 4.1. Average attendance in the Polish league over the 2013/14–2017/18 seasons.](http://www.90minut.pl)

$^3$Data source: [http://www.90minut.pl](http://www.90minut.pl), last access date 29 January 2019.
Yet another important factor in designing domestic leagues is the equality in the number of matches that teams play at home and away against one another. This is a relevant aspect, since home ground gives an edge over the visiting team as discussed in Chapter 2. The majority of league formats conform to this rule. However, for example in the case of a 3RR tournament, this requirement cannot be satisfied as the number of games each team plays against one another is odd. Schedules balanced in this sense are another feature of a league design.

4.2 Experiment setup

This section presents the components of the simulation experiment for the evaluation of league designs. Namely, team strength models, the chosen league formats, team strength definition and evaluation metrics used to compare different league designs are discussed.

We begin with the discussion of team rating systems used.

4.2.1 Team strength modelling

In order to run the simulations, a method for generating league results is needed. A model for sampling individual match results is its key building block. We shall focus on rating-based models discussed in Chapter 2, i.e., the frameworks in which a team is described by a single or a pair of parameters indicating its strength. Such models allow us to specify the true ranking of teams based on their latent strength parameters as they are explicitly given. More specifically, two basic team strength models are used in this application: the rating model based on the ordinal logistic regression (Section 2.2.2) and the Poisson regression-based approach (Section 2.2.4). The two models are referred to as OLR and PR_{(c,h)} for short, respectively. In principle, we could use the optimal parameters estimated in the previous chapter to run simulations here. However, to keep the exposition self-contained and to provide an additional evaluation of the methods, we recompute the models below.

Actual parameter setting. We now discuss how parameters (c, h, λ) are set in the simulation. The team abilities are estimated as well but their application in the simulation is discussed later. The choice of the regularisation parameter λ is driven by minimising prediction error for future game outcomes. To determine this parameter, we employ the sliding window procedure detailed in Section 2.1.3. To recall, starting from round k – accounting for approximately 40% of all the matches in a given season – the model is estimated and the predictions are generated for round k + 1. Next, the model is estimated again using first k + 1 rounds and the predictions are generated for round k + 2
until the predictions are generated for all rounds from the validation sample (accounting for about 60% of all matches). The predictions are evaluated using logloss and accuracy. The models are estimated for three different leagues – the German, Polish and Scottish – for the 2015/16 season.

Table 4.2 presents the parameters computed. Each entry in the table presents an optimal parameter triple \((c, h, \lambda)\). The regularisation parameter was found using the grid search. For \(PR^{\rho}_{(a,d)}\), the correlation is set to \(\rho = 0.45\), which roughly equals to the correlation observed in historical data (see Section 2.2.5). Finally, parameters \((c, h)\) reported are the estimates obtained on the sample of all matches in a given season.

<table>
<thead>
<tr>
<th>Model</th>
<th>Germany</th>
<th>Poland</th>
<th>Scotland</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLR_0</td>
<td>(0.591, 0.535, 3.0)</td>
<td>(0.555, 0.322, 11.5)</td>
<td>(0.601, 0.177, 8.0)</td>
</tr>
<tr>
<td>(PR^{\rho}_{(a,d)}) ((\rho = 0.45))</td>
<td>(0.085, 0.371, 14.5)</td>
<td>(0.063, 0.37, 27.5)</td>
<td>(0.124, 0.196, 13.0)</td>
</tr>
</tbody>
</table>

To set the parameters \((c, h)\) in the simulation, we averaged their values across the three leagues. This yields \((0.582, 0.345)\) for OLR_0 and \((0.091, 0.312)\) for \(PR^{\rho}_{(a,d)}\). This means that for equally rated teams the prediction determined by Equations (2.11) and (2.14) yields \((0.441, 0.275, 0.284)\) and \((0.464, 0.258, 0.277)\) for the two models, respectively. These results can be compared with the empirical result frequency (Table 2.1). As expected, we observe that a prediction for equally rated teams roughly corresponds to the overall frequency of particular results.

**Model diagnostic.** We perform a model diagnostic by investigating its predictive power. Given the optimised parameter value \(\lambda\) for the 2015/16 season, the predictions are generated in a sliding window manner as described above using a different sample of matches from the next 2016/17 season. Thus, the number of matches used for evaluation accounts for approximately 60% of all matches for these leagues in the 2016/17 season. The models’ performance is presented in Table 4.3 for both logloss and accuracy. For reference, the results are compared to the benchmark forecasts derived from betting odds (Section 2.1.3). In terms of logloss, the \(PR^{\rho}_{(a,d)}\) model achieves better results than the OLR_0 model, which was also found in Section 2.4. In the case of the German league, the prediction results are close to that of the betting odds. Overall, the models produce relatively accurate predictions bearing in mind their simplicity.

---

4Data source: [http://www.football-data.co.uk](http://www.football-data.co.uk)

5At the time of computing these results, the season 2017/18 was yet to end.
Table 4.3. Model performance for the 2016/2017 season.

<table>
<thead>
<tr>
<th>Model</th>
<th>Germany Logloss</th>
<th>Germany Acc.</th>
<th>Poland Logloss</th>
<th>Poland Acc.</th>
<th>Scotland Logloss</th>
<th>Scotland Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLR₀</td>
<td>1.0223</td>
<td>0.5056</td>
<td>1.0396</td>
<td>0.4716</td>
<td>0.9739</td>
<td>0.5147</td>
</tr>
<tr>
<td>PR⁺(a,d) (ρ = 0.45)</td>
<td>1.0104</td>
<td>0.4944</td>
<td>1.0227</td>
<td>0.5170</td>
<td>0.9440</td>
<td>0.5588</td>
</tr>
<tr>
<td>Bookmaker odds</td>
<td>1.0004</td>
<td>0.5333</td>
<td>0.9800</td>
<td>0.5341</td>
<td>0.9121</td>
<td>0.5588</td>
</tr>
</tbody>
</table>

**Team strength distribution.** To run league simulations, we need to start with team ratings that allows to sample match results. To assign teams’ strengths parameters we use Bayesian interpretation of the regularisation component in the penalised log-likelihood functions in Equations (2.12) and (2.15). For the OLR₀ model, exponentiating the equations and rewriting the penalty term we obtain \( \exp(-\|\mathbf{r}\|_2^2/2\sigma^2) \) with \( \sigma^2 = \lambda^{-1} \). For the PR⁺(a,d) model, we have \( \sigma^2 = (\lambda(1 - \rho^2))^{-1} \) as shown in Equation (2.16). We use these relations and sample initial team ratings independently for each team from the normal distribution, \( r_i \sim \mathcal{N}(0, \sigma^2) \) and from a bivariate normal distribution with mean zero and correlation structure as discussed in Section 2.2.4 for the two models, respectively. Using values reported in Table 4.2 we find that – for the German, Polish and Scottish league, respectively – in the case of OLR₀, optimal values for parameter \( \sigma \) are 0.577, 0.295, and 0.354 while for PR⁺(a,d) they are equal to 0.294, 0.214, and 0.311. In the simulations, parameter \( \sigma \) is varied on a grid of points 0.1, 0.2, 0.3, \ldots, 0.7 and 0.1, 0.15, 0.2, \ldots, 0.4 for the two models, respectively. This extends the range of parameter values obtained when tuning the model.

We may also refer to parameter \( \lambda \) as an alternative measure of competitive balance of teams in a league [Koning, 2000]. The higher the value of the parameter, the tighter competition within a league is (and, in turn, the results become less predictable). On the other hand, lower values of \( \lambda \) push toward higher discrepancy in teams’ strengths. In our analysis, it turns out that the optimal regularisation parameter \( \lambda \) is higher for the Polish league than for the other two leagues studied. This may indicate that the competition is more balanced in the Polish league.

