Assessing the skill of football players using statistical methods

Łukasz Szczepański

Salford Business School,
University of Salford, Salford, UK

Submitted in Partial Fulfilment of the Requirements
of the Degree of Doctor of Philosophy,
January 2015

Contents

List of Figures

List of Tables

Acknowledgements

Many people have contributed directly or indirectly to my Ph.D. research. I want to express my deepest gratitude to them.

I would like to thank my supervisor Dr Ian McHale for challenging my ideas, detailed feedback on this thesis and the papers we have written together, as well as his continuous support, encouragement and guidance during my research.

This research would not have been possible without my employer Smartodds Ltd. who provided funding and flexible working hours throughout the whole period of my candidature. I want to thank them as a whole but in particular my colleagues from the quantitative analysis team for creating a stimulating working environment, in which I could grow as a researcher. Among them William Fletcher deserves a special mention for proof reading this thesis and making many useful suggestions, Harry Hill for transforming raw data files into an easily accessible format for me and Benoit Jottreau for pointing me towards several interesting articles on performance analysis in football.

I owe my gratitude to Opta for creating a dataset that is a dream of any football analyst and giving me a permission to use it for this research.

I am also indebted to Professor Rose Baker and Professor Philip Scarf for their suggestions and insightful comments on an earlier draft of this dissertation.

Finally, I want to thank my parents Emilia and Sławomir for their love and all kinds of support in every stage of my life leading to this point. Last but not the least, I thank my beloved wife Kinga for her boundless patience and unextinguished encouragement during the whole period of my candidature. Thanks for believing in me.

Declaration

This thesis is a presentation of my original work. I declare that no part of it has been taken from existing published or unpublished material without due acknowledgement and that all secondary material used therein has been fully referenced.

Abstract

Professional football is a business worth billions of pounds a year. Player recruitment is a key aspect of the business with expenditures directly related to it (in the form of transfer fees and wages) accounting for the majority of clubs' budgets. The purpose of this study is to propose methods to assist player evaluation based on statistical modelling that could be used to support recruitment decisions. In this thesis we argue that if such methods are to serve as the basis of player valuation, they need to have predictive utility, since it is players' future performance that clubs benefit from and thus should be paying for. We present examples of how simplistic approaches to quantifying a footballer's skill lack such predictive character.

The original contribution of this thesis is a framework for evaluating footballers' worth to a team in terms of their expected contribution to its results. The framework attempts to address one of the key difficulties in modelling the game of football, i.e. its free-flowing nature, by discretising it into a series of events. The evolution of the game from one event to another is described using a Markov chain model in which each game is described by a specific transition matrix with elements depending on the skills of the players involved in this game. Based on this matrix it is possible to calculate game outcome related metrics such as expected goals difference between the two teams at the end of the game. It enables us to establish a link between a specific skill of a given player and the game outcome. The skill estimates come from separate, location specific, models, e.g. the shooting skill for each player is estimated in a model of converting shots to goals given the shot location.

We demonstrate how recognising the involvement of random chance in individual performance, together with accounting for the environment in which the evaluated performance occurred, gives our statistical model a predictive advantage when compared to naive methods which simply extrapolate past performance. This predictive advantage is shown to be present when passing and shooting skills are evaluated in isolation, as well as when measures of passing and shooting skills are combined in the proposed comprehensive metric of player's expected contribution to the success of a team.

Chapter 1

Introduction

The aim of this thesis is to propose a framework for assessing the skill of football players that could be used as a basis for their valuation. In this introductory chapter we present the background of the problem and motivation for studying it. We argue that it is yet to be solved in a satisfactory way by referring to the academic literature and quoting professionals working in the football industry. We briefly describe the data available for this research and outline the methods we use to solve the problem.

1.1 Background

Football as a business

According to the Annual Review of Football Finance (Deloitte, 2013), in season 2011/12 the size of the European football market was approximately €19.4 billion (£15.7 billion). The top divisions of the five biggest markets: England, Germany, Spain, Italy and France accounted for almost 50% of this amount. Of these leagues only the English Premier League and the German Bundesliga managed to generate an operating profit (before accounting for player trading and the cost of financing) despite these impressive revenues. The revenue of €2.9 billion (£2.3 billion) made the Premier League the biggest competition in club football, but even in its case the operating profit of €121m (£98m) was equivalent to a margin of just 4.2%. One might wonder what the rest of the money was spent on.

By far the most significant expenditure of professional football clubs is wages. In season 2011/12 they consumed 65% of the revenue in the five big leagues. In the Premier League, for example, the total wage bill of all the clubs was €2 billion (£1.66 billion) with two thirds of the amount paid to players. What is more, the Premier League clubs

spent a further €0.74 billion (£0.6 billion) gross on player transfers contributing to the net loss of €303m (£245m) for the whole league.

Expert judgement and statistical methods

Given the dominant position of player wages and transfer fees in the budget of a football club, it is safe to say that recruitment is among the most important decisions a football club must make. Evaluating players for this purpose has traditionally been based on opinions of football experts known as scouts. The process was summarised by James Smith, a member of the coaching staff of one of the Premier League's clubs Everton FC, in his interview for the Financial Times (Kuper, 2013):

> "Watching players is a very subjective thing, an inexact science. There are all kinds of inputs: live player reports, extensive video analysis, speaking to people who have worked with them, and data is one of those layers. Data plays a role - not a massive role at the moment."

In recruitment and other fields

The evidence from different fields suggests that traditional methods of recruitment based on expert opinions do not always yield good results. Laszlo Bock, senior vice president of people operations at Google, told The New York Times (Bryant, 2013):

> "Years ago, we did a study to determine whether anyone at Google is particularly good at hiring. We looked at tens of thousands of interviews, and everyone who had done the interviews and what they scored the candidate, and how that person ultimately performed in their job. We found zero relationship."

In his book *Thinking, fast and slow* Daniel Kahneman, a Nobel Prize laureate in economics, recalls his personal experience as a psychologist in the Israeli army as an example of how a simple formula helped him improve an interview process supposed to assess soldiers' fitness for combat (Kahneman, 2011). He devotes two chapters of this book to comparing the performance of expert opinions and algorithms. In the chapter entitled *Intuitions vs formulas* he writes (Kahneman, 2011, p. 223):

> "The number of studies reporting comparisons of clinical and statistical predictions has increased to roughly two hundred, but the score in the contest between algorithms and humans has not changed. About 60% of the studies

> have shown significantly better accuracy for the algorithms. The other comparisons scored a draw in accuracy, but a tie is tantamount to a win for the statistical rules, which are normally much less expensive to use than expert judgement. No exception has been convincingly documented."

He gives a long list of fields where such comparisons have been made: from the longevity of cancer patients, to the future career satisfaction of workers, to the future prices of Bordeaux wine. As for the reasons why experts are inferior to algorithms, Kahneman states that experts "try to be clever, think outside the box and consider complex combinations of features in making their predictions" as well as the fact that "humans are incorrigibly inconsistent in making summary judgements of complex information. When asked to evaluate the same information twice, they frequently give different answers." (Kahneman, 2011, p. 224).

In the chapter *Expert intuition: when can we trust it?* Kahneman admits that the predictive performance of experts varies between fields. He names two conditions for an environment to promote validity of experts' judgements. One is "regularity" in the sense that the closed system of the game of chess is more "regular" than the world of political forecasting with many unknowns (and unknown unknowns) having a potential impact on the developments. The other condition is "an opportunity to learn these regularities through prolonged practice". He adds that the effect of the practice depends on the speed and quality of feedback, for example anaesthesiologists receive immediate feedback on their decisions in the form of their patient's reaction to the drugs they order. They are contrasted with radiologists who receive much less information about the effect of their diagnoses (especially about the false negatives) and are therefore in a worse position to develop intuitive skills.

In player evaluation in sports

Whether these favourable conditions apply to the environment of evaluating sport participants is questionable. Most importantly the quality of feedback is rather poor. A player brought into a club based on an expert's judgement of his ability has his performance assessed by an expert again. This compares unfavourably to, say, the aforementioned anaesthesiologist, whose decisions can be evaluated more objectively (the patient sleeps or not), and may lead to wrong opinions on whether the player is worth signing being reinforced by similarly wrong ones made once the player joins the club. This may be particularly true if the same person is involved in both assessments as he may be anchored to his original view. Such "set views and prejudices" are named by Carling et al.

(2006, p.11) among the factors negatively affecting an expert's recollection of a match, which is the basis for a later assessment. The other ones are: viewing environment (e.g from the dugout on the pitch level rather than from higher up in the stadium), limitations of human memory and effects of emotions.

Due to the aforementioned difficulties, recent years have witnessed a growth in interest surrounding the application of statistical methods to player evaluation. This theme has even found its way to the popular culture as it was depicted in a film entitled *Moneyball* (Miller, 2011) based on a book by the same title (Lewis, 2004b). They tell the story of Billy Bean, the General Manager of a Major League Baseball franchise Oakland Athletics, who employs a statistician to help him identify undervalued talent in the players market. Recently this approach has slowly begun to enter the world of professional football as admitted by Steven Houston, the Head Analyst at Hamburger SV of the German Bundesliga, in his interview for Sky Sports (Bate, 2012):

> "It's definitely something that a lot of clubs are now paying attention to. I think the Premier League remains very innovative. If you look at the conferences now most of them have full attendance from Premier League clubs and from my point of view that's really exciting to see that kind of movement compared to 2008."

1.2 Problem statement

The movement is still in its infancy though. The previously quoted James Smith of Everton FC admits that the tools clubs currently use are at the level of "GCSE maths as opposed to PhD maths" (Kuper, 2013). At the same time it is likely that football actually requires more sophisticated analytical tools than baseball. The latter is a stop-and-go sport largely comprising of a series of duels between a pitcher of one team and a batter of the opposite one, whereas football is a free-flowing game with constant interactions between 22 players of both teams.

According to Gavin Fleig, the Head of Performance Analysis at Manchester City Football Club, most clubs have no resources to advance the field themselves (Slaton, 2012):

> "The reality is most clubs have a performance analysis department, but the very demands of the day-to-day requirements around the team, if you are playing around a 40 to 58 game season, (mean that) it's really tough for clubs to spend real time developing and working analytics."

On the other hand, those clubs that have invested the resources are naturally reluctant to share their findings with the outside world as admitted by Mike Forde, Chelsea's former Performance Director (Kuper, 2011).

Questions from practitioners

Fortunately for independent researchers interested in this area, football clubs are at least willing to reveal the challenges they face, perhaps in an attempt to attract proposed solutions. Here is what some analysts working for football clubs have to say about the gap in the knowledge in the field of player evaluation.

- Steve Brown, First Team Performance Analyst at Everton (Chang, 2012):

  > "Looking at the basic entry-level data (number of headers won, tackles made, final third and penalty area entries, passes attempted and received) is becoming less relevant to us. We are striving to put the data in greater context: pitch location and efficiency of actions attempted."

- Gavin Fleig the Head of Performance Analysis at Manchester City Football Club (Slaton, 2012):

  > "We're looking into the patterns and the relationships between the players via the X and Y data. I think that's where a lot of the future lies. The X and Y data gives you a positional sense and contextualization around the pitch. I think only then will we be able to take player analysis and team analysis to the next level."

- Steven Houston Head Analyst at Hamburger SV (Bate, 2012):

  > "It's no longer a case of saying a player has scored X amount of goals or a midfielder has created X amount of assists. You only have to look at something simple like a goal. There are so many types of goals - the difficulty of the goal, the quality of the goal. And with passes there are passes and then passes in the final third. We are now able to break down into it. The hardest thing is to work out what is important and what isn't important - at a team level but also for individual players."

The above quotes could be paraphrased as:

When assessing players, the challenge currently faced by football clubs is to evaluate individual performance in the context of the circumstances where it occurs (e.g. the location on the pitch) whilst accounting for the importance of various aspects of the individual performance to game outcomes.

State of art in the academic literature

Due to data limitations most of the academic literature on football statistics has been concentrated on modelling of goals (e.g. Reep et al., 1971; Maher, 1982; Dixon and Coles, 1997; Karlis and Ntzoufras, 2003; Baio and Blangiardo, 2010; Owen, 2011; Koopman and Lit, 2014) or shots (McHale and Scarf, 2007) on the team level. The use of quantitative methods in assessing a player's worth has been largely limited to quoting simple statistics such as the pass completion percentage, or, following recent technological advances, the distance a player covers in a game. A more comprehensive approach can be found in McHale and Scarf (2005) and McHale et al. (2012) who assume a game result to be a deterministic function of goals scored by both teams, which in turn are modelled based on the number of shots and their effectiveness. The shot count is regressed at the team level on statistics such as passes, dribbles, tackles and interceptions, while the shot effectiveness is assumed to depend deterministically on the rates of: shots on target, blocks and saves. Based on this regression, a marginal effect of each statistic on the number of points awarded for a given result can be calculated. A player's contribution is calculated by summing the marginal contributions of his individual statistics.

An alternative approach is taken by Duch et al. (2010) who, for each game of the European Championships in 2008:

- Construct a network with nodes representing players in this game as well as shots on target and wide, connected by arcs representing passes and shot attempts;
- For each player in the game calculate: pass success rate; the proportion of times their shots do not miss the goal; and the number of balls recovered in a match;
- Combine the above pieces of information into a measure of each player's performance in this match.

Duch et al. (2010) aggregate their performance metric on the team level for both sides competing in the game and demonstrate that it is a good predictor of its outcome.

Both McHale and Scarf (2005) and Duch et al. (2010) use players' individual statistics at their face value, i.e. without accounting for the fact that they are likely to have

been affected by factors beyond the players' control, in addition to their inherent skill. Such an approach seems right if the purpose of the analysis is to retrospectively evaluate players' performances. On the other hand, predicting players' future performance based on this approach may be problematic. For instance, McHale and Scarf (2005) find that their performance index "has too much noise to allow its use for reporting on a weekly basis". Similarly, Duch et al. (2010) have to limit their analysis to "those players that passed the ball at least 5 times in a match" presumably because their performance metric is otherwise dominated by random chance.

If one is supposed to determine a player's value to a club, it is crucial that the focus is on future performance as this alone is what will drive future revenue streams. Past performance can only be considered a useful source of information about future performance to the extent that it helps to evaluate a players' skill, and a player's future performance depends on that skill level. However, performance and skill are not synonymous. The former depends on the latter as well as external factors. Therefore, the best route to predicting future performance is not a direct extrapolation of past performance but, instead, should involve determining the player's inherent skill level, which is more stable in time than the performance itself. This is the general philosophy of our approach, since our aim is player assessment for the purpose of his valuation (rather than, for example, for recognising past performance with individual awards).

1.3 Contribution

Research aims

Based on what the analysts working in the industry have to say about challenges faced by the field as well as the current state of the academic literature, the aim of this thesis is to propose a framework for assessing skill of football players which:

- recognises that past performance is a source of information about, but not synonymous to, a player's skill level;
- accounts for importance of various aspects of individual performance, occurring in different locations of the pitch, to the outcome of a football match.

Such a framework could be useful for player valuation (more on this in chapter 3) but the valuation process itself is not a subject of this research.

Data

The data for the 2006/07 and 2007/08 seasons of the English Premier League was provided by Opta (www.optasports.com). The dataset includes information for every event during a game, including the event type (goal, pass, tackle, etc.), whether the action was a success, the location of the event (for passes, for example, information on the origin and destination of the pass is given), the player(s) involved in the event and the timing of the event.

Synopsis of the methods

The framework for evaluating football players proposed in this thesis attempts to address one of the key difficulties in modelling the game of football, i.e. its free-flowing nature, by discretising it into a series of events. The evolution of the game from one event to another is described using a Markov chain model in which:

- Each game is described by a specific transition matrix with elements depending on the skills of the players involved in this game. Based on this matrix it is possible to calculate game outcome related metrics such as expected goals difference between the two teams at the end of the game. It enables us to establish a link between a specific skill of a given player and the game outcome.

- The skills come from separate location-specific models, e.g. the shooting skill for each player is estimated in a model of converting shots to goals given the shot location. In these models players' skills are represented by random effects to reflect the fact that extremely good and extremely bad players are less common than "average" ones. As a result, the approach is suitable for evaluating skill regardless of how many times the related performance was observed. The estimate of a given skill, for players observed fewer times, is just regressed to the league mean more heavily as desired.

Software

We use `R` statistical programming language (R Core Team, 2012) to implement the methods and obtain the results described in this thesis. The packages utilised most extensively are the following:

- the `lme4` package (Bates and Maechler, 2010) for Generalized Mixed Linear Models in chapters 5 and 7;

- the `mgcv` package (Wood, 2006) for Generalized Additive Mixed Models in chapter 6;
- the `ggplot2` package (Wickham, 2009) to produce most of the figures.

Implications for practitioners

Our ambition is for the methods of player evaluation proposed in this thesis to extend the arsenal of tools used by decision makers at football clubs. It would be unrealistic to expect statistical methods to replace the work of scouts in the near future. Both methods of evaluating players have their strengths and weaknesses and a combination of them is expected to work best. This view is in agreement with the opinion of Steven Houston from Hamburger SV, who said that (Bate, 2012):

> "Stats are central to how we work but you need great scouts. You need them to watch players live. What technical scouting can do is allow you to be more efficient. Scouts can't watch every game; they can't watch every team. There are only so many resources you have and it's about trying to use those resources efficiently."

Outline of the thesis

In chapter 2 we review the literature on player evaluation in football and other selected team sports with a longer tradition of statistical analysis of players' performance. In chapter 3 we argue more formally than in this introduction that, if player valuation is the aim of this assessment, the method used to conduct it must:

- recognise that individual performance depends on the player's underlying skill as well as factors beyond his control, and
- link the individual performances to the team's success.

Chapter 4 describes the data available for this research. In chapters 5 and 6 we propose models for goalscoring and passing as examples of how the first of the above properties of an assessment method is beneficial for predicting future performance. These chapters are based on papers McHale and Szczepański (2014) and Szczepański and McHale (2015) respectively. Such models can be useful on their own for evaluating shooting and passing skills but also as components of a more comprehensive model of the game of football. This is demonstrated in chapter 7 where simpler versions of passing and

shooting models are integrated, along with a few others, into a Markov chain model of the game. Chapter 8 concludes by summarising key findings, listing limitations of the proposed approach and making recommendations for future work.

Chapter 2

Player evaluation in team sports

In this chapter we review the literature on player evaluation in team sports. Rating players in individual sports has also been a subject of academic research, for instance in tennis (e.g. Glickman, 1999; McHale and Morton, 2011) or in golf (e.g. Connolly and Rendleman, 2008; Baker and McHale, 2014), however it is not directly relevant to the subject of this thesis.

Dawson et al. (2000) suggested to model team production W (e.g. in terms of wins) as a function of player performance, L, direct coaching input, C, and other determinants of team performance, X:

$$W = f(L,C,X) \tag{2.0.1}$$

and the player performance in turn to be a function of his talent T, indirect coaching influence, C, and other factors, Y:

$$L = f(T,C,Y) \tag{2.0.2}$$

Despite some ambiguity in the notation, this general model captures the essence of the relationship between the skill of individual players and the match result of the team they play for, i.e. the fact that the former is reflected only indirectly in the latter with the individual performance serving as a link between the two. Even though most of the literature reviewed in this chapter is not explicitly concerned with the economic problem of sporting production functions, the Dawson model is general enough for all the works to relate to it in some way. Therefore, we will use it as a common denominator for the literature review in order to add to its coherence.

In the next section we define two general characteristics of a player evaluation method, each corresponding to one of the equations of the Dawson model. Based on these characteristics, we outline several general approaches for evaluating players in the context of the Dawson model. These approaches are referred to in sections 2.2 and 2.3 where

we review the literature on various team sports as well as in chapter 3 where we outline the original contribution of this thesis.

2.1 Characteristics of player evaluation metrics

2.1.1 Definitions

Let us define two characteristics of a method of evaluating players in team sports:

Number I The extent to which the relationship between the given aspect of individual performance and the overall team performance is specified. It corresponds to equation (2.0.1) of the Dawson model.

Number II The extent to which the method accounts for how much value of the metric for a given player depends on his skill as opposed to factors beyond his control (like the performance of his team mates and opponents or random chance). It corresponds to equation (2.0.2) of the Dawson model.

It is worth emphasising that these are characteristics of the evaluation methods and not of the performance they are based on nor the skill they are supposed to assess. In order to illustrate this, let us take ball juggling as an example of performance with little impact on the team result in football. A method which recognises this weak link between the individual performance at ball juggling and the overall team performance, for example by awarding few points to players performing it, would have a high characteristic number I. This characteristic is the capacity of the method to establish the relationship between the individual performance and the team result and not the strength of the relationship itself.

Similarly, a method for evaluating performance can have a high value of the characteristic number II regardless of the extent to which the performance itself depends on skill. It is sufficient that the method attempts to capture the strength of the relationship between skill and performance rather than treating them as equivalent.

Furthermore, while it could be argued that it is generally desirable to establish a link between individual and team performance (i.e. use a method with a high characteristic number I), the choice of a method on the spectrum number II depends more on the application. If the aim of the analysis is to retrospectively evaluate players' performance, e.g. in order to distribute annual awards, then there is not as much need to recognise the randomness involved in this performance compared to when the aim is to use the performance to assess the pure skill that generates it.

The primary application for methods designed to do the latter is player valuation. The reason why it is important to focus on skill rather than past performance when valuing players is that it is future performance that football clubs should be paying for in wages and transfer fees and the relationship between past and future performance can be determined more accurately if one recognises that there is some inherent skill behind both. We will be more specific about this statement in chapter 3, where the original contribution of this work is outlined, and it will be illustrated with examples in chapters 5 and 6. For now let us leave possible applications and prepare to review the literature in the context of the aforementioned characteristics and the Dawson model.

2.1.2 Approaches to player evaluation

In order to provide some intuition on the characteristics of player evaluation methods defined in the previous section, its four corner cases are listed below and put in the context of the Dawson model. They are referred to later in sections 2.2 and 2.3 where we review specific methods from the literature.

Low I - low II The most basic and historically the earliest approach is to evaluate players based on their individual statistics which are *believed* to be related to the team success to some degree. Examples of it are: number of goals per season and pass completion rate in football or batting average (the number of hits divided by the number of attempts) in baseball. The relationship between the individual performance and the team production is assumed to exist based on common sense but is unverified and unquantified. Furthermore, these individual statistics are taken at their face value without consideration for how much chance was involved in achieving particular values and, as a result, how likely it is that similar values will be obtained in future. It is as if only equation (2.0.2) was used in isolation and the impact of external factors was ignored reducing the Dawson model to:

$$L = T \tag{2.1.1}$$

High I - low II This approach addresses the *low I* deficiency of the above methods by establishing the relationship between individual performance and team success. In the context of the Dawson model this approach could be represented by:

$$W = f(L,C,X) \quad \text{and} \quad L = T \tag{2.1.2}$$

Low I - high II This approach addresses the other aspect of the *low I - low II* methods. It recognises that factors beyond players' control can affect their performance and

explicitly models the impact of covariates and random luck. As a result, a better estimate of a particular skill can be obtained, however, its importance for the team is unknown. In this approach the Dawson model is reduced to its second equation:

$$L = f(T, C, Y) \tag{2.1.3}$$

High I - high II Finally, the most complete approach combines the advantages of the two previous ones by (1) recognising that talent and performance are not identical and (2) linking the individual and the team performance. It can be represented by the Dawson model in its most general form.

We refer to the above corner cases in the next section when reviewing the literature on player evaluation methods in football as well as in section 2.3 where we look at the other sports.

2.2 Football research

Game results Due to data limitations most of the academic literature on football statistics has been concentrated on the modelling of goals at the team level. Reep et al. (1971) showed that the negative binomial distribution provides a good fit to the aggregate goal counts from 706 matches of the English Football League First Division. Maher (1982) modelled the goals distribution on a game by game basis conditional on parameters representing attacking and defensive ability of the competing teams. He used independent Poisson distributions for the goals scored by both teams but also considered a bivariate Poisson model to capture correlation between the scores. The departure from the assumption of independence between the scores in a football match was also recognised by Dixon and Coles (1997) who modified the joint distribution directly to address this problem. Karlis and Ntzoufras (2003) extended the bivariate Poisson model by inflating the draw probability and examined many variations of this model in great detail. More recently Baio and Blangiardo (2010) argued that the correlation between the scores of both competing teams is implicitly taken into account in a hierarchical model which they fit in the Bayesian framework. Owen (2011) and Koopman and Lit (2014) modelled game results in a dynamic framework, i.e. they allowed team parameters to evolve in time. McHale and Scarf (2007) modelled shots on goal. They found negative correlation between shots for the opposing teams and used copulas to model the joint distribution.

Individual performance The use of quantitative methods in assessing a player's worth has been largely limited to quoting simple statistics such as the pass completion percentage, or, following recent technological advances, the distance a player covers in a game, which are typical *low I - low II* methods. A more comprehensive approach can be found in McHale and Scarf (2005) who proposed a complex team production function, in which the game result is a deterministic function of goals scored by both teams, which in turn are modelled based on the number of shots and their effectiveness (see McHale et al. (2012) for a more detailed description). The shot count is regressed on the team level on statistics such as passes, dribbles, tackles and interceptions while the shot effectiveness is assumed to depend deterministically on the rates of shots on target, blocks and saves. Based on this, a marginal effect of each statistic on the number of points awarded for a given result can be calculated. A player's contribution is calculated by summing the marginal contributions of his individual statistics. Since the purpose of the model is to retrospectively evaluate players' performances, their individual statistics are used at their face value, i.e. without accounting for the fact that they are likely to have been affected by factors beyond the player's control, in addition to his inherent skill. Probably for this reason the authors find that such a performance index "has too much noise to allow its use for reporting on a weekly basis" (McHale et al., 2012, p. 346).

Similarly, Tiedemann et al. (2010), who weight players' goal and assist totals together with tackle and pass completion ratios to calculate players' efficiency, recognises that "problems may arise with football players who have only played a few games so that their performance may be affected by external factors or, simply, luck" (Tiedemann et al., 2010, p. 585). Szczepański (2008), Duch et al. (2010) and Oberstone (2011) are other examples of methods which, for the same reason, are suitable for retrospective evaluation of individual performance but may lack predictive utility. They can be classified as *high I - low II* type methods.

Applications for skill estimates A separate branch of research in the academic literature concerns possible applications for estimates of players' skill. In these studies an estimate of the skill is treated as given from an external source rather than derived by the authors.

Perhaps the most obvious application is player valuation. Tunaru and Viney (2010) as well as Gulbrandsen and Gulbrandsen (2011) proposed methods to conduct it conditionally on players' talent level approximated by a proprietary metric of player performance called the Opta Index. Dawson et al. (2000) explored the relationship between skill and value in the opposite direction. They approximated the talent of an individ-

ual player by his transfer value predicted from a model using covariates such as: age, goalscoring record, league experience, number of previous clubs, etc. Then they modelled team production (measured in win percentage) as a log-linear function of the sum of the talent of the individual players available to the team. A similar approach can be found in Gerrard (2001).

Another application can be found in Garcia-del Barrio and Pujol (2009) who modelled the proportion of minutes players are designated to spend on the pitch out of all the possible minutes depending on factors such as: age, nationality, market value and past performance . They did not attempt to come up with a measure of the latter but use journalists ratings instead.

2.3 Research in other sports

Due to the abundance of data, the application of statistical methods to evaluating players has a much longer history in sports other than football. This is particularly true for the American professional league sports such as baseball, basketball and ice hockey. It is worth reviewing the research done in these sports in order to learn from the experience of the researchers who have been analysing individual performance using statistical methods for a longer time.

2.3.1 Baseball

If a researcher was allowed to pick a team sport to analyse, it is likely that baseball would be his choice. Consisting largely of a series of duels between a batter and a pitcher, it is probably one of the easiest team sports to model due to the limited amount of interaction between players. For this reason, and due to the availability of detailed games data going back to the 19-th century, baseball has a long history of analysing player performance by academics and other sport enthusiasts. Research done by the latter is referred to as *sabermetrics*[1]. Despite the fact that it does not undergo formal peer review in the academic sense, it deserves a separate paragraph due to its popularity and impact on the world of professional sport. In this section we review both the sabermetric and the academic literature on evaluating baseball players.

[1]From the acronym SABR, which stands for the Society for American Baseball Research.

Sabermetrics research

Batters Baseball is the best known of all sports for the application of statistical methods to evaluating players. This theme has even found its way to the popular culture as it was depicted in a film entitled *Moneyball* by Miller (2011) based on a book by the same title by Lewis (2004b). They both tell a story of Billy Bean, the General Manager of a Major League Baseball franchise Oakland Athletics, who employs a statistician to help him identify undervalued talent in the players market. One of the tools they use for this purpose is said to be a statistic called On Base Percentage which improves on metrics commonly used to evaluate batters at the time, such as Batting Average (the number of hits divided by the number of opportunities). OBP recognises that drawing walks is an important part of a batter's skill set for getting on base. Up till then, it had been largely overshadowed by the more spectacular skill of hitting the ball. The OBP statistic itself has been actually known at least since Branch (1954) but the Oakland Athletics under Billy Bean are said to be among the first to implement it within an MLB organisation to recruit players.

Evaluating players based on BA or OBP can be thought of as a *low I - low II* approach with the advancement to OBP corresponding to a search for a statistic that is better linked to the team success than BA (i.e. improving the characteristic number I). Another step in this direction was made by the introduction of a weighted On-Base Average. The wOBA statistic weights the result of each at bat (a walk, a single, a double, etc.) by how much, on average, it affects the team runs expectancy compared to an out (Tango et al., 2008) rather than treating all the hits equally like the OBP does. It can be classified as a *high I - low II* type method.

Pitchers As far as evaluating pitchers is concerned, the most basic statistic is the number of runs (R) the other team scores while they are pitching. These can be split into earned runs (ER), for which the pitcher is held accountable, and unearned ones, which occur due to fielding errors. The runs and the earned runs are often scaled by the number of innings pitched (IP) to calculate (earned) runs average: $RA = R/IP$ and $ERA = ER/IP$.

The runs based statistics can be classified as *high I* type methods because of their direct relationship with the team result. The distinction between earned and unearned runs is introduced in order to separate the runs statistics into components which pitchers have control over and the rest. It corresponds to an advancement to a *high II* type approach. However, this distinction is arbitrary since it depends on the judgement of the official

scorer as to whether an offensive player would have been allowed to advance a base had the fielder acted as expected. A lot of research in the sabermetric community has been devoted to making this distinction more objective. The problem was first recognised by McCracken (2001) who argued that pitchers have little control of the outcome of their pitches that are hit into the field of play and that, as a result, there is little persistence in time in values of the statistic which measures this aspect of performance (called Batting Average on Balls in Play) for a given player. This led to the introduction of metrics based on "Defense Independent Pitching Statistics" such as the Fielding Independent Pitching (FIP) statistic (Tango et al., 2008), Quik ERA (Silver, 2006) and Skill-Interactive ERA (Swartz and Seldman, 2010), which were shown to be better predictors of future values of ERA than its values from previous seasons (Swartz and Seldman, 2010).

Wins Above Replacement An interesting concept is calculating players' contribution to team wins above what a hypothetical replacement player would do. Such a statistic is called Wins Above Replacement (WAR) and there are several ways of doing the calculations (Baseball-Reference.com, 2013). For example, one of the proposed methods for evaluating pitchers starts from the Fielding Independent Pitching statistic as an estimator of ERA and through a series of calculations puts it on the scale of Runs Average and then wins, adjusting for factors which are thought to unfairly affect the performance such as the ballparks the players pitched in (Cameron, 2009). For offensive players, WAR is calculated by combining their production in terms of batting (with the wOBA statistic outlined above a possible starting point), base running and fielding. The numbers are adjusted for the difficulty of playing in a given position (Cameron, 2008). WAR is a good example of a *high I* type method with player's performance expressed directly in terms of his contribution to the success of his team. Because some of its components (e.g. FIP) are designed to minimise the impact of external factors on their value, it can also be classified somewhere in the middle of the characteristic number II spectrum. We would not classify it as a *high II* approach as it does not involve explicit modelling of the influence of random chance nor of the covariates. Instead, attempts are made to correct for these factors *post hoc*, e.g. by scaling the values of component metrics (obtained by an individual) by the average value for the ballpark played in.

Projection systems The main purpose of the metrics outlined above is to describe the performance in a given season as well as possible and they may be flawed as predictors of future performance when used on their own (DuPaul, 2012). Several *projection systems* have been proposed in order to address this problem. Their methodology varies, and

some of them are proprietary, but they normally involve some form of weighting of (potentially regressed to the league mean) values of a given statistic in past seasons adjusted for ballpark effects, league difficulty, expected playing time as well as an ageing effect (Slowinski, 2011).

Academic research

Markov chains Analysing baseball players has also been a subject of interest of the academic community. For example, Cover and Keilers (1977) proposed a statistic they called Offensive ERA to evaluate batters. For each individual it is the expected number of runs scored in nine innings by a batting lineup consisting of players of his ability. The metric can be calculated from a Markov chain model of a baseball game based on a set of individual probabilities to hit for a single, a double, a triple, a home run and advance base on balls. Sueyoshi et al. (1999) combined the Offensive ERA with Data Envelopment Analysis to evaluate Japanese players. Bennett and Flueck (1983) introduced Expected Runs Production metric which weights a number of individual batters' statistics with coefficients estimated by regressing team season total runs on the equivalent team statistics. Bukiet et al. (1997) used a Markov chain model similar to that of Cover and Keilers (1977). They suggested evaluating the difference in the contribution of two players to winning for a particular team by substituting one for another in the batting line up and comparing the expected number of wins calculated based on the Markov chain model. All these are *high I - low II* type methods.

Overlap with sabermetrics There is some overlap in the analysis done by the academics and the sabermetricians. Most notably, the idea behind the wOBA statistic, so popular among the sabermetricians, was first introduced by Lindsey (1963) who suggested that:

> "A new approach to the assessment of batting effectiveness could be based on three assumptions:
>
> (a) that the ultimate purpose of the batter is to cause runs to be scored
>
> (b) that the measure of the batting effectiveness of an individual should not depend on the situations that faced him when he came to the plate (since they were not brought about by his own actions), and
>
> (c) that the probability of the batter making different kinds of hits is independent of the situation on the bases.

> It is generally believed that the third assumption is not true, but that there are so-called 'clutch hitters' who are particularly successful in critical situations. Evidence on this point is difficult to secure."

On the other hand, Bennett and Flueck (1984) proposed a modified version of the method which recognised players differently depending on the situation (in terms of the game stage, the current result, the number of outs and the occupied bases) thus following only the first of the Lindsey's postulates. As such, their method evaluated players also based on how well they performed in the clutch. A similar approach was taken in cricket by Lewis (2004a) and Scarf et al. (2011) who reward players for their contribution to winning and not losing differently depending on the stage of the match in which they score or concede runs. The extent to which clutch performance can be attributed to individual skill has itself been a subject of statistical studies. For example Albert (2007) fitted a series of Beta-binomial hierarchical models to individual walk, strikeout, home run and in-play hit rates under the assumption that there is no clutch specific skill. He then compared simulations from these models to empirical data and concluded that, although the performance of most of the players is in line with the models, there are some outliers (in particular for the walk and the strikeout rates) which could suggest that a very few players may indeed perform particularly well in clutch situations. This is an example of a *high II* type study with low characteristic number I (since it focuses on all the aspects of performance in isolation).

Regression to the mean The aforementioned study of Albert (2007) exemplifies the benefits of using formal statistical methods for modelling the relationship between skill and performance. It illustrates how exceptional performance in some aspect of the game one season tends not to be repeated the next year if it depends to little extent on real skill. Such regression to the mean of players' performance has been recognised and accounted for by other *low I - high II* type methods for evaluating baseball players. Efron and Morris (1975) used Stein's estimator to shrink observed batting averages and show the positive impact of the shrinkage on predictions. More recently, Albert (2006) disaggregated the ability of pitchers from luck in their observed performance using a binomial random effects model and Null (2009) did the same for batters using a nested Dirichlet-multinomial model. One of the first applications of random effects models to modelling individual performance in sport is due to Albert (1992), who employed a Poisson random effects model to analyse home run hitters. Related methods were applied to analysing other aspects of the game with Jensen et al. (2009) focusing on

fielding performance and Loughin and Bargen (2008) investigating pitcher and catcher influence on base stealing.

2.3.2 Invasion sports

Whilst baseball has probably the longest tradition of evaluating players, its character is too dissimilar from football for its methods to be directly transferable. Research done for invasion sports which share football's continuous nature and the amount of interaction between players may be more relevant for the topic of this thesis. In this section we review the literature on evaluating players in basketball and ice hockey as examples of such disciplines.

Basketball

Individual statistics The basketball research in player evaluation has been dominated by two general approaches. The more traditional approach (of the *low I - low II* type) has been to assess players based on their individual game statistics such as points scored, assists, rebounds, turnovers, etc. They can be aggregated in different ways and be used to calculate various rates in order to represent the aspect of performance of interest. Kubatko et al. (2007) provide a good summary of ideas of this sort conceived by basketball enthusiasts and professionals in an attempt to introduce them to the academic community. In a move to improve the characteristic number I, Berri (1999) proposed a method to combine simple versions of such individual statistics into a metric of total player contribution to team success. His approach is similar to the one Bennett and Flueck (1983) used for baseball players in that it weights individual statistics with estimates of coefficients from a regression of team wins on corresponding team statistics. A similar model was proposed earlier by Zak et al. (1979), although the primary purpose of this research was to evaluate team production efficiency and it only mentioned a possible application to player evaluation in the conclusions. It also used a multiplicative Cobb-Douglas model rather than the linear additive one that Bennett and Flueck (1983) later found to be more appropriate.

Plus/minus The second general approach to evaluating basketball players is to ignore individual game statistics and attempt to directly capture how players on court affect point scoring. It is by definition a *high I* approach. The basic version of this method, known as plus/minus, rewards players with points when they are on court at the time their team scores and takes points away when their team concedes. Rosenbaum (2004)

extended the method to account for the team mates and the opponents present on court by regressing the points per possession ratio on fixed effects representing player abilities. Ilardi and Barziali (2008) modified the method to evaluate offensive and defensive contribution separately and used five seasons of (weighted) data to obtain more robust estimates. Fearnhead and Taylor (2011) improved its characteristic number II by using random effects to represent player abilities in a Gaussian linear mixed model fit in the Bayesian framework. They also allowed player abilities to evolve between seasons. Finally, they provided a bridge between the two approaches to evaluating baseball players by regressing the estimates of player abilities from the mixed effects model on individual game statistics. They found that whereas the individual statistics can explain a considerable amount of variation in the offensive ability of players, they provide very little information about players contribution to preventing points for the opposite team.

Ice hockey

The return of plus/minus The first use of the plus/minus statistic is said to be in the 1950s in ice hockey but it was not until the recent developments in basketball that more advanced versions of the statistic found its way back there. The adjusted plus/minus metric of Rosenbaum (2004) was applied to ice hockey by Macdonald (2011). Macdonald (2012) extended it by using ridge regression instead of the ordinary least squares method in order to introduce regularisation (and improve its characteristic number II in the words of this chapter) of parameter estimates with a similar effect as Fearnhead and Taylor (2011) achieved earlier in basketball.