**Towards a dynamic model.** The model discussed in the previous paragraph is static in the sense that a team’s shape does not change throughout the season at all. In this section, we extend the models by introducing rating perturbations in consecutive league rounds. Such dynamic models are more realistic as they allow the team strength to vary during the season due to, for example, player injuries or form breakdown.
For conciseness, the discussion below focuses on the dynamic extension of the correlated Poisson model only. The extension of the logistic model is derived analogously with slightly different parameters. In the PR\textsubscript{\textit{a,d}} model, in consecutive rounds of play, for each team the attack and defence ratings are updated by adding to them a sample from a bivariate Gaussian distribution with mean zero, standard deviation \( \sigma_i \) (a team-specific drift parameter) and correlation \( \rho = 0.45 \). More precisely, in round \( k \), \( a_i^{(k)} = a_i^{(k-1)} + \epsilon_i^{(k-1)} = a_i^{(1)} + \sum_{j=1}^{k-1} \epsilon_i^{(j)} \), where the rating in the first round \( a_i^{(1)} \) is set as discussed in the previous paragraph. The updates \( \epsilon_i^{(j)} \) are sampled from a Gaussian distribution \( \mathcal{N}(0, \sigma_i^2) \) (analogously for the defence rating with a correlated update). We assume that the rating updates are independent over the rounds and teams. This means that a random walk model is assumed for the team strength evolution throughout the season (see, e.g., Glickman 2001, Rue and Salvesen 2000). The question is how to set parameter \( \sigma_i \). First, we focus on the overall season drift \( \bar{\sigma}_i \) for the ratings. It is sampled from the inverse Gamma distribution \( \Gamma^{-1}(\alpha, 1) \) with the density function

\[
  g(x|\alpha, 1) = \frac{x^{-\alpha-1}}{\Gamma(\alpha)} \cdot \exp \left( -\frac{1}{x} \right) \cdot 1(x > 0).
\]

Based on the properties of this distribution, the higher the shape parameter \( \alpha \), the lower the variation of team strength. In the special case \( \alpha = \infty \), we arrive at the static model in which a team’s shape remains constant throughout the season.

Once the overall season drift \( \bar{\sigma}_i \) is chosen, it should be distributed over the rounds of play to obtain \( \sigma_i \). We consider two cases. First, we employ constant drift across all league formats. That is, for team \( i \), we assume that its strength changes in every round by \( \sigma_i = \frac{1}{\sqrt{K_{\text{med}}}} \bar{\sigma}_i \), where \( K \) is the number of rounds played in a particular league format. Using \( a_i^{(K)} = a_i^{(1)} + \sum_{j=1}^{K-1} \epsilon_i^{(j)} \), it follows that the rating at the end of the season is normally distributed, \( a_i^{(K)}|a_i^{(1)} \sim \mathcal{N}(a_i^{(1)}, \bar{\sigma}_i^2) \), as a sum of independent Gaussian variables. As \( a_i^{(1)} \) is a normally distributed random variable itself (independent of all the updates) it also holds that \( a_i^{(K)} \sim \mathcal{N}(0, \sigma^2 + \bar{\sigma}_i^2) \). Second, to allow for a larger variation in team strength for the league formats involving a higher number of rounds, we assume that for each league format the drift rate is \( \sigma_i = \frac{1}{\sqrt{K_{\text{med}}}} \bar{\sigma}_i \), where \( K_{\text{med}} = 35 \) is the median length of the season measured by the number of rounds according to Table 4.6. Thus, \( a_i^{(K)}|a_i^{(1)} \sim \mathcal{N}(a_i^{(1)}, \frac{K-1}{K_{\text{med}}-1} \bar{\sigma}_i^2) \) and \( a_i^{(K)} \sim \mathcal{N}(0, \sigma^2 + \frac{K-1}{K_{\text{med}}-1} \bar{\sigma}_i^2) \). This way, the league formats with more rounds are going to produce higher variation in team strength in the end of the season.

A question that now arises is how to choose \( \alpha \) for sampling the teams’ overall drift distribution \( \bar{\sigma}_i \). The lower the value of this parameter, the higher is the variation of team strength at the end of a league in comparison to its prior rating. To set it, we look at Kendall’s \( \tau \) correlation coefficient (see Section 4.2.4) between the probability of becoming
champion before the start of the season derived from betting odds and the final league rankings for three league seasons: 2013/14, 2014/15 and 2015/16. The correlation values are given in Table 4.4.

Table 4.4. Kendall’s \( \tau \) correlation between the probability of outright winner derived from betting odds and the final league position.

<table>
<thead>
<tr>
<th>Season</th>
<th>Germany</th>
<th>Poland</th>
<th>Scotland</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013/14</td>
<td>0.499</td>
<td>0.333</td>
<td>0.788</td>
</tr>
<tr>
<td>2014/15</td>
<td>0.569</td>
<td>0.700</td>
<td>0.364</td>
</tr>
<tr>
<td>2015/16</td>
<td>0.464</td>
<td>0.346</td>
<td>0.515</td>
</tr>
</tbody>
</table>

These values serve as a proxy for the change of the prior and end distribution of an overall team’s strength defined as the sum of its attack and defence ratings. The parameter \( \alpha \) is varied on the geometric scale: 10, 20, 50, 100, 200, 500. Moreover, the special case \( \alpha = \infty \) is considered. This produces the following correlations for a grid of parameter values presented in Table 4.5. It also includes prior team strength distribution \( \sigma \) discussed in the previous section. The table was produced by sampling 10,000 values for every parameter combination for the prior and final ratings for a given season.

Table 4.5. Kendall’s \( \tau \) correlation coefficient between the initial and the final team strength for different parameter settings \((\alpha, \sigma)\).

<table>
<thead>
<tr>
<th>( \alpha ) ( \backslash ) ( \sigma )</th>
<th>0.1</th>
<th>0.15</th>
<th>0.2</th>
<th>0.25</th>
<th>0.3</th>
<th>0.35</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.082</td>
<td>0.120</td>
<td>0.163</td>
<td>0.199</td>
<td>0.239</td>
<td>0.268</td>
<td>0.303</td>
</tr>
<tr>
<td>20</td>
<td>0.169</td>
<td>0.245</td>
<td>0.315</td>
<td>0.376</td>
<td><strong>0.431</strong></td>
<td>0.477</td>
<td>0.519</td>
</tr>
<tr>
<td>50</td>
<td>0.384</td>
<td>0.509</td>
<td>0.599</td>
<td>0.666</td>
<td>0.712</td>
<td>0.749</td>
<td>0.778</td>
</tr>
<tr>
<td>100</td>
<td>0.602</td>
<td>0.715</td>
<td>0.779</td>
<td>0.821</td>
<td>0.849</td>
<td>0.871</td>
<td>0.887</td>
</tr>
<tr>
<td>200</td>
<td>0.781</td>
<td>0.852</td>
<td>0.888</td>
<td>0.909</td>
<td>0.924</td>
<td>0.935</td>
<td>0.944</td>
</tr>
<tr>
<td>500</td>
<td>0.909</td>
<td>0.940</td>
<td>0.954</td>
<td>0.964</td>
<td>0.970</td>
<td>0.974</td>
<td>0.977</td>
</tr>
<tr>
<td>( \infty )</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Naturally, the correlation approaches one as \( \alpha \to \infty \). On the other hand, for lower values of this parameter we observe a small correlation between the initial and end team strength. However, based on the reported correlation of betting odds and the final league standings, this parameter is relevant in practice. For example, the estimated correlation of 0.519 for Poland and 0.364 for Scotland in 2014/15 suggests a moderate relationship between betting odds and final league position.