Rarity of goals One problem with modelling ice hockey scores compared to basketball is the rarity of scoring events. It was recognised by Gramacy et al. (2013) who noted that goals are scored on less than 2% of shifts which led them to question the normality assumption of goals per minute in Macdonald (2011) and Macdonald (2012). In order to address this issue they focused only on goals, rather than all the shifts, as a unit of observation and modelled player's impact on "the odds that, given a goal has been scored, it is a goal for that player's team". They introduced regularisation with a Laplace prior distribution on player effects and also considered a version of their model with additional team effects. Another approach to deal with the rarity of scoring shifts was proposed by Thomas et al. (2013) who explicitly modelled the goal scoring process in a hockey game using competing Poisson processes and evaluated players based on their effect on the scoring rates.

Other game statistics in plus/minus framework There have been several attempts to use events other than just goals in the plus/minus framework. In addition to goals per minute, the previously mentioned work of Macdonald (2012) modelled the rates of shots on target (called simply "shots" in ice hockey), shots on target and missed (called the Fenwick statistic) and shots on target, missed and blocked (the Corsi statistic). Schuckers et al. (2011) proposed another version of the plus/minus method in which he:

1. estimated the probability that a goal will follow a given event type (a face off, a give away, a hit, etc.) within a fixed time interval;

2. subtracted this probability for each event from the indicator of whether a goal was actually scored;

3. regressed such a variable on indicators of the players present on the ice at the time of the event.

Schuckers and Curro (2013) removed the 2. step of the above procedure and used the probability calculated in 1. as the response variable modelled in the final step.

Chapter 3

Problem statement

In chapter 2 we quoted a team production model and defined two characteristics of player assessment methods each corresponding to one equation of this model. We reviewed the literature on player evaluation in football and other team sports in the context of the production model and these characteristics. This chapter attempts to define the team production model more precisely (section 3.1.1) and use it to argue that if player valuation is the purpose of the analysis, *high I - high II* methods of player assessment (corresponding to the full form of the team production model) are the most suitable (section 3.1.2). We refer to the literature review in chapter 2 in search of such methods and review the most promising approach used in other sports as a candidate for application in football (section 3.2.1). This method has several deficiencies in the context of football, so the problem of devising a *high I - high II* type method of footballers assessment remains open. A modelling framework to tackle this problem is proposed in the end of this chapter in section 3.2.2.

3.1 Team production model

Recall the Dawson production model from the previous chapter in which team performance depends on performance of individual players that in turn is a function of their skill and other factors. In section 3.1.1 we attempt to restate this general model in more detail in order to be able to highlight some of its implications (in section 3.1.2) and place the original contribution in its context more precisely (in section 3.2.2).

3.1.1 Individual and team performance

Let R_t be the result of a game played at time t and let the performance of the i-th player in this game be represented by an n element vector $\mathbf{p}_{i,t}$. Similarly to equation (2.0.1) let us assume:

$$R_t = f(\mathbf{p}_{1,t}, \mathbf{p}_{2,t}, \ldots, \mathbf{p}_{N,t}) \tag{3.1.1}$$

where

$$\mathbf{p}_{i,t} = \left[p_{i,t}^{(1)}, p_{i,t}^{(2)}, \ldots, p_{i,t}^{(n)}\right], i = 1, \ldots, N.$$

The specification is flexible enough to allow each element of the vector $\mathbf{p}_{i,t}$ to represent anything from the overall performance of the i-th player to just one aspect of it (like passing, shooting, tackling, etc.), or even a single action performed by this player in the game in question.

Similarly as in equation (2.0.2), the individual performance of the i-th player, or the k-th aspect of that player's performance in this case, is a function of his skill set at the time of the game and external factors including the skill of the other players in this game and random chance, $\varepsilon_{i,t}^{(k)}$:

$$p_{i,t}^{(k)} = g_k\left(\Pi_{1,t}, \Pi_{2,t}, \ldots, \Pi_{N,t}, \varepsilon_{i,t}^{(k)}\right) \tag{3.1.2}$$

where

$$\Pi_{i,t} = \left[\pi_{i,t}^{(1)}, \pi_{i,t}^{(2)}, \ldots, \pi_{i,t}^{(M)}\right], i = 1, \ldots, N$$

is the vector of skills of the i-th player at time t. For example, a player's performance at shooting can be thought to depend on

- his ability to create a good shooting position for himself and execute the shot;
- the ability of the players on the opposite team to close him down and block his shot as well as the shot saving skill of the goalkeeper;
- factors like pitch and weather conditions, which can be aggregated into the random component.

The individual shooting performance is studied in chapter 5 with the above in mind. In chapter 6 we study the relationship between skill and performance at passing in a similar way.

3.1.2 Player valuation in the context of the model

Two components of player's value Player valuation can be expressed in the context of the model (3.1.1)-(3.1.2) in the following way. At time t_0 the value of the i-th player V_{i,t_0} for a given football club can be assumed to consist of a sport related component, S_{i,t_0}, and a commercial component, C_{i,t_0}. The former is the benefit the club will receive directly due to the player's contribution on the pitch while the latter is what it will gain due to the player's celebrity appeal (think of David Beckham for instance):

$$V_{i,t_0} = h\left(S_{i,t_0}, C_{i,t_0}\right). \tag{3.1.3}$$

Total contribution to team performance The contribution on the pitch is the key component of the valuation (since the commercial value can be argued to depend on it) so it deserves a more specific definition. The total contribution of the i-th player to the team performance at time t can be expressed as a sum of the contributions from each of his M skills:

$$K_{i,t} = \sum_{j=1}^{M} \int \left(\pi_{i,t}^{(j)} \times \delta_{i,t}^{(j)}\right) \mathrm{d}\pi_{i,t}^{(j)} \tag{3.1.4}$$

where

$$\delta_{i,t}^{(j)} = \frac{\mathrm{d}R_t}{\mathrm{d}\pi_{i,t}^{(j)}} \tag{3.1.5}$$

is the marginal contribution from the j-th skill of the i-th player to the game result and the j-th integral is the total contribution from the j-th skill of the i-th player.

The sport value of player i at time t_0 can then be expressed in terms of his future total contributions to the results, for example, in the following way:

$$S_{i,t_0} = \sum_{t>t_0} \phi_s(t-t_0) \times K_{i,t} \tag{3.1.6}$$

where $\phi_s(t-t_0)$ is a function downweighting the later contributions to reflect the fact that results obtained this season are more important than those later in the future.[1] In this framework a player's future total contribution to the performance of a given team is the key to working out his value for this club. From equation (3.1.4), the contribution at a given time depends on the skill set at that time. Thus in order to predict future contributions we need to predict the evolution of the skill set. For example, it could be argued that the level of a given skill, at a given point in time, is a function of: its level in

[1] Similarly the commercial value at the time t_0 may be argued to be a downweighted aggregate of the future commercial gains from having the player i on the team (e.g. from selling merchandise with his name on it) which in turn will probably depend on his previous contribution on the pitch to some degree.

the past; the player's age; the amount and quality of the training he undertakes; and the injuries he suffers in between. Therefore we have:

$$\pi_{i,t_2}^{(m)} = r_m\left(\pi_{i,t_1}^{(m)}, age_{t_1}, injuries_{(t_1,t_2)}, training_{(t_1,t_2)}\right) \tag{3.1.7}$$

In this thesis the skills of a given player are assumed to be constant in time. This simplification is not expected to be too unrealistic over the course of the two seasons where data is available for this research.

Advantages of a *high I - high II* approach Equations (3.1.3)-(3.1.7) outline a framework for player valuation in the context of a team production model. The equations cover the full spectrum of relationships from a skill of a single player, to the individual performances of all the players, to the game outcome, and finally to the value of each player to the team. The procedure relies on the team production model being in its full form, i.e. with the team production equation (3.1.1) depending on individual performances which in turn are functions of some underlying skill in equation (3.1.2). As such, the procedure can only be implemented if a *high I - high II* approach to player evaluation, corresponding to the full team production model, is taken. *Low I* or *low II* methods reduce the team production model in one way or another (see section 2.1.2 for a discussion) meaning that some components of the valuation procedure in equations (3.1.3)-(3.1.7) are missing. Namely:

- Methods that do not specify a link between individual and team performance do not allow us to investigate how changes in the former influence the latter. This makes it impossible to determine importance of a given aspect of individual performance to the team.
- Ignoring the fact that performance of a player is a function of his skill, and factors beyond his control, has a negative impact on predictions of future performance, and, as a consequence, on predictions of total future contributions that ultimately matter for player valuation. Examples to support this statement will be provided in chapters 5 and 6.

3.2 *High I - high II* approach for evaluating football players

In the previous section we argued that a *high I - high II* approach is the most suitable basis for player valuation. Such methods have not been applied to football yet (see

section 2.2). In fact, the only approaches of this type used in any invasion sport up to date are the recently introduced regularised plus/minus methods in basketball and ice hockey (section 2.3.2). In section 3.2.1 we outline the idea behind the regularised plus/minus method emphasising some of its assumptions that reduce its appeal as a method for evaluating football players. In section 3.2.2 we propose an alternative *high I - high II* method.

3.2.1 Regularised plus/minus model

The general idea behind the plus/minus method is to model the observed variable, usually number of points scored by a team or points difference, as a function of the sum of parameters representing skills of the players present on the field of play during the unit of observation (for example a shift in ice hockey). Equation (3.2.1) presents the idea in its simplest form

$$y_t = \sum_{i \in A_t} \alpha_i - \sum_{i \in B_t} \alpha_i + \varepsilon_t \tag{3.2.1}$$

where y_t is the difference in the number of points scored by team A and team B during the observation unit t, α_i are parameters representing players' skill and ε_t is a noise component. A_t and B_t are sets of players present on the field of play during the observation unit t. The player parameters do not correspond to any specific skill like passing, shooting, skating, tackling, etc. Instead they are supposed to capture players' overall ability to contribute to team success. In the most recent applications (e.g. Fearnhead and Taylor, 2011; Macdonald, 2012; Gramacy et al., 2013; Thomas et al., 2013; Schuckers and Curro, 2013) they are estimated with regularization methods in order to induce some regression to the mean of parameter estimates.

There are at least three problems with using this method for evaluating football players:

- The method assumes that each player present on the field of play has the same impact on the observed variable. This might be approximately true for basketball where the distance of each player from the ball at any given time is constrained by the size of the court[2]. The ice hockey rink is bigger[3] in absolute terms but perhaps it could still be argued that because of high skating speed, all players have some effect on the outcome of each play. The argument already seems weak, especially for the goalies who neither skate very fast nor have any intention to participate in

[2] 94 ft by 50 ft (28.65 m by 15.24 m) in the NBA

[3] 200 ft by 85 ft (61 m by 26 m) in the NHL

offensive play most of the time. The violation of the equal contribution assumption is bigger yet in football which is played on a pitch of approximately 115 yd (105 m) by 74 yd (68 m) where a player at one end of the pitch can have very little effect on what happens at the other one.

- The second strong assumption of the plus/minus approach to player evaluation is the assumption that each player can be characterised successfully by a single skill. For example, the skills of offensive players in ice hockey and football at passing, dribbling and shooting, are all captured by the same number. Therefore, within this framework it is impossible for a coach to identify potential areas for improvement for a specific player which could addressed in training.

- Finally, the additive nature of the plus/minus method, together with the previous assumption, implies that the expected contribution (and hence the sporting value) of a given player is the same for every team. For example, the method suggests that a team that hardly ever allows its opponents to shoot at its goal, e.g. because its players are great at maintaining possession of the ball, would benefit to the same extent from upgrading their goalkeeper position as a weak team that is on the receiving end of many shots per game. Such implications are difficult to defend.

3.2.2 Introducing mixed effects Markov chain model for player evaluation

Before we outline the original contribution of this thesis, let us recap the line of thinking that motivates the research one last time.

In section 3.1 we presented a general team production model and argued that if player valuation is the aim of his assessment, the method used to conduct it must fulfil the criteria of such a production model in its most general form. This means that it needs to consist of the team production equation (3.1.1) depending on individual performances, which in turn are functions of some underlying skill in equation (3.1.2). We called such methods *high I - high II* in section 2.1.2. We have not come across an example of such an approach in the review of the literature on football player evaluation methods in section 2.2. In section 3.2.1 we listed drawbacks of the only *high I - high II* method used in other sports if it was to be applied to football. Finally, in this section we outline a new *high I - high II* framework devoid of these disadvantages. It will be presented in greater detail in chapter 7.

The general idea behind the proposed framework is to:

- Break down the game of football into states characterised by: which team is in possession of the ball; the game event (pass, shot, goal, etc.); the player executing it; and the location of the event;
- Use a Markov chain to model transitions between these states. This corresponds to the team production equation (3.1.1);
- Assume players are characterised by a vector of skills;
- Model elements of the transition matrix of the Markov chain conditional on the players' skills (represented by random effects) and state characteristics (represented by fixed effects). These models correspond to the player production equations (3.1.2).

Contrary to the plus/minus method, in this framework:

- The outcome of each event can depend to different extents on the skill of the player executing it and on the skills of other players on the pitch;
- Since each player is characterised by a vector of skills, the outcome of a given action depends on the skill relevant to performing it. For example players can be good or bad at passing and shooting the ball but it is only the latter skill that directly affects transition between a shot and a goal state.
- Interaction between players of various qualities is captured naturally through the Markov chain structure. For example a team with players who pass the ball very well will put its strikers in shooting positions more often than other teams, thus the shot converting skill of the strikers will have a bigger impact on its results. In other words, player's skill can affect not only how well he executes his actions, or how well other players execute their actions, but also which actions are actually executed by the players (e.g. good passers of the ball increase the number of shooting opportunities for their team). It is this indirect effect that can vary depending on the team set up, allowing the value of a given player to be different for various teams.

We believe this framework to be general enough to encompass very complicated models of a football game. Game states could be assumed to depend on many actions such as: passes, shots, dribbles, tackles, etc. and various player skills such as: passing, shooting, dribbling, tackling but also positioning and marking.[4] The Markovian property, which

[4]Data about positioning of players without the ball would be needed for this but is not available for this project.

assumes that the transition probability between states depends only on the current state, could be weakened by making states of the chain consist of the current game event and the previous one.[5]

The aim we set ourselves in this thesis is not to present a complete model of a football game covering all of its possible events depending on long vectors of player skills. It is rather to propose a general framework for player evaluation (outlined in this chapter), give a non-contrived example of a model that fits into it and demonstrate its usefulness. The latter is the topic of chapter 7 where we propose a method for evaluating players' total contribution to the team performance conditional on their skills, or $K_{i,t}$ in the notation of section 3.1.2. We demonstrate that a value of this statistic aggregated across players of a given team is a good indicator of its future success, hence can form a basis of assigning monetary value to individuals. However first, in chapters 5 and 6 we strengthen the argument of the final paragraph of section 3.1.2 by demonstrating the necessity of modelling individual performance based on an isolated latent skill ($\pi_{i,t}$ in the notation of this chapter) for the purpose of predicting future performance.

[5] See section 7.4.1 for why it may be beneficial.

Chapter 4

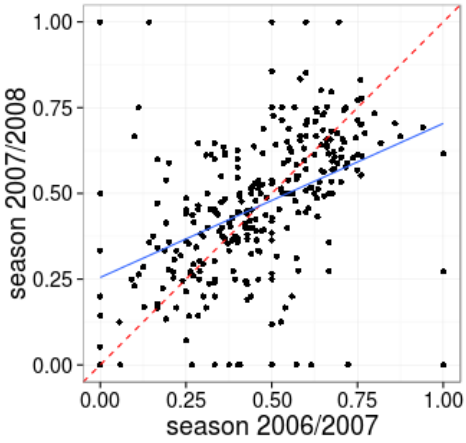
Data

Raw data specification The data for the 2006/07 and 2007/08 seasons of the English Premier League was provided by Opta (www.optasports.com). The dataset includes information for every event during a game including: the event type (goal, pass, tackle, etc.); whether the action was a success; the location of the event (for passes, for example, information on the origin and destination of the pass is given); the player(s) involved in the event; and the timing of the event. Table 4.1 shows a few selected columns of a sample of the data.

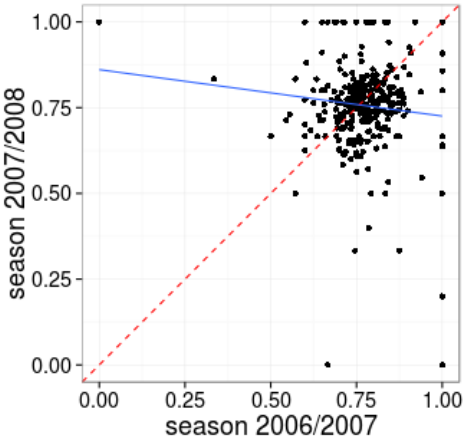
team_name	player_name	event_name	x	y
Manchester United	Giggs	Pass(Open play, Key pass) Successful - Short	0.75	0.52
Manchester United	Saha	Off target(Open play,Right foot)	0.78	0.49
Fulham	Niemi	Pass(Goal kick) Successful - Long	0.05	0.37
Fulham	Helguson	Duel Won (Aerial)	0.62	0.22
Manchester United	Evra	Duel Lost (Aerial)	0.39	0.86
Fulham	Helguson	Pass(Header) Unsuccessful - Short	0.62	0.22
Manchester United	Brown	Pass(Open play) Unsuccessful - Short	0.32	0.82
Fulham	Christanval	Clearance(Unsuccessful)	0.28	0.26
Manchester United	Rooney	Pass(Cross, Goal assist) Successful - Long	0.76	0.81
Manchester United	Ronaldo	Goal(Open play,Right foot)	0.94	0.40
Fulham	Helguson	Pass(Open play) Successful - Short	0.50	0.50

Table 4.1: Sample of the Opta events data. Variables x and y determine the location of the event. Covariate x can range from 0.00 to 1.00 as the length of the pitch where 0.00 is the defending goal line and 1.00 is the attacking goal line. Covariate y corresponds to the width of the pitch where the right hand touchline is 0.00 and the left hand touch line is 1.00.

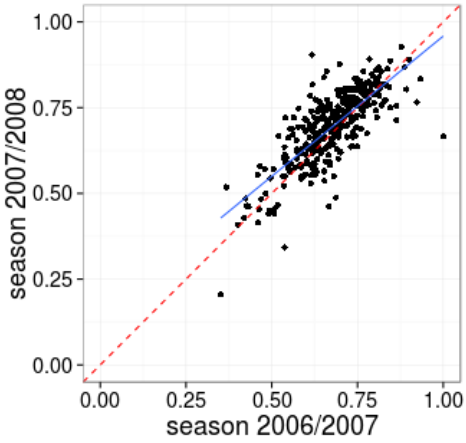
Example aggregate statistics Figure 4.1 compares players' success rate at performing four types of actions in season 2006/07 and 2007/08. Correlation between the observed



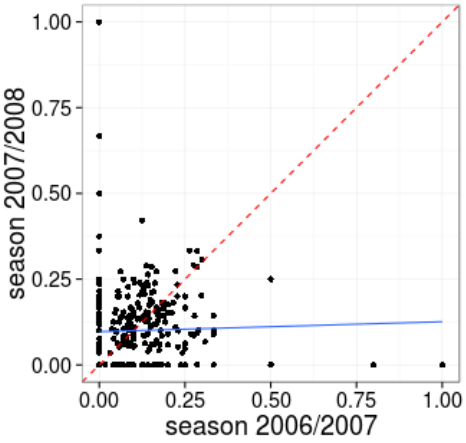
(a) Aerial duels won / all aerial duels



(b) Tackles won / all tackles



(c) Successful passes / all passes



(d) Goals / shots

Figure 4.1: Players' success rate at various actions in two consecutive seasons of the English Premier League. Each point represents a player. The dashed line is the identity function and the solid one is a linear regression fit.

performance in the two seasons varies greatly across the action types: from very weak for shots (figure 4.1d), or even negative for tackles (figure 4.1b), to strongly positive for passes (figure 4.1c). For the actions for which there is little correlation in the observed performance between seasons (e.g. tackles and shots), this suggests a few possible explanations:

- there is no skill involved in them and the observed performance is purely random;

- the skill exists but varies extremely between seasons, to the extent that it is pointless to predict future performance based on the past;
- the skill exists, is stable (relative to the observed performance) but in any given action, game, or even season, it is obscured by factors beyond the control of the executing player.

Any football fan would find the first two scenarios difficult to believe. The last one calls for the use of statistical methods in order to filter out the skill element from beneath the external factors. This is one of the challenges we set ourselves in this thesis: we assume the last scenario is true and design statistical models under this assumption. We verify predictions of future performance based on these models hoping for them to be more accurate than ones directly extrapolating past performance.

Statistical modelling can be useful not only for the aspects of performance that appear to exhibit no correlation between seasons. Take pass completion rate in figure 4.1c as an example. There seems to be a relatively strong correlation between its values for a given player in consecutive seasons. However, it can be argued that summarising players' passing ability with this statistic is inappropriate given that passes vary in terms of difficulty. One way to add some context to this statistic would be to split the sample according to factors believed to affect the difficulty of the pass. An example of this approach is demonstrated in figure 4.2 which presents the same relationship as figure 4.1c but split according to the location of the pass origin. Indeed, the pass success rate appears to depend on the location of its origin, e.g. the points in the last column tend to be positioned lower and more to the left than in the other columns. However, note that the relationship between performance in the two seasons in all the panels is now weaker than for the overall statistic in figure 4.1c.

The probable reason for this is that the empirical pass completion rates are now based on fewer observations. So what was gained in terms of accounting for pass difficulty, to some extent, was lost in terms of being able to predict future performance directly based on the past performance. In this thesis we attempt to avoid this trade off by using statistical modelling, which can offer a more comprehensive solution to the problem.

Player's position in tactical formation In addition to the variables available directly in the dataset, we intend to use information on players' positions in the tactical formation. We attempt to derive it based on the individual events data. More sophisticated algorithms could be used for this purpose but since this is not a primary focus of this study, we settle for the following crude method to anticipate the position of the k-th

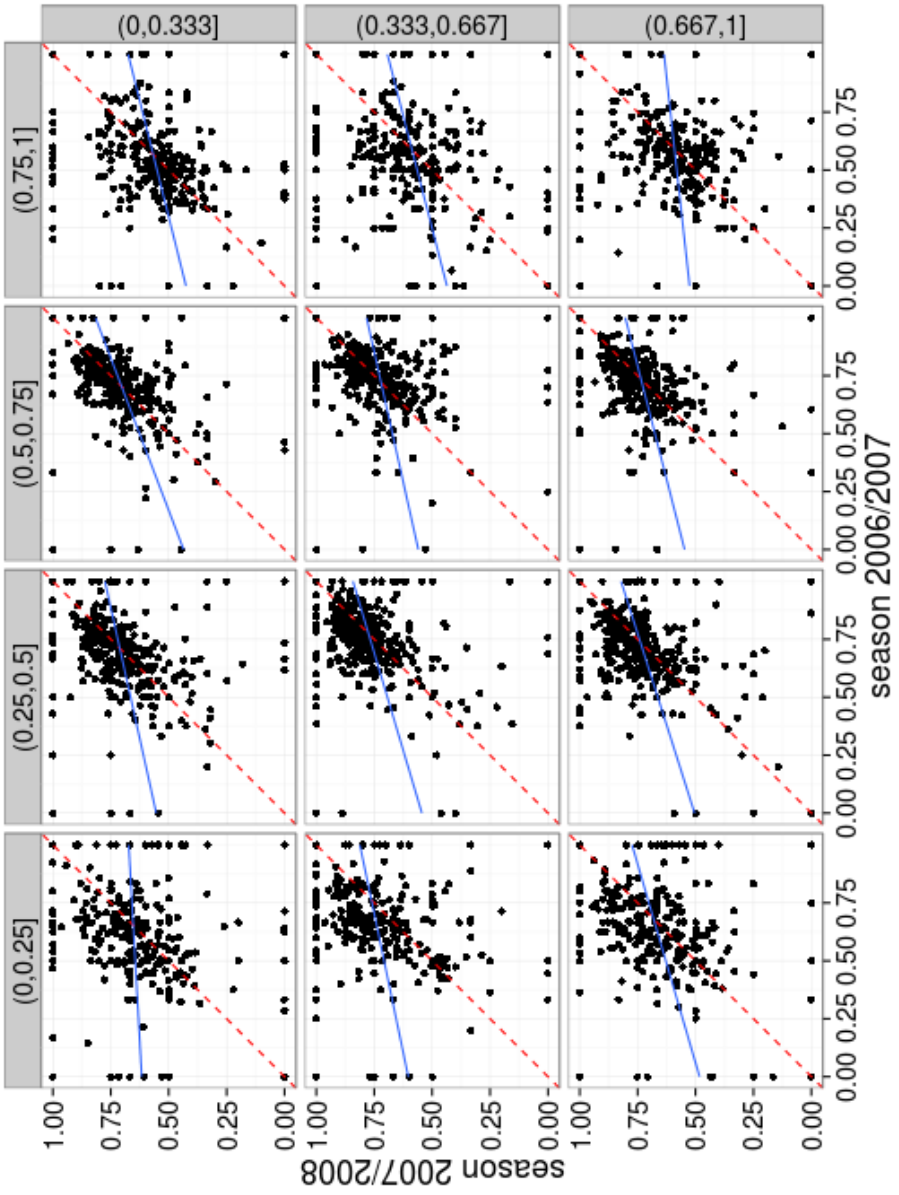


Figure 4.2: Players' pass success rate in two consecutive seasons of the English Premier League depending on the location of the pass origin. Pitch width, y, is split into thirds presented in rows whereas pitch length, x, is split into quarters presented in columns. Each point represents a player. The dashed line is the identity function and the solid one is a linear regression fit.

player in the j-th game:

- Assume that a player is a goalkeeper if he performed at least one goalkeeper specific action (like a save) in any of his games.

- Calculate the absolute value of the distance from the centre of the pitch in the width coordinate[1]:

$$\tilde{y} = |y - 1/2| .$$

  The absolute value of this distance is used to avoid cancelling of terms when we calculate the average distance during a match for players who switch wide positions during the game.

- Calculate $\left(\bar{x}_{k,j}, \bar{\tilde{y}}_{k,j}\right)$ as the weighted average coordinates of all the k-th player's events (shots, passes, tackles, duels, dribbles, etc.) for matches prior to the j-th game. The weights for events from a game played on day d_i depend exponentially on the number of days between that day and the day of the j-th fixture, and are given by $\exp[-\phi(d_j - d_i)]$. We set $\phi = 0.1$ which means that the coordinates from any one game contribute around half as much to the average as the coordinates from a game played a week later. This choice is entirely arbitrary.

- Categorise players to nominal positions in each game based on the $\left(\bar{x}_{k,j}, \bar{\tilde{y}}_{k,j}\right)$ values according to the rule illustrated in figure 4.3.

- Categorise players to nominal positions for the whole season based on how frequently they were assigned to each position in that season.

The boundary definitions given in figure 4.3 were subjective but this particular set of boundaries was found to be satisfactory in that players were, in general, assigned to the position one would expect them to be given knowledge of a player's expertise.

Finally, note that because players' positions are anticipated based on past games, there are missing data (for $\left(\bar{x}_{k,j}, \bar{\tilde{y}}_{k,j}\right)$) in the first game of each player in the sample. These observations are removed from the sample.

Sections 5.2, 6.2 and 7.2 provide more detail on the aspects of data relevant to the analysis presented in the respective chapters.

[1]The pitch coordinates are: $x \in < 0, 1 >$ for the pitch length (0 being the coordinate of the team's on the ball goal) and $y \in < 0, 1 >$ for its width (0 for the right sideline).

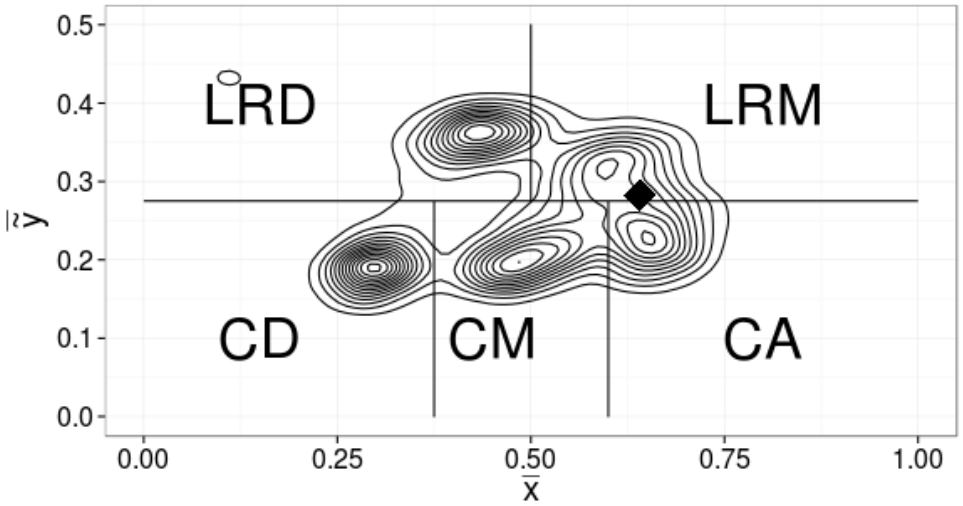


Figure 4.3: Contour plot of anticipated player positions in games of the 2006/07 season. The x-axis represents the sideline, $\bar{x} = 0.0$ is the attacking team's goal line, $\bar{x} = 1.0$ is the defending team's goal line. The y-axis is the distance from the axis going through the centre of the goals so that $\bar{\tilde{y}} = 0.0$ is the centre of the goals and $\bar{\tilde{y}} = 0.5$ corresponds to the two throw-in lines. Players' nominal positions are based on the boundary definitions shown: LRD = Left/Right Defender, CD = Central defender, LRM = Left/Right Midfielder, CM = Central Midfielder, CA = Central Attacker.

Example: The centre of ♦ marks $\left(\bar{x}_{k,j}, \bar{\tilde{y}}_{k,j}\right)$ coordinates for Cristiano Ronaldo going into his second game of the 2006/07 season. The values of $\left(\bar{x}_{k,j}, \bar{\tilde{y}}_{k,j}\right) \approx (0.64, 0.28)$ are average coordinates x_i and $\tilde{y}_i$ of the events involving Ronaldo in the first game of the season (as this is his only earlier game in the sample). Notably, in that game the average of Ronaldo's plain y_i coordinates was approximately 0.43 implying that on average he played a very central position. However, the average of y_i is misleading in this case as Ronaldo played wide but was switching sides from left to right and back again. Using the average of the distance from the centre, $\tilde{y}_i$, instead, accounts for this fact correctly classifying Ronaldo as a Left/Right Midfielder based on this game.

Chapter 5

Signal and noise in goalscoring statistics

In chapter 3 we argued that, if player valuation is the aim of his assessment, the method used to conduct it must:

1. recognise that individual performance depends on the player's underlying skill as well as factors beyond his control, and
2. link the individual performance to the team success.

The argument for the first point is that:

- it is future performance of a player that counts for the club and, thus, has direct impact on his value for them. This part of the argument was justified in section 3.1.2.
- predictions of future performance are more accurate if they are based on a model with the first of the above properties than if they are made by a direct extrapolation of past performance. We provide an example to support this part of the argument in the current chapter, focusing on a selected aspect of players' performance, namely scoring goals.

The chapter is structured as follows. In section 5.1 we provide some background and motivate the study of goals using mixed effects models. Section 5.2 presents the data and some descriptive statistics. Section 5.3 presents the model of goal scoring which comes in two parts: a model for the number of shots a player has in a game and a model for the number of goals a player scores in a game conditional on the number of shots he has in that game. In section 5.4 we present the results of the model and assess its performance

as a forecasting model for the number of goals a player will score in a season. Section 5.5 concludes with some closing remarks. The code used to fit the models of this section can be found in appendix D.

This chapter is based on the paper of McHale and Szczepański (2014).

5.1 Background and motivation

This chapter concentrates on what is arguably the single most important statistic in football: goals. Goals decide the outcome of matches so having an insight into which players have a greater capacity to score goals is clearly of great worth to football analysts. Such information could be used by coaches and managers in making team selection choices, and aid in decision making when identifying which player a club should buy and how much the player is worth.

Modelling of goals has, of course, been the subject of statistical research in the past. Due to data limitations most of the academic literature on football statistics has been concentrated on the modelling of goals at the team level. A review of these studies can be found in section 2.2. In contrast with that research, our work focuses on modelling goals scored by individual players rather than by teams.

Typically a player's ability to score goals has been measured by the total goals in a season and his goals per game or per minute ratio. However these statistics will often not represent true goal scoring ability. For example, many players will appear in only a handful of games so that the statistic would be based on a small sample size meaning lucky (or unlucky) breaks will play a significant role in a player's statistics. Further, team-based effects are not taken into account (a player playing in a top team will likely have more opportunities to score goals than a player playing for a low quality team).

Figure 5.1 presents a plot of the goals per 100 minutes played for the 2006/07 season versus the goals per 100 minutes played for the 2007/08 season, for players who featured in both seasons. Each point represents a single player. The dashed line is the identity function and the solid line is a linear regression fit. As expected there is a positive relationship between the players' performances in both seasons as represented by the solid line. However, there is also some evidence of bias in that players who were perceived to be very good (bad) in the first season generally tend to decrease (increase) their performance in the next one. This is represented by the fact that the solid line does not cover the dashed identity line but lies above it for the low arguments and below it for the higher ones.

The reason behind this *regression to the mean* phenomenon is likely to be the ex-

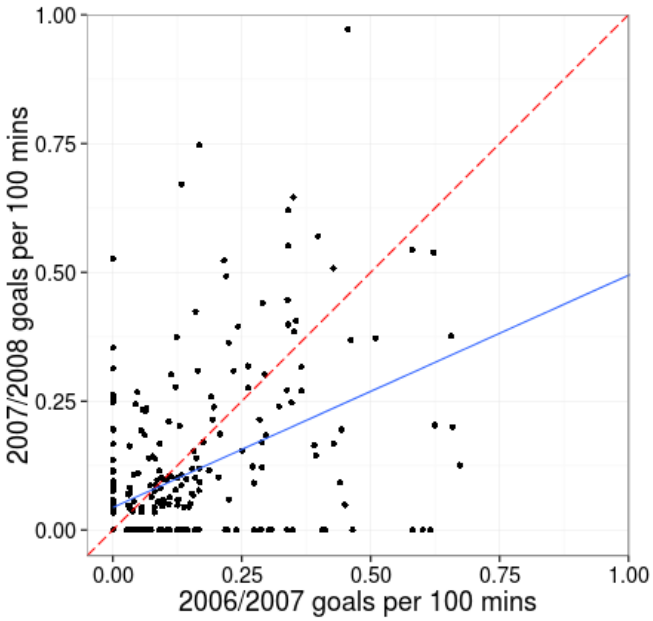


Figure 5.1: Goals per 100 minutes played for 2007/08 versus goals per 100 minutes played for 2006/07. Each point represents a player. The dashed line is the identity function and the solid line is a linear regression fit.

istence of a chance component in the observed goals per minute ratio. Some part of what we observe may be due to the inherent skill of a player, and it probably is in this case given the positive relationship between performances across seasons. However, a lucky bounce of the ball or a last second clearance from the goal line will also have an impact on a player's historical record. In fact, the players who find themselves at the top of the observed performance ranking are the ones that are more likely to have benefited from luck than the ones sitting at the bottom of it. By its definition, chance is not consistent in time, thus the observed performance of the players who have been lucky in one season is likely to deteriorate in the future when they may not benefit from it. In other words, if variation in past performance consists of variation in player abilities and random chance, it is desirable to base predictions about future performance only on the former component. Determining how much of a player's ability is reflected in goal scoring performance and how to filter it out is the subject of this study.

Regression to the mean of players' performance has been recognised and accounted for in models for other sports in the past. Refer to sections 2.3.1 and 2.3.2 for examples from baseball and invasion sports (basketball and ice hockey).

In this chapter we present a model for the process of goal scoring that can be used to identify a player's true goal scoring ability. The model takes into account the sample

size of each player's observations, as well as team-based effects and other covariates, and breaks down the process of scoring goals into shot generation and conversion of shots into goals.

5.2 Data

There are 20257 shots in the whole sample (seasons 2006/07 and 2007/08 of the English Premier League) of which 15489 have not been blocked. Quoting the data provider's website, a blocked shot is: *Any goal attempt heading roughly on target toward goal which is blocked by a defender, where there are other defenders or a goalkeeper behind the blocker*. Ignoring blocked attempts is consistent with earlier studies of shots in football (Pollard et al., 2004; Ensum et al., 2005), as well as shooting accuracy metrics used in the business[1]. We have rerun our analysis including blocked shots and found none of the conclusions change.

Penalties are also ignored due to their specific nature, resulting in 15289 shot observations to be used in the analysis. The models are fit to the 2006/07 sample leaving the second season for model validation. In the fitting sample there are 804 goals scored from 7678 shots (0.105 goals per shot) attempted in 380 games (20.2 shots per game).

One factor that may have a big impact on the number of shots a player attempts is his position in the tactical formation, for example goalkeepers obviously do not attempt nearly as many shots as the centre forwards. Information on a player's playing position is not available in the dataset, however, we may attempt to derive it based on the individual events data. A simple algorithm to do it was presented in chapter 4.

5.3 Methods

The process of scoring goals can be broken down into a player's ability to create shots and his ability to convert the shots into goals. By splitting the process into these two components we can measure the extent to which both depend on a player's ability, the strength of his team and the opposing team, as well as other factors including random chance. Letting n and y be the number of shots and the number of goals, we model the distribution of goals, $p(y)$, as

$$p(y) = \sum_n p(y|n)p(n) \tag{5.3.1}$$

[1] http://www.optasports.com/about/news/feature-opta's-football-action-definitions.html

which gives us two models to fit: one for the distribution of shots and another one for the conditional distribution of goals given shots.

Each of the two processes is modelled using generalized linear mixed effects models (GLMMs) which extend the generalized linear models methodology of Nelder and Wedderburn (1972) by addition of random effects. Mixed effects models have been applied to a wide range of problems in, for example.: education (Tekwe et al., 2004), medicine (Patton et al., 2002), politics (Fowler et al., 2008), law (Rigotti and DiFranza, 1997), environmental studies (Réale et al., 2003) and social studies (Moore and Gould, 2005; Diez-Roux et al., 2000). The particular strategy of breaking down a problem into a hierarchy of sub-problems and applying mixed effects models to each sub-problem our approach most similar to one that has been used in insurance at least since Pinquet (1997). He models the frequency of claims using a hierarchical Poisson model with Gamma distributed effects for policy holders and combines it with a hierarchical Gamma (or log-normal) model for the severity (cost) of claims with another set of Gamma effects for the individuals. Further levels of hierarchy can be added to this approach to flexibly model the specifics of the process in question as demonstrated more recently by Frees and Valdez (2008).

There are no well established models for our problem which we can compare our approach to, and a comparison only with the naive method may not be considered demanding enough. For these reasons, we choose to present results of two versions of each model for each process varying with regards to the initial set of covariates considered. Hopefully such a comparison will enable the reader to draw conclusions about where the strength of the whole modelling approach lies. The two versions of each model are:

- a *basic* model in which the covariates included only team specific information and a home field indicator; a specification which we considered minimal to satisfactorily represent the given process.