\(^6\text{Data obtained from } \text{https://www.sts.pl}, \text{https://www.efortuna.pl} \text{ and } \text{http://www.oddschecker.com} \text{ (last accessed 29 January 2019).}\)
0.431 for a parameter pair \((\alpha, \sigma) = (20, 0.3)\) is close to 0.499 which is the median empirical correlation given in Table 4.4. In this case, the prior team strength drift is roughly equal to 0.294 and 0.311 which are the optimal values of this parameter for German and Scottish leagues for the 2015/16 season, respectively (see the discussion above).

**Figure 4.2.** Difference in simulation for the correlated (left) and uncorrelated (right) attack and defence ratings for a single team throughout a 35-rounds season.

Finally, Figure 4.2 exposes the benefits of using correlated Poisson model over its basic version. In the consecutive rounds, the ratings are sampled from the correlated Gaussian distribution rather than the independent one. This allows to maintain rather than fade away the prior correlation between them.

### 4.2.2 League formats

Table 4.6 presents nine league designs chosen in our comparative study. In addition to \(kRR\) tournaments \((k = 1, 2, 3)\), it contains two-stage designs. To recall, \(2RR+(1RR/1RR)\) denotes a league design in which the first round comprises \(2RR\) after which the league table is split into two groups. The brackets indicate that there is an extra \(1RR\) played in each of the groups after the table split. Recall that prefix \(“1/2”\) denotes possible division of points by two after the first round. Additionally, each tournament format is described by the total number of rounds and matches it requires to be played. The league formats given in Table 4.6 are the most popular league designs overall. They account for about 70% of all league formats employed in the countries considered. The rows of the table are sorted according to the number of matches played in a given league design.

One of the parameters of a league format is the number of teams involved. Here we decided to study the league designs comprising either 12 or 16 teams – such settings cover almost half of the UEFA countries.

We move on to implementation details. First of all, the algorithm for generating a \(1RR\) tournament schedule given in (de Werra, 1981) was used. Moreover, for breaking possible ties in ranks in the end of the season, head-to-head match results between tied teams were
Table 4.6. League formats under investigation. Columns *rounds* and *matches* denote the total number of rounds and matches in a league with 12 or 16 teams, respectively.

<table>
<thead>
<tr>
<th>Format</th>
<th>Format short</th>
<th>Rounds 12</th>
<th>Matches 12</th>
<th>Rounds 16</th>
<th>Matches 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>$3RR + (1RR/1RR)$</td>
<td>$a_1$</td>
<td>38</td>
<td>52</td>
<td>228</td>
<td>416</td>
</tr>
<tr>
<td>$\frac{1}{2} \cdot 3RR + (1RR/1RR)$</td>
<td>$a_2$</td>
<td>38</td>
<td>52</td>
<td>228</td>
<td>416</td>
</tr>
<tr>
<td>$3RR$</td>
<td>$b$</td>
<td>33</td>
<td>45</td>
<td>198</td>
<td>360</td>
</tr>
<tr>
<td>$2RR + (2RR/2RR)$</td>
<td>$c_1$</td>
<td>32</td>
<td>44</td>
<td>192</td>
<td>352</td>
</tr>
<tr>
<td>$\frac{1}{2} \cdot 2RR + (2RR/2RR)$</td>
<td>$c_2$</td>
<td>32</td>
<td>44</td>
<td>192</td>
<td>352</td>
</tr>
<tr>
<td>$2RR + (1RR/1RR)$</td>
<td>$d_1$</td>
<td>27</td>
<td>37</td>
<td>162</td>
<td>296</td>
</tr>
<tr>
<td>$\frac{1}{2} \cdot 2RR + (1RR/1RR)$</td>
<td>$d_2$</td>
<td>27</td>
<td>37</td>
<td>162</td>
<td>296</td>
</tr>
<tr>
<td>$2RR$</td>
<td>$e$</td>
<td>22</td>
<td>30</td>
<td>132</td>
<td>240</td>
</tr>
<tr>
<td>$1RR$</td>
<td>$f$</td>
<td>11</td>
<td>15</td>
<td>66</td>
<td>120</td>
</tr>
</tbody>
</table>

used (considering only the win-draw-loss results, without referring to the exact number of goals scored). This is one of possible methods employed as a first choice rule for tie breaking, for example, in Montenegro, Poland (based on the first stage results), Romania (based on the second stage results), Slovakia or Spain. If the teams are still tied after considering mutual match results, ties are resolved randomly.

Finally, we note that in the case of two stage league formats and *RR* tournaments with odd number of rounds in the second stage, teams play an uneven number of home and away matches against one another. This is the case for $2RR + (1RR/1RR)$ and $3RR + (1RR/1RR)$ formats (and their variations by points division after the first stage) for the pairs of teams which compete against each other only during the second and the first stage, respectively. In case of the latter format, after three rounds of matches in the first stage, the fourth match in the second stage was set so that the teams play against each other two matches home and away in total. To obtain a match schedule for the second stage for $2RR + (1RR/1RR)$ format we follow the rules that have been applied in the Polish league since the 2013/14 season (i.e., when a two-stage league format was introduced). More precisely, Tables 4.7 and 4.8 details the schedules for the final round in this format which is employed in the championship and the relegation group in the case of 12 and 16 teams, respectively. The integer codes represent a team’s rank after the first stage of the tournament. The schedule given in 4.8 was originally applied in the Polish league in the 2013/14 season. The schedule given in Table 4.7 is its modification for 12 teams. Notably, according to these schedules, the top half teams after the initial stage of the competition play one more match at their home field than the bottom half. Moreover,
the match between the first and the second team after the first phase is played at the first team’s home field.

**Table 4.7.** Schedule of the second stage of $2RR + (1RR/1RR)$ for 12 teams.

<table>
<thead>
<tr>
<th>Round</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>6 - 1 2 - 5 3 - 4</td>
</tr>
<tr>
<td>24</td>
<td>1 - 3 6 - 2 5 - 4</td>
</tr>
<tr>
<td>25</td>
<td>1 - 5 2 - 4 3 - 6</td>
</tr>
<tr>
<td>26</td>
<td>4 - 1 2 - 3 5 - 6</td>
</tr>
<tr>
<td>27</td>
<td>1 - 2 3 - 5 4 - 6</td>
</tr>
</tbody>
</table>

**Table 4.8.** Schedule of the second stage of $2RR + (1RR/1RR)$ for 16 teams.

<table>
<thead>
<tr>
<th>Round</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>1 - 6 2 - 5 3 - 8 4 - 7</td>
</tr>
<tr>
<td>32</td>
<td>8 - 1 6 - 2 7 - 3 5 - 4</td>
</tr>
<tr>
<td>33</td>
<td>1 - 5 2 - 7 3 - 6 4 - 8</td>
</tr>
<tr>
<td>34</td>
<td>7 - 1 8 - 2 3 - 5 4 - 6</td>
</tr>
<tr>
<td>35</td>
<td>1 - 3 2 - 4 5 - 7 6 - 8</td>
</tr>
<tr>
<td>36</td>
<td>4 - 1 2 - 3 6 - 7 8 - 5</td>
</tr>
<tr>
<td>37</td>
<td>1 - 2 3 - 4 5 - 6 7 - 8</td>
</tr>
</tbody>
</table>

### 4.2.3 Definition of a team’s strength

With varying teams’ strength parameters there is a need for an aggregation method to produce the overall season strength $r_i$ in order to be able to compare it with the final league standings. We suggest that the team ratings are averaged throughout the season. For $\text{OLR}_0$ model this means that $r_i = \frac{1}{K} \sum_{k=1}^{K} r_i^{(k)}$, with $K$ being the number of rounds in a league. Analogously, for $\text{PR}^\rho_{(a,d)}$, $r_i = \frac{1}{K} \sum_{k=1}^{K} a_i^{(k)} + d_i^{(k)}$. That is, the overall team strength is taken to be the sum of its attack and defence capabilities. We also investigated an aggregation scheme based on the median team strength. However, we noted that the results obtained were virtually identical with the same qualitative conclusions applying to them.