- an *extended* model in which the covariates included information on player positions, the time he spent on the pitch (for the shot count model) and the number of shots a player has (for the shots to goals conversion model).

In section 5.3.1 the theory of Generalized Linear Mixed Models is presented, then, in sections 5.3.2 and 5.3.3, we describe the shot count and the shot conversion models both of which are examples of a GLMM. Section 5.3.4 discusses how to make predictions based on both models and, finally, how to combine them into predictions of the future number of goals that can be expected of a given player in a game.

5.3.1 Generalized Linear Mixed Model

A Generalized Linear Mixed Model (GLMM) is an extension of a GLM, outlined in appendix A.1, such that the response y depends on a vector of random effects b, in addition to some fixed effects β, through the linear predictor:

$$\eta_i = X_i\beta + U_i b\,. \tag{5.3.2}$$

The vector of random effects b is assumed to come from a multivariate normal distribution:

$$b \sim \mathcal{N}(0, G_\theta) \tag{5.3.3}$$

where G_θ is a covariance matrix depending on some unknown parameters θ.

Approximating marginal likelihood

Fitting the above model requires estimation of the fixed parameters β and θ. This is done by the Maximum Likelihood method. In order to obtain the marginal likelihood of the fixed parameters we integrate the random effects out of the joint distribution:

$$L_{\theta,\beta} = \int p(y,b|\theta,\beta)\,\mathrm{d}b = \int p(y|\beta,b)p(b|\theta)\,\mathrm{d}b \tag{5.3.4}$$

$$\propto |G_\theta|^{-1/2} \int \exp\{\log p(y|\beta,b) - \frac{1}{2}b'G_\theta^{-1}b\}\,\mathrm{d}b\,. \tag{5.3.5}$$

The above integral is intractable in general but can be evaluated approximately. We represent the exponent as

$$f_{\theta,\beta}(b) = \log p(y|\beta,b) - \frac{1}{2}b'G_\theta^{-1}b \tag{5.3.6}$$

and can approximate the above integral using the Laplace's method (appendix A.4) as an adjusted profile likelihood (e.g. Lee et al., 2006, p.103-104):

$$|G_\theta|^{-1/2} \int \exp\{f_{\theta,\beta}(b)\}\,\mathrm{d}b \simeq |G_\theta|^{-1/2} \times (2\pi)^{-k/2}|-f''_{\theta,\beta}(\tilde{b})|^{-1/2} \times \exp\{f_{\theta,\beta}(\tilde{b})\} \tag{5.3.7}$$

$$\propto |G_\theta|^{-1/2} \times |-f''_{\theta,\beta}(\tilde{b})|^{(-1/2)} \times \exp\{f_{\theta,\beta}(\tilde{b})\} \tag{5.3.8}$$

where $\tilde{b} = \tilde{b}(\theta,\beta)$ maximises $f_{\theta,\beta}(b)$:

$$\tilde{b} = \arg\max_b f_{\theta,\beta}(b)\,. \tag{5.3.9}$$

The problem in (5.3.9) is solved using a *penalized iteratively re-weighted least squares* (PIRLS) algorithm described in appendix A.3.

In summary, we can approximately evaluate the log-likelihood of (θ, β) at $\tilde{b}$ as:

$$\ell_{\theta,\beta} \propto -\frac{1}{2}\log|G_\theta| - \frac{1}{2}\log|-f''_{\theta,\beta}(\tilde{b})| + f_{\theta,\beta}(\tilde{b}) \tag{5.3.10}$$

and maximise the above with respect to (θ, β) with each evaluation involving maximisation of $f_{\theta,\beta}(b)$ with respect to b for the given (θ, β) as in (5.3.9).

Two of the possible improvements of the basic Laplace approximation are to (i) base it on higher order series expansions (Raudenbush et al., 2000) or (ii) evaluate the integrand on more points near the mode in the adaptive Gauss-Hermite quadrature procedure (Liu and Pierce, 1994; Pinheiro and Bates, 1995). We have employed the latter of these methods, in addition to the basic Laplace approximation, to fit the models of sections 5.3.2 and 5.3.3 but found both methods to give practically the same estimates for both models.

In the Bayesian framework this type of model can be estimated using MCMC sampling (Fahrmeir and Lang, 2001) or the more recently developed integrated nested Laplace approximations method (Rue et al., 2009).

Predicting random effects

The random effects are of little interest in some applications. In this chapter, however, they represent player abilities which are the central point of the analysis. Interestingly, $\tilde{b}(\theta, \beta)$ obtained from solving (5.3.9) for given θ and β also maximises $p(b|y, \theta, \beta)$ (e.g. Jiang, 2007, p.136-137), i.e. the conditional density of the random effects given the observed data and the parameters. We use this mode of the conditional distribution at the parameter estimates, $\hat{b} = \tilde{b}(\hat{\theta}, \hat{\beta})$, as the estimate of players' random parameters in what is sometimes called maximum *a posteriori* or penalised likelihood estimation (Lee et al., 2006, p.106).

5.3.2 Shot counts

The model for shot counts is a Poisson GLMM. The basic and extended versions are presented in this section. The results of fitting them to the data from season 2006/07 of the English Premier League are presented in section 5.4.1.

Basic model

There are $K = 534$ players and $T = 20$ teams in the fitting sample which consists of 380 games played in the 2006/07 season of the English Premier League. Let $\beta^{(att)}$ and $\beta^{(def)}$

denote the row vectors of T team attack and defence ability parameters respectively and b the column vector of K player ability parameters. The player abilities are assumed to be random and come from a common multivariate normal distribution:

$$\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I) \tag{5.3.11}$$

where I is a $K \times K$ identity matrix and σ^2 is a common variance parameter to be estimated.

One single observation n_i corresponds to the number of shots generated by a particular player in a given game. If there were only observations for the two starting elevens in each game, in total there would have been $2 \times 11 \times 380 = 8360$ such data points, however because we also observe the number of shots made by the substitute players there are 10230 observations in the sample. Finally, note that because players' positions are anticipated based on past games, there are missing data for the first game of each player in the sample. Since one of the variations of the shots count model will use the information about tactical position, we remove these observations for all the shots count models, in order to make the model fits comparable, leaving $N = 9744$ data points.

The number of shots made by a particular player in a given game is assumed to have a Poisson distribution:

$$n_i \sim Poisson(\exp \eta_i) \tag{5.3.12}$$

with the linear predictor depending on the player and team abilities:

$$\eta_i = \gamma^{(n)} + \nu^{(n)} h_i + \beta_{l(i)}^{(att)} + \beta_{m(i)}^{(def)} + b_{k(i)} \tag{5.3.13}$$

where $k(i)$, $l(i)$ and $m(i)$ denote respectively the indices of the player, the team he plays for and the opposition. The parameter $\nu^{(n)}$ represents the home advantage effect (h_i is a dummy variable set equal to 1 if the i-th observation corresponds to a home team player, 0 otherwise) and $\gamma^{(n)}$ is the intercept term. The superscript (n) is used to indicate that the number of shots is modelled.

The above relationship can be written in the matrix representation for all the N observations :

$$\eta = \mathbf{Z}\mathbf{b} + \mathbf{X}\beta \tag{5.3.14}$$

where η is a column vector of length N of the linear predictors. $\mathbf{Z}$ is an $N \times K$ design matrix which for each observation row selects the element of vector $\mathbf{b}$ corresponding to the

player observed in that row. The vector of fixed parameters $\beta = [\gamma^{(n)}, \nu^{(n)}, \beta^{(att)}, \beta^{(def)}]^T$ is multiplied by a block design matrix:

$$\mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{h} & \mathbf{X}^{(att)} & \mathbf{X}^{(def)} \end{pmatrix}$$

where:

- $\mathbf{1}$ is column vector of length N of ones (corresponding to the intercept term);
- $\mathbf{h}$ is an N long column vector with ones for the home team players and zeros for the away team players;
- $\mathbf{X}^{(att)}$ is an $N \times T$ design matrix which for each row selects the attack parameter of the team the observed player plays for;
- $\mathbf{X}^{(def)}$ is an $N \times T$ design matrix which for each row selects the defence parameter of the team the observed player plays against.

Note that in the fitting procedure, in order to ensure model identifiability, we constrain $\beta_1^{(att)} = \beta_1^{(def)} = 0$, where 1 is the index of Manchester United the Premier League champions of the 2006/07 season, leaving effectively $2(T-1)$ team parameters to estimate. As a result the column corresponding to the first team drops out of both the design matrices $\mathbf{X}^{(att)}$ and $\mathbf{X}^{(def)}$ leaving $(T-1)$ columns in each of them and as a result $2+2(T-1)$ columns in $\mathbf{X}$.

Figure 5.2 shows the histogram of the number of shots per player per game. Also shown is the fitted Poisson distribution with mean 0.75 (the average of the fitting sample). The data is clearly over-dispersed. The Chi-square test for the goodness of fit of the single mean Poisson distribution to the empirical data rejects the null hypothesis considerably below the 0.001 significance level. This might suggest that the negative binomial distribution could be more suitable here. Note, however, that the single mean Poisson distribution in figure 5.2 ignores the variance of the mean parameter due to player and team effects and the effects of other covariates. Rather than using the negative binomial distribution to model the over-dispersion explicitly, we proceed with the Poisson distribution based model and verify the fit of the full model in section 6.4.

Extended model

In the extended version of this model we additionally consider the number of minutes the shooting player (k) played in the game (j) that the i-th observation corresponds to (divided by 100), $t_{[k,j](i)}$, and his position in the tactical formation in this game, $\pi_{pos_{[k,j](i)}}$

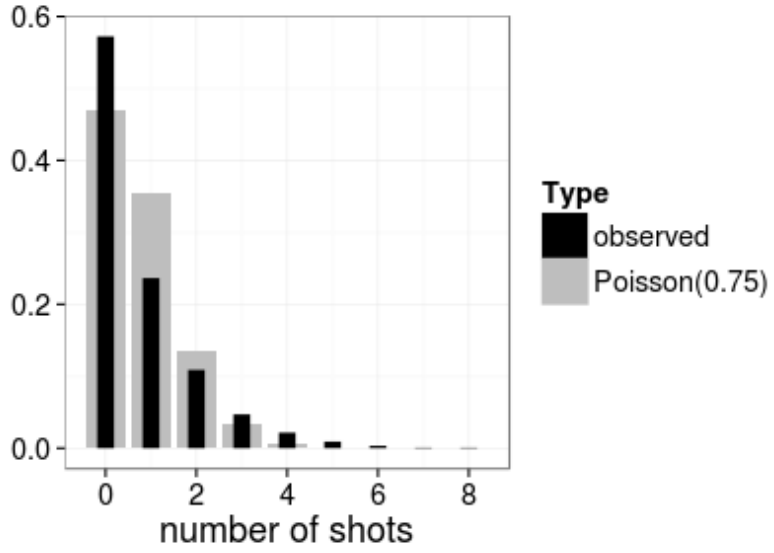


Figure 5.2: Histogram of shots per player per game (black bars), with fitted Poisson distribution.

(as estimated using the procedure described in chapter 4), in the initial set of covariates leading to the following linear predictor:

$$\eta_i = \gamma^{(n)} + \nu^{(n)} h_i + \tau \log t_{[k,j](i)} + \beta^{(att)}_{l(i)} + \beta^{(def)}_{m(i)} + \pi_{pos_{[k,j](i)}} + b_{k(i)}\,. \tag{5.3.15}$$

5.3.3 Shots to goals conversion rates

The model for shot conversion is a binomial GLMM. Its basic and extended versions are presented in this section. The results of fitting them to the data from season 2006/07 of the English Premier League are presented in section 5.4.2.

Basic model

Of the 10230 player-game observations analysed in the shots count model, $J = 4347$ have a positive number of shots ($n_i > 0$). Only these observations provide information about players' abilities to convert shots to goals. Therefore we drop the observations with no shots recorded and index the remaining ones $j = 1, 2, ..., 4347$. Each j-th observation contains the number of goals, y_j, player $k(j)$ scored from n_j shots playing on team $l(j)$ against team $m(j)$.

We adopt a binomial mixed effects model for the number of goals, y, converted from n shots:

$$y_j | n_j \sim \text{Bin}\left(\frac{\exp(\psi_j)}{1+\exp(\psi_j)}, n_j\right) \tag{5.3.16}$$

with the following linear predictor:

$$\psi_j = \gamma^{(y)} + \nu^{(y)} h_j + \alpha^{(att)}_{l(j)} + \alpha^{(def)}_{m(j)} + a_{k(j)} \tag{5.3.17}$$

where $\gamma^{(y)}$ is the intercept term and $\nu^{(y)}$ is the home advantage parameter. The superscript (y) is used to indicate that the number of goals scored is modelled.

Whereas the b, $\beta^{(att)}$ and $\beta^{(def)}$ in equation(5.3.13) represented players' abilities to take shots and teams' abilities to create and prevent shooting opportunities, the a, $\alpha^{(att)}$ and $\alpha^{(def)}$ terms represent the abilities of players to convert shots into goals as well as the effect the player's team and the opposition have on this process. The team parameters are stored in vectors $\alpha^{(att)}$ and $\alpha^{(def)}$. The player parameters are stored in a vector a and are assumed to be random effects with the multivariate normal distribution:

$$\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \phi^2 I) \tag{5.3.18}$$

where I is a $K \times K$ identity matrix and ϕ^2 is a common variance parameter to be estimated.

The linear predictor for all the J observations can be written in matrix representation:

$$\psi = \mathbf{Z}\mathbf{a} + \mathbf{X}\alpha \tag{5.3.19}$$

where ψ is a column vector of length J of the linear predictors. $\mathbf{Z}$ is a $J \times K$ design matrix for which each observation row selects the element of the vector $\mathbf{a}$ corresponding to the player observed in that row. The vector of fixed parameters $\alpha = [\gamma^{(y)}, \nu^{(y)}, \alpha^{(att)}, \alpha^{(def)}]^T$ is multiplied by a block design matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{h} & \mathbf{X}^{(att)} & \mathbf{X}^{(def)} \end{pmatrix}.$$

The design matrices $\mathbf{Z}_{J\times K}$ and $\mathbf{X}_{J\times(2+2(T-1))}$ used here are defined in the same way as those in the shots model but have fewer rows since $J < N$ as explained in the beginning of this section.

Extended model

In the extended version of this model we additionally consider the total number of shots, n_j, the player attempted in a given game j, in the initial set of covariates, leading to the following linear predictor:

$$\psi_j = \gamma^{(y)} + \nu^{(y)} h_j + \theta \log n_j + \alpha^{(att)}_{l(j)} + \alpha^{(def)}_{m(j)} + a_{k(j)}\,. \tag{5.3.20}$$

The hypothesis behind this extension is that the more shots a player decides to take in a game the less prepared they tend to be when shooting, leading to a lower conversion rate. A player who waits for good shooting opportunities will make fewer attempts but the chance of scoring from each of them will tend to be higher.

5.3.4 Predicting future performance

In order to make predictions about players' performances, parameters in equations (5.3.13) and (5.3.17), or their extended equivalents, are substituted with their estimates obtained in the procedure outlined in the previous section.

We make two types of predictions for season 2007/08:

- *Complete* predictions $\breve{y}_{k,j}$ for player k in his j-th game for which the real observed values of the covariates (opposing team, time played, nominal position, etc.) in this sample are used. Note that at the time of the game all the covariates are either known (venue and the competing teams) or can be controlled by the manager (player's position and the time he spends on the pitch).

  Since we want to make the predictions comparable among players who played different number of minutes, we scale the total expected goals by the total time played in the predicted sample by the given player:

$$\breve{y}_k = \frac{\sum_{j=1}^{N_k} \breve{y}_{k,j}}{\sum_{j=1}^{N_k} t_{k,j}} \times 100\,. \tag{5.3.21}$$

- *Averaged* predictions in which values of the covariates in the predicted sample are projected based on what they where in season 2006/07 before being plugged into the model equations (5.3.13) and (5.3.17).

The motivation behind producing the averaged predictions is to emulate the situation in which we would have been before the season started, i.e. ignore the information about which game featured which player, how long they played, in what position they played, etc. This procedure enables a manager, for example, to compare all players on an equal footing. In order to do this we average all the information about venue, team, position and time played out of the predictions in the following procedure:

1. Calculate the expected number of shots player k makes when playing in position pos_z:

$$\tilde{n}_k^{pos_z} = E_n(n_k^{pos_z}) = \exp(\tilde{\eta}_k^{pos_z}) \tag{5.3.22}$$

   where

$$\tilde{\eta}_k^{pos_z} = \hat{\gamma}^{(n)} + \frac{1}{2}\hat{v}^{(n)} + \hat{\tau}\log\bar{t}_{k,pos_z} + \frac{1}{T}\sum_{t=1}^{T}\hat{\beta}_t^{(att)} + \frac{1}{T}\sum_{t=1}^{T}\hat{\beta}_t^{(def)} + \hat{\pi}_{pos_z} + \hat{b}_k$$

   and $\bar{t}_{k,pos_z}$ is the average time played per game by player k in position pos_z.

2. Given the above, calculate the expected number of goals player k scores when playing in position pos_z using the law of total expectation:

$$\breve{y}_k^{pos_z} = E(y_k^{pos_z}) = E_n(E_y(y_k|n^{pos_z})) = \sum_{n=0}^{\infty} p_{\tilde{n}_k^{pos_z}}(n) \times \tilde{y}_k(n) \tag{5.3.23}$$

where $p_{\tilde{n}_k}(n)$ is the Poisson probability with the expected value of $\tilde{n}_k$. In practice we only sum up to $n = 20$, but this seems more than enough given the distribution of n (figure 5.2). $\tilde{y}_k(n)$ is the goals expectation conditional on the number of shots based on the conversion model:

$$\tilde{y}_k(n) = E_y(y_k|n) = n\frac{\exp[\tilde{\psi}_k(n)]}{1+\exp[\tilde{\psi}_k(n)]} \tag{5.3.24}$$

where

$$\tilde{\psi}_k(n) = \hat{\gamma}^{(y)} + \frac{1}{2}\hat{v}^{(y)} + \hat{\theta}\log n + \frac{1}{T}\sum_{t=1}^{T}\hat{\alpha}_t^{(att)} + \frac{1}{T}\sum_{t=1}^{T}\hat{\alpha}_t^{(def)} + \hat{a}_k\,.$$

3. Finally, we weight the player's goals expectancies in all the positions by the number of times they played in each of position in the fitting sample. Then we scale this number by the total number of minutes played, to obtain values per 100 minutes comparable among all the players:

$$\breve{y}_k = \frac{\sum_{z=1}^{6} N_k^{pos_z}\breve{y}_k^{pos_z}}{\sum_{z=1}^{6} N_k^{pos_z}\bar{t}_{k,pos_z}} \times 100\,. \tag{5.3.25}$$

5.4 Results

In this section we present the results of fitting the shots count model and the shots to goals conversion model. Both models are combined in section 5.4.4 to make predictions for the 2007/08 season, which are then compared with the naive predictions and the empirical data.

Recall that penalty shot goals are excluded from the analysis and note that we are only able to make predictions for the 2007/08 season, and as a result compare the methods, for players who featured in both the 2006/07 and 2007/08 seasons of the English Premier League.

5.4.1 Model fit for shot counts

We fitted several versions of both the basic and the extended model varying the set of fixed effects used, i.e with or without: home advantage parameter, attacking team ability

and defending team ability. Based on the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) the best version of the shots count model included the home advantage and the defending ability of the opposing team as fixed effects for the basic model and additionally the position of the player and the time played for the extended model. The attacking ability of a player's team was dropped from each version of the model. We comment on this in the discussion section of this chapter.

Table 5.1 presents parameter estimates for the two specifications of the shots count model. Note that in both cases the estimates of team ability to prevent shots are with respect to the 2006/07 champions Manchester United. The lower the parameter value, the fewer shots the given team allows.

On the natural scale the global mean and home advantage parameter estimates for the basic model imply that an average player makes $\exp(-1.06+0.25) \approx 0.45$ shots per game against Manchester United on the home field and $\exp(-1.06) \approx 0.35$ away from home with a statistically significant difference between the two numbers. These numbers are in agreement with empirical observations with the average players (players with estimated b_k within the 25th and 75th percentiles) producing roughly 0.4 shots per game against Manchester United during that season.

For the extended model, the estimates of the effect a player's position has on the number of shots he makes are made with respect to the central attackers. For example, the parameter value of -0.27 for the central midfielders means that they shoot on average $\exp(-0.27) \approx 76\%$ as often as the centre forwards. The fact that we estimate $\tau = 0.86 < 1$ is interesting as it implies that players tend to shoot less with each additional minute they spend on the pitch, which may be due to fatigue.

Note also how much lower the estimate of the players' random effects variance (σ_p^2) is for the extended model. This stems from the fact that, in this model, a considerable proportion of between player variability in the number of shots is explained by the players' position and the number of minutes they spend on the pitch.

Table 5.2 presents results of the likelihood ratio test with $H_0 : \tau = \pi_G = \pi_{CD} = \pi_{LRD} = \pi_{CM} = \pi_{LRM} = 0$. The null hypothesis is easily rejected even at very low significance levels indicating that the extended model provides a better fit for the data.

Recall figure 5.2 in which we have observed over-dispersion of the empirical distribution of the number of shots relative to the Poisson distribution with single mean parameter. The left panel of figure 5.3 contrasts it with the model implied distribution which assumes different values of the mean parameter for different player-game observations due to various player, opponent and home advantage configurations. The model distributions were obtained from 1000 simulations of the shot counts for each player-

	Basic model	Extended model
$\gamma^{(n)}$	-1.06 $(0.08)^{***}$	-0.09 (0.08)
$\nu^{(n)}$	0.25 $(0.02)^{***}$	0.25 $(0.02)^{***}$
Liverpool	-0.31 $(0.09)^{***}$	-0.25 $(0.09)^{**}$
Chelsea	-0.08 (0.08)	-0.08 (0.08)
Arsenal	-0.03 (0.08)	0.01 (0.08)
Wigan Athletic	0.01 (0.08)	0.00 (0.08)
Manchester City	0.03 (0.08)	0.04 (0.08)
Tottenham Hotspur	0.06 (0.08)	0.09 (0.08)
Aston Villa	0.15 (0.08)	0.16 $(0.08)^{*}$
Bolton Wanderers	0.15 $(0.08)^{*}$	0.18 $(0.08)^{*}$
Sheffield United	0.15 (0.08)	0.17 $(0.08)^{*}$
Portsmouth	0.17 $(0.08)^{*}$	0.18 $(0.08)^{*}$
Middlesbrough	0.19 $(0.08)^{*}$	0.20 $(0.08)^{**}$
Blackburn Rovers	0.20 $(0.08)^{*}$	0.22 $(0.08)^{**}$
Newcastle United	0.23 $(0.08)^{**}$	0.25 $(0.08)^{**}$
West Ham United	0.26 $(0.08)^{***}$	0.27 $(0.08)^{***}$
Fulham	0.27 $(0.08)^{***}$	0.27 $(0.08)^{***}$
Reading	0.29 $(0.08)^{***}$	0.32 $(0.08)^{***}$
Everton	0.32 $(0.08)^{***}$	0.32 $(0.08)^{***}$
Charlton Athletic	0.34 $(0.08)^{***}$	0.36 $(0.08)^{***}$
Watford	0.34 $(0.08)^{***}$	0.32 $(0.08)^{***}$
τ	-	0.86 $(0.03)^{***}$
π_G	-	-4.23 $(0.28)^{***}$
π_{CD}	-	-1.06 $(0.08)^{***}$
π_{LRD}	-	-0.88 $(0.07)^{***}$
π_{CM}	-	-0.27 $(0.05)^{***}$
π_{LRM}	-	-0.21 $(0.05)^{***}$
σ_p^2	0.95	0.32
AIC	20841.03	19545.20
BIC	20999.09	19746.36
Num. obs.	9744	9744
Num. players	506	506

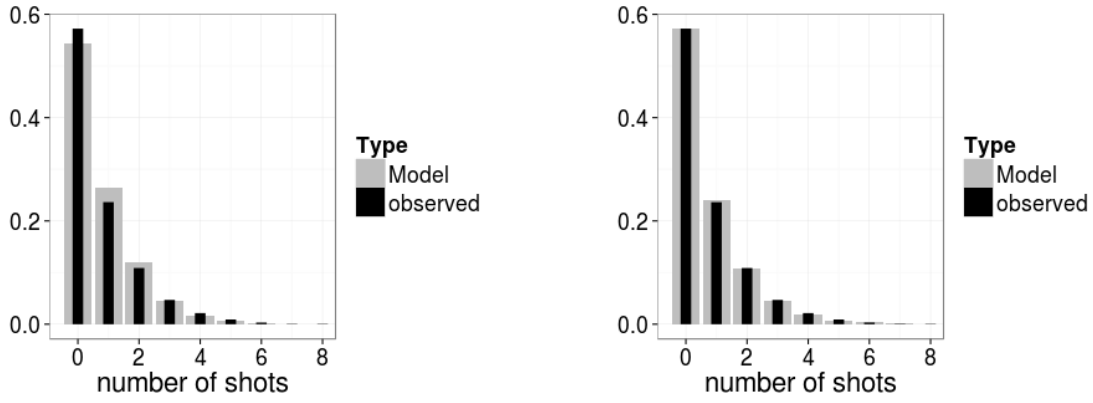
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.1: Parameter estimates (with standard errors) of the shots count models.

	Df	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
Basic model	22	-10398.51	20797.03			
Extended model	28	-9744.60	19489.20	1307.83	6	< 0.0001

Table 5.2: Likelihood ratio test for the basic and extended shot count models.

game observation given the fitted values of the mean parameter. The fitting sample mean and variance of the number of shots per player per game are 0.75 and 1.29 respectively indicating a considerable over-dispersion relative to the Poisson distribution with a single mean parameter.



Basic model Extended model

Figure 5.3: Histogram of shots per player per game (black bars) and model based frequencies.

In our basic GLMM the extra variance from the random effects and the covariates increases the dispersion of the model distribution relative to the one with a single parameter with the simulated sample mean and variance of 0.75 and 1.22 respectively meaning that approximately 95% of the over-dispersion is accounted for by the model.

There does still appear to be some excess of zero counts left unexplained by the basic model. This may be due to the fact that some players, e.g. goalkeepers or those playing very few minutes, have very little chance to take any shot. These two factors are taken into account by the extended model (presented in the right panel of figure 5.3) with the marginal distribution providing an excellent fit to the data (sample mean 0.75 and variance 1.27). The Chi-square test for the goodness of fit does not find evidence to reject the hypothesis that the data is generated by the model (p-value $= 0.4497$). In the absence of any residual over-dispersion (which would have suggested a more complicated model based on the negative binomial distribution would be appropriate), we conclude that it was reasonable to assume a mixed effects Poisson model.

5.4.2 Model fit for shots to goals conversion

The model selection procedure suggested that team and opponent fixed effects give worse fits as measured by the AIC and BIC for both versions of the shots to goals conversion model. As for the home advantage, even though its parameter estimate in the basic model was statistically significant only at 0.08 (see table 5.3), the AIC was slightly better for the model including it and so it remained in the model.

	Basic model	Extended model
$\gamma^{(y)}$	$-2.30\ (0.06)^{***}$	$-2.20\ (0.08)^{***}$
$\nu^{(y)}$	$0.13\ (0.08)$	$0.15\ (0.08)$
θ		$-0.18\ (0.07)^{**}$
ϕ_p^2	0.13	0.15
AIC	4147.7	4142.8
BIC	4166.9	4168.3
Num. obs.	4347	4347
Num. players	432	432

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.3: Parameter estimates (with standard errors) of the conversion rate model.

The estimates of the global mean and the home advantage parameters for the basic model imply that an average player is expected to score $\frac{\exp(-2.30+0.13)}{1+\exp(-2.30+0.13)} \approx 10.2\%$ of their shots on the home pitch and about 9.1% away from home. In the case of the extended model the average conversion rate decreases with the number of shots taken by the player in the game, e.g. it is $\frac{\exp(-2.20+0.15-0.18\times log(1))}{1+\exp(-2.20+0.15-0.18\times log(1))} \approx 11.4\%$ for a single shot per game on the home field, 10.2% for two shots and 9.5% for three. The more shots the player decides to take, the less prepared they may be for any given shot.

Table 5.4 presents results of the likelihood ratio test with $H_0 : \theta = 0$. The p-value for the test is 0.0083 indicating that the extended model fits the data better than the basic one.

	Df	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
Basic model	3	-2070.87	4141.73			
Extended model	4	-2067.38	4134.77	6.97	1	0.0083

Table 5.4: Likelihood ratio test for the basic and extended shot conversion models.

Figure 5.4 compares average conversion rate residuals against the number of shots for the basic and the extended model. The basic model underestimates the conversion rate in situations when a player takes only one shot per game and tends to overestimate the conversion rate when more shots are observed per player per game. The extended model corrects this deficiency quite well.

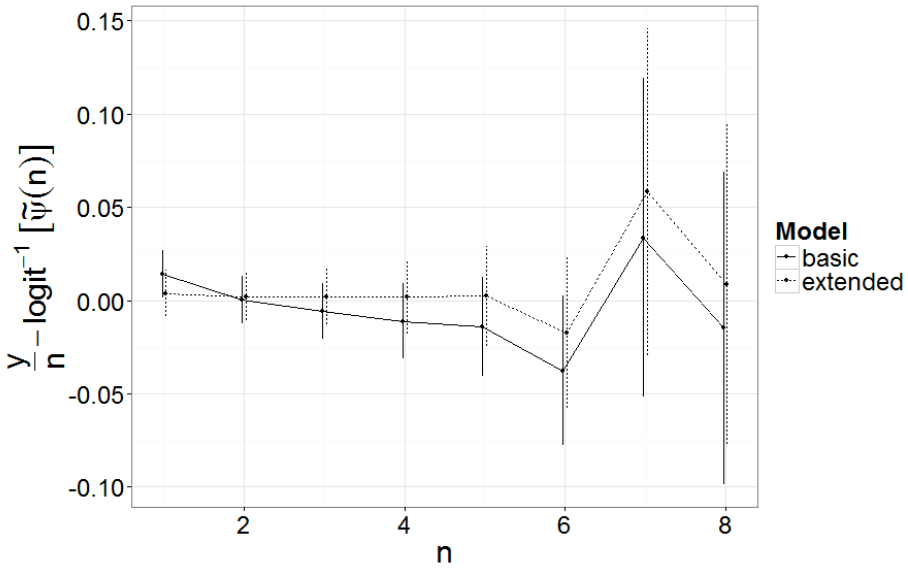


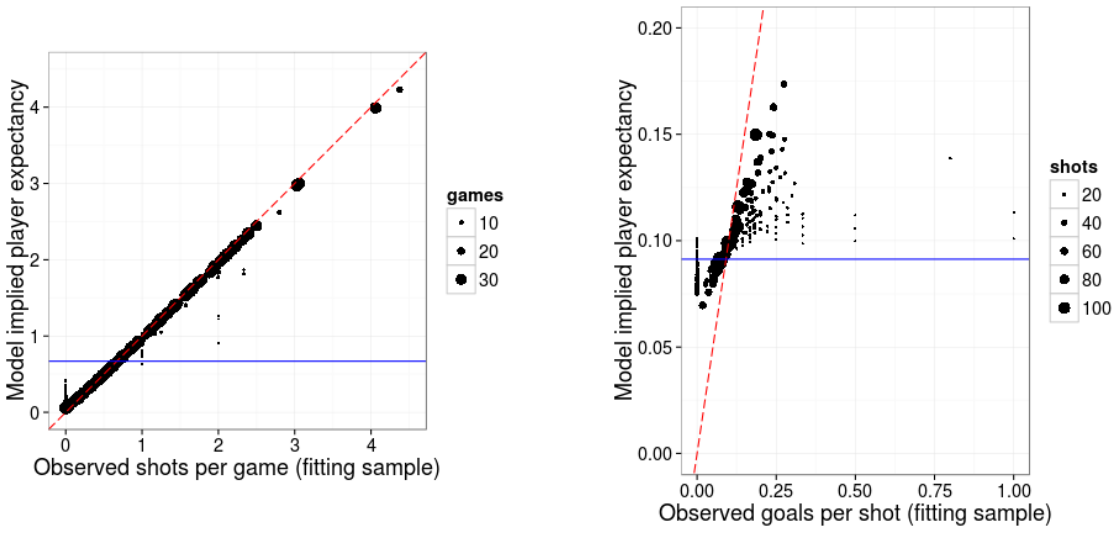
Figure 5.4: Average conversion rate residuals against the number of shots for the basic and the extended model with approximate normal 95% confidence intervals.

5.4.3 Comparing characteristics of the model fits

Figure 5.5 presents model implied player expectancies for the number of shots per game and the conversion rate against the empirical rates they are based on. Each point represents a player and the size of the point indicates the number of games or shots the rate was obtained from. The solid horizontal line is the average number of shots per game per player and the average conversion rate respectively. The dashed line is the identity function. For the sake of brevity only the basic models' predictions are presented since they do not vary noticeably from the extended models' predictions.

Predictions of shots (left of figure 5.5) are close to the empirical data they are based on with a little bit of the regression to the mean, which is stronger, the fewer games the rate was achieved on (the bigger points are closer to the identity line and the smaller ones divert slightly towards the solid mean line). This represents the fact that the uncertainty about a player's shots per game ratio is higher for players who appear in fewer games.

The regression to the mean effect is much stronger for the shot conversion model (right of figure 5.5). An extremely high rate reveals very little about the true player



Shot generation Shot conversion

Figure 5.5: Average (basic) model implied predictions of shot generation and conversion versus the average observed values per player in the fitting sample. The dashed line is the identity function and the solid horizontal line is the average number of shots per game per player and the average conversion rate respectively.

ability if it was achieved on only a small number of shots (take the points far to the right hand side as an example). The same applies to very low conversion rates, i.e. scoring no goals does not impact a player's evaluation too much if he only took one shot. In fact, these cases tell us almost nothing about his true ability so the best estimate we can come up with is close to the overall average. As a result, despite the fact that the observed conversion rate ranges from 0% to 100% the model assigns expectations from about 7% to 17%.

5.4.4 Goals predictions

We now use the fitted models to make predictions for players' number of shots and conversion rate and combine them into predicted players' goals per game ratios, as outlined in section 5.3.4. We have presented two specifications of each model (a basic specification and an extended specification for both the model for shots and the conversion rate model), and these can be combined to produce four models for predicting players' goals per game ratios. Presenting the performance of all four models allows us to identify where the performance of the models is coming from, when compared to the naive model. Recall that the naive model simply uses players goals per game ratio from the

2006/07 season to predict that same players goals per game ratio for the 2007/08 season. In order to be able to compare the model predictions to the naive ones on a level playing field we use the *averaged* model predictions, so that our predictions for the 2007/08 season are based only on information from the 2006/07 season.

Table 5.5 presents three measures of performance for the four models, and the naive model: Pearson's correlation coefficient between predicted and observed goals per game ratios, Spearman's rank correlation between predicted and observed goals per game ratios and the root mean square error (RMSE). Using the extended shot generation model (that accounts for players' positions and time on the pitch) improves goal prediction quality for both the basic and the extended shots conversion models. This is shown by the higher Pearson's and Spearman's correlations with the observed values and a lower root mean squared error (bottom two rows compared to the top two rows).

shots count	shots conversion	Pearson	Spearman	RMSE
basic	basic	0.604	0.515	0.119
basic	extended	0.595	0.515	0.120
extended	basic	0.644	0.530	0.115
extended	extended	0.637	0.530	0.116
naive		0.472	0.514	0.158

Table 5.5: Performance of the models and the naive method (338 players) in predicting goals.

Conversely, the extended shot conversion model does not yield any improvement over the basic version. This might be expected since, the quality of the fit as measured by AIC and BIC is actually very similar for the basic and the extended model (table 5.3). The extended model does offer a better fit conditionally on the observed number of shots (figure 5.4). However, note that the actual number of shots a player is going to make in a game is not known for prediction and we need to average over all the possibilities (equation (5.3.23)). This is probably where any relative advantage of the extended conversion rate model is lost.

To conclude, the extended shot generation / basic shot conversion combination performs the best predictively and is also consistently better than the naive method across the presented diagnostics.

Finally, note that whereas the best model considerably improves the Pearson correlation coefficient and reduces the prediction error relative to the naive method, the naive method performs reasonably well evaluated by the Spearman's rank correlation. Thus, if one is only concerned with ranking players, as opposed to predicting their output, then

the naive model has similar (albeit slightly worse) predictive power when compared to our model.

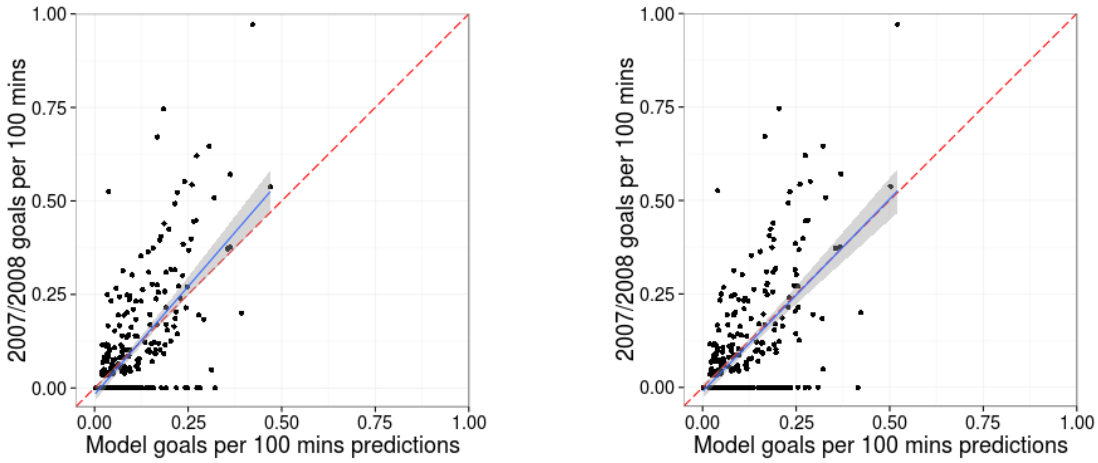
Table 5.6 lists: the top 15 goal scorers per minute according to the model; the teams the 15 players played for; the 2007/08 predictions; and the corresponding observed values. Comparing the model predictions to the naive ones, we can observe how the model regresses the empirical values to the mean as the model predictions for these top performers are all lower than the naive predictions. This appears to be the right thing to do as the model is closer to the actual 2007/08 numbers for 10 out of the 15 players on the list. Of course this is only a limited sample and just exemplifies the overall model superiority for all the players reported previously in table 5.5.

		Team				
Rank	Name	2006/07	2007/08	Model	Naive	Actual
1	Drogba	Chelsea	Chelsea	0.47	0.62	0.54
2	Ronaldo	Man. Utd.	Man. Utd.	0.43	0.46	0.97
3	Van Persie	Arsenal	Arsenal	0.40	0.66	0.20
4	Rooney	Man. Utd.	Man. Utd.	0.36	0.40	0.57
5	Viduka	Middlesbrough	Newcastle	0.36	0.66	0.38
6	Crouch	Liverpool	Liverpool	0.36	0.51	0.37
7	Vaughan	Everton	Everton	0.32	0.58	0.00
8	Berbatov	Tottenham	Tottenham	0.32	0.43	0.51
9	Kuyt	Liverpool	Liverpool	0.31	0.45	0.05
10	Defoe	Tottenham	Tott./Portsm.	0.31	0.35	0.65
11	Saha	Man. Utd.	Man. Utd.	0.29	0.30	0.18
12	Cole	Portsmouth	Sunderland	0.28	0.41	0.00
13	McCarthy	Blackburn	Blackburn	0.28	0.44	0.19
14	Zamora	West Ham	West Ham	0.27	0.41	0.00
15	Adebayor	Arsenal	Arsenal	0.27	0.34	0.62

Table 5.6: Predicted and actual 2007/08 goals per 100 minutes for the top 15 model predicted goals per 100 minutes scorers based on season 2006/07 data.

Figure 5.6 presents the relationship between the best model predictions and the 2007/08 actual values which can be compared to the corresponding relationship for the naive predictions presented in figure 5.1. The dashed line is the identity function and the solid line with the shade is a linear regression fit plus a 95% confidence interval. Focusing on the *averaged* predictions in the left panel first, the relationship with the empirical data is stronger for the model predictions than for the naive method which confirms the results in table 5.5. Also note that the bias present in predictions made from the naive

model, is considerably reduced, as shown by the linear fit being much closer to the identity function. Actually, it could be argued that this effect is over corrected for by the *averaged* model predictions which now appear to be underestimating good goal scorers and slightly overestimating the worse ones. It appears that some part of the players' ability to score is not captured by these model predictions.



Averaged predictions *Complete* predictions

Figure 5.6: Goals scored per 100 minutes in season 2007/08 versus the (extended/basic) model player predictions. The dashed line is the identity function and the solid one is the linear model fit.

One of the reasons for this small bias may be the fact that opponent strength is averaged out from the predictions. Good players tend to play on better teams, which in turn means that the average level of the opponents that they face will generally be lower than for worse players. The strength of the opponents is contained in the model so ignoring this information in the *averaged* predictions, as we did for the reasons outlined in section 5.3.4, leads to lower numbers for good players and higher numbers for worse players than taking the opponents information into account would give. Similarly, it may be wrong to assume that players would feature in various tactical positions with the same frequency as they did in the fitting sample, like we did when making the *averaged* predictions. It may well be the case that good scorers in one season tend to be moved to positions favouring goal scoring even more in the consecutive one and accounting for this fact would result in boosting their predictions. In other words, this slight prediction bias does not necessarily indicate a flaw in the model itself but results from a compromise we settled for at the prediction step so as not to give the model an unfair advantage

of "future knowledge" when comparing its performance with the naive method. The right panel of figure 5.6 appears to confirm this as any bias exhibited in the left panel (by the solid line) has been considerably reduced when the predictions are made based on all the data available to (or controlled by) the manager at the start of the game.

5.5 Discussion

The model presented in this chapter can serve as a useful tool for predicting players' productivity in terms of what is arguably the most important aspect of the game - goals. The model provides more accurate out-of-sample predictions of a player's number of goals in a season than a naive method of using the previous year's goals scored per minute ratio. Furthermore, the model corrects systematic biases present in the naive method whereby the ability of good performers is overestimated whilst the ability of poor performers is underestimated.

The improvement in predictions comes from two sources. First, player abilities are assumed to come from a common distribution in which extreme values are unlikely. The effect of this is that it takes many observations of unusual performance for the model to believe that the underlying skill is really that exceptional. Second, the model breaks down the goal scoring process into a shot generation process and a shots to goals conversion process. These processes are allowed to consist of both the inherent player ability and random chance.

The analysis presented in this chapter serves two purposes:

- to support the argument of chapter 3 that, if player valuation is the aim of player assessment, then the method used to conduct such assessment must recognise that individual performance depends on the player's underlying skill as well as factors beyond his control, and

- as a standalone tool for evaluating players' goalscoring abilities. Parts of the model presented here will also be used in chapter 7 as components of a bigger model of the football game in which a player's shooting skill is responsible for a part of his total contribution to the team success.

An interesting secondary finding here is that the player's team attacking ability does not appear to be a predictor of the number of shots a player has. This may seem counter-intuitive in that one might expect better attacking teams to generate more shots for players on that team. However, the model suggests this is not the case. A player of fixed

skill will generate as many shots playing for a top team as he will playing for a bottom team. He will be a relatively less important player on a better team thus executing a smaller percentage of a bigger total number of shots as opposed to a bigger proportion of a smaller total he would account for on a lesser side. A similarly interesting result is that the process of converting shots to goals turns out to be a less predictable skill, so that when predicting future performance, the observed performance is regressed to the mean more heavily than in the case of the shot generating process.

One natural extension to the methods proposed here would be to account for correlation between the random effects in the specification of the two models. A tentative study in this direction revealed no statistically significant relationship between the shots count predictions and the residuals from the conversion model aggregated on a player level nor the other way around. This suggests that it is not the case that players who are predicted to shoot a lot tend to have their shot conversion rate over- or underestimated by the conversion model. If that was the case (i.e. if there was some significant pattern), it would indicate that there is some common variability in the two processes at the player level which is not captured by the current model structure. In the absence of such clear relationships we conclude that increasing the model complexity in this direction is not the priority at the current stage and leave it for subsequent studies to investigate. Another area of advancement would be to include the identity (and hence ability) of goalkeepers in the shots to goals conversion part of the model. Finally, the shot conversion part of the model could be improved by accounting for the shot location or defensive pressure on the executing player. How such additional covariates can improve a model will be demonstrated in chapter 6 in which we analyse passing performance.

The application of the methodology proposed here is not limited to goal scoring statistics. It seems reasonable to believe that inference about player skill from any other statistic could benefit from such a random effects formulation.

Chapter 6

Adding context to passing analysis

In chapter 5 we demonstrated how statistical modelling can be useful for evaluating an isolated aspect of individual performance (shooting). In particular, we showed that capturing players' skill with random effects proves beneficial for predictive purposes by regressing predictions to the mean (relative to the past performance) more when players were observed fewer times in the past. These findings are used in this chapter to study a different aspect of performance, namely passing the ball to another player. Additionally, the main focus of this chapter is to analyse the performance in as much context as possible. This means accounting for many covariates describing the situation in which the performance occurred. This was done to a small extent in chapter 5, for example, by accounting for the home field advantage in the models. In this chapter we take this much further and use slightly more sophisticated methods to capture the relationship between the covariates describing the environment in which an action was performed and the outcome of that action.

The chapter is structured as follows. In section 6.1 we provide some background and motivate the need for statistical modelling of passing. Section 6.2 outlines the data employed in this analysis and section 6.3 presents a model of passing. In section 6.4 the results of fitting the model are presented and then used to make various types of predictions, which are later compared with the empirical data. We conclude with some discussion in section 6.5. The code used to fit the models of this section can be found in appendix D.

This chapter is based on the paper of Szczepański and McHale (2015).

6.1 Background and motivation

In football passing is the most common way to move the ball around the pitch towards the opposition's goal before a shot can be made in an attempt to score. As a result, passing the ball is one of the key skills of football players. It has been a subject of statistical analysis at least since Reep and Benjamin (1968), who note that

> "There are a number factors affecting the likelihood of a successful rth pass:
>
> 1. the positions of the players between whom the pass is attempted and the defending players who try to intercept;
> 2. the relative skills of the players and the effectiveness and confidence with which those skills are applied at this particular stage of the game."

Perhaps partially because their data is not broken down by players, they claim that "In evenly matched teams playing under the conditions normally obtaining in good class football (...) the second of these factors does not vary widely from one attempted pass to another (...)" and proceed to analyse length of a passing sequence as a random variable without consideration for individual skill. More recent academic literature appears to approach the problem from the other extreme by attributing all the variation in observed pass outcomes to individual skills of executing players. Passing statistics are included as a key component of some recently developed player rating systems (for example Duch et al., 2010; Oberstone, 2011; McHale et al., 2012). However, the statistics most commonly used to represent player's passing skill are: the number of completed passes and the completion rate (which is simply the number of successfully completed passes divided by the total number of attempted ones). These are particularly crude metrics with many flaws outlined very well by Steven Houston, head of technical scouting at a German football club, HSV Hamburg, in his interview with Sky Sports (Bate, 2012):

> "I think if you just looked at the players with the highest pass completion you would just be getting defensive midfielders like (John Obi) Mikel (at Chelsea) who tend to make shorter and less incisive passes. Passes in the final third are much more difficult to make and through-balls are passes that create higher quality chances for forwards rather than, say, a cross or something like that."

Putting empirical pass completion rates in the context of the part of the pitch they were executed from or their direction is certainly an improvement over quoting a single value

per player. However, it leads to some new problems: why for example, should we focus on passes in the final third of the pitch and not, say, the final quarter? Making such an arbitrary decision is clearly not attractive from a scientific standpoint. A more appropriate metric to measure passing ability would not simply be a player's pass completion percentage in the final third of the pitch, but include information on the origin and intended destination of the pass.[1] To do this, and to measure passing ability properly, it is clear there needs to be a more scientific approach to addressing the problem of identifying passing ability.

6.2 Data

The fitting sample consists of passes attempted in the English Premier League in season 2006/07 (season 2007/08 is left for model validation) provided by Opta. This database contains characteristics of each pass such as: the executing player, game time of the event in minutes and seconds, pitch coordinates of the pass origin, pitch coordinates of the pass target, a header pass indicator, and a success indicator among many others.

In order to focus our attention on what can be expect to be a roughly homogeneous group of events we select all open play passes between two outfield players giving us $I = 253090$ events to analyse. There are $K = 481$ outfield players among $T = 20$ teams in the fitting sample.

We use information on player k's playing position in game j anticipated according to the algorithm given in chapter 4 in model fitting (variables $\left(\bar{x}_{k,j}, \bar{\bar{y}}_{k,j}\right)$). In presentation of the results we group players by positions to which they were assigned most frequently.

6.3 Methods

In this section we first motivate the use of the particular set of covariates in the model (section 6.3.1). Then we outline the theory of Generalized Additive Mixed Models (section 6.3.2) before specifying the model used for estimating passing ability of players (section 6.3.3). Finally, we describe several types of predictions used to validate the model (section 6.3.4).

[1] For example, recall figure 4.2 and the discussion around it for why just splitting pass completion rate by the zone of pass origin is problematic.

6.3.1 Factors influencing pass success and their proxies

Our approach to estimating each player's passing ability is to model the probability of each pass being successful, given information on the environment in which the pass was made and, of course, the identity of the player making the pass. We thus want to create a set of covariates describing the situation of the pass from the data we have available to us.

There are a number of factors which potentially influence the outcome of a pass. We believe these include:

- The inherent **skill** of the player passing the ball.
- The degree of **control** the executing player has on the ball when attempting the pass. For example, a ball bouncing at waist height is more difficult to pass than a ball that is stationary on the ground.
- The level of **pressure** the opposition team put on the executor of the pass.
- The **distance** of the attempted pass.
- The level of **pressure** the opposition team put on the player receiving the ball.
- **Familiarity** with the type of situation the pass is attempted in. For example, home team players may know the surface better. Similarly a winger is more likely to make a successful cross from a wide area of the field than a central forward who finds himself in this area only occasionally.

Of these factors only pass distance can be derived directly from our dataset. The information about the other factors is not directly available, however, we can develop proxies for these factors using the data. For instance, we do not have information on how much pressure is being placed on the player receiving the pass, but we can hypothesise that it will generally be more the closer he is to the opposition's goal yet may be less as the opposition players tire due to fatigue towards the end of the match. This leads us to the idea of using the intended destination of the pass and the timing of the pass as proxies for pressure on the player receiving the ball and to experiment with including them as covariates in our model to estimate the probability of a pass being successful.

Continuing with this mode of thought, we create several variables derived from the data to proxy factors influencing pass success. Each of these variables can be considered to influence the success of the pass in a number of ways.

- The origin of the pass (x and y coordinates on the pitch) and the intended destination of the pass (which we denote by the x_{end} and y_{end} coordinates on the pitch) proxy the pressure on the passing player, the pressure on the receiving player and the difficulty of the pass in terms of distance.

- The time since the previous pass (which we denote δt) and the pass number in the current sequence of passes for that team (e) proxy the control the passing player may have of the ball and the pressure the opposition players are placing him and the receiving player under.

- The game time (t, in minutes) proxies the pressure the passing player and the receiving player might be under. In addition, it may reflect the fatigue of the player with the ball and effect the passing player negatively.

- Whether the pass is performed with the player's head or foot, or indeed whether the previous pass was executed with the head or foot serve as proxies of how well the player is in control of the ball. We give this covariate the symbol a.

- From the dataset we can extract information on whether each action followed a duel[2] and for pass events this serves as a proxy for the pressure the passing player might be under. We denote duels as d_a for an aerial duel and d_t for a duel on the ground. These variables equal 1 if the pass immediately follows a duel. A third duel variable, d_s, indicates that the pass was made by the player involved in the duel. For example, if player A successfully tackles an opponent, takes control of the ball and makes a pass, $d_t = 1$, $d_a = 0$ and $d_s = 1$ for this pass. If following the tackle the ball falls to his team mate B instead, $d_s = 0$ for B's pass that follows.

- Whether the player is at his home ground (h) serves as a proxy for familiarity with the conditions.

- Lastly, we consider the player's position as a proxy for whether he is under pressure from the opposition and whether the player he is passing to is under pressure. We denote this in terms of the average x-y coordinates for the k-th player in games before the j-th match as $\bar{x}_{k,j}$ and $\bar{\bar{y}}_{k,j}$. We discuss the definition of and the meaning of this variable in more detail below.

The resultant set of covariates are defined, with their symbols in table 6.1. Also included is the factor or factors that each covariate is serving as a proxy for and whether

[2] According to the data provider's definition: "A duel is a 50-50 contest between two players of opposing sides in the match."

we include a lag of the variable in the model. Whether or not these covariates carry information about pass success rate can be verified when including them in our statistical model described in section 6.3.3.

				Approximated factor						
Type	Covariate	Symbol	Lags	Control	Passing player pressure	Distance	Receiving player pressure	Familiarity		
continuous	origin and destination	x, y, x_{end}, y_{end}	0		✓	✓		✓		
	time since previous pass (in seconds)	δt	0, 1	✓	✓			✓		
	pass number in this sequence of passes	e	0	✓	✓			✓		
	game time (in minutes)	t	0	✓	✓			✓		
	player position in the game	$\bar{x}_{k,j}, \bar{\bar{y}}_{k,j}$	0		✓			✓		✓
ind.	headed pass	a	0, 1	✓						
	duel (aerial, tackle, same player)	d_a, d_t, d_s	1	✓	✓					
	home advantage	h	0							✓

Table 6.1: Covariates used to proxy factors influencing pass success. Lags indicate whether the value corresponding to the executed pass ($lag = 0$) or the previous pass ($lag = 1$) is considered.

6.3.2 Generalized Additive (Mixed) Model

The model proposed for the outcome of a pass in a football game belongs to the class of Generalized Additive Mixed Models. Before we specify its exact form, in this section we outline the idea of the Generalized Additive Model and the Generalized Additive Mixed Model.

Generalized Additive Model

Generalized Additive Model (Hastie and Tibshirani, 1986) is an extension of the GLM, outlined in appendix A.1, such that the response y can depend on some smooth functions of covariates (in addition to their linear combination) through the linear predictor:

$$\eta_i = X_i\beta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots . \tag{6.3.1}$$

The f_j terms are smooth functions. Such functions offer a great deal of flexibility in specifying the relationship between covariates and a response variable. In the estimation

procedure each of them is represented by a sum of some known basis functions weighted by regression coefficients that need to be estimated:

$$f(x) = \sum_{k=1}^{q} b_k(x)\alpha_k \tag{6.3.2}$$

The basis functions b_k form a basis of a space the smooth functions belong to. Note that substituting (6.3.2) into (6.3.1) gives a Generalized Linear Model.

The risk attached to the flexibility of this approach is that, given a sufficiently large basis, the smooth functions can overfit the observed data with a shape that is unlikely to represent the underlying data generating process. There is a trade off between the smoothness of a function and the extent to which it fits the observed data. The general idea to express this trade off is to penalise a loss function L used in the estimation, e.g.:

$$L_p = L + \lambda \alpha^T \mathbf{S} \alpha \tag{6.3.3}$$

and optimise the penalised function L_p instead of L. $\mathbf{S}$ is a matrix of known coefficients implied by the choice of the basis. The smoothness of the resulting function depends on λ: the bigger the weight assigned to the penalty, the smoother the function. The value of λ can be selected using cross validation methods (see Wood, 2006, p.172-189).

There are many ways of choosing the basis, which is a set of basis functions that define the space supposed to contain an approximation of the target function. In this chapter for smooth functions of a single covariate we use *thin plate regression splines*, which approximate *thin plate splines*. *Thin plate splines* can be shown to provide an optimal solution to a smoothing problem given a formula for the penalty for non-smoothness (or "wiggliness"). This formula can be defined quite flexibly and leads to a set of basis functions that can be split into completely smooth functions, for which the "wiggliness" penalty is always zero, and remaining "wiggly" functions. In all, *thin plate splines* are a very neat theoretical concept, however, the computational cost involved in their calculation makes their use impractical. *Thin plate regression splines* provide a useful approximation at a lower computational cost. The general idea behind this approximation is to truncate the space of the "wiggly" component[3] of the *thin plate spline* based on the eigen-decomposition of the matrix consisting of corresponding basis function values. See Wood (2006, p.154-158) for details.

One of the properties of the *thin plate regression splines* (and the *thin plate splines* in general) is that, when used as smooths of multiple variables, they treat the smoothness

[3]Consisting of basis functions that are not completely smooth in the sense of the specified "wiggliness" penalty.

of the fitted spline equally in all dimensions. In our application there is no reason to believe that such isotropy exists. For example, the smoothness along the pitch may well be different than across it, even if we scale both dimensions to the same real scale (e.g. metres). Therefore, for smooth functions of more than one variable we use *tensor product smooths* which are not isotropic in general. The general idea here is to start with a smooth of a single covariate, e.g. like f for the covariate x in equation 6.3.2 and turn them into smooths of an additional covariate, say z, by allowing the coefficients α_k to vary smoothly with that covariate. The resulting function is smooth with respect to x and z. For details see Wood (2006, p.162-167).

Generalized Additive Mixed Model

Generalized Additive Mixed Model (GAMM) is an extension of the GLM combining features of GAM from section 6.3.2 and GLMM from section 5.3.1. In GAMM the response variable y depends on a linear combination of fixed effects, their smooth functions as well as a vector of random effects through the linear predictor:

$$\eta_i = X_i\beta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \cdots + U_i b\,. \tag{6.3.4}$$

The vector of random effects b is assumed to come from a multivariate normal distribution:

$$b \sim \mathcal{N}(0, G_\theta) \tag{6.3.5}$$

where G_θ is a covariance matrix depending on some unknown parameters θ.

The Generalized Additive Mixed Model has an equivalent Generalized Linear Mixed Model representation in which regression splines (of the form in equation (6.3.2)), representing the smooth functions f_s, are re-parametrised as fixed and random effects and absorbed into corresponding components of a mixed model (Wood, 2006, p.316-318). The parameters governing the smoothness of the smooth functions (like λ in equation (6.3.3)), are treated as variance parameters corresponding to the random effects and can be estimated using the methods of section 5.3.1.

6.3.3 Pass outcome model

Let $W = \left(1, a^{(n)}, a^{(n-1)}, d_a^{(n-1)}, d_t^{(n-1)}, d_s^{(n-1)}, h^{(n)}\right)$ be a matrix of fixed effects with columns consisting of all the indicator variables listed in table 6.1 and let W_i be a row of this matrix corresponding to the i-th pass. The superscript $n-L$ indicates a value lagged by L, i.e. corresponding to the event L before the current one, e.g. $d_a^{(n-1)} = 1$

for all the passes preceded by an aerial duel and 0 for the rest. Furthermore, let b be a vector of random effects with the first $K = 456$ elements representing the passing ability of players and the remaining $2 \times T = 2 \times 20$ elements corresponding to the ability of the passing player's team and the ability of the opposition facilitating and hampering pass execution respectively, so that

$$\mathbf{b}_{(K+2T)\times 1} = \left((\mathbf{b}^{(p)})^T \quad (\mathbf{b}^{(t)})^T \quad (\mathbf{b}^{(o)})^T \right)^T .$$

We assume a Generalized Additive Mixed Model (GAMM) (Lin and Zhang, 1999) in which the outcome of the i-th pass ($o_i = 1$ for a successful pass, 0 otherwise) has a Bernoulli distribution with the probability of success represented by the inverse logit function of the linear predictor η_i. A GAMM is an extension of a Generalized Linear Model in which the linear predictor is allowed to involve smooth functions of covariates as well as random effects:

$$(o_i|\eta_i) \sim \text{Bernoulli}\left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right) \tag{6.3.6}$$

where

$$\begin{aligned}\eta_i =& W_i\beta + Z_i b + s_1\left(t_i^{(n)}\right) + s_2\left(\tilde{e}_i^{(n)}\right) + s_3\left(\delta t_i^{(n)}\right) + s_4\left(\delta t_i^{(n-1)}\right) + \\ &+ s_5\left(\bar{x}_{[k,j](i)}^{(n)}, \bar{\bar{y}}_{[k,j](i)}^{(n)}\right) + \\ &+ s_6\left(x_i^{(n)}, x_{end,i}^{(n)}, \left|y_i^{(n)} - 0.5\right|, \left|y_{end,i}^{(n)} - 0.5\right|\right) + \\ &+ s_7\left(x_i^{(n)}, x_{end,i}^{(n)}, \left[y_i^{(n)} - 0.5\right] \times \left[y_{end,i}^{(n)} - 0.5\right]\right).\end{aligned} \tag{6.3.7}$$

Recall that indices j and k mean the j-th game of the k-th player and $\bar{x}_{[k,j](i)}^{(n)}$, $\bar{\bar{y}}_{[k,j](i)}^{(n)}$ are average coordinates of all the events of the player executing the i-th pass in his previous games (see figure 4.3). Z_i is a row of a design matrix selecting the elements of the random effects vector b corresponding to the player executing the i-th pass, the team he plays for and the opposition. $s_1, ..., s_7$ are smooth functions. We note that we truncate e, the pass number in the sequence of passes, so that the covariate we use in the model is actually $\tilde{e} = \min(e, 15)$. This is because the shape of the fitted smooth function corresponding to this covariate suggests that it is fitting noise for values above 15. Finally, for the random effects we assume:

$$\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \Sigma(\sigma)) \tag{6.3.8}$$

where $\Sigma(\sigma) = \Sigma(\sigma_p, \sigma_t, \sigma_o)$ is a $(K+2T)$ dimensional diagonal skill covariance matrix with the first K elements on the diagonal equal to the player skill variance, σ_p^2, the

next T elements equal to the player's team skill variance, σ_t^2, and the final T elements equal to the opposite team ability variance σ_o^2. This reflects a belief that extremely good (and bad) players and teams are less common than the *average* ones. The values of the random effects are themselves of interest here since they can be interpreted as the passing ability of the players, and the abilities of each team to facilitate and hamper passing.

The pass location (origin and target) component is described with two functions: s_6 and s_7. To some extent we want to impose symmetry on pass completion with respect to the left and the right (along the y axis) side of the pitch. For example, everything else being equal, passes 10 metres left from the axis going through the centre of both goals can be expected to have the same chance of success as passes 10 metres right from it. The same for passes to this point. This belief is reflected in the use of the absolute values in the s_6 function. However, we want to distinguish passes from a point 10 metres right from the axis played 1 metre to the right and 21 metres to the left (for the same x). The s_6 function does not allow it (the values of $|y_i^{(n)} - 0.5|$ and $|y_{end,i}^{(n)} - 0.5|$ are the same for both these passes). For this reason we introduce the $[y_i^{(n)} - 0.5] \times [y_{end,i}^{(n)} - 0.5]$ term which is positive for passes played to the same side of the pitch and negative for ones crossing the axis of the pitch. We put it in function s_7 together with x and x_{end} covariates in order to allow its effect to differ with the distance of the pass origin and target from both ends of the pitch.

We use *thin plate regression splines* for smooth functions of a single covariate, $s_f, f = 1,\ldots,4$ and *tensor product smooths* for functions of multiple covariates: s_5, s_6 and s_7.

6.3.4 Prediction types

Given the model fit, it is possible to calculate several pass completion rate predictions which will be of interest in the analysis:

- **full predictions**, $\hat{p}_i^{(f)}$, by substituting the fixed parameters in equation (6.3.7) with their estimates $\hat{\boldsymbol{\beta}}$, the random effects with their predictions $\left(\hat{\mathbf{b}}^{(p)}, \hat{\mathbf{b}}^{(t)}, \hat{\mathbf{b}}^{(o)}\right)$ and using the fitted smooth functions of the remaining covariates. We also calculate the average $\bar{\hat{p}}_{k,s}^{(f)}$ of this value for all the k-th player's passes in both seasons s.

- **average player predictions**, $\hat{p}_i^{(e)}$, calculated the same way as the **full** predictions except that players' random effects $\mathbf{b}^{(p)}$ are set to zero. This value is predicted pass completion probability by an average player and can be thought of as a proxy for the ease of pass. We also calculate the average $\bar{\hat{p}}_{k,s}^{(e)}$ of this value for all the k-th

player's passes in both seasons s.

- **prediction for an "average" difficulty pass**, in season 2006/07 by the k-th player $\hat{p}^{(av)}_{k,2006/07}$ calculated using the following procedure:

  1. For each i-th pass the linear predictor η_i is calculated in the same way as for the average player predictions, i.e. by setting players' random effects $\mathbf{b}^{(p)}$ to zero and all the other parameters to their estimates.
  2. We calculate the average of such linear predictor for all the passes in season 2006/07.
  3. The above averaged linear predictors are added to the players' random effect predictions $\hat{\mathbf{b}}^{(p)}$.
  4. Finally, we put the values on the probability scale by calculating the inverse logit function of the above adjusted linear predictors.

  These predictions are used as measures of players' passing ability. Of course, we can use just the player's random effect predictions $\hat{\mathbf{b}}^{(p)}$ instead for this purpose. We use this transformation to put them on the scale of pass completion rate just for the ease of interpretation.

- **fixture specific prediction for an "average" difficulty pass**, $\hat{p}^{(pto)}_{k,j}$, for player k in fixture j calculated using the following procedure[4]:

  1. First, the average linear predictor for passes in season 2006/07 is calculated in a similar way as for the **prediction for an "average" difficulty pass**, except that for each pass we set all the random effects: for players, their teams and the opposition, to 0 (and all the other parameters to their estimates).
  2. For each player k in each fixture j in season 2007/08 the above averaged linear predictor is added to the predictions of random effects for players, their teams and the opposition.[5]
  3. We put the values on the probability scale by calculating the inverse logit function of the above adjusted linear predictors.

  For each fixture j we calculate the average of these predictions for the home, $\bar{\hat{p}}^{(pto)}_{h,j}$, and the away team, $\bar{\hat{p}}^{(pto)}_{a,j}$. We also calculate corresponding averages of

[4]The (pto) abbreviation stands for "player, team, opponent".

[5]For the teams newly promoted to the league in season 2007/08, which do not have their own random effect predictions, we use averages of the respective random effects of the teams relegated from the league in season 2006/07.

naive predictions, $\bar{\delta}_{h,j}$ and $\bar{\delta}_{a,j}$, according to which in the j-th fixture of season 2007/08 the k-th player is expected to complete passes at his average rate in the fitting sample (season 2006/07). These two sets of averages are used as predictors of the score in the j-th fixture in order to evaluate the utility of the model in comparison to the raw pass completion rate as a measure of player skill.

6.4 Results

6.4.1 GAMM estimation results

Table 6.2 presents estimates of the parametric model terms contained in the vector β.

Covariate	Name	Estimate (Std. Error)
1	Intercept	$1.28\ (0.03)^{***}$
$a^{(n)}$	Headed pass	$-1.22\ (0.02)^{***}$
$a^{(n-1)}$	Previous pass was headed	$-0.21\ (0.02)^{***}$
$d_a^{(n-1)}$	Previous event was an aerial duel	$-0.51\ (0.05)^{***}$
$d_t^{(n-1)}$	Previous event was a tackle	$0.22\ (0.04)^{***}$
$d_s^{(n-1)}$	Previous event was a duel involving the pass executor	$0.13\ (0.04)^{**}$
$h^{(n)}$	Pass executor plays for the home team	$0.09\ (0.01)^{***}$
	σ_p	0.16
	σ_t	0.11
	σ_o	0.06
	AIC	214538.59
	BIC	221565.21
	Num. obs.	242478

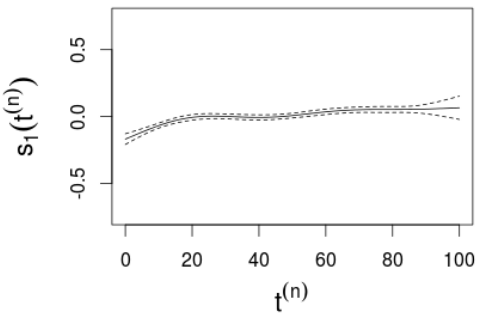
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6.2: Estimates of the parametric terms (respective elements of vector β) in the pass success model.

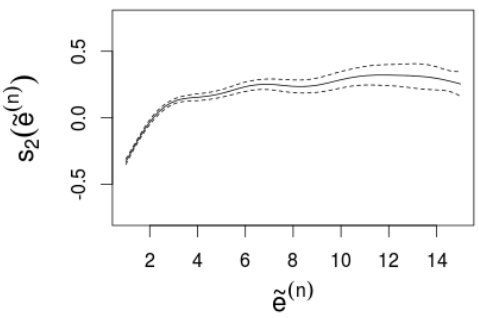
As expected, headed passes ($a^{(n)} = 1$) are less accurate than ones played with a foot and they also have a negative impact on the following pass ($a^{(n-1)} = 1$), perhaps because they force the receiver to either head it again or take more time to control the ball and bring it down to his foot. Headed passes are generally less accurate as the executing player has less control on the ball than when passing with a foot. If a pass is a direct result of winning an aerial duel ($d_a^{(n-1)} = 1$), the chance of its completion drops further but this effect is somewhat compensated for if the same player wins the duel and makes the pass ($d_s^{(n-1)} = 1$). Passes made immediately after interrupting the opponent's

possession with a tackle are generally more likely to be completed, perhaps because the opposition needs some time to reorganise themselves (for instance, the tackled player may still be on the ground when the pass is made).

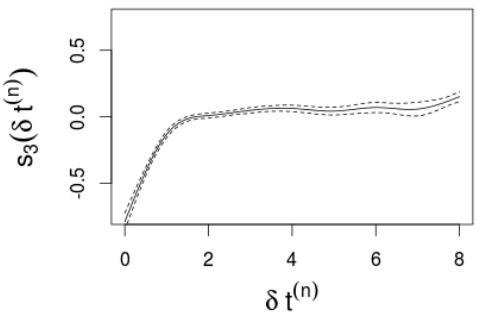
Figure 6.1 presents the estimated smooth functions of time related covariates on the scale of the linear predictor. Passes made under time pressure (bottom-left panel) have a



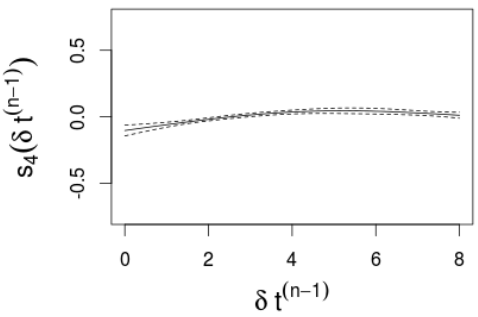
(a) Game time (in minutes)



(b) Pass number in a given possession



(c) Time since the previous pass (in seconds)



(d) Time between the previous pass and the one before

Figure 6.1: Time related component smooth functions on the scale of the linear predictor of pass success. The dashed lines are Bayesian 95% credible intervals.

relatively low probability of success as do those made before teams establish possession having exchanged only a few passes (top-right). Interestingly, it is generally easier to pass later in the game (top-left) as teams get tired and are not able to apply pressure on the passer as effectively as they might have done in the early stages of the match, however, the effect is rather small.

The success of a pass is also related to the executing player's tactical position in that game as evidenced in figure 6.2. Controlling for everything else, defenders seem to have it easier than all the other players, followed by wingers and central midfielders. Central forwards are usually faced with the toughest task. To appreciate why this may be true, consider a pass from near one's own goal. The possession phase tends to be different

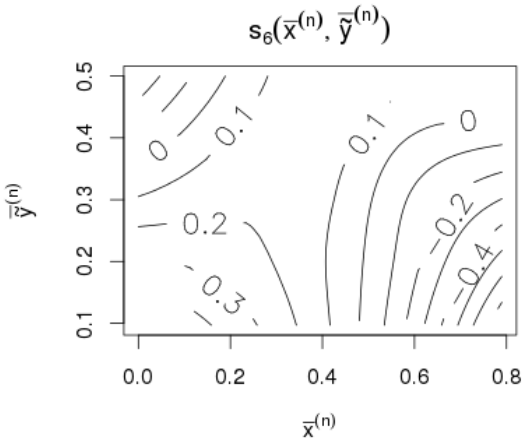


Figure 6.2: Linear predictor of the pass completion rate as a smooth function of the executing player's anticipated position (average position of all events involving him in previous games). $\bar{\bar{y}}^{(n)}$ averages the distance from the axis going through the centre of both goals so as not to allow values of y to cancel out for players who switch sides of the pitch.

if such a pass is performed by an offensive player (e.g. may be a clearance following a defensive corner) than if it is made by a defender (for whom it is a standard pass location in an established possession). Generally, there will be fewer opposition players in the area of the pitch surrounding the pass executing player in the latter case.

Presenting the impact of pass origin and target on the probability of pass success is a bigger challenge since in the model the linear predictor for the latter depends on the pitch coordinates through two multidimensional functions. Figure 6.3 is an attempt to address this challenge. The idea is to fix

- the location of the pass origin at a certain point (the •'s in the figure);
- the indicator variables at the most commonly occurring values;
- the continuous variables at the closest observed value to the median.

and draw contours of the linear predictor against the location of the pass target.

Firstly, note that the designed symmetry with respect to the axis going through the centre of both goals. Apart from this, passes played towards the opponent's goal (along the horizontal axis of the •'s) tend to have a smaller chance of success than ones played sideways or, in particular, backwards. Furthermore, passing to either of the wings is more likely to succeed than passing straight ahead. This is because the defending team tends to concentrate their efforts on not allowing the team on the ball to get into convenient shooting positions straight ahead of goal. Finally, the probability of success tends to dip just ahead of the passing player (assuming he is facing the opponent's goal). This

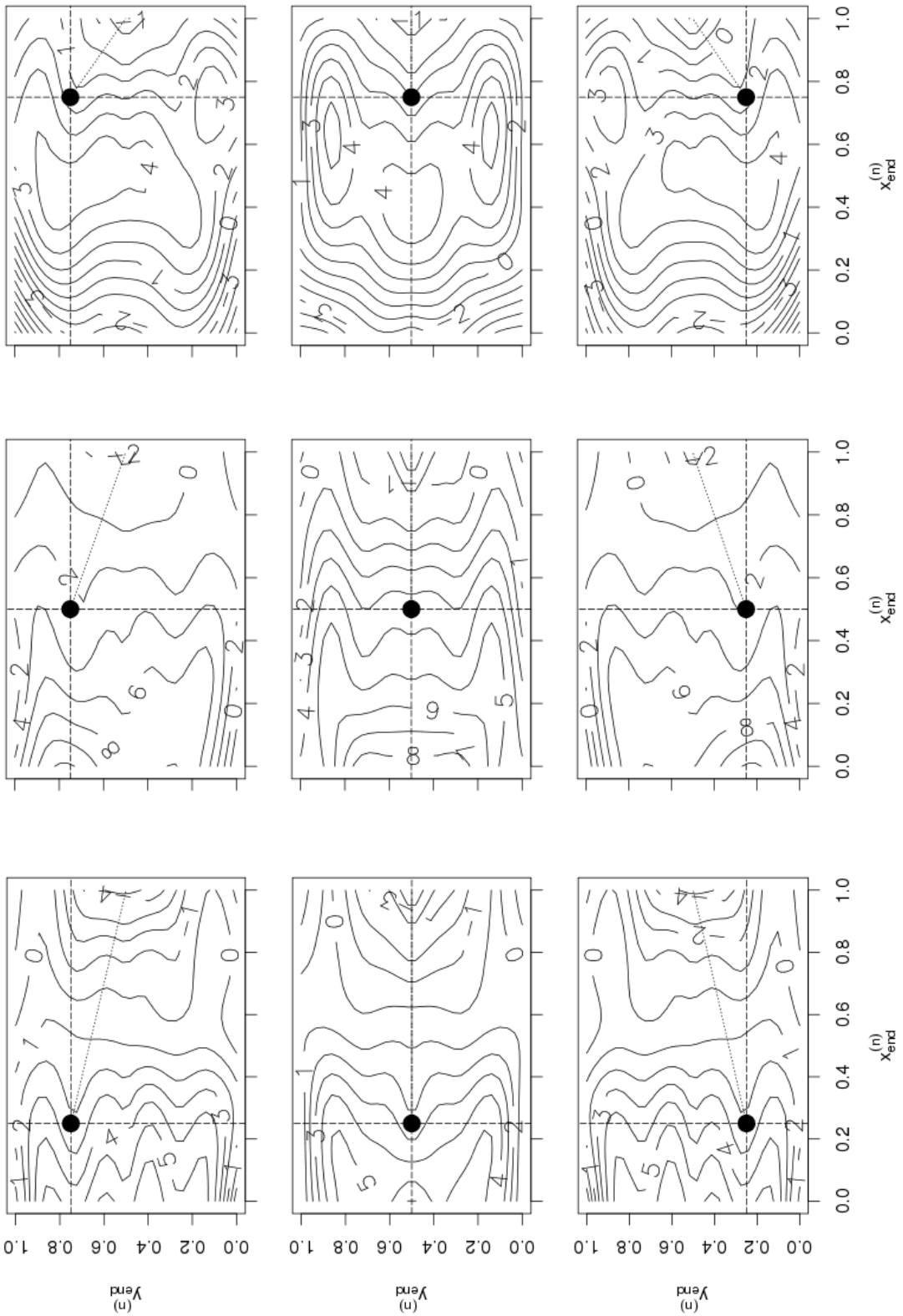


Figure 6.3: Value of the linear predictor of pass success with respect to the location of the pass origin and its target for a team attacking left to right. The contours are values of the linear predictor for the pass target as on the horizontal ($x_{end}^{(n)}$) and the vertical ($y_{end}^{(n)}$) axes and the pass origin variables ($x^{(n)}$ and $y^{(n)}$) fixed at the values indicated by a •. The origins are selected from a $(0.25, 0.50, 0.75) \times (0.25, 0.50, 0.75)$ grid. The dotted line is the direct route to the goal.

may be because there is usually an opponent in front of the passing player obstructing the most direct route to the goal (the dotted line).