### 4.2.4 Evaluation methods

To evaluate the results, the true team ranking needs to be compared with the one produced at the end of a tournament. We suggest to compare the rankings in three ways: based on
the Kendall’s τ correlation, Spearman’s Footrule distance and also the fraction of the best team wins. We describe these metrics in detail. For team \( i \), let \( u_i \) denote its rank based on the theoretical strength discussed above and \( v_i \) its final league standing. Since the team ratings are drawn from continuous distributions, ties in the ranks of latent team strength occur with probability zero. The ties in the final league standings are resolved according to the tie-breaking rule discussed in Section 4.2.2. Hence, no ties are possible in the lists of ranks \( u_i \) and \( v_i \).

Kendall’s τ is defined as a normalised difference between the number of concordant pairs and disconcordant pairs in both lists. A pair of teams \((i, j)\) is said to be concordant if \( u_i > u_j \) and \( v_i > v_j \) or \( u_i < u_j \) and \( v_i < v_j \). It is called disconcordant otherwise. Normalisation by \( \left( \frac{n}{2} \right) \) assures the metric values are in the interval \([-1, 1]\). Spearman’s Footrule distance is defined as \( \sum_{i=1}^{n} |u_i - v_i| \). Finally, in a single simulation, the best team wins means that both \( u_i = 1 \) and \( v_i = 1 \) for some team \( i \). We decide to include this metric for comparison due to its simplicity and direct interpretation.

The metrics presented have been popular tools for evaluation of tournament structures (Appleton, 1995; Langville and Meyer, 2012; Lee, 1997; Mendonca and Raghavachari, 2000; Scarf et al., 2009). In the following part, we simulate the tournament for a large number of times and compute average tournament metrics over all runs.

To provide some context for the metric values, it is informative to derive them in the case of a random outcome of the league (with respect to true teams’ strengths). In such a case, we see that Kendall’s correlation and the best team win fraction are 0 and \( \frac{1}{n} \) (0.083 or 0.063 in our application with \( n = 12 \) or \( n = 16 \)), respectively. In the case of Spearman’s Footrule distance, due to linearity of expectation, the calculation boils down to computing \( \mathbb{E}|U - V| \), where \( U \) and \( V \) are independently and identically distributed random variables on integer values \( 1, 2, \ldots, n \) (corresponding to a team’s true latent rank and its final league position). It follows that \( |U - V| \in \{0, 1, \ldots, n - 1\} \) and \( \mathbb{P}(|U - V| = k) = \frac{2}{n^2} (n-k) \) for \( k \geq 1 \). Hence, \( \mathbb{E}|U - V| = \frac{2}{n^2} \sum_{k=1}^{n-1} k \cdot (n-k) = \frac{(n-1)(n+1)}{3n} \). For \( n = 12 \) and \( n = 16 \), this yields 3.972 and 5.313, respectively.

### 4.3 Results

This section presents the simulation results under the various settings discussed above. To start with, we show the results for 12 teams and the total drift in team strength equal across all the league formats considered (see Section 4.2.1). That is, the variance in team strength at the end of a season is kept equal among all the formats.
4.3.1 Special case analysis

First, we present the results for the special case of parameter settings \((\alpha, \sigma) = (20, 0.3)\). This case turned out to be realistic based on our analysis in Section 4.2.1. Table 4.9 presents the results for 100,000 simulation runs. This many simulations appeared to produce sufficiently stable measurements of the average evaluation metric values (convergence was observed).

Table 4.9. Average tournament metrics for the parameter setting \((\alpha, \sigma) = (20, 0.3)\).

<table>
<thead>
<tr>
<th>Format</th>
<th>Number of matches</th>
<th>Kendall’s (\tau)</th>
<th>Spearman’s Footrule</th>
<th>The best team win</th>
</tr>
</thead>
<tbody>
<tr>
<td>(3RR + (1RR/1RR))</td>
<td>228</td>
<td>0.730</td>
<td>1.256</td>
<td>0.646</td>
</tr>
<tr>
<td>(\frac{1}{2} \cdot 3RR + (1RR/1RR))</td>
<td>228</td>
<td>0.721</td>
<td>1.290</td>
<td>0.631</td>
</tr>
<tr>
<td>(3RR)</td>
<td>198</td>
<td>0.714</td>
<td>1.322</td>
<td>0.621</td>
</tr>
<tr>
<td>(2RR + (2RR/2RR))</td>
<td>192</td>
<td>0.704</td>
<td>1.364</td>
<td>0.625</td>
</tr>
<tr>
<td>(\frac{1}{2} \cdot 2RR + (2RR/2RR))</td>
<td>192</td>
<td>0.696</td>
<td>1.397</td>
<td>0.607</td>
</tr>
<tr>
<td>(2RR + (1RR/1RR))</td>
<td>162</td>
<td>0.686</td>
<td>1.434</td>
<td>0.598</td>
</tr>
<tr>
<td>(\frac{1}{2} \cdot 2RR + (1RR/1RR))</td>
<td>162</td>
<td>0.680</td>
<td>1.462</td>
<td>0.584</td>
</tr>
<tr>
<td>(2RR)</td>
<td>132</td>
<td>0.662</td>
<td>1.536</td>
<td>0.563</td>
</tr>
<tr>
<td>(1RR)</td>
<td>66</td>
<td>0.558</td>
<td>1.951</td>
<td>0.455</td>
</tr>
</tbody>
</table>

For all the metrics the differences between the values reported in Table 4.9 are statistically significant (at the 0.05 level; see the discussion in Section 4.5 at the end of this chapter). First, we observe that \(3RR + (1RR/1RR)\) is the most efficacious format according to all three criteria. Notably, the division of points makes the results worse though not by a large margin. Interestingly, while the relative order of tournament agrees for Kendall’s \(\tau\) and Spearman’s Footrule distance, \(2RR + (2RR/2RR)\) design turned out to be superior in selecting the best team in a league to \(3RR\) structure. Moreover, the first two metrics produced the same ranking of tournaments as the total number of matches played in a league. All in all, the number of matches appears to be an important factor in determining tournament efficacy.

4.3.2 Overall analysis

To analyse different settings we aggregate their results by averaging standardised results for all 49 parameter settings \((\alpha, \sigma)\) from Table 4.5. Standardising was introduced to account for the fact that different parameter settings introduce a different scale for the metrics considered. The standardisation performed is the \(z\)-score. That is, each result
was transformed to \((x_i - \bar{x})/sd(x)\), where \(\bar{x}\) and sd are the empirical mean and standard deviation for the nine tournament formats considered. Table 4.10 presents the results. The exact values of different metrics for a subset of parameter settings are presented in Section 4.5.