Predictions of the team random effects are presented in figure 6.4. The vertical axis

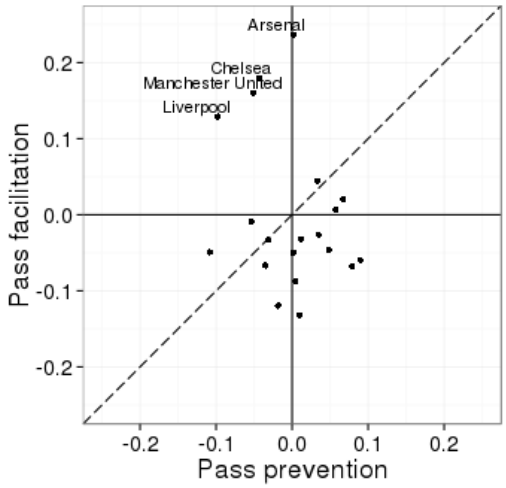


Figure 6.4: Team random effects prediction in the pass success model.

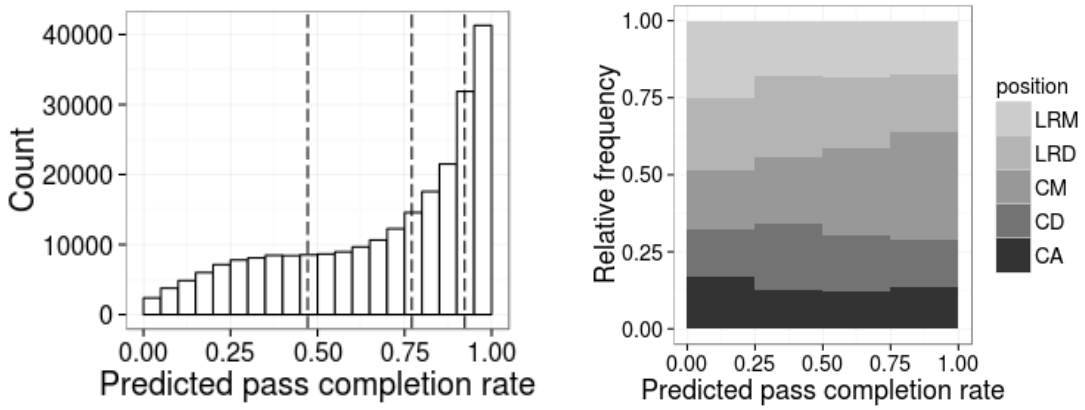
corresponds to the terms $b^{(t)}$ representing each team's ability to facilitate passing (e.g. by clever off the ball movement). The higher it is the better the team is. The horizontal axis contains the terms $b^{(o)}$ capturing each team's ability to prevent passing of their opponents (e.g. by aggressive pressing and close marking). The lower the number the better. Distance from the diagonal dashed line can be viewed as a summary of the team ability in these two aspects. There are four clear outliers in the plot: Arsenal, Chelsea, Manchester United and Liverpool who dominated the league particularly in terms of the ability to facilitate passing. Arsenal are an extreme example here as they were the best at facilitating passing but only average at preventing it. Liverpool, on the other hand, were almost equally good in both areas.

6.4.2 Ease of pass

We approximate ease of each pass in the fitting sample with the probability of the pass being completed had it been played by the average player, $\hat{p}_i^{(e)}$. The left panel of figure 6.5 shows the resulting density. The further to the right the easier the pass (the more likely it is to be completed). Interestingly, the distribution is highly skewed towards the easy passes: half of the passes have expected completion probability of more than about 76% and a quarter of the passes are 90% or more likely to be successfully executed. On the other hand, only about a quarter of the passes are less likely to be completed than not.

The pass difficulty information can be broken down by the nominal position of the

players. This is done in the right panel of figure 6.5. A relatively high proportion of the easy passes (the furthest to the right) are attempted by players playing in the central midfield (CM). The more difficult the passes, the lower the proportion of them are executed by this group of players and, conversely, the more that are attempted by the offensive players in the central (CA) and the wide positions (LRM). The proportion of the passes made by the defenders (CD and LRD) is fairly constant with respect to the ease of pass.



(a) All positions. Vertical dashed lines cut off consecutive quartiles.

(b) Relative frequency of the ease of passes made by players from given nominal positions.

Figure 6.5: Ease of pass.

We argue that one of the reasons why raw pass completion rate is a poor measure of players' passing ability is that it is polluted by the difficulty of the attempted passes. In other words, this simple metric can fluctuate purely due to changes in the type of attempted passes rather than the inherent skill level of the executing player. If that is the case, then we may expect the completion rate to increase from one season to another for players who attempted easier passes in the second one and *vice versa*. This is what is analysed in figure 6.6. It compares the average 2007/08 completion rate with the 2006/07 one (left panel) and the average of the full model predictions, $\bar{\hat{p}}^{(f)}_{k,2007/08}$, (right panel). Focusing on the left panel first, there is some correlation between the empirical values from one season to another. However, it is also clear that many of the deviations could be explained by the ease of passes attempted as players whose performance increased (above the dashed identity line) tended to be faced with an easier task in season 2007/08 than in the previous one, $\bar{\hat{p}}^{(e)}_{k,2007/08} - \bar{\hat{p}}^{(e)}_{k,2006/07} > 0$. Conversely, the completion rate of the players who attempted more difficult passes in the second season, tended to drop. Since the model is able to control for the pass difficulty, the relationship between its

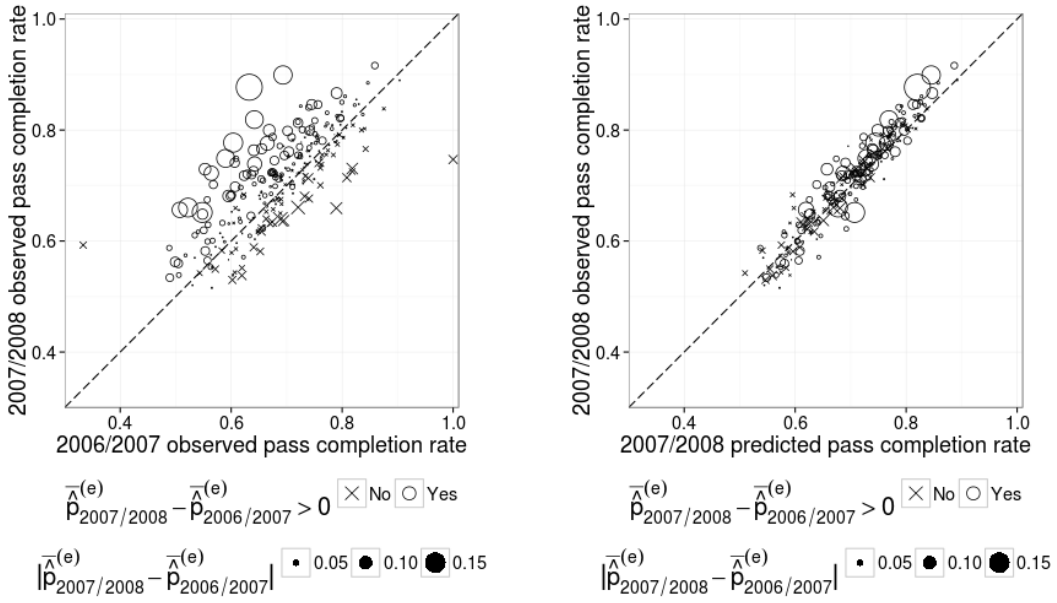


Figure 6.6: Average observed 2007/08 pass completion rate ($\bar{o}_{k,2007/08}$) against the naive ($\bar{o}_{k,2006/07}$) and the model ($\bar{\hat{p}}^{(f)}_{k,2007/08}$) predictions. The quantity $\bar{\hat{p}}^{(e)}_{k,2007/08} - \bar{\hat{p}}^{(e)}_{k,2006/07}$ is the change in the value of the proxy for pass ease from season 2006/07 to 2007/08 for the k-th player. It is marked by $\circ$ if positive and by $\times$ if negative. The absolute value of the change is indicated by the size of the points. The dashed line is the identity function.

predictions and the 2007/08 empirical values is much stronger (the right panel) with the Pearson correlation coefficient 0.92 for the model and 0.72 for the naive predictions.

Table 6.3 lists the top 5 players for each position according to the model together with their predictions and empirical values. The list is limited to players who made at least 100 passes in the season 2007/2008 to allow reliable comparison between model predictions and the observed values in the validation sample. Table 6.3 reveals some specific examples of how the model incorporates, and accounts for, pass difficulty in making predictions.

For example, Carlos Tevez's pass completion rate ($\bar{o}$) jumped by a few percentage points from season 2006/2007 to the next one (0.74 to 0.80). However, the model anticipates it very well ($\bar{\hat{p}}^{(f)}_{2007/2008} = 0.80$) as a big proportion of the improvement can be explained by the fact the ease of the passes attempted was much higher in the second season ($\bar{\hat{p}}^{(e)}_{2007/2008} = 0.78$ compared to $\bar{\hat{p}}^{(e)}_{2006/2007} = 0.71$). In the case of Tevez this has a lot to do with the fact that he played with players of better quality in 2007/2008 after he was transferred from West Ham United, a team threatened with relegation in 2006/2007, to Manchester United, the Premier League champions in 2007/2008.

			2006/07					2007/08				
Pos	Forename	Surname	Team	n	$\bar{o}$	$\bar{\hat{p}}^{(e)}$	$\hat{p}^{(av)}$	Team	n	$\bar{\hat{p}}^{(f)}$	$\bar{o}$	$\bar{\hat{p}}^{(e)}$	$\bar{o} - \bar{\hat{p}}^{(f)}$
CD	John	Terry	Chelsea	1058	0.88	0.81	0.83	Chelsea	613	0.84	0.84	0.79	0.00
CD	William	Gallas	Arsenal	730	0.86	0.81	0.81	Arsenal	884	0.89	0.92	0.87	0.03
CD	Sami	Hyypia	Liverpool	1088	0.75	0.72	0.79	Liverpool	648	0.78	0.79	0.76	0.01
CD	Ricardo	Carvalho	Chelsea	1193	0.83	0.80	0.79	Chelsea	621	0.82	0.86	0.81	0.03
CD	Chris	Riggott	Middlesbrough	142	0.72	0.65	0.79	Middlesbrough	220	0.67	0.72	0.65	0.05
LRD	Pascal	Chimbonda	Wigan/Tottenham	1351	0.75	0.72	0.80	Tottenham Hotspur	1119	0.77	0.79	0.74	0.02
LRD	Fabio	Aurelio	Liverpool	435	0.71	0.66	0.80	Liverpool	592	0.72	0.70	0.69	-0.03
LRD	Andrew	Taylor	Middlesbrough	1174	0.68	0.65	0.79	Middlesbrough	453	0.67	0.68	0.64	0.01
LRD	Steve	Finnan	Liverpool	1311	0.75	0.73	0.79	Liverpool	631	0.76	0.74	0.74	-0.02
LRD	Stephen	Carr	Newcastle United	858	0.73	0.70	0.79	Newcastle United	297	0.72	0.70	0.70	-0.02
CM	Paul	Scholes	Manchester United	1898	0.90	0.84	0.85	Manchester United	1341	0.89	0.89	0.84	-0.00
CM	Stilian	Petrov	Aston Villa	996	0.79	0.74	0.80	Aston Villa	562	0.78	0.79	0.74	0.01
CM	Michael	Essien	Chelsea	1637	0.84	0.81	0.80	Chelsea	1220	0.82	0.80	0.79	-0.01
CM	Michael	Carrick	Manchester United	1762	0.82	0.79	0.80	Manchester United	1280	0.82	0.82	0.79	0.01
CM	Didier	Zokora	Tottenham Hotspur	1105	0.82	0.79	0.80	Tottenham Hotspur	1032	0.83	0.84	0.81	0.01
LRM	Mikel	Arteta	Everton	1292	0.73	0.68	0.81	Everton	877	0.69	0.72	0.65	0.03
LRM	Alexander	Hleb	Arsenal	1455	0.80	0.78	0.79	Arsenal	1349	0.80	0.82	0.78	0.02
LRM	Gareth	Barry	Aston Villa	1451	0.68	0.65	0.79	Aston Villa	1203	0.71	0.72	0.69	0.01
LRM	Kevin	Kilbane	Everton/Wigan	831	0.62	0.59	0.79	Wigan Athletic	725	0.64	0.57	0.62	-0.07
LRM	Cristiano	Ronaldo	Manchester United	1200	0.78	0.76	0.78	Manchester United	1017	0.76	0.76	0.74	0.00
CA	John	Carew	Aston Villa	309	0.56	0.52	0.79	Aston Villa	723	0.56	0.55	0.54	-0.01
CA	Darren	Bent	Charlton Athletic	679	0.65	0.63	0.78	Tottenham Hotspur	160	0.59	0.58	0.57	-0.01
CA	Carlos	Tevez	West Ham United	518	0.74	0.71	0.78	Manchester United	950	0.80	0.80	0.78	0.00
CA	Nicolas	Anelka	Bolton Wanderers	813	0.66	0.64	0.78	Bolton /Chelsea	620	0.68	0.66	0.66	-0.01
CA	Nwankwo	Kanu	Portsmouth	902	0.72	0.70	0.78	Portsmouth	436	0.76	0.75	0.75	-0.01

Table 6.3: Top 5 passers by position based on 2006/07 season performance. Column n is the number of attempted passes, $\bar{o}$ is the average observed completion rate, $\bar{\hat{p}}^{(e)}$ is the average of the a proxy for the ease of the passes, $\hat{p}^{(av)}$ is the passing rating, $\bar{\hat{p}}^{(f)}$ is the average of the full model predictions.

6.4.3 Evaluating players

Figure 6.7 plots model derived players' passing ability ($\hat{p}^{(av)}$) against the observed pass completion rate ($\bar{o}$) in season 2006/07. The dashed line is the identity function. Specific player examples can be examined in table 6.3.

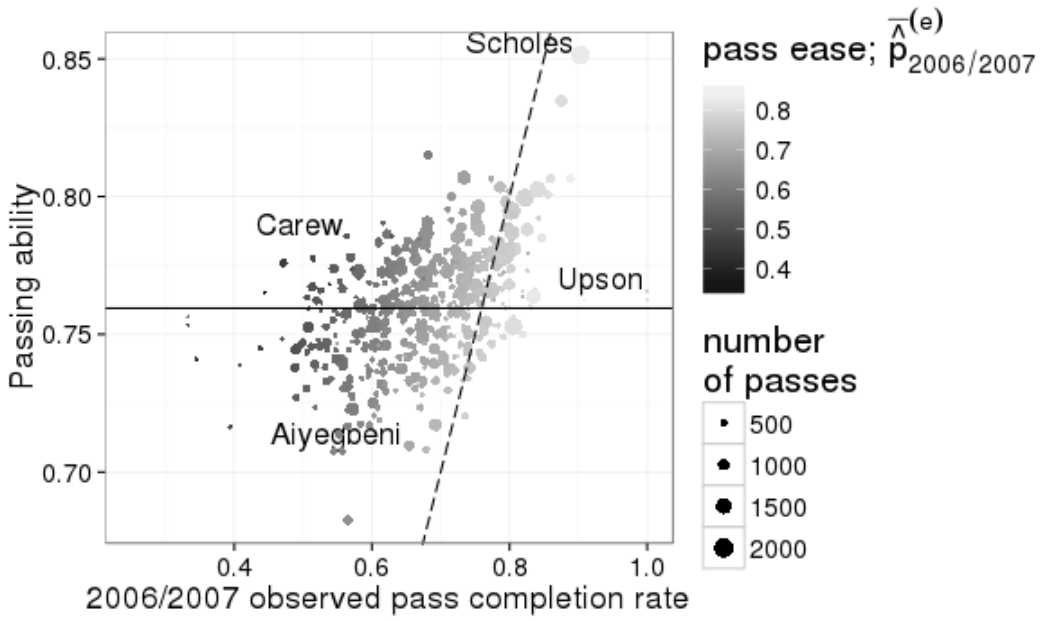


Figure 6.7: Estimated players' passing ability ($\hat{p}^{(av)}_{k,2006/07}$) against the observed pass completion rate ($\bar{o}_{k,2006/07}$), a proxy for the ease of pass ($\bar{\hat{p}}^{(e)}_{k,2006/07}$) and the number of passes in the fitting sample. The points corresponding to the labelled players are the ones to the bottom right of their names (see main text for discussion). The dashed line is the identity function and the solid horizontal one is the passing ability of an average player.

Naturally, there is positive correlation between the empirical completion rate in the fitting sample ($\bar{o}$) and the model based passing ability ($\hat{p}^{(av)}$) as players who complete passes at a higher rate are generally considered to be better at this skill by the model. There are, however, considerable departures from this naive rule which are summarised below.

First of all, the circumstances from which the players attempted passes differ. Some of them passed in easier situations and/or chose easier options which boosted their observed completion rate above what could be expected simply based on their passing ability. Conversely, some were faced with an unusually difficult task which made their empirical completion rate look worse than they deserve when compared on a level playing field. This is reflected in the positive correlation between the average pass ease (the brightness of the points) and the observed success rate (the horizontal axis). To illustrate

how the model takes pass difficulty into account when rating players' skill consider the pair of centre forwards: John Carew and Yakubu Aiyegbeni. The former had a lower empirical pass completion rate in the fitting sample, however, his skill is rated higher by the model as the passes he attempted were generally more difficult (a darker point). Similarly, in table 6.3 Sami Hyypia's passing skill is rated almost identically to Ricardo Carvalho's ($\hat{p}^{(av)} \approx 0.79$ for both) despite his observed completion rate ($\bar{o}$) being much lower because the passes he attempted were considerably more difficult (lower $\bar{\hat{p}}^{(e)}$) on average.

Secondly, some players success rates are based on few observations making their numbers less reliable. The model recognises this fact by regressing the individual performance to the overall mean, represented by the solid horizontal line, the effect being stronger for fewer passes. As an extreme example, consider Matthew Upson who had a 100% completion rate but achieved it on just 6 passes. The model recognises that there is very little information contained in such small samples. On the other hand, Paul Scholes completed many more passes (a bigger point) in the fitting sample and, as a result, is rated much higher despite his empirical completion rate being lower. Similarly, in table 6.3 Chris Riggott's passing skill is rated about the same as Ricardo Carvalho's ($\hat{p}^{(av)} \approx 0.79$ for both) despite the difference between the empirical completion rate ($\bar{o}$) and the difficulty of passes ($\bar{\hat{p}}^{(e)}$) being much bigger for Riggot. This is because Carvalho proved his unusually high completion rate on many more passes.

In all, in order to be recognised for his empirical passing success in the model framework a player would have to pass at a higher rate than an average player in these circumstances and provide sufficient evidence for this.

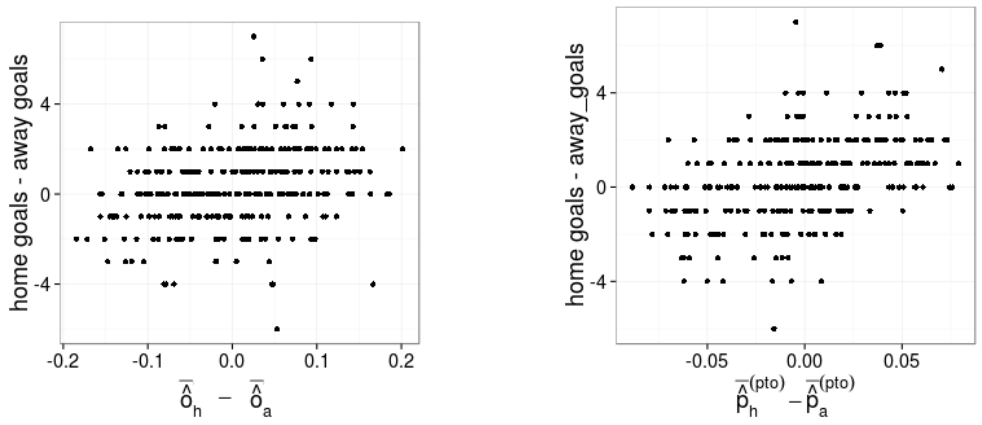
6.4.4 Comparing predictive utility

The ultimate test of a rating method is its predictive utility. Verifying it is complicated in this case because what we are trying to rate, i.e. the passing skill of footballers, is not observable. For instance, we could not just use the observed pass completion rate in season 2007/08 as a benchmark for predictions, since the very essence of the argument is that it is a poor indicator of the passing skill.

One objective measure that exists is team success. If the talent pool of a team as evaluated by one index is a better predictor of the team's future results than one based on another index, then the former should be preferred over the latter. In other words, football clubs should assess players based on methods that are informative about future team results. A "good" player is one that helps his team win. With this in mind, for

every fixture of the 2007/08 season we calculated two statistics supposed to capture the general level of the passing skill in both competing teams: one based on the raw pass completion rate in season 2006/07, $\bar{\delta}$, and one based on the model fitted on that season, $\bar{\hat{p}}^{(pto)}$. They are defined in the end of the list of prediction types in section 6.3.4. We check how well a difference in values of these indices for competing teams predicts the result of a fixture.

Firstly, in figure 6.8 for each fixture the difference in the home and away team goals is plotted against the difference in the indices for both teams. The Pearson correlation coefficient for the pass completion based index (figure 6.8a) with the home team goals supremacy in season 2007/08 is 0.268 with a 90% confidence interval of $(0.177, 0.354)$, whereas its value for the model based predictor (figure 6.8b) is 0.414 with a $(0.332, 0.490)$ 90% confidence interval.



(a) Previous season's raw completion rates used as predictors of players' completion rate in the given fixture

(b) Fixture specific "average" difficulty pass predictions used as predictors of players' completion rate in the given fixture

Figure 6.8: Home team goals supremacy in fixtures of 2007/08 season against the difference in the average predictor of the pass completion rate for the home and the away team players in that fixture.

Secondly, we fit two ordered logit regression models of the game outcome (home win, draw or away win) with the difference in the average passing index for the home and away team as the only covariate: one model for the index based on the raw pass completion rate, $\bar{\delta}$, and one for the model based index, $\bar{\hat{p}}^{(pto)}$. The latter model offers a better fit with log likelihood of -290.58 compared to -304.72 for the pass completion rate based model (both models have the same number of parameters).

We confirmed that the above results are not sensitive to the minimum number of

players with a passing skill rating in a given fixture required (before calculating the average passing skill indices).

6.5 Discussion

In this chapter we presented a method which can be used to evaluate passing skill of football players controlling for the difficulty of their attempts. We combine proxies for various factors influencing the probability of a pass being successful in a statistical model and evaluate the inherent player skill in this context. The measure of player passing skill has a natural interpretation in this framework, as does the metric proposed for pass difficulty. Finally, we are able to comprehensively handle all the players in the observed sample with the same procedure without a need to arbitrarily discard players who have been observed *too few* times to be reliably evaluated. The reduced reliability of empirical passing rates based on a small number of observations is naturally taken into account within the proposed framework.

When comparing the utility of the proposed method against the raw pass completion rate for predicting fixture results, we used model predictions conditional on the estimates of the abilities of the players as well as the teams involved in each fixture. This is because team ability is confounded with player ability in the pass completion rate statistic. Ignoring team abilities in model predictions would give the naive method an unfair advantage since most of the players play for the same team in the fitting and the prediction sample. It might be argued that the approach we took, in turn, gives our method an advantage because some players do change teams between the two seasons. However, we regard the fact that our method is able to disentangle player abilities form team abilities (and other factors) and put them back together in a different configuration in order to make useful and accurate predictions, to be one of the strengths of our approach. Therefore we do not believe the advantage is unfair.

One important point that needs to be made about player evaluations produced by this model is that they are most useful when compared among players performing similar types of passes. Breaking down the results by position is a step in this direction. To suggest, for example, that a central defender would maintain his passing rating when transferred to a winger position without at least a period of transition would be naive.

Speaking of positions, players are classified to only a few categories and based only on the location of their actions on the ball in their previous games. As any football fan will know, this is a very simplistic approach as there are more possible positions and other factors that determine which of them a player belongs to. Classifying players

to positions based on the actions they perform could itself be an interesting research problem but is not a goal of this research. Therefore we settled for this simplistic classification approach just to highlight some potentially interesting aspects of the results (in figure 6.5 and table 6.3) but it is not a component of the model.

Another caveat for the results presented here is that while the general team ability to facilitate successful pass completion is accounted for, the individual skill of the pass receiver is not factored in. Therefore it may still be possible that the latter may be confounded in a rating of a player who tends to play an unusually high proportion of passes towards certain team mates. For example, John Terry's rating may be inflated if he frequently played long passes which are normally difficult to complete but perhaps less so if Didier Drogba, who is known to be a strong receiver, is the target man. Including pass receiver in the equation could be a potential model extension. The challenge would be to identify the receiver for unsuccessful passes and this is not currently collected in the Opta data we use.

Further work in this area could also involve evaluating passes based on their value for the team rather than the difficulty. It may be the case that some players are able to add value with their passing above what could be expected by the difficulty of their passes, while others tend to attempt unnecessarily difficult passes, which is not recognised in the framework proposed here. Further, our model, as specified here, may reward players for attempting difficult passes that have no positive effect, and possibly even a negative effect, on the team. However, despite this possibility we believe the results demonstrate that the model is valuable, and is certainly a step in the right direction if statistical modelling is to be used to measure passing ability of players in football.

Chapter 7

Individual player contribution to team success

7.1 Introduction

In chapter 3 we argued that, if a player's valuation is the aim of his assessment, the method used to conduct it must:

- recognise that individual performance depends on the player's underlying skill as well as factors beyond his control and
- link the individual performance to the team success.

In this chapter we propose a model that fulfils these criteria and fit it to some real data. The model attempts to address one of the key difficulties in modelling the game of football, i.e. its free-flowing nature, by discretising it into a series of events. The evolution of the game from one event to another is described using a Markov chain model.

The use of Markovian models to describe the game of football is not new. Historically the earliest attempts are due to Hirotsu and Wright (2002), who model the game as a Markov process with 4 states: 2 for either team being in possession of the ball and 2 for goals for each team. They construct transition probabilities to depend on team specific parameters for shot conversion and stopping as well as possession regaining and retention. Furthermore, they allow the transition probabilities to vary depending on the tactical approach (attacking or defending) of the teams. They study the optimal times to switch between these two modes depending on the score in order to optimise the expected points reward from the game. In Hirotsu and Wright (2003) they fit the models to some real data from the English Premier League. More recently, Pena (2014) uses a

simple Markovian model of the game consisting of a possession retention state, "Keep", and two possession ending states: "Shot" and "Loss", to model the length of possession in a football game. He demonstrates that even such a simple bottom-up approach offers a better fit to the empirical possession lengths than Poisson, Negative Binomial and Pareto distributions fitted to the aggregated data. He also claims that team specific parameter estimates reflect to some extent teams' style of play.

The models of Hirotsu and Wright (2002) and Pena (2014) are not appropriate for our purposes since they do not depend on abilities of individual players, hence they would not enable us to evaluate them. It this sense the approach of Jarvandi et al. (2013) is the most similar one to ours. They propose a Markovian model of the game with the transition probability matrix depending on players' decisions (to pass short, pass long, dribble or shoot) and a success rate at executing them. Transitions between states corresponding to different players are also influenced by "dependency matrices" which, for example, contain the probability that a pass is aimed at a given team mate, if a short pass is attempted. They evaluate a player's expected contribution to a given team by substituting him for a player of the same position on that team and comparing the difference in the expected goals scored and conceded before, and after, such a substitution. The main differences between our approach and theirs are that:

- we attempt to account for the pitch location of the events;
- instead of using empirical percentages from historical data to directly represent players' skills (e.g. at passing) we derive the latter from skill specific sub-models which, in addition to accounting for the location of individual actions, recognise the existence of chance in their outcome.

The rest of this chapter is organised as follows. A simple example of a Markovian model of a football game is presented in section 7.3.2 and it is extended in section 7.3.3 to depend on players' skills. This basic model of section 7.3.2 is not player dependent, hence it cannot be used for player evaluation, and getting acquainted with it is not necessary for the understanding of the player specific model of section 7.3.3. We are presenting it, nevertheless, as we believe it may be helpful in providing basic intuition about the mechanism of a Markovian model of the game of football. Both models are estimated in section 7.4 based on the data from season 2006/07 of the English Premier League and the player specific model is used to evaluate players who played in this season. Section 7.5 ends the chapter with some discussion of the strengths and weaknesses of the proposed approach. Some model extensions to address possible weaknesses are

also mentioned there. Structure of the code used to implement the ideas of this chapter can be found in appendix D.

7.2 Data

The data for the 2006/07 and 2007/08 seasons of the English Premier League was provided by Opta. The dataset includes information for every event during a game, including the event type (goal, pass, tackle, etc.), whether the action was a success, the location of the event (for passes, for example, information on the origin and destination of the pass is given), the player(s) involved in the event and the timing of the event. In this analysis we focus on the pass and the shot events.

7.3 Methods

In this section we propose a mixed effects time-homogeneous Markov chain model of a football game. In section 7.3.1 we recall the definition of the time-homogeneous Markov chain and we present its basic application to modelling a football game in section 7.3.2. It is extended in section 7.3.3 to account for players' skills.

7.3.1 Time-homogeneous Markov chain with a finite state space

Let S be a finite state space and $\mathbb{N}$ be a set of natural numbers including zero. A time-homogeneous Markov chain with a finite state space is a stochastic process $\{X_n : n \in \mathbb{N}\}$ so that

$$\begin{aligned}\forall s_i, s_j, s_0, s_1, \ldots, s_{n-1} \in S \\ P(X_{n+1} = s_j | X_n = s_i, X_{n-1} = s_{n-1}, ..., X_1 = s_1, X_0 = s_0) \\ = P(X_{n+1} = s_j | X_n = s_i) = p_{ij} \quad (7.3.1)\end{aligned}$$

This means that the probability of the process moving between two states from one step to the next one:

- does not depend on the evolution of the process before this step and
- is constant for all the steps (does not depend on n).

Such conditional probabilities are called one step transition probabilities and can be organised in a transition matrix $\mathbf{P} = \left[p_{ij}\right]$ for all $i, j \in S$.

7.3.2 Basic model of the game

In this section we present a basic Markov chain model of the game of football. The states correspond to game events such as passes, shots and goals and are characterised further by the team executing them and the location of the execution. Because the model is not player specific, it cannot be used for player assessment but will hopefully prove useful as an introduction to the main model presented in section 7.3.3.

In the next paragraph the states of the model are presented in detail and then the structure of the transition matrix is shown. The results of fitting the model to the data are in section 7.4.1.

Definition of the states Let us define a state of a basic Markov chain of a football game as a triplet $s = (s^{(t)}, s^{(e)}, s^{(l)})$ with elements corresponding to the following characteristics of the game events:

- The team (*home* or *away*) in possession of the ball at the time of the event: $s^{(t)} \in \{h, a\}$.
- Type of the event: $s^{(e)} \in \{pass, shot, goal\}$;
- Pitch location of the event, $s^{(l)}$.

Values of the location attribute are best presented in a graphical form directly in the context of a pitch scheme in figure 7.1.

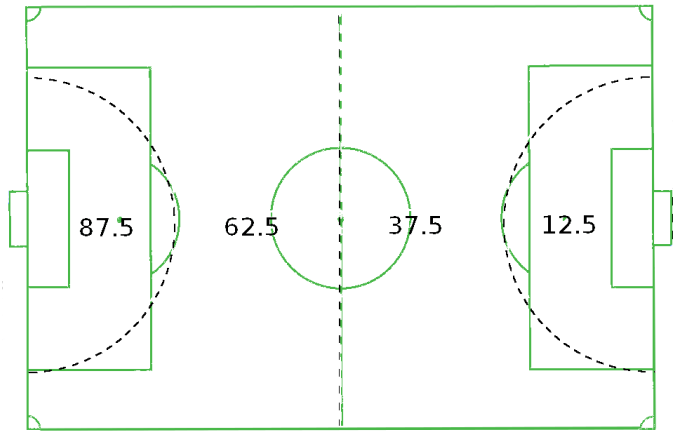


Figure 7.1: State location, $s^{(l)}$, for the pass events (teams attacking left to right) in the Markov chain model. The zone labels 87.5, 62.5, 37.5 and 12.5 are described in the main text.

In words:

- Passes originating from within 25 metres of the centre of the opponent's goal are assigned location value of $12.5 = \frac{0+25}{2}$. Those from the remaining part of the opponent's half have location $37.5 = \frac{25+50}{2}$. Passes from within 25 metres of the team's own goal have location $87.5 = \frac{75+100}{2}$ and those from the rest of their own half have location $62.5 = \frac{50+75}{2}$.
- The location of shots is defined in a similar way except that all shots from beyond 25 metres of the opponents goal are grouped together in the 37.5 category (regardless of the true point of origin).
- Goals have no location or, in other words, an empty location attribute, $\varnothing$.

This choice of the pitch division is entirely arbitrary. We have explored alternative schemes as well and examine the sensitivity of the results on this choice in appendix C.

In all, the state space of the basic Markov model is a Cartesian product of the three sets of attributes: team, event type and location. It can be written as:

$$S = \{\{h,a\} \times A\} \tag{7.3.2}$$

where

$$A = \{\{pass\} \times \{12.5, 37.5, 62.5, 87.5\} \cup \{shot\} \times \{12.5, 37.5\} \cup \{(goal, \varnothing)\}\}$$

Event type and location are grouped together in set A just for the ease of presentation. An example element of this state space $s = (h, pass, 87.5) \in S$ is a pass made by the home team from within 25 metres of the centre of their own goal.

Table 7.1 illustrates how Opta game events are translated to the states of the basic Markov model. Note that:

- only passes and shots from the left part of the table have a corresponding model state on the right hand side of it;
- goal states do not appear explicitly as Opta events (they are shot outcome attributes instead) but are added to the Markov chain.

Structure of the transition matrix In a Markov chain a transition between states from one step (n) to the next ($n+1$) is defined by a transition matrix. In the basic model of a football game, this matrix is assumed to be generic for all the fixtures. The structure of such a matrix is presented next.

team_name	event_name	x	y	team	event_type	location
Manchester United	Pass(Passed free kick taken) Successful - Short	0.36	0.49	*a*	*pass*	62.50
Manchester United	Pass(Open play) Successful - Long	0.33	0.21	*a*	*pass*	62.50
Fulham	Duel Lost (Challenge)	0.27	0.95	-	-	-
Manchester United	Duel Won (Take on)	0.70	0.06	-	-	-
Manchester United	Pass(Open play) Successful - Short	0.71	0.24	*a*	*pass*	37.50
Manchester United	Pass(Open play, Key pass) Successful - Short	0.75	0.52	*a*	*pass*	37.50
Manchester United	Off target(Open play,Right foot)	0.78	0.49	*a*	*shot*	12.50
Fulham	Pass(Goal kick) Successful - Long	0.05	0.37	*h*	*pass*	87.50
Fulham	Duel Won (Aerial)	0.62	0.22	-	-	-
Manchester United	Duel Lost (Aerial)	0.39	0.86	-	-	-
Fulham	Pass(Header) Unsuccessful - Short	0.62	0.22	*h*	*pass*	37.50
Manchester United	Pass(Open play) Unsuccessful - Short	0.32	0.82	*a*	*pass*	62.50
Fulham	Clearance(Unsuccessful)	0.28	0.26	-	-	-
Manchester United	Pass(Cross, Goal assist) Successful - Long	0.76	0.81	*a*	*pass*	37.50
Manchester United	Goal(Open play,Right foot)	0.94	0.40	*a*	*shot*	12.50
-	-	-	-	*a*	*goal*	∅
Fulham	Pass(Open play) Successful - Short	0.50	0.50	*h*	*pass*	37.50

Table 7.1: Translation of a sample of Opta events to states (the basic Markov chain model).

Let $\{X_n : n \in \mathbb{N}\}$ be a Markov chain with elements $X_n = (X_n^{(t)}, X_n^{(e)}, X_n^{(l)})$ taking values from the state space S. The structure of the corresponding one step transition matrix is presented in figure 7.2. The rows of the matrix correspond to the state at the step n and the columns correspond to the state at the step $n+1$. Some transitions, e.g. from a pass to a goal, are not possible in one step, hence the zeros in some elements. The remaining elements need to be filled in. In section 7.4.1 they will be estimated using sample frequencies, e.g. if 60% of passes of the home team from location 12.5 were followed by a pass of the home team from location 37.5, the element in the first row and the second column of the transition matrix would be assigned the value of 60%.

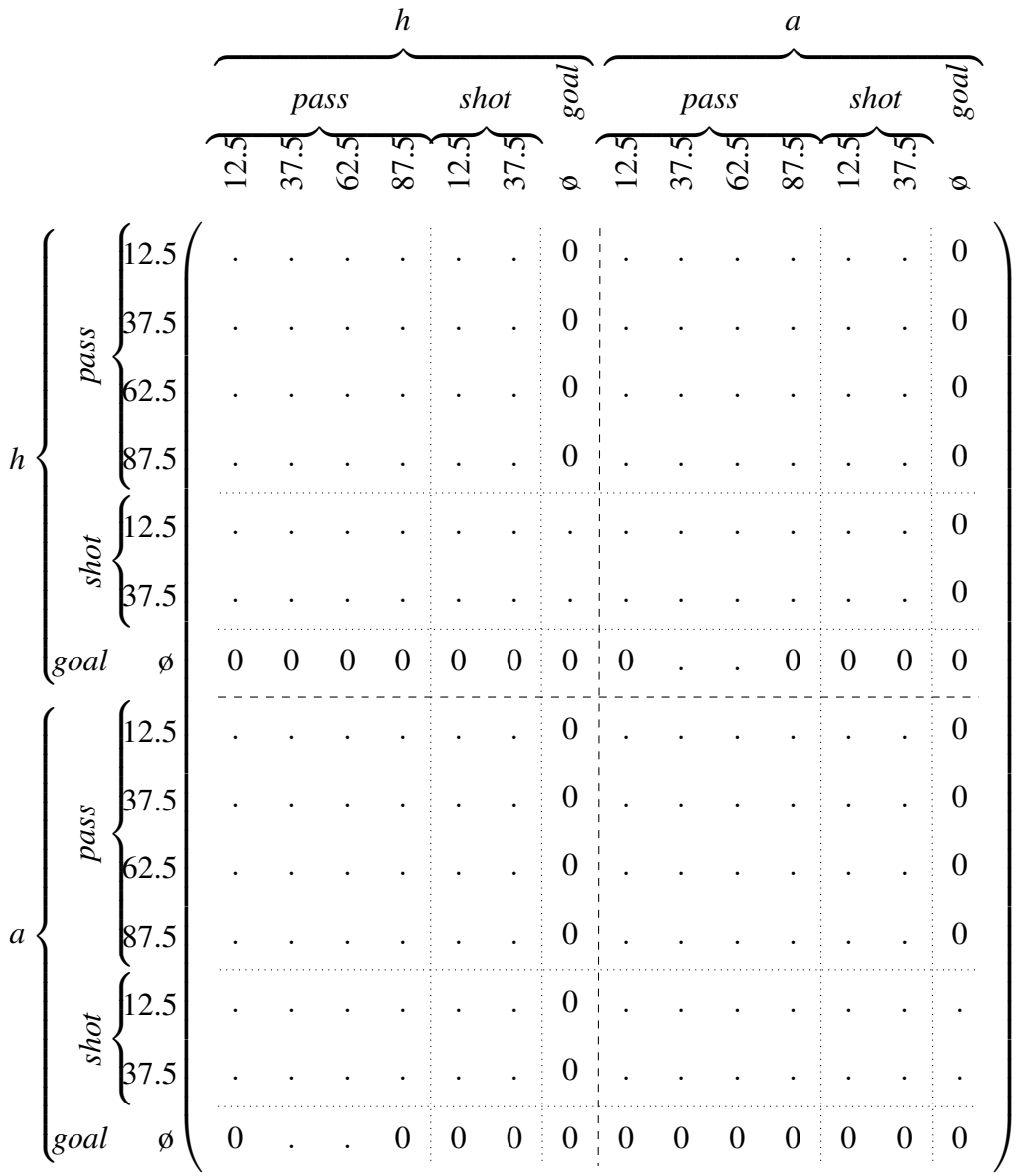


Figure 7.2: Structure of the transition matrix **P** of the basic Markov model of the game.