### Table 4.10. Average \(z\)-scores of tournament metrics across all simulation settings.

<table>
<thead>
<tr>
<th>Format</th>
<th>Number of matches</th>
<th>Kendall’s (\tau)</th>
<th>Spearman’s Footrule</th>
<th>The best team win</th>
</tr>
</thead>
<tbody>
<tr>
<td>(3RR + (1RR/1RR))</td>
<td>228</td>
<td>0.932</td>
<td>-0.946</td>
<td>0.946</td>
</tr>
<tr>
<td>(\frac{1}{2} \cdot 3RR + (1RR/1RR))</td>
<td>228</td>
<td>0.778</td>
<td>-0.785</td>
<td>0.681</td>
</tr>
<tr>
<td>(3RR)</td>
<td>198</td>
<td>0.612</td>
<td>-0.618</td>
<td>0.514</td>
</tr>
<tr>
<td>(2RR + (2RR/2RR))</td>
<td>192</td>
<td>0.367</td>
<td>-0.363</td>
<td>0.581</td>
</tr>
<tr>
<td>(\frac{1}{2} \cdot 2RR + (2RR/2RR))</td>
<td>192</td>
<td>0.224</td>
<td>-0.215</td>
<td>0.298</td>
</tr>
<tr>
<td>(2RR + (1RR/1RR))</td>
<td>162</td>
<td>0.026</td>
<td>-0.019</td>
<td>0.069</td>
</tr>
<tr>
<td>(\frac{1}{2} \cdot 2RR + (1RR/1RR))</td>
<td>162</td>
<td>-0.087</td>
<td>0.097</td>
<td>-0.161</td>
</tr>
<tr>
<td>(2RR)</td>
<td>132</td>
<td>-0.456</td>
<td>0.466</td>
<td>-0.567</td>
</tr>
<tr>
<td>(1RR)</td>
<td>66</td>
<td>-2.395</td>
<td>2.384</td>
<td>-2.361</td>
</tr>
</tbody>
</table>

First, we note that the most efficacious format with respect to all three metrics considered is \(3RR + (1RR/1RR)\). The table again reveals that there is a high correlation between the metric values and the number of matches played in a particular format. Moreover, optional dividing of points by two after the first round of play produces inferior results as compared to awarding three points for a win in each match. However, the results are not considerably worse as can be seen from the relative ordering of different tournament formats.

We also observed that there is almost perfect agreement between Kendall’s \(\tau\) and Spearman’s Footrule in terms of the relative ordering of different tournament across various parameter settings. Therefore, we henceforth discuss the results on the basis of Kendall’s \(\tau\) statistic. Again, the fraction of the best team wins is higher in the case of \(2RR + (2RR/2RR)\) than for \(3RR\) as opposed to the other two criteria considered. This was also observed in the analysis of the special case above.

To investigate the influence of given parameters on the results, we compare their values against values of a given metric on a plot, see Figure 4.3. For clarity, the designs with the points division were omitted. Their performance is analogous to their versions without it. We observe that as the discrepancy of the prior strength distribution \(\sigma\) increases, the separation between the teams increases as measured by Kendall’s \(\tau\). The effect of decreasing the seasonal drift parameter \(\alpha\) is analogous but less pronounced. Usually,
the prior team strength is a major determinant of its final league rank (see the discussion in Section 4.2.1). Finally, different designs are superior with respect to all parameter settings as can be seen by that the lines do not cross.

### 4.3.3 Influence of particular factors

In the simulations, different parameter settings were used. We investigated different numbers of teams involved: 12 and 16. We note that similar qualitative conclusions applied in both cases. With respect to quantitative differences in different metrics, we observed that in the case of 16 teams, higher values for Kendall’s $\tau$ and the fraction of the best teams wins were observed by a narrow margin. We conclude that a larger number of teams increases efficacy in terms of the probability of the best team to win and the rank correlation. On the other hand, the formats for 16 teams produced higher Spearman’s Footrule distance values. A larger number of teams possibly introduces wider gaps between the true teams’ ranks and their final league position. We also note that the range of this metric depends on the number of teams involved, whereas the other two metrics are normalised to the intervals $[-1, 1]$ and $[0, 1]$. The analysis proceeds with 12 teams henceforth.

Finally, as far as the normalisation of the drift parameter is considered, the conclusions were similar. We did not observe changes in the overall performance of the leagues for low values of $\sigma$ and high variation of team strength $\alpha$. For every parameter setting, the relative ordering of tournament formats measured by both Kendall’s $\tau$ and Spearman’s Footrule distance was identical to the ranking presented in Table 4.10.
4.3.4 Influence of the number of matches

It is interesting to model the relation between the number of matches (rounds) played in a league and the tournament metrics. We perform such an analysis in the case of the RR structures and an example simulation setting \((\alpha, \sigma) = (0.3, 20)\). The different metrics considered are estimated using 10,000 simulation runs for \(kRR\) design for \(k = 1, 2, \ldots, 10\) and 12 teams. The results are presented in Figure 4.4 for the number of rounds played (matchdays) in \(kRR\) format equal to \(11k\). We analysed the relation between metrics using the linear regression model with the metric values as the dependent variable and the logarithm of the number of rounds as the explanatory variable. We found that this logarithmic model fits data very well \((R^2 \approx 0.99)\). The conclusion is that the impact of additional rounds exhibit diminishing improvements which may be approximated by a logarithmic dependency.

![Graphs showing the relation between metrics and number of rounds](image)

**Figure 4.4.** The three metrics considered (from the left): Kendall’s \(\tau\) correlation, Spearman’s Footrule distance and the fraction of the best team wins \((y\text{-axis})\) as a function of the number of rounds \((x\text{-axis, on the logarithmic scale})\) in \(kRR\) tournament, \(k = 1, 2, \ldots, 10\).

4.3.5 Enhancing the 3RR + (1RR/1RR) format

An interesting question that arises is whether the most efficacious league design \(3RR + (1RR/1RR)\) can be further improved. The basic modification of this format would be to introduce a different number of points awarded for a particular result. Since we are considering a league format in which the points for the results in a series of matches are summed, this may be investigated by changing only the number of points allocated for a win, setting the number of points allocated for a draw and a loss to one and zero, respectively – any other point allocation rule obeys such a representation. Figure 4.5 presents the values of different metrics for the modified \(3RR + (1RR/1RR)\) format by awarding 1.5, 2, 2.5, \ldots, 5 points for a win in a match.
We observe that the efficacy of the format can be improved by allocating two points for a win in terms of Kendall’s \( \tau \) and Spearman’s Footrule distance. The differences in average values of these metrics are small but significant for different allocations of points for winning a match. Awarding two points for a win was the official rule applied in most of the European leagues until it was replaced by three-points-for-a-win standard after FIFA introduced it. On the other hand, there are no significant differences in awarding two or more points for a win for the fraction of the best team wins. With respect to this metric, in the case of 1.5 points for a win, the results are significantly inferior.

It should be noted that the analysis comes with certain limitations. The number of points awarded for a particular outcome may influence a team’s attitude and style of playing. For example, if four points are awarded for a win, a team may impose a more attacking style of play in case of a draw in the end of the game as there is a relatively large pay-off for winning it as compared to a single point for a draw. The opposite effect may be observed if the number of points for a win is set to two. It should be noted that the data used in the analysis stem from three-points-for-a-win system, which may produce some bias when studying different points allocation rules. We leave a detailed analysis of such effects as a part of further work in this area.

### 4.4 Discussion

From the experiments (in particular, Table 4.10) we conclude that \( 3RR + (1RR/1RR) \) is the most efficacious league format when the agreement between the competitors ranking it produces and their latent abilities is considered.