7.3.3 Player specific model of the game

While providing some intuition about the way in which the game of football can be modelled using Markov chains, the model of the previous section is too simplistic to serve as a player assessment tool. In particular it does not depend on players' skills. In this section we present a model with the extensions necessary for it to serve the purpose of player evaluation:

- The state space is extended by another dimension corresponding to the player executing the given event;
- The transition matrix becomes fixture specific, rather than generic for all the games.
- Instead of estimating each transition matrix by sample frequencies, we impose a general structure on all the matrices conditional on players' skills. Crucially this point means the we will be able to use the model to predict alternative scenarios different to the ones that actually occurred in the fitting sample, e.g. games with a subset of the players who played in a given fixture replaced by some other players.

The next paragraph presents the first of these extensions and the next two come in the remaining part of this section.

Definition of the states

Let Q be a set of player ids. Define a state of the player specific model of a football game as a quadruplet

$$\begin{aligned} &s = (s^{(t)}, s^{(e)}, s^{(l)}, s^{(q)}), \text{ where} \\ &s^{(t)} \in \{h, a\} \wedge s^{(l)} \in \{12.5, 37.5, 62.5, 87.5, \varnothing\} \wedge s^{(q)} \in Q \cup \{\varnothing\} \end{aligned} \tag{7.3.3}$$

The empty element is included in the set of player ids to be used in the goal state since it is not assigned to any particular player[1]. The set of event types is extended by an intermediate *choice* type preceding all the *pass* and *shot* states: $s^{(e)} \in \{choice, pass, shot, goal\}$. The *choice* state represents the state of a game when a player is in possession of the ball and is faced with a decision whether to pass the ball to a team mate or take a shot. The introduction of the intermediate *choice* state is shown in table 7.2, which illustrates how Opta game events are translated to the states of the player specific Markov model.

[1]The shot that leads to it is.

team_name	player_name	event_name	x	y	team	event_type	location	player_id
-	-	-	-	-	*a*	*choice*	37.50	275
Manchester United	Giggs	Pass(Open play, Key pass) Successful - Short	0.75	0.52	*a*	*pass*	37.50	275
-	-	-	-	-	*a*	*choice*	12.50	286
Manchester United	Saha	Off target(Open play,Right foot)	0.78	0.49	*a*	*shot*	12.50	286
-	-	-	-	-	*h*	*choice*	87.50	203
Fulham	Niemi	Pass(Goal kick) Successful - Long	0.05	0.37	*h*	*pass*	87.50	203
Fulham	Helguson	Duel Won (Aerial)	0.62	0.22	-	-	-	-
Manchester United	Evra	Duel Lost (Aerial)	0.39	0.86	-	-	-	-
-	-	-	-	-	*h*	*choice*	37.50	196
Fulham	Helguson	Pass(Header) Unsuccessful - Short	0.62	0.22	*h*	*pass*	37.50	196
-	-	-	-	-	*a*	*choice*	62.50	268
Manchester United	Brown	Pass(Open play) Unsuccessful - Short	0.32	0.82	*a*	*pass*	62.50	268
Fulham	Christanval	Clearance(Unsuccessful)	0.28	0.26	-	-	-	-
-	-	-	-	-	*a*	*choice*	37.50	285
Manchester United	Rooney	Pass(Cross, Goal assist) Successful - Long	0.76	0.81	*a*	*pass*	37.50	285
-	-	-	-	-	*a*	*choice*	12.50	284
Manchester United	Ronaldo	Goal(Open play,Right foot)	0.94	0.40	*a*	*shot*	12.50	284
-	-	-	-	-	*a*	*goal*	-	-
-	-	-	-	-	*h*	*choice*	37.50	196
Fulham	Helguson	Pass(Open play) Successful - Short	0.50	0.50	*h*	*pass*	37.50	196

Table 7.2: Translation of a sample of Opta events to states (the player specific Markov chain model).

The state space of the player specific Markov model is a Cartesian product of the four sets of attributes: team, event type, location and player. It can be written as:

$$S = \{\{h,a\} \times [(A \times Q) \cup G]\} \tag{7.3.4}$$

where

$$A = \{\{pass, choice\} \times \{12.5, 37.5, 62.5, 87.5\} \cup \{shot\} \times \{12.5, 37.5\}\} \tag{7.3.5}$$

$$G = \{(goal, \varnothing, \varnothing)\} \tag{7.3.6}$$

Event type and location are grouped together in the set A for the *pass*, *choice* and *shot* states just for the ease of presentation. The two empty set elements in the set G reflect the fact that goal states have neither a specific location nor a player attached to them. An example element of this state space $s = (h, pass, 87.5, 1) \in S$ is a pass made by player 1 of the home team from within 25 metres of the centre of their own goal.

Model structure

For the m-th fixture let $\{X_{m,n} : n \in \mathbb{N}\}$ be a Markov chain with elements

$$X_{m,n} = (X^{(t)}_{m,n}, X^{(e)}_{m,n}, X^{(l)}_{m,n}, X^{(q)}_{m,n})$$

taking values from the state space S. The structure of the one step transition matrix (or more specifically a part of it for transitions between the states of selected player k from team h) is presented in figure 7.3.

One way to estimate the elements of the fixture specific transition matrices would be to use sample frequencies like we did for the generic transition matrix of the basic model in section 7.3.2 (The results of this procedure are presented in a graphical form in figure 7.4 in section 7.4.1.). The problem with this approach here is that such estimates would be based on very small sample sizes. For example, consider the transition between the choice state of a given player on a given team in a specific location to a shot state of that player in that location. The probability of this transition is equal to the probability that the player decides to shoot, rather than pass, when in possession of the ball in this location. For a given fixture we could estimate it as the proportion of times this player decided to shoot in that location. However, there may not be many such observations in a single fixture. Therefore, it may be helpful to assume that there are some patterns in the transition matrix. For example, the frequency of shots from a given zone for a given player may be somehow linked to the frequency of his shots from the adjacent zone (i.e. there may be some player specific "tendency to shoot" behind both) or the frequency

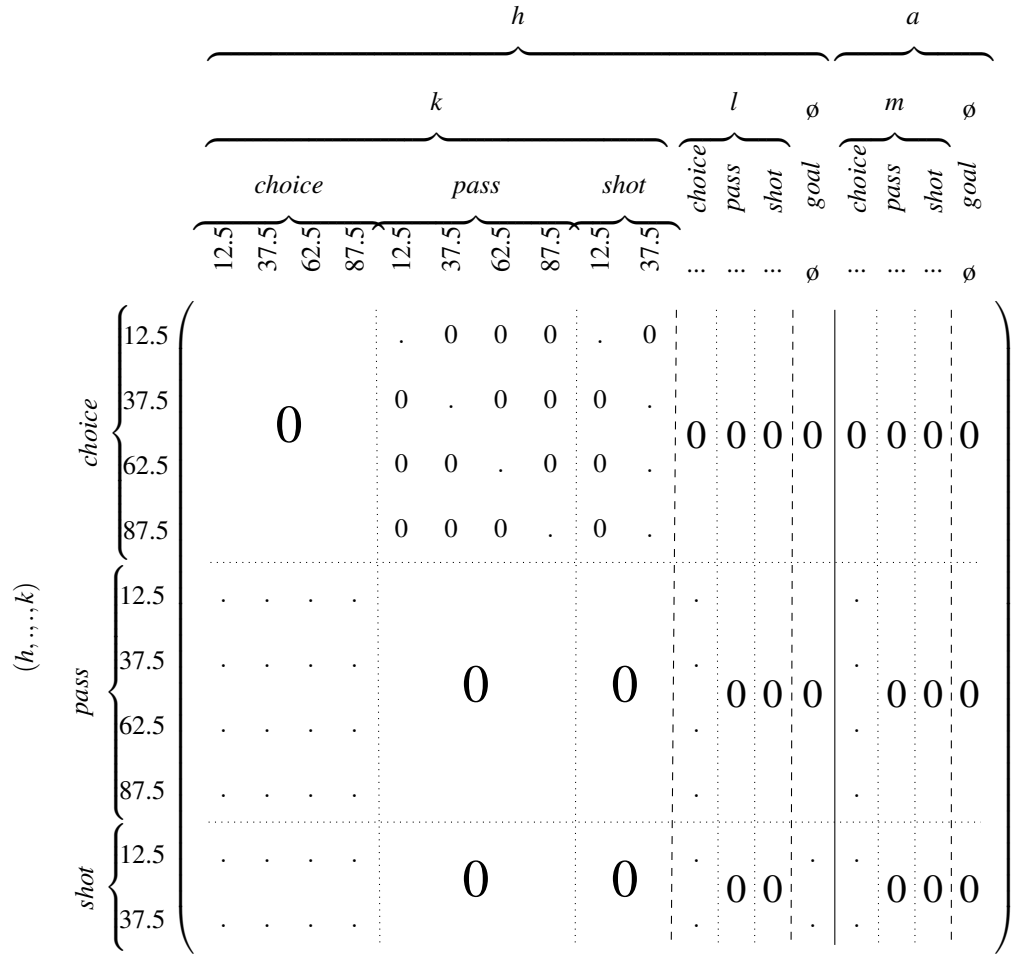


Figure 7.3: Structure of the transition matrix $\mathbf{P}_m$ of the player specific Markov model of the game (from the states of player k from team h).

at which other players decide to shoot from this zone. Imposing some model structure on the transition matrices would effectively lead to their elements being estimated from bigger samples thus improving the efficiency of the estimation. It would also enable us to make predictions of alternative scenarios for a given fixture, e.g. what if some other player featured for the home team instead of their star man.

We can propose a natural regression model for the transition probability between the *choice* and the *shot* state of a given player in a given location considered above (e.g. see equation (7.3.21)). Unfortunately, this is not necessarily the case for all elements of the transition matrices. For example, consider a transition between a *pass* of player i of the home team from zone 37.5 to a *choice* state in zone 12.5 of player j of the away team. It is not obvious how to design a model for this probability that would depend on some players' skills in a natural way. The approach we propose is to write such a transition probability as a product of conditional probabilities which can be modelled

more naturally conditional on player skills. The transition probability in the example just given, is equivalent to the probability that:

- when passing the ball from the 37.5 zone of the home side, player i chooses to direct it to the 87.5 zone, e.g. back passing the ball to his own goalkeeper, and
- such a pass is unsuccessful resulting in the away team gaining possession of the ball in their 12.5 zone (which corresponds to the 87.5 zone of the home team), and
- player j ends up in possession of the ball given the previous two conditions.

The above relationship can be written in terms of the elements of the Markov chain $\{X_{m,n} : n \in \mathbb{N}\}$ and the state space S. Before we do that, we just need to define $L_{m,n}$ as the location "targeted" by the pass at the step n:

$$L_{m,n} = \begin{cases} X^{(l)}_{m,n+1} & \text{if } X^{(t)}_{m,n+1} = X^{(t)}_{m,n} \\ 100 - X^{(l)}_{m,n+1} & \text{if } X^{(t)}_{m,n+1} \neq X^{(t)}_{m,n} \end{cases} \tag{7.3.7}$$

In our example $X^{(t)}_{m,n} = h$, $X^{(t)}_{m,n+1} = a$ and $X^{(l)}_{m,n+1} = 12.5$ so $L_{m,n} = 87.5$.

Equivalently to the plain English description above, the transition probability from the example can be factored as

$$\begin{aligned} P(X_{m,n+1} = (a,&choice, 12.5, j) | X_{m,n} = (h, pass, 37.5, i)) \\ &= P(L_{m,n} = 87.5 | X_{m,n} = (h, pass, 37.5, i)) \\ &\quad \times P(X^{(t)}_{m,n+1} \neq X^{(t)}_{m,n} | L_{m,n} = 87.5, X_{m,n} = (h, pass, 37.5, i)) \\ &\quad \times P(X^{(q)}_{m,n+1} = j | L_{m,n} = 87.5, X^{(t)}_{m,n+1} \neq X^{(t)}_{m,n}, X_{m,n} = (h, pass, 37.5, i)) . \end{aligned} \tag{7.3.8}$$

Later in this section we propose regression models for the first two of these factors. For example, in equation (7.3.18) we model the second factor, which gives the probability that "a pass from one zone to another attempted by a given player is unsuccessful", using a logistic mixed effects regression with random effects representing players passing skill.[2] As for the last factor, note that based on equation (7.3.7) it is equivalent to

$$P(X^{(q)}_{m,n+1} = j | X^{(l)}_{m,n+1} = 12.5, X^{(t)}_{m,n+1} = a, X_{m,n} = (h, pass, 37.5, i)) .$$

We make a simplifying assumption that the distribution of the players in possession of the ball for a given team in a given zone does not depend on the previous action, i.e.

[2] This is simply a pass completion model, a more elaborate version of which was described in chapter 6.

that:

$$P(X^{(q)}_{m,n+1} = j | X^{(l)}_{m,n+1} = 12.5, X^{(t)}_{m,n+1} = a, X_{m,n} = (h, pass, 37.5, i)) \\ = P(X^{(q)}_{m,n+1} = j | X^{(l)}_{m,n+1} = 12.5, X^{(t)}_{m,n+1} = a) = P(X^{(q)}_{m,n} = j | X^{(l)}_{m,n} = 12.5, X^{(t)}_{m,n} = a) \tag{7.3.9}$$

and estimate this quantity using sample frequencies.

In appendix B we factor the transition probabilities for the *pass* node in the general case and do the same for the *choice* and the *shot* nodes.

Sub-models

In this section we propose several models for the conditional probabilities needed to estimate the elements of the transition matrix described in the previous section (and appendix B).

Shot conversion model The *shot conversion model* is used for the shot conversion probability in equations (B.3.2) and (B.3.4), i.e.

$$P(X^{(e)}_{m,n+1} = goal | X_{m,n} = (s^{(t)}_b, shot, s^{(l)}_c, s^{(q)}_d))$$

and

$$P(X^{(e)}_{m,n+1} = choice | X_{m,n} = (s^{(t)}_b, shot, s^{(l)}_c, s^{(q)}_d)) .$$

Let $\mathbf{g}$ be a vector of all the elements of $X^{(e)}_{m,n+1}$ for which $X^{(e)}_{m,n} = shot$. We adopt a binomial mixed effects model for such shot outcomes (*goal* being a success):

$$g_{m,n+1} | \eta^{(g)}_{m,n} \sim \text{Bernoulli}\left(\frac{\exp(\eta^{(g)}_{m,n})}{1+\exp(\eta^{(g)}_{m,n})}\right) \tag{7.3.10}$$

with the following linear predictor:

$$\eta^{(g)}_{m,n} = \beta^{(g)}_0 + \beta^{(g)}_{37.5} \mathbb{1}_{\{X^{(l)}_{m,n}=37.5\}} + b^{(g)}_{X^{(q)}_{m,n}} \tag{7.3.11}$$

where $\beta^{(g)}_0$ is the intercept term and $\beta^{(g)}_{37.5}$ modifies the linear predictor for shots taken from zone 37.5. Finally, $b^{(g)}_{X^{(q)}_{m,n}}$ is a random effect representing the ability of the player $X^{(q)}_{m,n}$ to convert shots into goals. The random effects for all the Q players are stored in a vector $\mathbf{b}^{(g)}$ and are assumed to have the multivariate normal distribution:

$$\mathbf{b}^{(g)} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I) \tag{7.3.12}$$

where I is a $Q \times Q$ identity matrix and σ^2 is a common variance parameter to be estimated.

Pass direction model The pass direction model describes the first factor in equation (B.2.4), i.e.

$$P(L_{m,n} = s_i^{(l)} | X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)}))\,.$$

Let $\tilde{d}_{L_{m,n}}$ be the underlying continuous latent variable for $L_{m,n}$ and let $\{\zeta_0, \zeta_{25}, \zeta_{50}, \zeta_{75}, \zeta_{100}\}$ with

$$\zeta_0 = -\inf < \zeta_{25} < \zeta_{50} < \zeta_{75} < \zeta_{100} = \inf$$

be a set of thresholds separating pitch zones that $\tilde{d}_{L_{m,n}}$ can fall into. We assume a cumulative link model:

$$P(\tilde{d}_{L_{m,n}} \leq k) = F(k) = F(\zeta_k - \eta^{(d)}) \tag{7.3.13}$$

where

$$\eta^{(d)} = \begin{pmatrix} \mathbb{1}^T_{X^{(l)}_{m,n}=37.5} \\ \mathbb{1}^T_{X^{(l)}_{m,n}=62.5} \\ \mathbb{1}^T_{X^{(l)}_{m,n}=87.5} \\ \mathbb{1}^T_{pos(X^{(q)}_{m,n})=G} \\ \mathbb{1}^T_{pos(X^{(q)}_{m,n})=CD} \\ \mathbb{1}^T_{pos(X^{(q)}_{m,n})=LRD} \\ \mathbb{1}^T_{pos(X^{(q)}_{m,n})=CM} \\ \mathbb{1}^T_{pos(X^{(q)}_{m,n})=LRM} \end{pmatrix}^T \begin{pmatrix} \beta^{(d)}_{37.5} \\ \beta^{(d)}_{62.5} \\ \beta^{(d)}_{87.5} \\ \beta^{(d)}_{G} \\ \beta^{(d)}_{CD} \\ \beta^{(d)}_{LRD} \\ \beta^{(d)}_{CM} \\ \beta^{(d)}_{LRM} \end{pmatrix} \tag{7.3.14}$$

For instance, a conditional probability that a pass will be targeted to zone 37.5

$$\begin{aligned} P(L_{m,n} =& 37.5 | X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \\ =& P(25 < \tilde{d}_{L_{m,n}} \leq 50 | X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \\ =& P(\tilde{d}_{L_{m,n}} \leq 50 | X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \\ & - P(\tilde{d}_{L_{m,n}} \leq 25 | X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \\ =& F(\zeta_{50} - \eta^{(d)}) - F(\zeta_{25} - \eta^{(d)})\,. \end{aligned} \tag{7.3.15}$$

We choose $F(x) = logit^{-1}(x)$ which leads to a *proportional odds* model.

Pass completion model The *pass completion model* is used for the second factors in equations (B.2.4) and (B.2.6), i.e.

$$P(X^{(t)}_{m,n+1} = X^{(t)}_{m,n} | L_{m,n} = s_i^{(l)}, X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \tag{7.3.16}$$

and

$$P(X^{(t)}_{m,n+1} \neq X^{(t)}_{m,n} | L_{m,n} = s^{(l)}_i, X_{m,n} = (s^{(t)}_b, pass, s^{(l)}_c, s^{(q)}_d)) \,. \tag{7.3.17}$$

$L_{m,n}$ is defined in equation (7.3.7).

Let us define a successful pass as one for which the team maintains possession of the ball for the next event ($X^{(t)}_{m,n+1} = X^{(t)}_{m,n}$). All the other passes (with $X^{(t)}_{m,n+1} \neq X^{(t)}_{m,n}$) are considered unsuccessful. Let $\mathbf{w}$ be the vector of pass outcomes defined this way.

We adopt a binomial mixed effects model for the outcome of a pass at the n-th step of the m-th fixture:

$$w_{m,n+1} | \eta^{(w)}_{m,n} \sim \text{Bernoulli}\left(\frac{\exp(\eta^{(w)}_{m,n})}{1+\exp(\eta^{(w)}_{m,n})}\right) \tag{7.3.18}$$

The linear predictor for all the passes is:

$$\eta^{(w)} = \begin{pmatrix} \mathbf{1}^T \\ \mathbb{1}^T_{L_{m,n}=37.5} \\ \mathbb{1}^T_{L_{m,n}=62.5} \\ \mathbb{1}^T_{L_{m,n}=87.5} \\ \mathbb{1}^T_{X^{(l)}_{m,n}=37.5} \\ \mathbb{1}^T_{X^{(l)}_{m,n}=62.5} \\ \mathbb{1}^T_{X^{(l)}_{m,n}=87.5} \\ \mathbb{1}^T_{X^{(l)}_{m,n}=37.5 \wedge L_{m,n}=37.5} \\ \mathbb{1}^T_{X^{(l)}_{m,n}=37.5 \wedge L_{m,n}=62.5} \\ \mathbb{1}^T_{X^{(l)}_{m,n}=37.5 \wedge L_{m,n}=87.5} \\ \mathbb{1}^T_{X^{(l)}_{m,n}=62.5 \wedge L_{m,n}=37.5} \\ \mathbb{1}^T_{X^{(l)}_{m,n}=62.5 \wedge L_{m,n}=62.5} \\ \mathbb{1}^T_{X^{(l)}_{m,n}=62.5 \wedge L_{m,n}=87.5} \\ \mathbb{1}^T_{X^{(l)}_{m,n}=87.5 \wedge L_{m,n}=37.5} \\ \mathbb{1}^T_{X^{(l)}_{m,n}=87.5 \wedge L_{m,n}=62.5} \\ \mathbb{1}^T_{X^{(l)}_{m,n}=87.5 \wedge L_{m,n}=87.5} \\ \mathbb{1}^T_{pos(X^{(q)}_{m,n})=G} \end{pmatrix}^T \begin{pmatrix} \beta^{(w)}_0 \\ \beta^{(w)}_{.,37.5} \\ \beta^{(w)}_{.,62.5} \\ \beta^{(w)}_{.,87.5} \\ \beta^{(w)}_{37.5,.} \\ \beta^{(w)}_{62.5,.} \\ \beta^{(w)}_{87.5,.} \\ \beta^{(w)}_{37.5,37.5} \\ \beta^{(w)}_{37.5,62.5} \\ \beta^{(w)}_{37.5,87.5} \\ \beta^{(w)}_{62.5,37.5} \\ \beta^{(w)}_{62.5,62.5} \\ \beta^{(w)}_{62.5,87.5} \\ \beta^{(w)}_{87.5,37.5} \\ \beta^{(w)}_{87.5,62.5} \\ \beta^{(w)}_{87.5,87.5} \\ \beta^{(w)}_{G} \end{pmatrix} + \mathbf{Z}\mathbf{b}^{(w)} \tag{7.3.19}$$

Rows of the $\mathbf{Z}$ design matrix select the elements of the random effects vector $\mathbf{b}^{(w)}$ corresponding to the player executing the given pass. Finally, for the random effects we

assume

$$\mathbf{b}^{(w)} \sim \mathcal{N}(\mathbf{0}, \phi^2 I) \tag{7.3.20}$$

where ϕ^2 represents the variance of the passing skill among players and I is the identity matrix. The values of the random effects can be interpreted as the passing ability of the players.

This model can be viewed as a simplified version of the pass completion model of chapter 6. The simplification is forced by the fact that here the pool of covariates is limited to the elements of the state vector at the pass execution, $X_{m,n}$, or derived variables, like $L_{m,n}$ (see equation (7.3.16)).

The estimated model is used to predict the pass completion probability in equations (B.2.4) and (B.2.6).

Choice model The following model is used to predict the action choice probability in equations (B.1.3) and (B.1.4), i.e.

$$P(X^{(e)}_{m,n+1} = pass | X_{m,n} = (s^{(t)}_b, choice, s^{(l)}_c, s^{(q)}_d))$$

and

$$P(X^{(e)}_{m,n+1} = shot | X_{m,n} = (s^{(t)}_b, choice, s^{(l)}_c, s^{(q)}_d)) .$$

Let $\mathbf{z}$ be the vector of all the elements of $X^{(e)}_{m,n+1}$ for which $X^{(e)}_{m,n} = choice$. We adopt a binomial mixed effects model for the choice ($X^{(e)}_{m,n+1} = shot$ being a success and $X^{(e)}_{m,n+1} = pass$ corresponding to a failure) for the n-th step of the m-th fixture:

$$z_{m,n+1} | \eta^{(z)}_{m,n} \sim \text{Bernoulli}\left(\frac{\exp(\eta^{(z)}_{m,n})}{1+\exp(\eta^{(z)}_{m,n})}\right) \tag{7.3.21}$$

where

$$\eta^{(z)}_{m,n} = \begin{pmatrix} \mathbf{1} & \mathbb{1}_{X^{(l)}_{m,n}=37.5} \end{pmatrix} \begin{pmatrix} \beta^{(z)}_0 \\ \beta^{(z)}_{37.5} \end{pmatrix} \tag{7.3.22}$$

The estimated model is used to predict the action choice probability in equations (B.1.3) and (B.1.4).

Next player The conditional probability in equations (B.2.5) and (B.2.7), and in the first factor in equation (B.3.3), is the probability of the given player being in possession

of the ball at the step $n+1$. We assume it does not depend on the state at the previous step

$$\begin{aligned}
&P(X^{(q)}_{m,n+1} = s^{(q)}_k | X^{(t)}_{m,n+1} = s^{(t)}_i, X^{(l)}_{m,n+1} = s^{(l)}_j, X_{m,n} = (s^{(t)}_b, shot, s^{(l)}_c, s^{(q)}_d)) \\
=&P(X^{(q)}_{m,n+1} = s^{(q)}_k | X^{(t)}_{m,n+1} = s^{(t)}_i, X^{(l)}_{m,n+1} = s^{(l)}_j, X_{m,n} = (s^{(t)}_b, pass, s^{(l)}_c, s^{(q)}_d)) \\
=&P(X^{(q)}_{m,n+1} = s^{(q)}_k | X^{(t)}_{m,n+1} = s^{(t)}_i, X^{(l)}_{m,n+1} = s^{(l)}_j) \\
=&P(X^{(q)}_{m,n} = s^{(q)}_k | X^{(t)}_{m,n} = s^{(t)}_i, X^{(l)}_{m,n} = s^{(l)}_j)
\end{aligned} \tag{7.3.23}$$

and estimate it separately for each fixture using sample frequencies.

Fixture state distributions

For each m-th fixture it only makes sense to consider a subset $S^{(m)}$ of the whole state space with the players who played that game:

$$S^{(m)} = \left\{ s^{(m)}_k : k = 1, 2, ..., N_{S^{(m)}} \right\} \tag{7.3.24}$$

where $N_{S^{(m)}}$ is the number of states in the state space of the m-th game. Having estimated the transition probabilities $P\left(X_{m,n+1} = s_j | X_{m,n} = s_i\right)$, for each fixture m, according to the appendix B and the sub-models of the previous section, we build a transition matrix $\mathbf{P}_m$ of the probabilities corresponding to the relevant subset of the whole state space. Now the unconditional distribution of the states can be calculated at each step of the Markov chain for each fixture:

$$d_{m,n} = \left[P(X_{m,n} = s^{(m)}_1), P(X_{m,n} = s^{(m)}_2), ..., P(X_{m,n} = s^{(m)}_{N_{S^{(m)}}}) \right] \tag{7.3.25}$$

as

$$d_{m,n} = d_{m,n-1} \times \mathbf{P}_m = d_{m,1} \times (\mathbf{P}_m)^{n-1} \tag{7.3.26}$$

where $d_{m,1}$ is the distribution of the states at the kick off. For example, $d_{10,17}$ is the distribution of the states from $S^{(10)}$ at the 17-th step of the chain for the 10-th fixture.

Two elements of the unconditional distributions $d_{m,n}$ are of particular interest to us because they refer directly to the currency in which team success is measured.

$$P(X_{m,n} = (h, goal, \varnothing, \varnothing)) \quad \text{and} \quad P(X_{m,n} = (a, goal, \varnothing, \varnothing))$$

are probabilities of home and away goals at the n-th step of the m-th fixture. Aggregating these quantities across all the steps of the given fixture, we can calculate the expected goals supremacy for the fixture. Let $g^{(m)}$ be the observed goal supremacy in the m-th fixture calculated with respect to the home team (i.e. the home team goals minus the

away team goals). The corresponding estimate can be calculated by aggregating the elements of the $d_{m,n}$ distributions corresponding to the goal states for all the steps in this fixture

$$\hat{g}^{(m)} = \sum_{n=1}^{N_m} [P(X_{m,n} = (h, goal, \varnothing, \varnothing)) - P(X_{m,n} = (a, goal, \varnothing, \varnothing))] \tag{7.3.27}$$

where N_m is the number of steps of the chain in the m-th fixture. We can use $g^{(m)}$ and $\hat{g}^{(m)}$, for example, to calculate the average observed and expected goal supremacy for each i-th team during the season (38 games per team and 380 in total):

$$\begin{aligned} \bar{g}_i &= \sum_{m=1}^{380} \left(\mathbb{1}_{\{i \text{ plays at home in } m\}} - \mathbb{1}_{\{i \text{ plays away in } m\}} \right) g^{(m)}/38 \\ \hat{g}_i &= \sum_{m=1}^{380} \left(\mathbb{1}_{\{i \text{ plays at home in } m\}} - \mathbb{1}_{\{i \text{ plays away in } m\}} \right) \hat{g}^{(m)}/38 \end{aligned} \tag{7.3.28}$$

and compare these quantities to verify the model fit. Finally, we can also calculate a contribution of a given player to the expected goals supremacy of his team. A way to do this is proposed in the next section.

Note that in the above procedure we are not simulating games based on randomly generated numbers but are calculating the unconditional probabilities of the states at each step exactly from the transition matrices. This is done under the assumption that the number of steps of the chain is fixed for each game at the true value from the observed data. In order to actually simulate games without this assumption the model would need to be extended to a Markov process by an addition of a random variable governing the time between steps of the built in chain. This is beyond the scope of this investigation.

Comparison with an *average* player

Recall that the aim of the whole exercise is to evaluate each player in terms of how much he contributes to the success of his team. In chapter 3 we showed in general how a model linking individual skill vectors to game outcomes through vectors of individual performance can be used to assess player's contribution to the performance of a team in a given game. The considerable complexity of the model proposed in the current chapter makes direct use of that theoretical formula impractical. Therefore in the next paragraph we propose a method to approximate the player's contribution to the outcome of a game.

First of all, $\hat{g}^{(m)}$ is used as a measure of how well the teams competing in the m-th fixture were expected to perform with the players both of them fielded for it. Additionally, we could calculate the expected goal supremacy for this fixture assuming any set of

players by substituting them for the ones who actually played. In an attempt to isolate the impact of an individual player, say j, only he is substituted and all the other players are kept as they actually were in the given fixture. Any player could be used to substitute player j, however, we will use an *average* player of his position here, i.e. one with average individual random effects predictions among players of this position, and call the resulting supremacy expectation $\hat{g}_{-j}^{(m)}$.

Finally, the average of $\hat{g}^{(m)}$ and $\hat{g}_{-j}^{(m)}$ is calculated for all the fixtures in which the j-th player originally played

$$\begin{aligned} \hat{G}_j &= \frac{\sum_{m=1}^{380} \left(\mathbb{1}_{\{j \text{ plays for hosts in } m\}} - \mathbb{1}_{\{j \text{ plays for guests in } m\}} \right) \hat{g}^{(m)}}{\sum_{m=1}^{380} \left(\mathbb{1}_{\{j \text{ plays for hosts in } m\}} + \mathbb{1}_{\{j \text{ plays for guests in } m\}} \right)} \\ \hat{G}_{-j} &= \frac{\sum_{m=1}^{380} \left(\mathbb{1}_{\{j \text{ plays for hosts in } m\}} - \mathbb{1}_{\{j \text{ plays for guests in } m\}} \right) \hat{g}_{-j}^{(m)}}{\sum_{m=1}^{380} \left(\mathbb{1}_{\{j \text{ plays for hosts in } m\}} + \mathbb{1}_{\{j \text{ plays for guests in } m\}} \right)} \end{aligned} \tag{7.3.29}$$

and use $\hat{G}_j - \hat{G}_{-j}$ to measure the average expected contribution to the goal supremacy of player j above what an average player of his position would be expected to produce. We call it *expected goal supremacy above average*.

7.4 Results

In this section we present results of fitting the models of section 7.3 to the data from the 2006/07 season of the English Premier League. First, in section 7.4.1 the transition matrix of the basic model is estimated. In section 7.4.2 player-dependent fixture-specific transition matrices are estimated and used to compare players based on their contribution to the expected goal supremacy above average. The predictive utility of this metric is evaluated in section 7.4.3. In appendix C we study the sensitivity of the results to the choice of the pitch division scheme.

7.4.1 Basic model of the game

Estimate of the transition matrix Figure 7.4 presents the estimate of the transition matrix introduced in figure 7.2 based on sample frequencies. Unsurprisingly, there is some positive correlation between the location of consecutive passes as passes played further up the pitch, with lower $s^{(l)}$, tend to be followed by other passes with low $s^{(l)}$. Similarly, shots are more likely to follow such passes and are more likely to lead to a goal the closer the shot origin is to the opponents goal. However, giving up possession to the other team is still the most probable outcome of a shot as many of them miss the target and result in a goal kick.

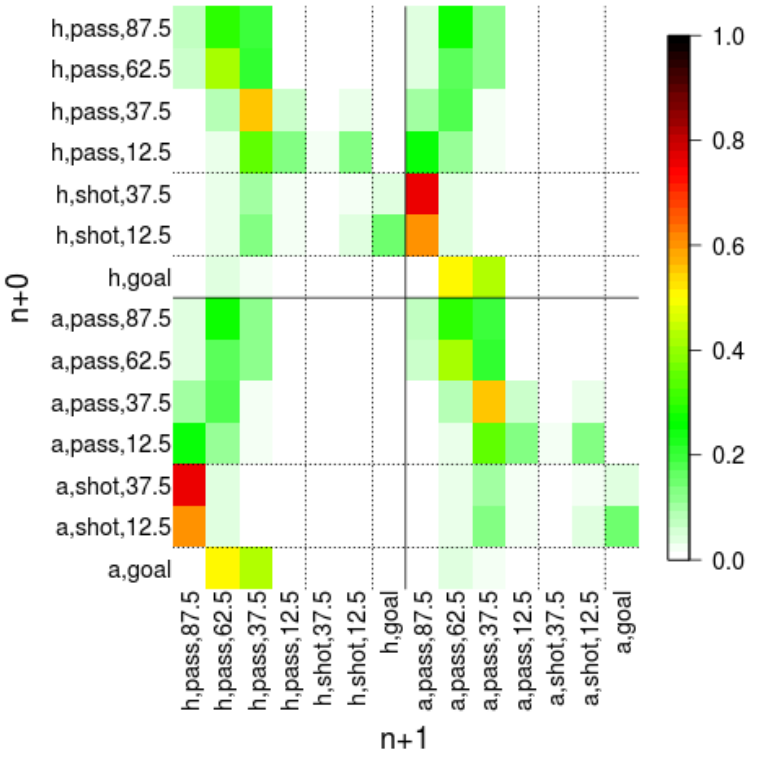


Figure 7.4: Estimate of the one step transition matrix in the basic Markov chain model.

Testing of the Markovian property It may be interesting to investigate whether the Markovian property from equation 7.3.1 holds for a football game and if not, in what situations and to what extent it is violated. Figure 7.5 graphically explores a weaker version of this assumption in which transition probabilities from step $n+1$ to $n+2$ are compared depending only on the *team* element of the state triplet to check that:

$$P(X_{n+2} = s_j | X_{n+1} = s_i, X_n^{(t)} = h) = P(X_{n+2} = s_j | X_{n+1} = s_i, X_n^{(t)} = a) \qquad (7.4.1)$$

Each panel of figure 7.5 corresponds to a single row in figure 7.4, however, instead of a single estimate, two sets of values are compared here depending on the team in possession at the step $n+0$. The distributions appear different visually and the difference was confirmed by Chi-square tests for equality of distributions meaning that the assumption in equation 7.4.1 does not hold, hence neither does the one in equation 7.3.1, and hence the Markovian property is violated. Including such dependencies could improve the model in the future.

7.4.2 Player specific model of the game

In this section the player specific model is fitted to the data from the 2006/07 season of the English Premier League. First, the shot conversion, the pass direction, the pass

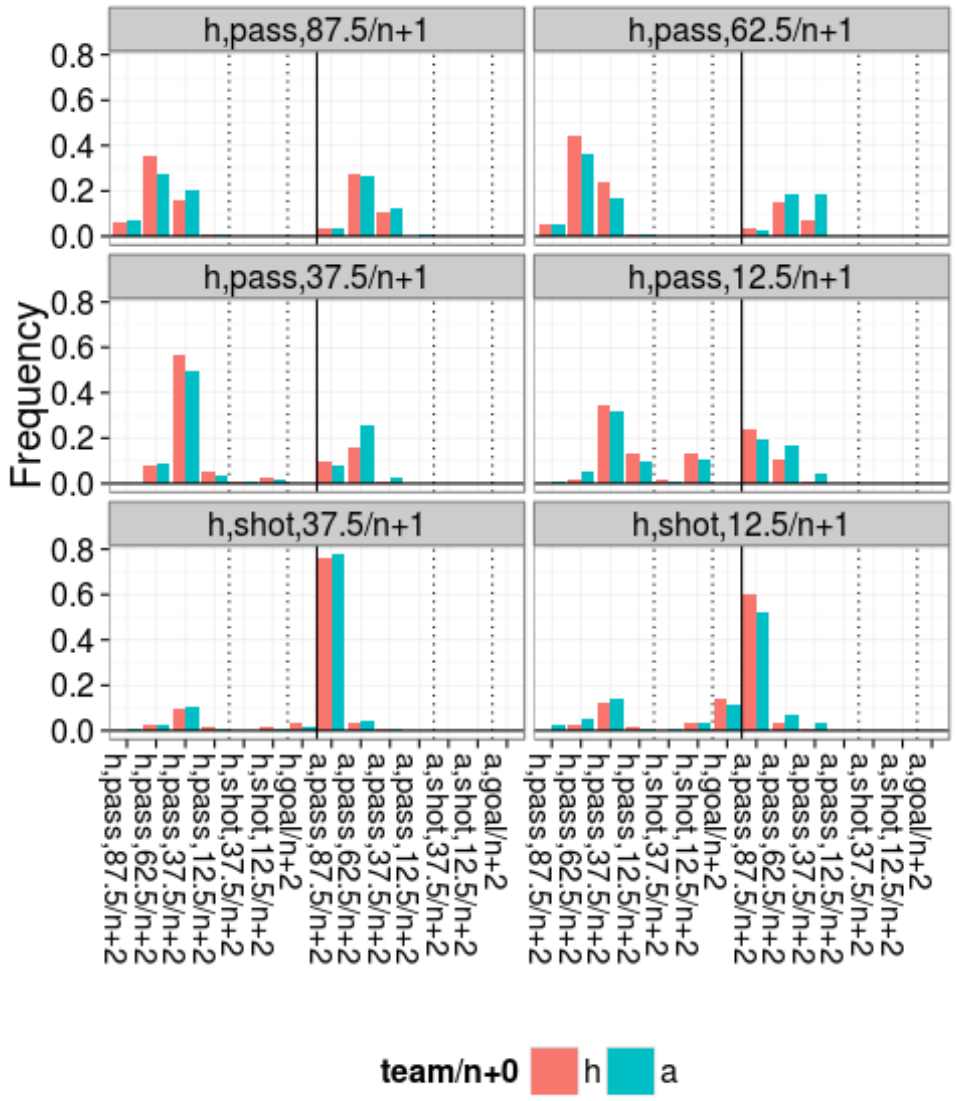


Figure 7.5: State transition from $n+1$ to $n+2$ depending on the team in possession at $n+0$.

completion and the action choice models are estimated. They are combined to work out two sets of fixture specific transition matrices for the following scenarios:

- with all the players who actually played in the 2006/07 season fixtures and
- with specific players replaced one by one, in turn, by an average player of their position.

Average expected goal supremacy in the season is then calculated based on the two sets of transition matrices. For a given player the comparison between the two supremacies (one with him in the team and one with him replaced by an average player) is used to assess his value for that team (see equation (7.3.29)).

Sub-models

Shot conversion model Table 7.3 presents estimates of the shot conversion model.

$\beta_0^{(g)}$	-1.861	$[-1.948$	$-1.774]$
$\beta_{37.5}^{(g)}$	-1.524	$[-1.776$	$-1.271]$
σ^2	0.105		
AIC	5317.145		
BIC	5338.027		
N	7790		

Table 7.3: Summary of the shot conversion model fit (with 95% confidence intervals for parameter estimates).

The values of the fixed effect estimates imply that $\frac{\exp(-1.861)}{1+\exp(-1.861)} \approx 13.4\%$ of shots from the 12.5 zone and $\frac{\exp(-1.861-1.524)}{1+\exp(-1.861-1.524)} \approx 3.3\%$ outside of it are expected to lead to a goal when executed by an average player. Figure 7.6 shows the histogram of the player random effects predictions to give an idea about the estimated distribution of player shooting skills.

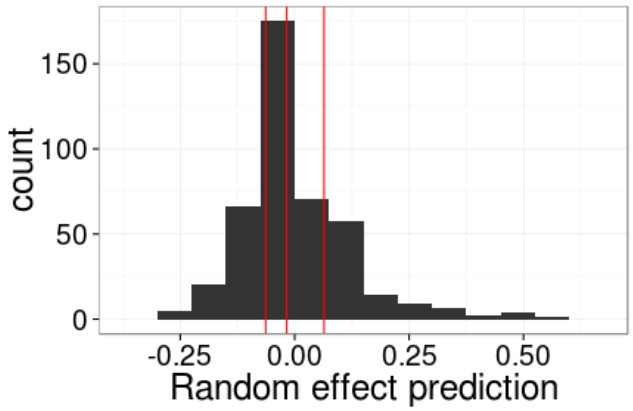


Figure 7.6: Histogram of the predicted values of player random effects in the shot conversion model. The vertical red lines mark the 25%, 50% and 75% quartiles of the distribution.

Since the values are difficult to interpret on this scale, figure 7.7 combines them with the fixed effects estimates from table 7.3 to give player specific shot conversion predictions. Even the worst players are expected to convert a higher percentage of shots from the 12.5 zone than even the best players when shooting from outside of it.

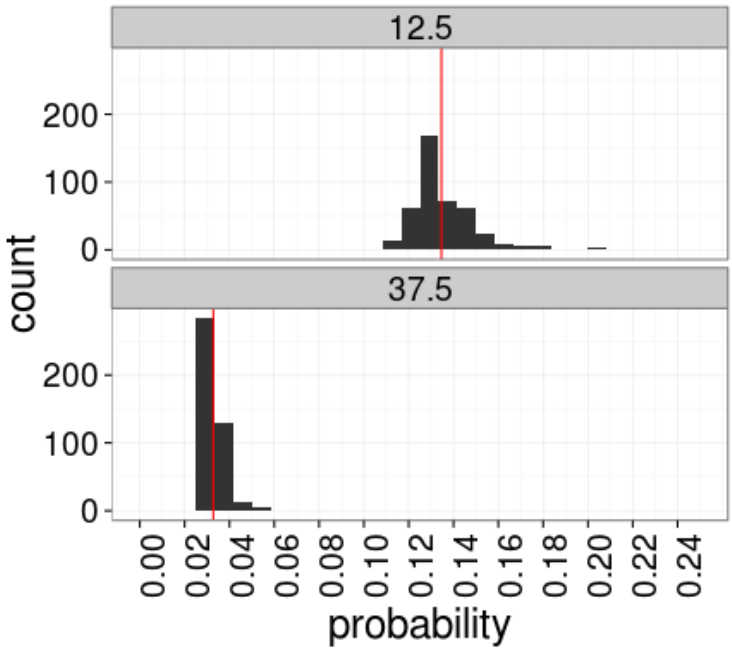


Figure 7.7: Histograms of player specific shot conversion rates predictions depending on where the shot is taken from. The vertical red lines correspond to average player predictions, i.e. the values of 13.4% and 3.3% from the last page.

Pass direction model Table 7.4 presents parameter estimates of the pass direction model and figure 7.8 shows some example predictions to aid interpretation. In general, for each zone the pass can originate from, the most likely destination tends to be the same zone, followed by the adjacent zones. The exception are passes originating close to ones own goal (zone 87.5) which are most often played one zone ahead. Goalkeepers generally pass more forward than other players which makes sense since they tend to execute goal kicks or kick the ball long following a catch.

$\beta^{(d)}_{37.5}$	1.807	[1.772	1.842]
$\beta^{(d)}_{62.5}$	4.093	[4.055	4.132]
$\beta^{(d)}_{87.5}$	5.282	[5.231	5.334]
$\beta^{(d)}_{G}$	−2.203	[−2.249	−2.158]
$\beta^{(d)}_{CD}$	0.067	[0.037	0.097]
$\beta^{(d)}_{LRD}$	0.063	[0.036	0.090]
$\beta^{(d)}_{CM}$	−0.046	[−0.072	−0.020]
$\beta^{(d)}_{LRM}$	−0.122	[−0.150	−0.095]
ζ_{25}	0.244	[0.209	0.279]
ζ_{50}	3.750	[3.711	3.788]
ζ_{75}	6.853	[6.809	6.896]
AIC	560263.957		
BIC	560381.287		
N	316911		

Table 7.4: Summary of the pass direction model fit (with 95% confidence intervals for parameter estimates).

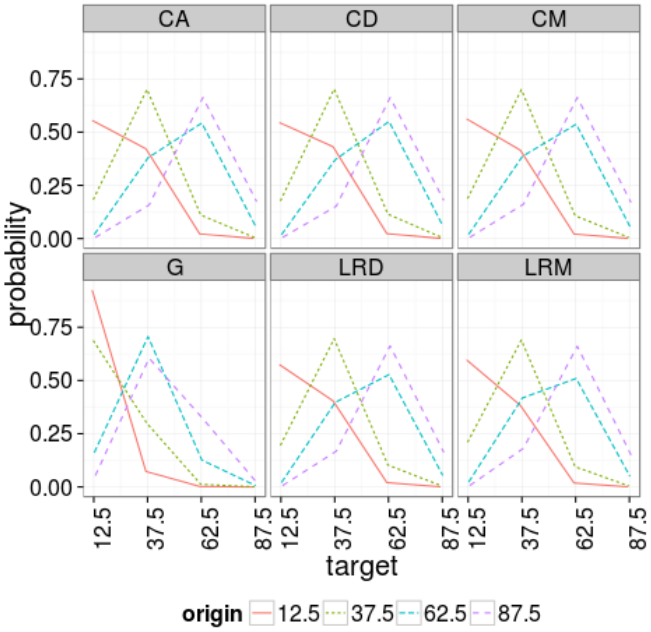


Figure 7.8: Predicted probability of a pass being directed to a given zone depending on its origin and the nominal position of the executing player.

Pass completion model Table 7.5 presents estimates of the pass completion model. Some example predictions for an average goalkeeper and an outfield player are shown

$\beta_0^{(w)}$	0.063	[0.001	0.125]
$\beta_{.,37.5}^{(w)}$	1.163	[1.078	1.249]
$\beta_{.,62.5}^{(w)}$	0.726	[0.492	0.959]
$\beta_{.,87.5}^{(w)}$	1.38	[0.588	2.172]
$\beta_{37.5,.}^{(w)}$	−0.278	[−0.339	−0.218]
$\beta_{62.5,.}^{(w)}$	−1.31	[−1.395	−1.226]
$\beta_{87.5,.}^{(w)}$	−2.102	[−2.260	−1.944]
$\beta_{37.5,37.5}^{(w)}$	0.201	[0.111	0.291]
$\beta_{62.5,37.5}^{(w)}$	1.308	[1.067	1.548]
$\beta_{87.5,37.5}^{(w)}$	0.265	[−0.566	1.096]
$\beta_{37.5,62.5}^{(w)}$	0.3	[0.192	0.408]
$\beta_{62.5,62.5}^{(w)}$	1.679	[1.436	1.922]
$\beta_{87.5,62.5}^{(w)}$	2.758	[1.957	3.559]
$\beta_{37.5,87.5}^{(w)}$	0.132	[−0.035	0.299]
$\beta_{62.5,87.5}^{(w)}$	1.995	[1.719	2.271]
$\beta_{87.5,87.5}^{(w)}$	2.388	[1.576	3.200]
$\beta_G^{(w)}$	0.458	[0.348	0.568]
ϕ^2	0.092		
AIC	363515.567		
BIC	363707.863		
N	322244		

Table 7.5: Summary of the pass completion model fit (with 95% confidence intervals for parameter estimates).

in figure 7.9. For example, an average outfield player is expected to complete about 55% of passes advancing the ball from zone 62.5 to zone 37.5 (marked by the ♦ symbol) and just about 45% of passes when advancing the ball from zone 37.5 to zone 12.5 (the ▲ symbol). Generally the further up the pitch a pass is targeted to, the lower the chance that it will be completed because of the intensified defensive effort of the opposite team. Even though the plot for the goalkeepers shows passes from all the zones, the curve corresponding to the zone closest to his goal (87.5) is most relevant as they will very rarely pass from anywhere else. It is shifted upwards with respect to the corresponding curve of an outfield player which reflects the fact that normally a goalkeeper has got full

control of the ball (following a catch or during a goal kick) and is given a lot of time to execute a pass. This advantage tends to be worth around 5-10 percentage points of the expected pass completion rate according to the model estimates (regardless of the target zone).

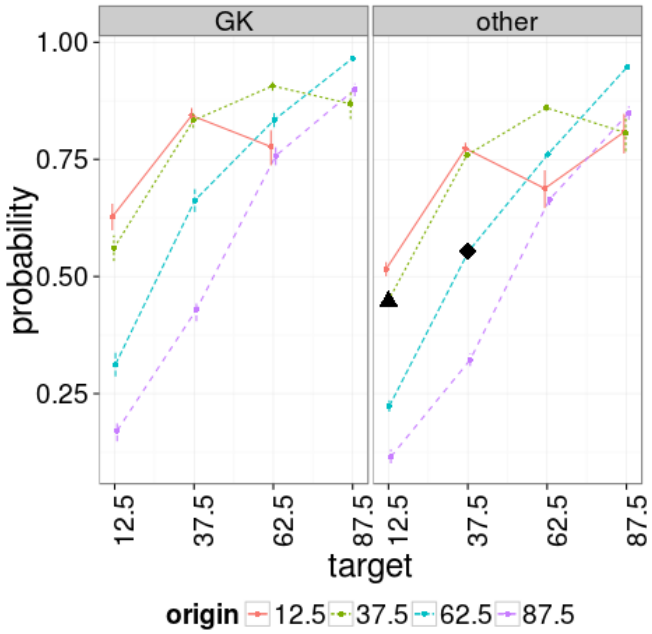


Figure 7.9: Predicted pass completion rate of an average player depending on the zone of origin and the targeted zone with approximate normal confidence intervals (based only on the fixed effects uncertainty). Symbols ▲ and ♦ are explained in the main text.

Figure 7.10 shows the histogram of the player random effects predictions to give an idea about the estimated distribution of the passing skill.

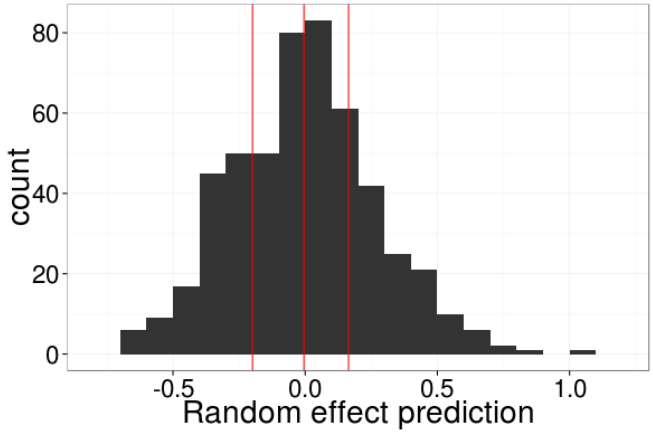


Figure 7.10: Histogram of the predicted values of player random effects in the pass completion model. The vertical red lines mark the 25%, 50% and 75% quartiles of the distribution.

The values are difficult to interpret on this scale so figure 7.11 puts them on the scale of the pass completion rate by showing distributions of the predictions given the zone of the pass origin (rows) and the targeted zone (columns). The vertical lines (solid

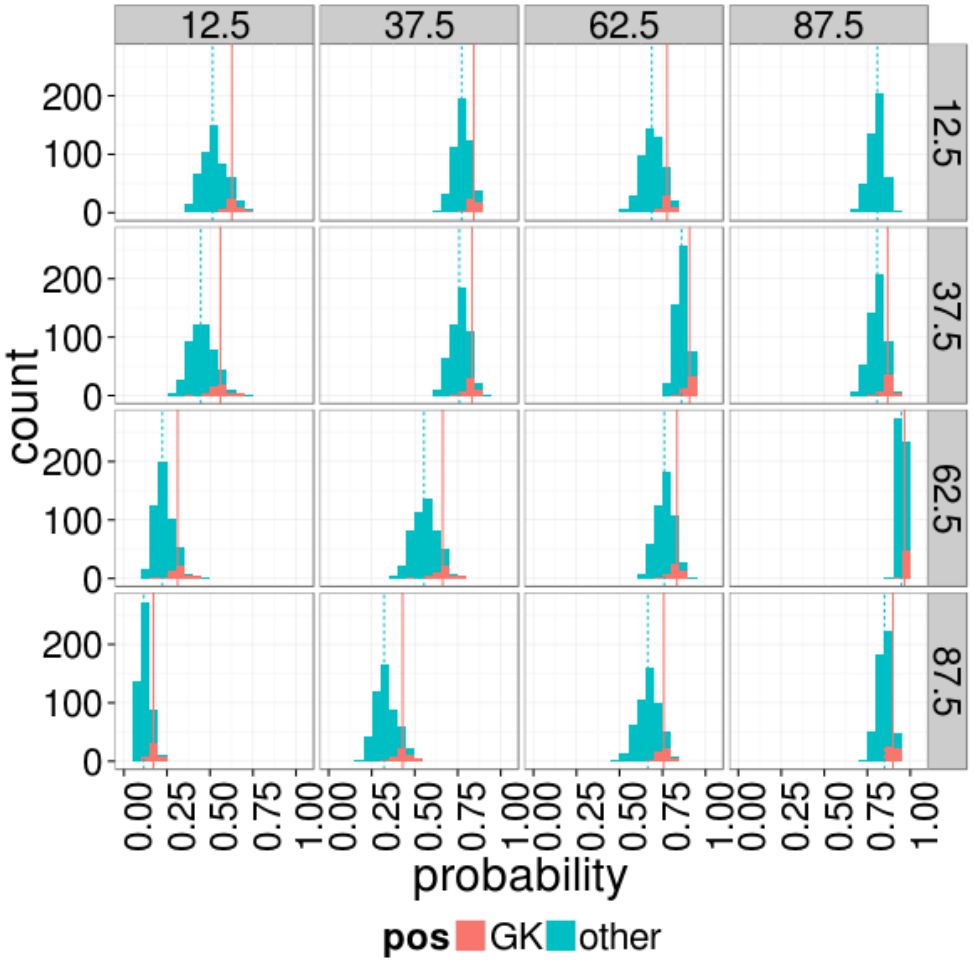


Figure 7.11: Distributions of the predicted player specific pass completion rate depending on the zone of origin (rows) and the targeted zone (columns) for goalkeepers and outfield players. The vertical lines indicate predictions for an average player: solid red for a goalkeeper and dashed blue for an outfield player.

red for the goalkeepers and dashed blue for the outfield players) are average player predictions and they correspond to the values plotted in figure 7.9. For example, the dashed blue vertical line at around 0.45 in the second row and the first column panel of figure 7.11, for passes from 37.5 to 12.5, corresponds to the point marked by the ▲ symbol in figure 7.9. It is interesting to compare the distribution of the predictions for different types of passes. For example, when passing from zone 62.5 to zone 37.5 (the third row and the second column of the figure 7.11), i.e. advancing the ball from one's

own half to the opponent's half of the pitch, the success rate of the best outfield player is expected to be about twice as large as that of the worst (70% vs. 35%) but there is hardly any difference expected between the same two players when passing backwards from the same zone to the 87.5 zone (the third row and the fourth column). This is because such a backwards pass is normally very easy to execute. There is little pressure or marking from the opposite team hence even an inaccurate pass may not lead to a loss of possession. In other words, there is not much room to demonstrate skill in such passes.

Choice model Table 7.6 presents estimates of the action choice model. The parameter estimates imply that players decide to shoot on approximately $\frac{\exp(-0.641)}{1+\exp(-0.641)} \approx 34.5\%$ of actions in zone 12.5 but only on $\frac{\exp(-0.641-3.676)}{1+\exp(-0.641-3.676)} \approx 1.3\%$ outside of it.

$\beta_0^{(z)}$	-0.641	$[-0.673$	$-0.609]$
$\beta_{37.5}^{(z)}$	-3.676	$[-3.730$	$-3.622]$
AIC	43450.103		
BIC	43492.935		
N	330362		

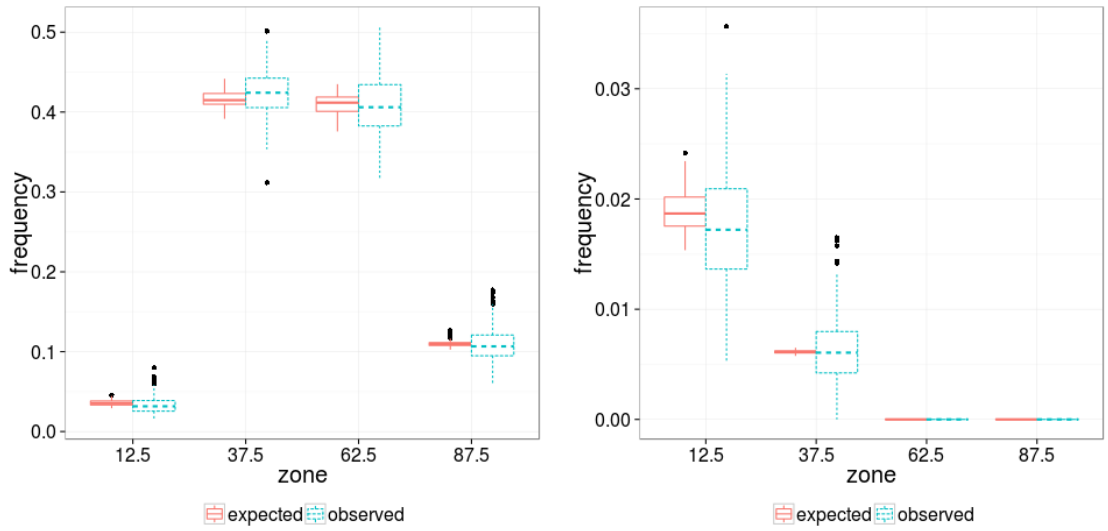
Table 7.6: Summary of the action choice model fit (with 95% confidence intervals for parameter estimates).

Fixture state distributions

Once the sub-models are estimated, they can be used to fill in elements of the fixture-specific transition matrices $\mathbf{P}_m$ as described in the discussion around equation (7.3.8). These in turn are used to calculate unconditional distributions, $d_{m,n}$, of the events in corresponding fixtures based on equation (7.3.26). Figure 7.12 attempts to verify whether this procedure gives sensible predictions by comparing the location attribute of the predicted distributions to the frequency of the observed events in the fitting period. The general pattern is captured by the model quite well.

In the proposed framework of player evaluation, the two elements of the unconditional distribution corresponding to the goal states

$$P(X_{m,n} = (h, goal, \varnothing, \varnothing)) \quad \text{and} \quad P(X_{m,n} = (a, goal, \varnothing, \varnothing))$$



(a) Passes

(b) Shots

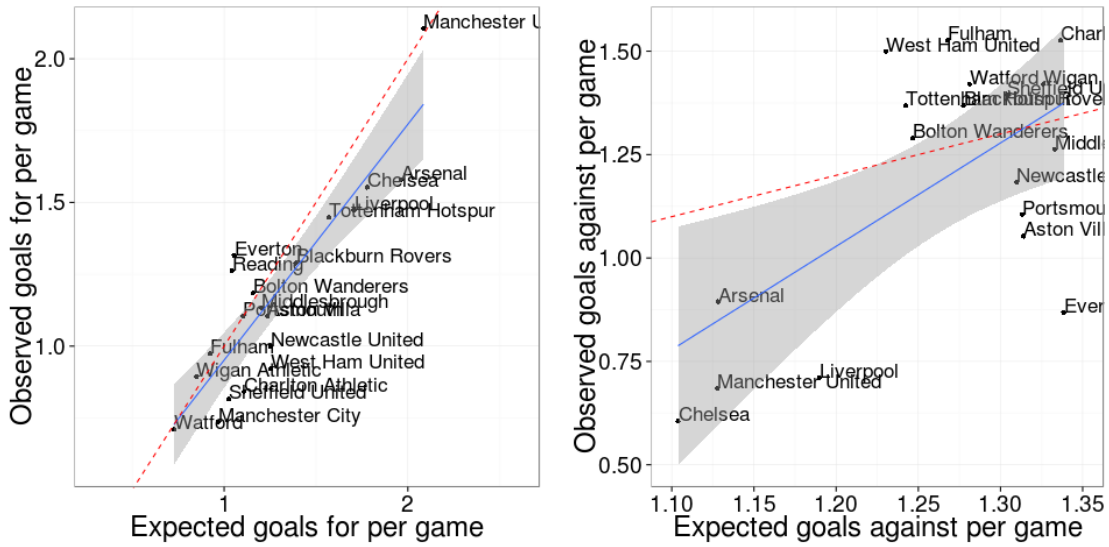
Figure 7.12: Expected (solid red) and observed (dashed blue) frequency of events per game in a given zone.

are of particular importance. Aggregating the two measures across all the event steps of the m-th fixture, gives the expected goals for both teams, for that fixture. The difference between the two aggregated measures is the expected goals supremacy $\hat{g}^{(m)}$ in the m-th fixture (see equation (7.3.27)). Figure 7.13 compares the model implied season average goals for and against (the horizontal axis) with their observed counterparts for all the teams in season 2006/07.

Figure 7.14 does the same for the goals supremacy: $\bar{g}$ against $\hat{g}$ (see equation (7.3.28)). The relationship between the expected and the observed goals for is stronger than for the goals against, which is expected since our model of the game does not account for many defensive skills. Most importantly, however, there is a good overall agreement between the expected and the observed goals supremacy per team (figure 7.14), which forms a basis of the proposed player evaluation method.

Comparison with an *average* player

One way to determine a player's value for his team is to compare expected results with and without him in the line up. The Markov chain model enables such a comparison since it is possible to conduct all the calculations from the previous section assuming an alternative line up to the one that was actually fielded in any given game. In particular, it is possible to calculate the average expected goals supremacy of a given team with a specific player replaced by one with average skills for his position. Comparing this value to



(a) Goals for

(b) Goals against

Figure 7.13: Expected and observed goals per game by team. The dashed red line is the identity function and the solid blue line is the linear model fit.

the corresponding expected supremacy for the actual player in the line up (see equation (7.3.29)) informs about his contribution to the average expected goals supremacy above the average player.

Figures 7.15 and 7.16 show how swapping an *average* player for Cristiano Ronaldo changes the expected distribution of the game states for his team, Manchester United, and their opposition.

These changes may appear small but translate to about 0.17 expected goal supremacy difference meaning that Ronaldo is expected to contribute that much for his team above an average player. To put this value into context, refer to figure 7.14 in which the total expected goal supremacy per game for Manchester United is almost 1 goal. According to the model almost half of this value is due to two best players of this team: Ronaldo and Paul Scholes (whose expected goal supremacy above average is 0.27).

Figure 7.17 shows the expected goal supremacy contribution above average for all the players. Despite the fact that the players are evaluated based on only one season of data, the list generally consists of players that would be recognised as solid performers by any football fan. The few surprises (like no Steven Gerrard) can probably be attributed to the small sample size or the fact the the model attempts to capture only two of a wide range of player skills. Furthermore, there is a good mix of player tactical positions in the top 20 list but it is to some extent by design of the metric which relates player's expected supremacy contribution to that of an average player of his position.

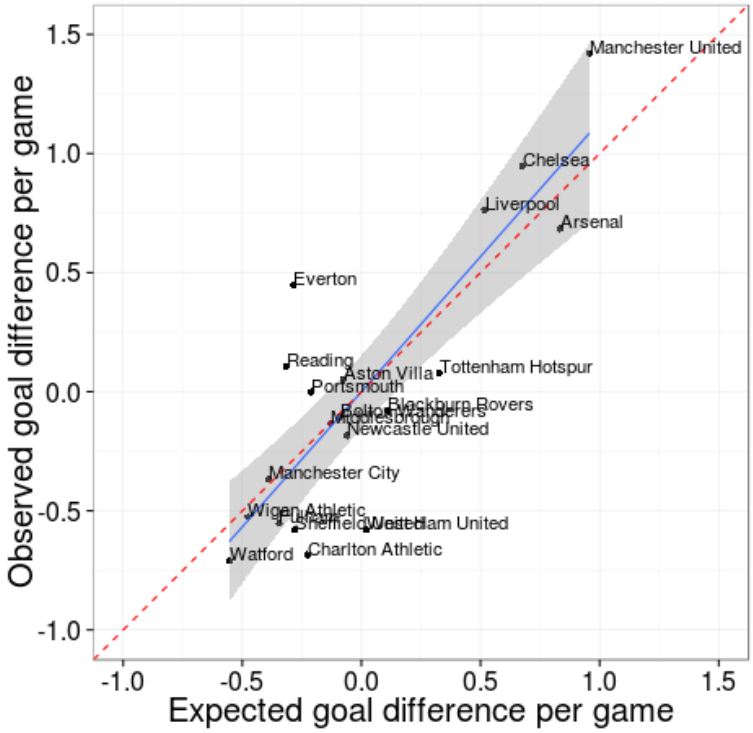


Figure 7.14: Expected and observed goal supremacy per game by team. The dashed line is the identity function and the solid line is the linear model fit.

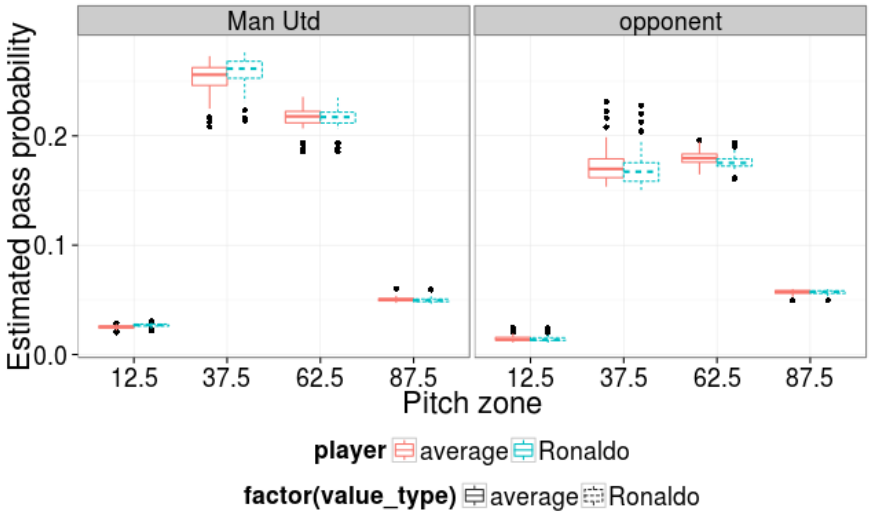


Figure 7.15: Estimated probability distribution of passes in Manchester United games with Cristiano Ronaldo (dashed blue) or his average replacement (solid red) in the team.

Finally, note that even though the shooting and passing models constituting the Markov chain model of the game are not quite the same as the shooting and passing models of chapters 5 and 6, there are some common names between figure 7.17 and the goalscoring

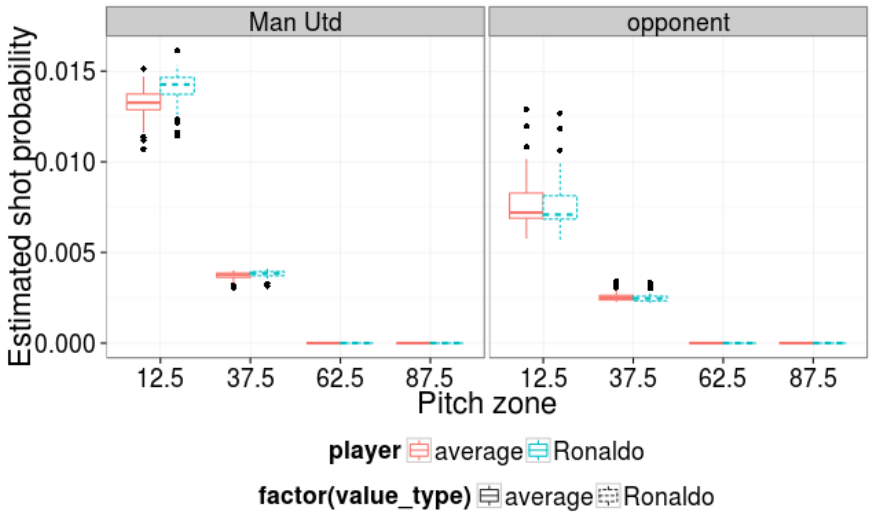
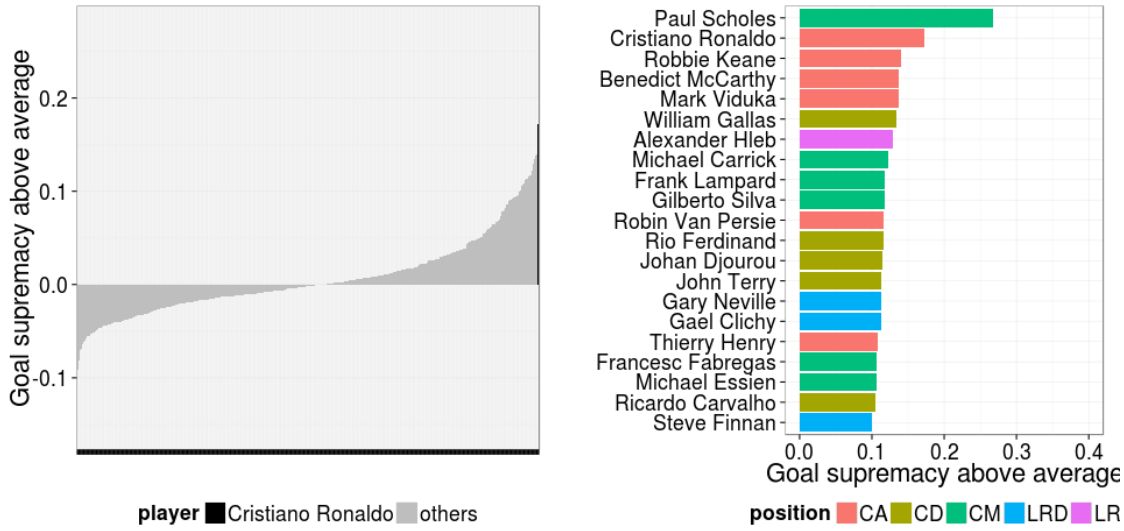


Figure 7.16: Estimated probability distribution of shots in Manchester United games with Cristiano Ronaldo (dashed blue) or his average replacement (solid red) in the team.



(a) All players

(b) Top 20 players

Figure 7.17: Expected goal supremacy above an average player in their position in season 2006/07 (left panel, with Ronaldo marked as a black bar to the far right) and the names of the top 20 of them (right panel).

rankings in table 5.6 (Ronaldo, Van Persie, Viduka, McCarthy) and passing rankings in table 6.3 (Scholes, Ronaldo, Gallas, Hleb, Carrick, etc.). From this comparison we may guess that the top position of Paul Scholes comes probably from his fantastic passing ability, whereas the second position of Cristiano Ronaldo is due to him being near the top of both the passing and the goalscoring charts. The appealing feature of the compre-

hensive model of the game presented in this chapter is that it combines these two skills into a single meassure of player contribution without any preconcieved knowledge of their relative importance.

7.4.3 Evaluating predictive utility

The ultimate test of a rating method is its predictive utility. It is future performance that teams are interested in and should be paying for. One way of assessing the predictive utility of a player evaluation method was proposed in section 6.4.4. It is based on the premise that teams should be interested in optimising their overall future performance when evaluating players. Therefore, what has to be verified is whether the proposed method of player evaluation can be turned into a good predictor of future team performance.

With this in mind, for every fixture of the 2007/08 season we calculate three statistics supposed to capture the general level of the skill in both competing teams:

- Average level of the expected goals supremacy above an average player among all the players (calculated based on the 2006/07 data);
- Average level of:
  - average pass completion rate in season 2006/07;
  - average shot conversion rate in season 2006/07.

For each of these indices we check how well the difference in their values for competing teams predicts the result of a fixture between those teams.

The Pearson correlation coefficient for the pass completion based index with the home team goals supremacy in season 2007/08 is 0.308 with a 90% confidence interval of $(0.228, 0.383)$ and $0.263(0.181, 0.341)$ for the shot conversion based index, whereas its value for the expected goals supremacy contribution is $0.383(0.308, 0.454)$.

Secondly, we fit ordered logit regression models of the game outcome (home win, draw or away win) with the difference in the average skill index for the home and away team as the only covariate: one model for the index based on the pass completion rate, one based on shot conversion rate and one for the expected goals supremacy contribution based index. The last of these models offers the best fit with an AIC of 730.39 compared to 741.84 for the pass completion rate based model, and 769.13 for the shot conversion based model. Even a model combining the pass completion and shot conver-

sion indices in an additive way has an AIC value of 736.27 suggesting that it is not as good at accounting for these skills in the expected goal supremacy contribution metric.

In appendix C we examined the sensitivity of the results of this chapter on the choice of the pitch location scheme and found that none of the conclusions change.

7.5 Discussion

In this chapter we proposed a model of a football match which attempts to capture the dynamics of the game events with a Markov chain. The transition matrix for each fixture depends on random effects representing skills of the players involved in this fixture. Based on this matrix we can calculate the expected goals supremacy between the two competing teams. This design enables us to also consider alternative scenarios, e.g. one in which a specific player is replaced by a different player. The transition matrix can be re-evaluated for this case and the expected goals supremacy can be recalculated. Comparing the expected goals supremacy with the two alternative players in the team can be a way to determine their relative value. This can be done over multiple fixtures against various opponents.

The proposed method has several weaknesses including:

1. The assumptions of the Markovian property and time homogeneity of the process are violated.

2. Defensive skills are not evaluated, at least not beyond the defensive contribution of the passing skill in terms of keeping the ball away from the opponents.

3. It ignores the impact of players other than the one executing a given action on its outcome.

4. The choice of the zones the pitch is split into is arbitrary.

In appendix C variations of the model with alternative pitch division schemes are studied and the choice between them is shown not to be critical to the results. The other issues apply more to the particular implementation of the proposed framework for player evaluation than to the framework itself. They could be addressed by extensions of the model explored here in the following ways (in the order of the above points):

1. The Markovian assumption could be relaxed by defining the states of the process as consisting of characteristics of multiple consecutive game events rather than a single one. For example, since in section 7.4.1 the transition probability was

found to depend on the team in possession of the ball before the event, the state could be extended to include the information about the team in possession during the previous event.

2. The model of the game could be extended to include actions that depend on other individual skills. For example,

   - Rather than having only a binary success-failure outcome, passes may be additionally assumed to result in a duel between players of the opposite teams. The outcome of such duels could depend on a duel winning skill of the competing players.
   - Running with the ball could be considered as another way of moving it around the pitch. The player with the ball could be challenged for its possession by a tackle from on opponent, the outcome of which could depend on a dribbling skill of the former and a tackling skill of the latter player.
   - Shot outcomes could be assumed to depend on the saving skill of the goalkeeper and not just the skill of the shooting player.

3. The outcome of each action could be naturally extended to depend on random effects representing skills of other players. For example, the outcome of a pass could, in addition to the passing skill of the executing player, depend on random effects representing off the ball movement of his team mates and the marking skill of the opponents. The impact of these skills could be weighted by the distance of these other players from the one passing the ball. Such an extension would require player positioning data not available for this research.

Even without these extensions the model presented here has the following advantages compared to alternatives reviewed in chapter 2:

- It evaluates players based on actions that they actually execute (unlike the plus-minus methods).
- It recognises that individual performance is not equivalent to skill and can be affected by factors beyond the player's control.
- It accounts for the context, such as location, of the individual performance when evaluating the skill behind it.
- It allows players to be characterised by a vector of multiple skills, rather than a single one, and captures the importance of each of them for team success "for

free" through the Markov chain structure. For instance, no additional regression model for team performance on the individual performance is required.

- It expresses a player's relative value in terms of team performance. In the example implementation presented in this chapter the expected goals supremacy was proposed, however, it could be translated into expected points gain using a higher level model. This in turn could be used to work out the expected probability of, say, a top four finish in the league and players could be compared based on their expected contribution to it. Such achievements could be easier to attach a monetary value to, which would be important if player valuation was the purpose of the analysis.

- It allows the value of a given player to vary for different teams depending on what other players they consist of. For example a team with players who pass the ball very well will put its strikers in shooting positions more often than other teams, thus the shot converting skill of the strikers will have a bigger impact on its results. As a result, players who shoot the ball very well will be particularly important for this team.