The simulations revealed that Kendall’s \( \tau \) and Spearman’s Footrule distance yield
the same qualitative conclusion as far as comparison between different designs are concerned. The ordering of the tournament formats for these two metrics are identical. On the other hand, the fraction of the best team wins provides mixed, yet very interesting results. In particular, we observe that it ranks $2RR + (2RR/2RR)$ over $3RR$ in certain cases. The reason may be that the second stage of the competition allows for a more refined selection of the best team while the table split after the first two round-robin rounds appears to be premature as for determining the whole ranking of teams. Moreover, based on the analysis of the points allocated for a win for the most efficacious $3RR+(1RR/1RR)$ league design, we found that it can be further enhanced in terms of Kendall’s $\tau$ correlation and Spearman’s Footrule distance by awarding two or two and a half points for a win instead of the current official three-points-setting. However, allocating a different number of points did not result in significant improvement in the fraction of the best team wins.

One of the most important findings is that the performance of a given league format highly depends on the total number of matches played. In fact, there is a perfect agreement between the number of matches played and Kendall’s $\tau$. This conclusion is in line with the intuition as well as the principle of statistics that more samples lead to better estimates. Moreover, the modification of the two-stage designs by dividing points after the first stage of the competition does not introduce a significant amount of noise in the efficacy of a league format.

The influence of extra round-robin rounds on the accuracy of the results was also investigated. We identified that the improvement in the efficacy of $kRR$ design is logarithmic in the number of rounds (matches) played. This relation was found empirically by fitting a linear regression model. In terms of further research, it would be interesting to come up with an analytical approximation of this finding.

We also note that in the strongest European leagues which also involve the highest number of teams (for example, English, French, German, Italian and Spanish) the $2RR$ tournament design is employed. Since these leagues operate on larger number of teams (18 or 20), it appears that the implementation of more complex league formats would be impractical due to the large number of matches required to complete them. This may justify the prevalence of this particular league design among the strongest leagues in the UEFA countries. This is also a recommended design in the case of leagues involving more teams. Moreover, among the league formats studied here it is the only design in which each team plays against one another exactly the same number of matches home and away. This may be a desirable feature of a tournament. Playing equal number of matches home and away against each team is also the feature of the $2RR + (2RR/2RR)$ design. In this case the teams play against one another two or four matches depending on the group they compete in during the second stage of the season (championship or
relegation). We also note that some countries, for example, in Latin America, employ 1RR tournament (organised in season splits referred to as *Apertura* and *Clausura*). This design was found to be significantly less efficacious in the comparison and therefore should be avoided.

It should be noted that the conclusions are based on a particular match result model and, of course, may have certain limitations. In particular, many factors can influence the matches’ outcomes. For example, international cup matches, players’ injuries or transfers may impact a team’s form. However, by studying a variety of parameter settings for a team’s strength and its fluctuations, the effort was made to arrive at robust conclusions. Another limitation stems from the fact that in the case of two stage league formats and the points division, teams may exhibit a different attitude toward the first stage of the competition since in practice each game is worth half of the points. Moreover, the decisive matches in domestic competition are played at the season’s end. These factors may influence a team’s attitude toward a match in different parts of the season. The assumed model does not include such psychological factors. On the other hand, the following implicit conclusion might be drawn. For a team to benefit from the efficacy of a particular format, it needs to play with all its might to win each match regardless of the league stage in order to reach a final rank reflecting its true strength (in particular, to claim the championship title).

Design of tournaments has many aspects. Among others there is fan excitement, profit from distribution of television rights and the tournament efficacy as discussed in depth in this work. In particular, based on the conclusions obtained in a simulation study, we note that the recent changes in, for example, Danish, Polish, Ukrainian or Serbian leagues to the extended league designs should have a positive impact on the efficacy of competition in these countries. It should be also emphasised that halving points decreases tournament efficacy in producing accurate team rankings. The question of how many points to award for a win is also worth revisiting. We observed that tournament efficacy may be slightly improved in certain aspects by setting it to two as in the previously applied standard.

The next final chapter summarises the main results of the thesis and discusses some area for future work.

### 4.5 Supplement: Tournament metrics for selected parameter combinations

Tables 4.11 and 4.12 present the estimates of different tournament metrics for the league designs studied. Each entry in the table corresponds to a different pair of parameters \((\alpha, \sigma)\) given in the rows and columns, respectively. The table is organised in blocks. Each
block corresponds to nine tournament designs studied. They are presented in the same order as in Table 4.6 – the last column gives the short name for the particular league design scheme.

We suggest that the differences between different formats are compared based on the confidence intervals resulting from the normal approximation. More precisely, the $1 - \bar{\alpha}$ confidence interval for a sample of observations $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is

$$\left( \bar{x} - \frac{z_{1-\bar{\alpha}/2} \cdot \text{sd}(\mathbf{x})}{\sqrt{n}}, \bar{x} + \frac{z_{1-\bar{\alpha}/2} \cdot \text{sd}(\mathbf{x})}{\sqrt{n}} \right),$$

where $z$ denotes a given quantile of the standard normal distribution and $\bar{x}$ and $\text{sd}(\mathbf{x})$ are the sample mean and the sample standard deviation, respectively. The width of this interval is $\frac{2}{\sqrt{n}} z_{1-\bar{\alpha}/2} \cdot \text{sd}(\mathbf{x})$. Assuming $\bar{\alpha} = 0.05$, we suggest the three given metrics considered – Kendall’s $\tau$, Spearman’s Footrule distance and the fraction of the best team wins – should be considered with error margins of ca. $\pm 0.001$, $\pm 0.004$ and $\pm 0.003$, respectively. Accordingly, the metric values are rounded to the third decimal place. These margins should be taken into account when considering the significance of differences between the reported numbers.
Table 4.11. Averages of tournament metrics over 100,000 simulations for different parameter settings for the ordinal logistic regression model and 12 teams.

<table>
<thead>
<tr>
<th>$\alpha \setminus \sigma$</th>
<th>Kendall’s $\tau$</th>
<th>Spearman’s Footrule</th>
<th>The best team wins</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>0.631</td>
<td>0.699</td>
<td>0.753</td>
<td>1.647</td>
</tr>
<tr>
<td>0.630</td>
<td>0.696</td>
<td>0.748</td>
<td>1.651</td>
</tr>
<tr>
<td>0.616</td>
<td>0.684</td>
<td>0.738</td>
<td>1.706</td>
</tr>
<tr>
<td>0.582</td>
<td>0.658</td>
<td>0.718</td>
<td>1.851</td>
</tr>
<tr>
<td>0.583</td>
<td>0.656</td>
<td>0.714</td>
<td>1.847</td>
</tr>
<tr>
<td>0.576</td>
<td>0.649</td>
<td>0.709</td>
<td>1.869</td>
</tr>
<tr>
<td>0.577</td>
<td>0.647</td>
<td>0.705</td>
<td>1.865</td>
</tr>
<tr>
<td>0.556</td>
<td>0.627</td>
<td>0.687</td>
<td>1.945</td>
</tr>
<tr>
<td>0.445</td>
<td>0.517</td>
<td>0.584</td>
<td>2.377</td>
</tr>
<tr>
<td>0.394</td>
<td>0.595</td>
<td>0.703</td>
<td>2.580</td>
</tr>
<tr>
<td>0.389</td>
<td>0.587</td>
<td>0.695</td>
<td>2.601</td>
</tr>
<tr>
<td>0.377</td>
<td>0.576</td>
<td>0.686</td>
<td>2.645</td>
</tr>
<tr>
<td>0.358</td>
<td>0.560</td>
<td>0.674</td>
<td>2.719</td>
</tr>
<tr>
<td>0.353</td>
<td>0.552</td>
<td>0.666</td>
<td>2.735</td>
</tr>
<tr>
<td>0.341</td>
<td>0.539</td>
<td>0.655</td>
<td>2.778</td>
</tr>
<tr>
<td>0.337</td>
<td>0.532</td>
<td>0.647</td>
<td>2.793</td>
</tr>
<tr>
<td>0.319</td>
<td>0.512</td>
<td>0.629</td>
<td>2.856</td>
</tr>
<tr>
<td>0.235</td>
<td>0.404</td>
<td>0.522</td>
<td>3.163</td>
</tr>
<tr>
<td>0.379</td>
<td>0.592</td>
<td>0.702</td>
<td>2.634</td>
</tr>
<tr>
<td>0.373</td>
<td>0.584</td>
<td>0.694</td>
<td>2.656</td>
</tr>
<tr>
<td>0.362</td>
<td>0.573</td>
<td>0.685</td>
<td>2.698</td>
</tr>
<tr>
<td>0.346</td>
<td>0.557</td>
<td>0.673</td>
<td>2.762</td>
</tr>
<tr>
<td>0.340</td>
<td>0.549</td>
<td>0.666</td>
<td>2.782</td>
</tr>
<tr>
<td>0.329</td>
<td>0.537</td>
<td>0.655</td>
<td>2.822</td>
</tr>
<tr>
<td>0.324</td>
<td>0.529</td>
<td>0.647</td>
<td>2.840</td>
</tr>
<tr>
<td>0.307</td>
<td>0.509</td>
<td>0.628</td>
<td>2.903</td>
</tr>
<tr>
<td>0.226</td>
<td>0.401</td>
<td>0.521</td>
<td>3.196</td>
</tr>
<tr>
<td>0.379</td>
<td>0.592</td>
<td>0.702</td>
<td>2.636</td>
</tr>
<tr>
<td>0.373</td>
<td>0.583</td>
<td>0.694</td>
<td>2.660</td>
</tr>
<tr>
<td>0.362</td>
<td>0.573</td>
<td>0.685</td>
<td>2.700</td>
</tr>
<tr>
<td>0.345</td>
<td>0.556</td>
<td>0.674</td>
<td>2.765</td>
</tr>
<tr>
<td>0.340</td>
<td>0.549</td>
<td>0.666</td>
<td>2.785</td>
</tr>
<tr>
<td>0.329</td>
<td>0.537</td>
<td>0.654</td>
<td>2.824</td>
</tr>
<tr>
<td>0.323</td>
<td>0.529</td>
<td>0.647</td>
<td>2.843</td>
</tr>
<tr>
<td>0.306</td>
<td>0.509</td>
<td>0.628</td>
<td>2.904</td>
</tr>
<tr>
<td>0.225</td>
<td>0.401</td>
<td>0.521</td>
<td>3.198</td>
</tr>
</tbody>
</table>
Table 4.12. Averages of tournament metrics over 100,000 simulations for different parameter settings for the Poisson model and 12 teams.

<table>
<thead>
<tr>
<th>$\alpha \backslash \sigma$</th>
<th>Kendall’s $\tau$</th>
<th>Spearman’s Footrule</th>
<th>The best team wins</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.25</td>
<td>0.4</td>
</tr>
<tr>
<td>0.463</td>
<td>0.692</td>
<td>0.789</td>
<td></td>
</tr>
<tr>
<td>0.458</td>
<td>0.685</td>
<td>0.781</td>
<td></td>
</tr>
<tr>
<td>0.445</td>
<td>0.676</td>
<td>0.775</td>
<td></td>
</tr>
<tr>
<td>0.421</td>
<td>0.662</td>
<td>0.768</td>
<td></td>
</tr>
<tr>
<td>0.418</td>
<td>0.654</td>
<td>0.760</td>
<td></td>
</tr>
<tr>
<td>0.405</td>
<td>0.644</td>
<td>0.753</td>
<td></td>
</tr>
<tr>
<td>0.403</td>
<td>0.638</td>
<td>0.747</td>
<td></td>
</tr>
<tr>
<td>0.382</td>
<td>0.618</td>
<td>0.731</td>
<td></td>
</tr>
<tr>
<td>0.285</td>
<td>0.510</td>
<td>0.639</td>
<td></td>
</tr>
<tr>
<td>0.423</td>
<td>0.686</td>
<td>0.787</td>
<td></td>
</tr>
<tr>
<td>0.416</td>
<td>0.678</td>
<td>0.780</td>
<td></td>
</tr>
<tr>
<td>0.405</td>
<td>0.669</td>
<td>0.773</td>
<td></td>
</tr>
<tr>
<td>0.385</td>
<td>0.657</td>
<td>0.766</td>
<td></td>
</tr>
<tr>
<td>0.380</td>
<td>0.649</td>
<td>0.759</td>
<td></td>
</tr>
<tr>
<td>0.369</td>
<td>0.638</td>
<td>0.751</td>
<td></td>
</tr>
<tr>
<td>0.365</td>
<td>0.631</td>
<td>0.745</td>
<td></td>
</tr>
<tr>
<td>0.345</td>
<td>0.612</td>
<td>0.729</td>
<td></td>
</tr>
<tr>
<td>0.255</td>
<td>0.503</td>
<td>0.637</td>
<td></td>
</tr>
<tr>
<td>0.410</td>
<td>0.685</td>
<td>0.787</td>
<td></td>
</tr>
<tr>
<td>0.403</td>
<td>0.676</td>
<td>0.779</td>
<td></td>
</tr>
<tr>
<td>0.392</td>
<td>0.668</td>
<td>0.773</td>
<td></td>
</tr>
<tr>
<td>0.375</td>
<td>0.656</td>
<td>0.766</td>
<td></td>
</tr>
<tr>
<td>0.368</td>
<td>0.647</td>
<td>0.758</td>
<td></td>
</tr>
<tr>
<td>0.357</td>
<td>0.636</td>
<td>0.751</td>
<td></td>
</tr>
<tr>
<td>0.352</td>
<td>0.630</td>
<td>0.744</td>
<td></td>
</tr>
<tr>
<td>0.334</td>
<td>0.610</td>
<td>0.728</td>
<td></td>
</tr>
<tr>
<td>0.247</td>
<td>0.502</td>
<td>0.637</td>
<td></td>
</tr>
<tr>
<td>0.409</td>
<td>0.685</td>
<td>0.787</td>
<td></td>
</tr>
<tr>
<td>0.402</td>
<td>0.676</td>
<td>0.779</td>
<td></td>
</tr>
<tr>
<td>0.392</td>
<td>0.668</td>
<td>0.773</td>
<td></td>
</tr>
<tr>
<td>0.375</td>
<td>0.655</td>
<td>0.767</td>
<td></td>
</tr>
<tr>
<td>0.369</td>
<td>0.647</td>
<td>0.759</td>
<td></td>
</tr>
<tr>
<td>0.357</td>
<td>0.636</td>
<td>0.751</td>
<td></td>
</tr>
<tr>
<td>0.352</td>
<td>0.630</td>
<td>0.745</td>
<td></td>
</tr>
<tr>
<td>0.334</td>
<td>0.610</td>
<td>0.729</td>
<td></td>
</tr>
<tr>
<td>0.247</td>
<td>0.502</td>
<td>0.637</td>
<td></td>
</tr>
</tbody>
</table>
Summary and conclusions

The focus of this thesis is to propose new team rating models for association football. In Chapter 2, we approached the problem in a top-down way by proposing methods for inferring team ratings from match results as well as several extensions of well-founded models. We studied the Elo rating system and described it as a special case of the gradient descent algorithm. This inspired us to propose new iterative rating systems that enjoy intuitive properties similar to those offered by the Elo model. Further, we aggregated individual player ratings from a video game to obtain a team rating as an example of the bottom-up approach. It turns out that this data-driven model yields accurate ratings as confirmed by the quality of its predictions. This prompted a focus on the methods for designing accurate player ratings using positional data. In Chapter 3, we therefore demonstrated how such data can be used to build movement models tailored to individual football players. In turn, such a model lays the groundwork for developing ratings for different player qualities.

As for applications, in Chapter 4, the rating models were used to evaluate different league formats employed in the UEFA countries. The idea is to use a rating system as the true team strength model and compare how team strengths conform to the final league table standings. This is an important issue since many domestic league championships are undergoing transformations with respect to their design today.

5.1 Verification of the research goals and hypotheses

In the light of the results presented, we conclude that data-driven approaches for devising team ratings lead to models of increased prediction accuracy. We conclude that auxiliary sources of data in the form of player ratings, as presented in Section 2.3, may be used to build accurate team rating systems. Moreover, given the availability of positional data, the methods for assessing individual player qualities may be improved using them. As a basic building block for such methods we proposed a data-driven movement model
in Section 3.2. We demonstrated its utility for match situation analysis and benchmarked its accuracy for the pass outcome prediction problem. We argue that the proposed model constitutes more realistic zones of control based on it than the previous methods based on physical models for player movement.

In Section 2.4, based on a theoretical analysis of the Elo model, we devised several new rating systems built upon the ordinal logistic regression and Poisson regression models. The proposed models are not overly complicated and they are governed by transparent update rules. We also found that they produce more accurate predictions than the baseline Elo model, which additionally reinforces their practical value.

We also conclude that rating systems are useful tools that help to make an informed decision as to which league format produces the most accurate team ratings. In Chapter 4, in a simulation study driven by selected team rating systems, we computed several metrics that reflect the accuracy of different league designs in estimating the true, unobservable team strengths.

Finally, we proposed a couple of extensions of the models used in the literature and demonstrated their improved accuracy for match outcome prediction. First, in Section 2.2.4 we proposed a model that incorporates the correlation between attack and defence team strengths in the Poisson regression model for football scores. Second, in Section 2.4.4 we suggested an improvement in the heuristic for generating three-way outcome probabilities from a binary prediction obtained from, e.g., the Elo model.

5.2 Future research directions

To conclude the thesis, we highlight some research directions for future work in the area of sport analytics. First of all, we note that many of the results presented can be adjusted for other sports. Variations of the stochastic gradient descent algorithm (Ruder, 2016) may be applied to develop new rating models in different sports. This may lead to transparent, interpretable rating systems of high accuracy. In Section 2.4, we presented several such methods and also experimented with a more advanced version of the gradient descent algorithm that includes momentum. At the same time, this will also prompt some further research questions. For example, an important issue to address is how many matches a team needs to play in order for its strength (rating) to be accurately estimated. The unofficial claim that “30 games suffice” for the Elo model to provide accurate ratings has been questioned by Aldous (2017) in a simulation experiment (this work also addresses a number of other issues related to the Elo model).

Rating systems have important real-world applications. For example, in July 2018, FIFA introduced a new methodology for rating teams based on the Elo model (FIFA).
The question remains, however, to what extent this rating system is accurate. For example, the potential issue is how to weights for different tournament types (e.g., friendly match or world cup qualifier) should be set. Currently, this is often done arbitrarily on a rule-of-thumb basis. We also note that the new FIFA ranking methodology does not account for the home-team advantage. While our intuition may tell us that the official ranking system is based on a solid basis, there may remain some room left to optimise it. Finally, the models proposed in Chapter 2 can be adapted as alternatives to the Elo model in association football. At the club level, analogous questions may be asked about the rating methodology that is applied by UEFA for seeding teams in qualifying rounds for European cups.

In terms of tournament design, many countries are undergoing changes today. At the international level, from 2022 a new World Cup design will be introduced with 48 teams competing instead of the 32 that do so now. It would be revealing to evaluate the new competition format in analogy to the study presented in Chapter 4. We also refer to (Scarf, 2017) for an overview of current issues in this area.

We have referenced several data sources throughout the thesis. Yet statistics are ubiquitous in sports and come from a wide variety of sources. In particular, the previous research suggests that shots on target are a strong indicator of team performance (Link et al., 2016). This suggests that shots on goal may be a useful data source to build a rating system. As for constructing prediction models, few approaches have employed shots on goal or other match statistics for generating predictions. The work of Stenerud (2015) is a study that did. Obviously, the data generated during a match cannot be used directly as an input to a model as such information is not available at the prediction time. However, such data may still be useful for building a prediction model. Multi-task learning algorithms (Caruana, 1997) can be used here. In this way, not only the final match result but also shots on target or possession data can be used to train a model.

In recent years we observed an increased use of positional data in sport analytics (Bornn et al., 2018; Link, 2018). This trend is likely to sustain. In Chapter 3, in particular in Section 3.3, we discussed some examples of the recent research on using positional data to design accurate player ratings. In particular, rating individual player actions (Bransen and Van Haaren, 2018; Decroos et al., 2017), or evaluating scoring opportunities using positional data (Eastwood, 2017; Harmon et al., 2016; Link et al., 2016; Lucey et al., 2014; Wagenaar et al., 2017) are among the main research directions related to this topic. Fulfilling these research goals may bring football closer to realising the success of sport analytics in baseball – recognised as an independent research area referred to as sabermetrics – for which Moneyball (Lewis, 2003) is one of the most prominent examples.
Notation and abbreviations

\( \mathbf{x}, \mathbf{X} \) A vector (in the column notation) and a matrix

\( \|\mathbf{x}\|_p \) \( L_p \) norm of a vector \( \mathbf{x} \in \mathbb{R}^n \)

\( \langle \mathbf{x}, \mathbf{y} \rangle \) Inner product of two vectors \( \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \)

\( \mathbf{x}^\top, \mathbf{X}^\top \) Transpose of a vector \( \mathbf{x} \) or a matrix \( \mathbf{X} \)

\( 1(A) \) Indicator function – equals to one if a predicate \( A \) is satisfied and zero otherwise

\( \mathbb{P}(\cdot) \) Probability measure

\( \mathbb{E}[X] \) Expected value of a random variable \( X \)

\( \text{Elo}_b, \text{Elo}_g \) Two versions of the Elo model – the basic and goal-based, respectively – discussed in Section 2.2.1

\( \text{OLR}_0 \) Ordinal logistic regression rating model (Section 2.2.2)

\( \text{OLR}_1 \) Ordinal logistic regression model introduced in Section 2.3.4

\( \text{OLR}_2 \) Ordinal logistic regression model introduced in Section 2.3.6

\( \text{PR}_{(a,d)} \) Basic Poisson regression model (Section 2.2.3)

\( \text{PR}^\rho_{(a,d)} \) Correlated Poisson regression model (Section 2.2.4)

\( \text{PR}_0 \) Single parameter Poisson regression rating model (Section 2.2.6)

\( \text{PR}_1 \) Poisson regression model introduced in Section 2.3.5

\( \text{PR}_2 \) Poisson regression model introduced in Section 2.3.6

\( \text{AUC} \) Area under the receiver operating characteristic (ROC) curve

\( \text{RPS} \) Ranked probability score

\( \text{KDE} \) Kernel density estimate

\( \text{RR} \) Round robin tournament

\( \text{IFAB} \) International Football Association Board

\( \text{FIFA} \) Fédération Internationale de Football Association

\( \text{UEFA} \) Union of European Football Associations
Bibliography


Hubáček, O., Šourek, G., and Železný, F. (2018). Learning to predict soccer results from


distribution. R package version 0.2.0.