The ultimate test of the rating method proposed in this chapter is an examination of its predictive utility. This was done in section 7.4.3 where the level of the contribution to the expected goals supremacy above an average player, among players of two competing teams, is shown to predict the outcome of a game between them relatively well.

Chapter 8

Conclusions

In this concluding chapter we summarise the findings of this research (in section 8.1), consider limitations of the proposed methods (in section 8.2) and give recommendations for future research (in section 8.3).

8.1 Summary of the findings

The aim of this thesis is to propose a framework for assessing skill of football players for the purpose of establishing their value to a club. In chapter 3 we argued that the desired properties of such a framework are that:

- it recognises that individual performance depends on the player's underlying skill as well as factors beyond his control, and
- it establishes a link between the individual performance and the team success.

In chapters 5 and 6 we provided examples to support the need for the first of these properties by analysing shots and passes in isolation. Additionally, the analysis presented there serves two purposes:

- As standalone tools for evaluating players' goalscoring and passing abilities.
- As components of a bigger model of the football game in which a player's shooting and passing skills are responsible for a part of his total contribution to the team success.

With respect to the first of these points we found that:

- In both cases mixed effects models provide superior out-of-sample predictions than simple methods extrapolating past performance metrics to the future.

- Home advantage, time spent on the pitch, position in the tactical formation as well as the opposition faced in the game influence the number of shots a player attempts per unit of time.
- Only home advantage was found to affect the rate at which players convert shots to goals.
- Pass completion rate was found to depend on:
  - its origin and destination;
  - part of the body used to execute the pass;
  - time since the previous pass;
  - its number in the current team possession;
  - game time of the execution;
  - whether or not it followed a duel (and the type of the duel);
  - position in the tactical formation of the executing player;
  - home advantage.

  These covariates served as proxies of:
  - The degree of control the executing player has on the ball when attempting the pass;
  - The player's orientation with respect to the direction of play;
  - The level of defensive pressure put on the executor of the pass;
  - The distance of the attempted pass;
  - The level of marking on the player receiving the ball;
  - Familiarity with the type of situation the pass is attempted in.

The application of the methodology proposed in chapters 5 and 6 is not limited to goal scoring and passing statistics. Inference about player skill from any other statistic could benefit from such a mixed effects formulation.

In chapter 7 we proposed a framework for evaluating football players having the two desired properties named at the beginning of this section. The proposed method attempts to address one of the key difficulties in modelling the game of football, i.e. its free-flowing nature, by discretising it into a series of events. The evolution of the game from one event to another is described using a Markov chain model in which each

game is described by a specific transition matrix with elements depending on the skills of the players involved in this game. Based on this matrix it is possible to calculate game outcome related metrics such as expected goals difference between the two teams. Thanks to this it is possible to establish a link between specific players and their skills and the game outcome. The skills come from separate, location specific, models like those in chapters 5 and 6.

The proposed framework has the following advantages compared to alternatives from the literature reviewed in chapter 2:

- It evaluates players based on actions that they actually execute (unlike the plus-minus methods).
- It recognises that individual performance is not equivalent to skill and can be affected by factors beyond the player's control.
- It accounts for the context, such as location, of the individual performance when evaluating the skill behind it.
- It allows players to be characterised by a vector of multiple skills, rather than a single one, and captures the importance of each of them for team success "for free" through the Markov chain structure.
- It expresses player's relative value in terms of team performance.
- It allows the value of a given player to vary for different teams depending on what other players they consist of.

Even though the primary motivation of the research was to construct metrics that could be used for player valuation (for example using a procedure of Tunaru and Viney (2010)), the application of the methods developed here is not limited to this purpose. For example, because it allows to evaluate various elements of players' skill set and determine their importance to game outcomes, the model proposed here could also be useful in designing player development strategies, training regimes or game tactics.

8.2 Limitations of the study

The research described in this thesis has several limitations that need to be considered when interpreting its findings.

- It is based only on two seasons of data from a single competition.

- Only two skills (passing and shooting) are analysed. In particular, defensive skills are not evaluated, at least not beyond the defensive contribution of the passing skill in terms of keeping the ball away from the opponents.
- Players' skills are assumed to be static in time.
- Correlation between different skills of an individual player is not considered.
- The assumption of the Markovian property and time homogeneity of the Markov chain model are likely to be violated.
- It ignores the impact of players other than the one executing a given action on its outcome.

8.3 Recommendations for future work

The above issues apply more to the particular implementation of the proposed framework for player evaluation than to the framework itself. They could be addressed by extensions of the model explored here in the following ways:

- The Markovian assumption could be relaxed by defining the states of the process as consisting of characteristics of multiple consecutive game events rather than a single one.
- The model of the game could be extended to include actions that depend on individual skills other than passing and shooting.
- The outcome of each action could be naturally extended to depend on random effects representing skills of other players. For example the shot stopping ability of a goalkeeper is likely to have an impact on shot conversion. Similarly, the skill of the pass receiver probably affects the pass completion rate.

Appendix A

Some results from the estimation theory

The results in this appendix are not ours but are quoted from other publications (e.g. Lee et al., 2006; Jiang, 2007). We have gathered them here for reader's convenience.

A.1 Generalized Linear Models

A Generalized Linear Model is an extension of the linear model such that:

- The mean μ of the response variable y depends on a set of explanatory variables $X = [x_1, ..., x_k]$ through a monotone function of the linear predictor $\eta = X\beta$:

$$\mu = g^{-1}(\eta) \tag{A.1.1}$$

  where β is a column vector of k parameters and $g()$ is called the link function.

- For a given i-th observation, conditionally on the linear predictor, the response variable y_i comes from the exponential family with the log-likelihood function given by:

$$\ell(\theta_i|y_i) = \{y_i\theta_i - b(\theta_i)\}/\phi + c(y_i, \phi)\,. \tag{A.1.2}$$

  Knowing that (e.g. Nelder and Wedderburn (1972)):

$$E\left(\frac{\partial \ell(\theta)}{\partial \theta}\right) = 0 \tag{A.1.3}$$

and

$$E\left(\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right) = -E\left(\frac{\partial \ell(\theta)}{\partial \theta}\right)^2 \tag{A.1.4}$$

we can derive:

$$\mu_i \equiv E(y_i) = b'(\theta_i) \tag{A.1.5}$$

and

$$V(\mu_i) \equiv \frac{1}{\phi} var(y_i) = b''(\theta_i) \tag{A.1.6}$$

where $V()$ is the variance function linking the variance of the response variable to its mean. Note that from (A.1.5), and (A.1.6):

$$V(\mu_i) = \frac{\partial \mu_i}{\partial \theta_i}. \tag{A.1.7}$$

A.2 Iteratively Weighted Least Squares

For a single observation the derivative of the log-likelihood with respect to the r-th element of the parameter vector β is:

$$\frac{\partial \ell(\theta_i|y_i)}{\partial \beta_r} = \frac{\partial \ell(\theta|y_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_r} \tag{A.2.1}$$

therefore for all the observations we can write:

$$u_r \equiv \frac{\partial \ell(\theta|y)}{\partial \beta_r} = \sum_i \frac{\partial \eta_i}{\partial \mu_i} W_i x_{r,i} (y_i - \mu_i) \tag{A.2.2}$$

where

$$W_i = \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 V^{-1}(\mu_i). \tag{A.2.3}$$

In matrix representation for all β's it becomes:

$$u = X^T \Sigma^{-1} \frac{\partial \eta}{\partial \mu} (y - \mu) \tag{A.2.4}$$

where Σ is a diagonal matrix with elements $\Sigma_{ii} = 1/W_i$.

Furthermore, the r-th row in the s-th column of the negative expectation matrix of the second derivative is:

$$\begin{aligned} A_{rs} \equiv -E \frac{\partial u_r}{\partial \beta_s} &= -E \sum_i \left\{ (y_i - \mu_i) \frac{\partial}{\partial \beta_s} \left[\frac{\partial \eta_i}{\partial \mu_i} W_i x_{r,i} \right] + \frac{\partial \eta_i}{\partial \mu_i} W_i x_{r,i} \frac{\partial}{\partial \beta_s} (y_i - \mu_i) \right\} \\ &= \sum_i W_i x_{r,i} \frac{\partial \eta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_s} = \sum_i W_i x_{r,i} x_{s,i} \end{aligned} \tag{A.2.5}$$

and the whole matrix can be written as

$$A = X^T \Sigma^{-1} X \,. \tag{A.2.6}$$

We use equations (A.2.4) and (A.2.6) in the Fisher scoring algorithm (see section A.5):

$$X^T \Sigma^{-1} X \beta^{(1)} = X^T \Sigma^{-1} X \beta^{(0)} + X^T \Sigma^{-1} \frac{\partial \eta}{\partial \mu} (y - \mu) \tag{A.2.7}$$

which leads to Weighted Least Squares (WLS) equations for updating the value of the parameter vector, $\beta^{(1)}$, given its previous value $\beta^{(0)}$:

$$X^T \Sigma^{-1} X \beta^{(1)} = X^T \Sigma^{-1} z \tag{A.2.8}$$

where $z = \eta + \frac{\partial \eta}{\partial \mu}(y - \mu)$.

We solve equation (A.2.8) for $\beta^{(1)}$, set $\beta^{(0)} = \beta^{(1)}$, recalculate η, μ and Σ given the updated value of the parameter vector and iterate until convergence.

A.3 Penalized Iteratively Weighted Least Squares

A Generalized Linear Mixed Model (GLMM) is an extension of the GLM, outlined in appendix A.1, such that the response y depends on a vector of random effects b, in addition to some fixed effects β, through the linear predictor:

$$\eta_i = X_i \beta + U_i b \,. \tag{A.3.1}$$

The vector of random effects b is assumed to come from a multivariate normal distribution:

$$b \sim \mathcal{N}(0, G_\theta) \tag{A.3.2}$$

where G_θ is a covariance matrix depending on some unknown parameters θ.

The Penalized Iteratively Weighted Least Squares algorithm is used in the estimation algorithm outlined in section 5.3.1 to find $\tilde{b}$ maximizing the penalized log-likelihood function $f_{\theta,\beta}(b)$ from equation (5.3.6) for given (θ, β). It is an extension of the Iteratively Weighted Least Squares algorithm (section A.2) used for fitting Generalized Linear Models outlined in section A.1.

Note that the equations (A.2.8) for β in each step of the IWLS algorithm for fitting GLM can be viewed as WLS equations for the linear model:

$$z = X\beta + \varepsilon \tag{A.3.3}$$

where $E(z) = X\beta$ and $var(z) = \Sigma$. If b in equation (A.3.1) were fixed effects like β, we would simply append them to the fixed effects vector and write

$$z = \begin{pmatrix} X & U \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} + \varepsilon, \quad var(\varepsilon) = \Sigma \tag{A.3.4}$$

and use IWLS to fit what would simply be a GLM.

However, b are random effects as stated in equation (A.3.2). We can build this additional assumption into the linearised model form by augmenting the data in the following way:

$$z^* = X^*\beta^* + \varepsilon^*, \quad var(\varepsilon^*) = \Sigma^* \tag{A.3.5}$$

where

$$z^* = \begin{pmatrix} z \\ 0 \end{pmatrix}, \quad X^* = \begin{pmatrix} X & U \\ 0 & I \end{pmatrix}, \quad \beta^* = \begin{pmatrix} \beta \\ b \end{pmatrix}, \quad \Sigma^* = \begin{pmatrix} \Sigma & 0 \\ 0 & G \end{pmatrix}. \tag{A.3.6}$$

The above system of equations has the following WLS solution at each iteration of the IWLS algorithm:

$$(X^*)^T(\Sigma^*)^{-1}X^*\hat{\beta}^* = (X^*)^T(\Sigma^*)^{-1}z^* \tag{A.3.7}$$

or equivalently

$$\begin{pmatrix} X^T\Sigma^{-1}X & X^T\Sigma^{-1}U \\ U^T\Sigma^{-1}X & U^T\Sigma^{-1}U + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} X^T\Sigma^{-1}z \\ U^T\Sigma^{-1}z \end{pmatrix}. \tag{A.3.8}$$

Since we seek the mode of b for fixed values of β and θ we only need to solve:

$$(U^T\Sigma^{-1}U + G^{-1})\hat{b}(\theta,\beta) = U^T\Sigma^{-1}(z - X\beta) \tag{A.3.9}$$

for $\hat{b}(\theta,\beta)$.

A.4 Laplace approximation

Laplace's method is used to approximate $\int \exp\{g(x)\}\,\mathrm{d}x$ using the second order Taylor series expansion at $\tilde{x}$:

$$g(x) \approx g(\tilde{x}) + g'(\tilde{x})(x - \tilde{x}) + \frac{1}{2}g''(\tilde{x})(x - \tilde{x})^2. \tag{A.4.1}$$

If $\tilde{x}$ maximizes $g(x)$ then $g'(\tilde{x}) = 0$, $g''(\tilde{x}) < 0$ and as a result we receive

$$\int \exp\{g(x)\}\,\mathrm{d}x \approx \sqrt{2\pi}|-g''(\tilde{x})|^{-1/2}\exp\{g(\tilde{x})\}. \tag{A.4.2}$$

A.5 Fisher scoring

Let $\ell(\beta)$ be a twice differentiable function we want to find the maximum, β^*, of and let $f(\beta) \equiv \frac{\partial \ell}{\partial \beta}$ be its score function. Expanding the score function at the current guess $\beta^{(0)}$ of β^*:

$$f(\beta^*) \approx f(\beta^{(0)}) + f'(\beta^{(0)})(\beta^* - \beta^{(0)}) \tag{A.5.1}$$

and knowing that $f(\beta^*) = 0$ we can get closer to the maximum by updating:

$$\beta^* = \beta^{(0)} + \frac{\partial \ell}{\partial \beta}_{|\beta^{(0)}} \left[-\frac{\partial^2 \ell}{\partial \beta^2}_{|\beta^{(0)}} \right]^{-1}. \tag{A.5.2}$$

The above is Newton-Raphson algorithm which is known as Fisher scoring when the observed Fisher information is replaced with its expectancy:

$$\beta^* = \beta^{(0)} + \frac{\partial \ell}{\partial \beta}_{|\beta^{(0)}} \left[-E\left(\frac{\partial^2 \ell}{\partial \beta^2} \right)_{|\beta^{(0)}} \right]^{-1}. \tag{A.5.3}$$

Appendix B

Breakdown of transition probabilities

In this appendix we write the transition probabilities for the *choice*, *pass* and *shot* nodes of the player specific Markov chain model as products of conditional probabilities. The factored probabilities can be modeled more naturally conditional on players' skills.

The following equations follow the notation of section 7.3.3.

B.1 Choice node

At this node, a player can choose to pass or shoot with the next state having the same team attribute and the appropriate location:

$$P(X_{m,n+1}|X_{m,n}=(s_b^{(t)},choice,s_c^{(l)},s_d^{(q)}))=$$
$$\begin{cases} P(X_{m,n+1}=(s_b^{(t)},pass,s_c^{(l)},s_d^{(q)})|X_{m,n}=(s_b^{(t)},choice,s_c^{(l)},s_d^{(q)})) \text{ if} \\ \qquad X_{m,n+1}^{(t)}=X_{m,n}^{(t)}\wedge X_{m,n+1}^{(e)}=pass\wedge X_{m,n+1}^{(l)}=X_{m,n}^{(l)}\wedge X_{m,n+1}^{(q)}=X_{m,n}^{(q)} \\ P(X_{m,n+1}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)})|X_{m,n}=(s_b^{(t)},choice,s_c^{(l)},s_d^{(q)})) \text{ if} \\ \qquad X_{m,n+1}^{(t)}=X_{m,n}^{(t)}\wedge X_{m,n+1}^{(e)}=shot\wedge X_{m,n+1}^{(l)}=X_{m,n}^{(l)}\wedge X_{m,n+1}^{(q)}=X_{m,n}^{(q)} \\ 0 \text{ otherwise.} \end{cases} \tag{B.1.1}$$

Based on the definition of conditional probability we can rewrite the first of the cases as the following product:

$$\begin{aligned} &P(X_{m,n+1}=(s_b^{(t)},pass,s_c^{(l)},s_d^{(q)})|X_{m,n}=(s_b^{(t)},choice,s_c^{(l)},s_d^{(q)})) \\ =&P(X_{m,n+1}^{(e)}=pass|X_{m,n}=(s_b^{(t)},choice,s_c^{(l)},s_d^{(q)})) \\ &\times P(X_{m,n+1}^{(t)}=s_b^{(t)},X_{m,n+1}^{(l)}=s_c^{(l)},X_{m,n+1}^{(q)}=s_d^{(q)}|X_{m,n+1}^{(e)}=pass,X_{m,n}=(s_b^{(t)},choice,s_c^{(l)},s_d^{(q)})) \end{aligned} \tag{B.1.2}$$

where the second factor is the probability that the team, location and the player at step $n+1$ will be the same as at step n for a transition from a state of type *choice* to one of type *pass*. By the construction of the Markov chain (see figure 7.3) it is equal to 1 so the above becomes:

$$\begin{aligned} P(X_{m,n+1} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)}) | X_{m,n} = (s_b^{(t)}, choice, s_c^{(l)}, s_d^{(q)})) \\ = P(X_{m,n+1}^{(e)} = pass | X_{m,n} = (s_b^{(t)}, choice, s_c^{(l)}, s_d^{(q)})) \end{aligned} \quad \text{(B.1.3)}$$

i.e. the probability that player $s_d^{(q)}$ from team $s_b^{(t)}$ chooses to pass in location $s_c^{(l)}$.

Similarly in the second case:

$$\begin{aligned} P(X_{m,n+1} = (s_b^{(t)}, shot, s_c^{(l)}, s_d^{(q)}) | X_{m,n} = (s_b^{(t)}, choice, s_c^{(l)}, s_d^{(q)})) \\ = P(X_{m,n+1}^{(e)} = shot | X_{m,n} = (s_b^{(t)}, choice, s_c^{(l)}, s_d^{(q)})) \\ = 1 - P(X_{m,n+1}^{(e)} = pass | X_{m,n} = (s_b^{(t)}, choice, s_c^{(l)}, s_d^{(q)})) . \end{aligned} \quad \text{(B.1.4)}$$

In all, estimating $P(X_{m,n+1}^{(e)} = pass | X_{m,n} = (s_b^{(t)}, choice, s_c^{(l)}, s_d^{(q)}))$ allows us to fill in all the non-zero elements in the first 4 rows of the transition matrix in figure 7.3.

We propose a model for this probability in section 7.3.3 and call it the *choice model*.

B.2 Pass node

Let $L_{m,n}$ be the location "targeted" by the pass at step n defined as:

$$L_{m,n} = \begin{cases} X_{m,n+1}^{(l)} & \text{if } X_{m,n+1}^{(t)} = X_{m,n}^{(t)} \\ 100 - X_{m,n+1}^{(l)} & \text{if } X_{m,n+1}^{(t)} \neq X_{m,n}^{(t)} . \end{cases} \quad \text{(B.2.1)}$$

States following passes are of the type *choice* (see figure 7.3) so the transition probability for passes can be written as:

$$\begin{aligned} & P(X_{m,n+1} | X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) = \\ & \begin{cases} P(X_{m,n+1} = (s_b^{(t)}, choice, s_i^{(l)}, s_j^{(q)}) | X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \text{ if} \\ \qquad X_{m,n+1}^{(t)} = X_{m,n}^{(t)} \wedge X_{m,n+1}^{(e)} = choice \wedge s_j^{(q)} \text{ plays for } s_b^{(t)} \\ P(X_{m,n+1} = ((s_b^{(t)})', choice, 100 - s_i^{(l)}, s_k^{(q)}) | X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \text{ if} \\ \qquad X_{m,n+1}^{(t)} \neq X_{m,n}^{(t)} \wedge X_{m,n+1}^{(e)} = choice \wedge s_k^{(q)} \text{ plays for } (s_b^{(t)})' \\ 0 \text{ otherwise.} \end{cases} \end{aligned} \quad \text{(B.2.2)}$$

where

$$(s_i^{(t)})' = \begin{cases} a & \text{if } s_i^{(t)} = h \\ h & \text{if } s_i^{(t)} = a \end{cases} \tag{B.2.3}$$

The first case can be factored as

$$\begin{aligned} P(X_{m,n+1} = (s_b^{(t)}, choice, s_i^{(l)}, s_j^{(q)}) | X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \\ = P(L_{m,n} = s_i^{(l)} | X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \\ \times P(X_{m,n+1}^{(t)} = X_{m,n}^{(t)} | L_{m,n} = s_i^{(l)}, X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \\ \times P(X_{m,n+1}^{(q)} = s_j^{(q)} | L_{m,n} = s_i^{(l)}, X_{m,n+1}^{(t)} = X_{m,n}^{(t)}, X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) . \end{aligned} \tag{B.2.4}$$

The first factor is the probability that location $s_i^{(l)}$ is targeted given the state at step n. The second factor is the probability that the team maintains possession given the state at step n and the fact that its player targeted location $s_i^{(l)}$. The third factor is the probability that the player $s_j^{(q)}$ will be in possession at step $n+1$ given the circumstances and it can also be written as:

$$\begin{aligned} P(X_{m,n+1}^{(q)} = s_j^{(q)} | L_{m,n} = s_i^{(l)}, X_{m,n+1}^{(t)} = X_{m,n}^{(t)}, X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \\ = P(X_{m,n+1}^{(q)} = s_j^{(q)} | X_{m,n+1}^{(t)} = s_b^{(t)}, X_{m,n+1}^{(l)} = s_i^{(l)}, X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) . \end{aligned} \tag{B.2.5}$$

Estimating these three factors allows us to fill in all the non-zero elements in rows 5-8 and columns 1-4 and 11 of the transition matrix in figure 7.3.

Similarly, the second case can be factored as

$$\begin{aligned} P(X_{m,n+1} = ((s_b^{(t)})', choice, 100 - s_i^{(l)}, s_k^{(q)}) | X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \\ = P(L_{m,n} = s_i^{(l)} | X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \\ \times P(X_{m,n+1}^{(t)} \neq X_{m,n}^{(t)} | L_{m,n} = s_i^{(l)}, X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \\ \times P(X_{m,n+1}^{(q)} = s_k^{(q)} | L_{m,n} = s_i^{(l)}, X_{m,n+1}^{(t)} \neq X_{m,n}^{(t)}, X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \end{aligned} \tag{B.2.6}$$

where the last factor can be rewritten as

$$\begin{aligned} P(X_{m,n+1}^{(q)} = s_k^{(q)} | L_{m,n} = s_i^{(l)}, X_{m,n+1}^{(t)} \neq X_{m,n}^{(t)}, X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) \\ = P(X_{m,n+1}^{(q)} = s_k^{(q)} | X_{m,n+1}^{(t)} = (s_b^{(t)})', X_{m,n+1}^{(l)} = 100 - s_i^{(l)}, X_{m,n} = (s_b^{(t)}, pass, s_c^{(l)}, s_d^{(q)})) . \end{aligned} \tag{B.2.7}$$

Estimating these three factors allows us to fill in all the non-zero elements in rows 5-8 and column 15 of the transition matrix in figure 7.3.

In section 7.3.3 we propose a model for the first two factors in both cases, calling them *pass direction* and *pass completion* models respectively, as well as a simplified method for estimating the third factor.

B.3 Shot node

Finally, a *shot* for a team can lead to a *goal* for it or a *choice* state:

$$P(X_{m,n+1}|X_{m,n}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)}))=$$
$$\begin{cases} P(X_{m,n+1}=(s_b^{(t)},goal,\varnothing,\varnothing)|X_{m,n}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)}))\text{if} \\ \qquad\qquad X_{m,n+1}^{(t)}=X_{m,n}^{(t)}\wedge X_{m,n+1}^{(e)}=goal \\ P(X_{m,n+1}=(s_i^{(t)},choice,s_j^{(l)},s_k^{(q)})|X_{m,n}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)}))\text{if} \\ \qquad\qquad X_{m,n+1}^{(e)}=choice\wedge s_k^{(q)}\text{ plays for }s_i^{(t)} \\ 0\text{ otherwise.} \end{cases} \tag{B.3.1}$$

The goal case can be simplified to:

$$\begin{aligned} P(X_{m,n+1}=(s_b^{(t)},goal,\varnothing,\varnothing)|X_{m,n}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)})) \\ =&P(X_{m,n+1}^{(e)}=goal|X_{m,n}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)})) \\ &\times P(X_{m,n+1}^{(t)}=s_b^{(t)}|X_{m,n+1}^{(e)}=goal,X_{m,n}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)})) \\ =&P(X_{m,n+1}^{(e)}=goal|X_{m,n}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)}))\,. \end{aligned} \tag{B.3.2}$$

Estimating $P(X_{m,n+1}^{(e)}=goal|X_{m,n}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)}))$ allows us to fill in all the non-zero elements in rows 9-10 and column 14 of the transition matrix in figure 7.3. A model we propose for this factor in section 7.3.3 is called *shot conversion* model.

The second case can be factored as:

$$\begin{aligned} &P(X_{m,n+1}=(s_i^{(t)},choice,s_j^{(l)},s_k^{(q)})|X_{m,n}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)})) \\ =&P(X_{m,n+1}^{(q)}=s_k^{(q)}|X_{m,n+1}^{(t)}=s_i^{(t)},X_{m,n+1}^{(e)}=choice,X_{m,n+1}^{(l)}=s_j^{(l)},X_{m,n}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)})) \\ &\times P(X_{m,n+1}^{(t)}=s_i^{(t)},X_{m,n+1}^{(e)}=choice,X_{m,n+1}^{(l)}=s_j^{(l)}|X_{m,n}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)})) \\ =&P(X_{m,n+1}^{(q)}=s_k^{(q)}|X_{m,n+1}^{(t)}=s_i^{(t)},X_{m,n+1}^{(e)}=choice,X_{m,n+1}^{(l)}=s_j^{(l)},X_{m,n}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)})) \\ &\times P(X_{m,n+1}^{(e)}=choice|X_{m,n}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)})) \\ &\times P(X_{m,n+1}^{(t)}=s_i^{(t)},X_{m,n+1}^{(l)}=s_j^{(l)}|X_{m,n+1}^{(e)}=choice,X_{m,n}=(s_b^{(t)},shot,s_c^{(l)},s_d^{(q)}))\,. \end{aligned} \tag{B.3.3}$$

The last factor is the probability that the team $s_i^{(t)}$ will be in possession in the location $s_j^{(l)}$ following a shot which does not lead to a goal and the first factor is the probability that its player $s_k^{(q)}$ will be in possession given these circumstances. The second factor is:

$$\begin{aligned} P(X^{(e)}_{m,n+1} = choice | X_{m,n} = (s_b^{(t)}, shot, s_c^{(l)}, s_d^{(q)})) \\ = 1 - P(X^{(e)}_{m,n+1} = goal | X_{m,n} = (s_b^{(t)}, shot, s_c^{(l)}, s_d^{(q)})) \end{aligned} \quad \text{(B.3.4)}$$

Estimating these three factors allows us to fill in all the non-zero elements in rows 9-10 and columns 1-4, 11 and 15 of the transition matrix in figure 7.3. In section 7.3.3 we propose a *shot conversion* model to evaluate the second factor as well as a simplified method for estimating the first factor. The third factor will be estimated based on sample frequencies.

Appendix C

Sensitivity analysis

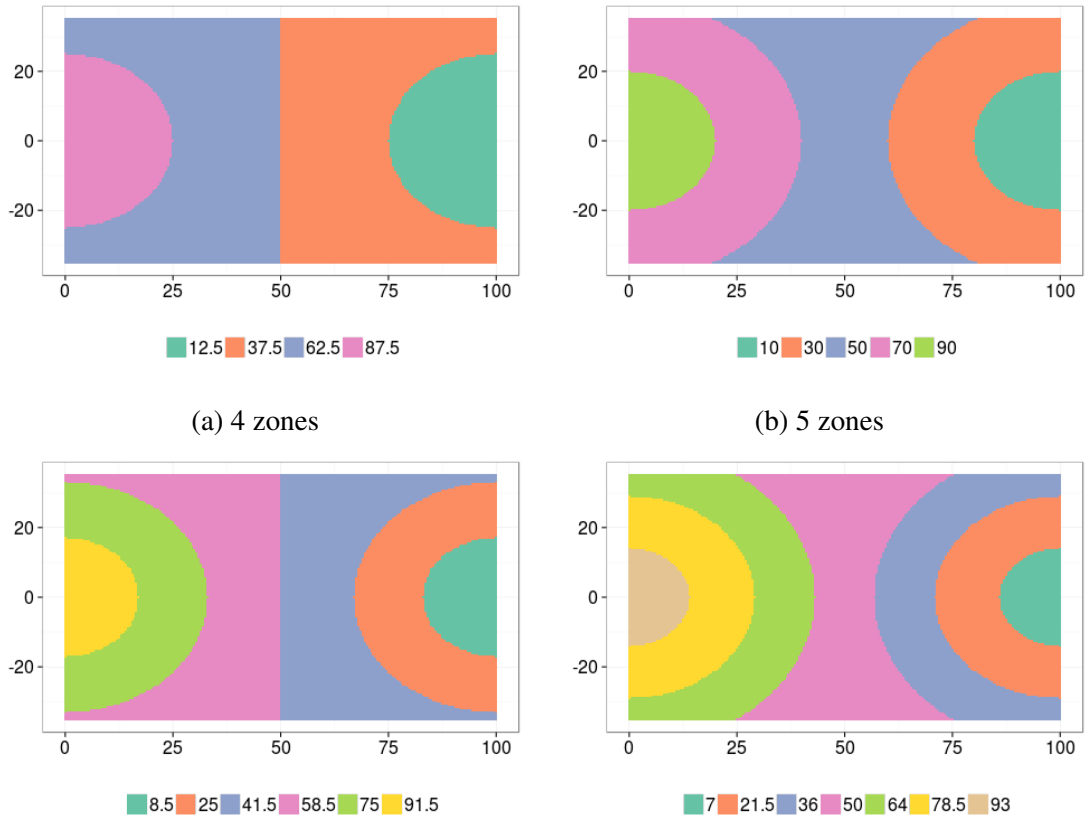
In chapter 7 we made several arbitrary assumptions in the construction of the model of a football game. One of them is the division of the pitch into 4 zones proposed in figure 7.1. It may be interesting to verify whether comparable results can be obtained for alternative pitch division schemes. In this appendix the analysis is repeated for a pitch split into 5, 6 and 7 zones as presented in figure C.1 and the results are compared with the default scheme.

Figure C.2 presents estimates of the one step transition matrix of the non player specific Markov chain model from section 7.3.2.

The patterns observed previously in the model with 4 zones are retained by more complex models. For example:

- passes from a given zone tend to be followed most frequently by passes from adjacent zones;
- shots tend to follow passes from zones closest to the opponents goal;
- goals are more likely to be scored from shots closer to the opponent's goal.

One difference can be noticed between models with an odd number of zones compared to the ones with an even number of locations. In the former goals are followed almost exclusively by passes by the opposite team from the central zone. In the latter the lack of a single central zone leads to the game restart being assigned to either of the two most central zones. This can be considered a slight flaw of the models with an even number of zones. Even in the case of the models with an odd number of zones it could make sense to extend the model to have a designated game restart state as such passes can be argued to be different than regular passes from the central zone.



(a) 4 zones

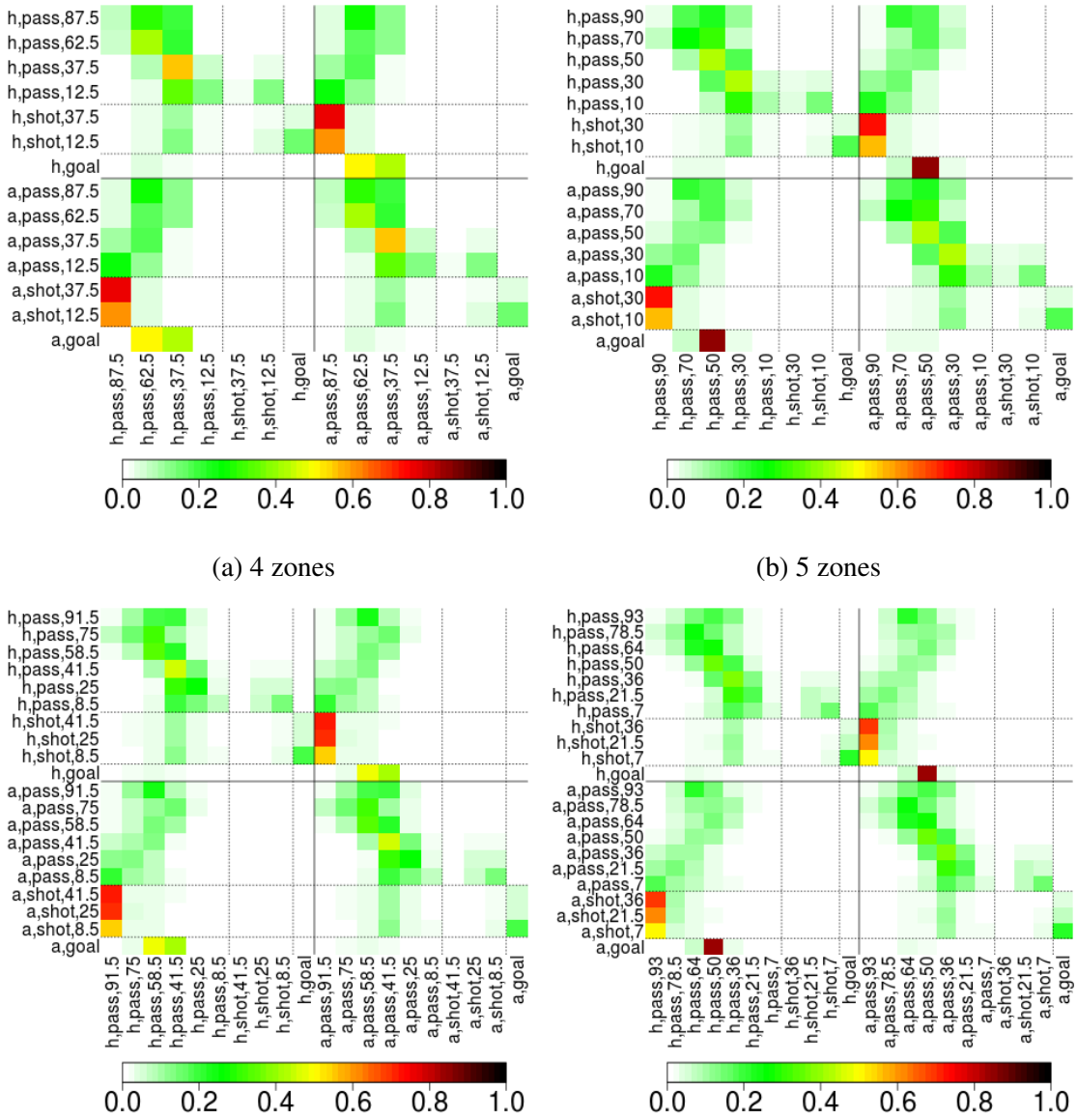
(b) 5 zones

(c) 6 zones

(d) 7 zones

Figure C.1: Pitch division into 4, 5, 6 and 7 zones.

Figures C.3 and C.4 present results of fitting pass direction and pass completion models respectively. The model predictions follow the same patterns seen in figures 7.8 and 7.9 with more complex models yielding generally smoother shapes.



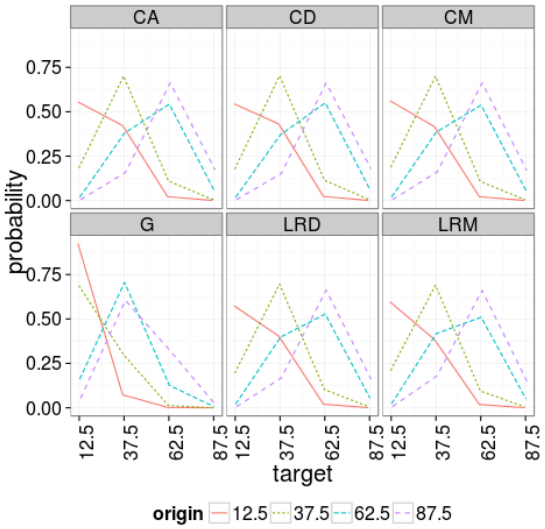
(a) 4 zones

(b) 5 zones

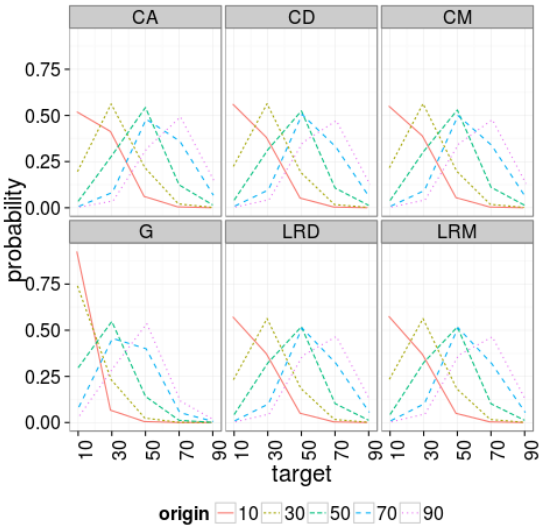
(c) 6 zones

(d) 7 zones

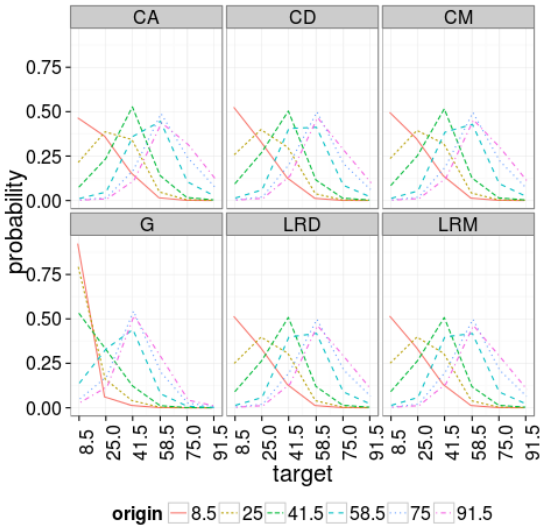
Figure C.2: Estimate of the one step transition matrix in the basic Markov chain model (different pitch divisions).



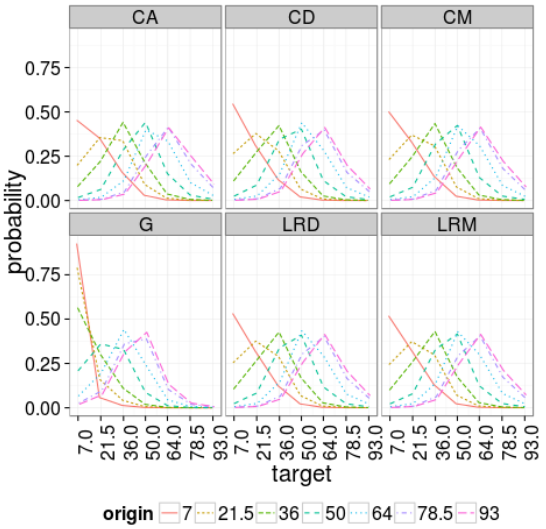
(a) 4 zones



(b) 5 zones

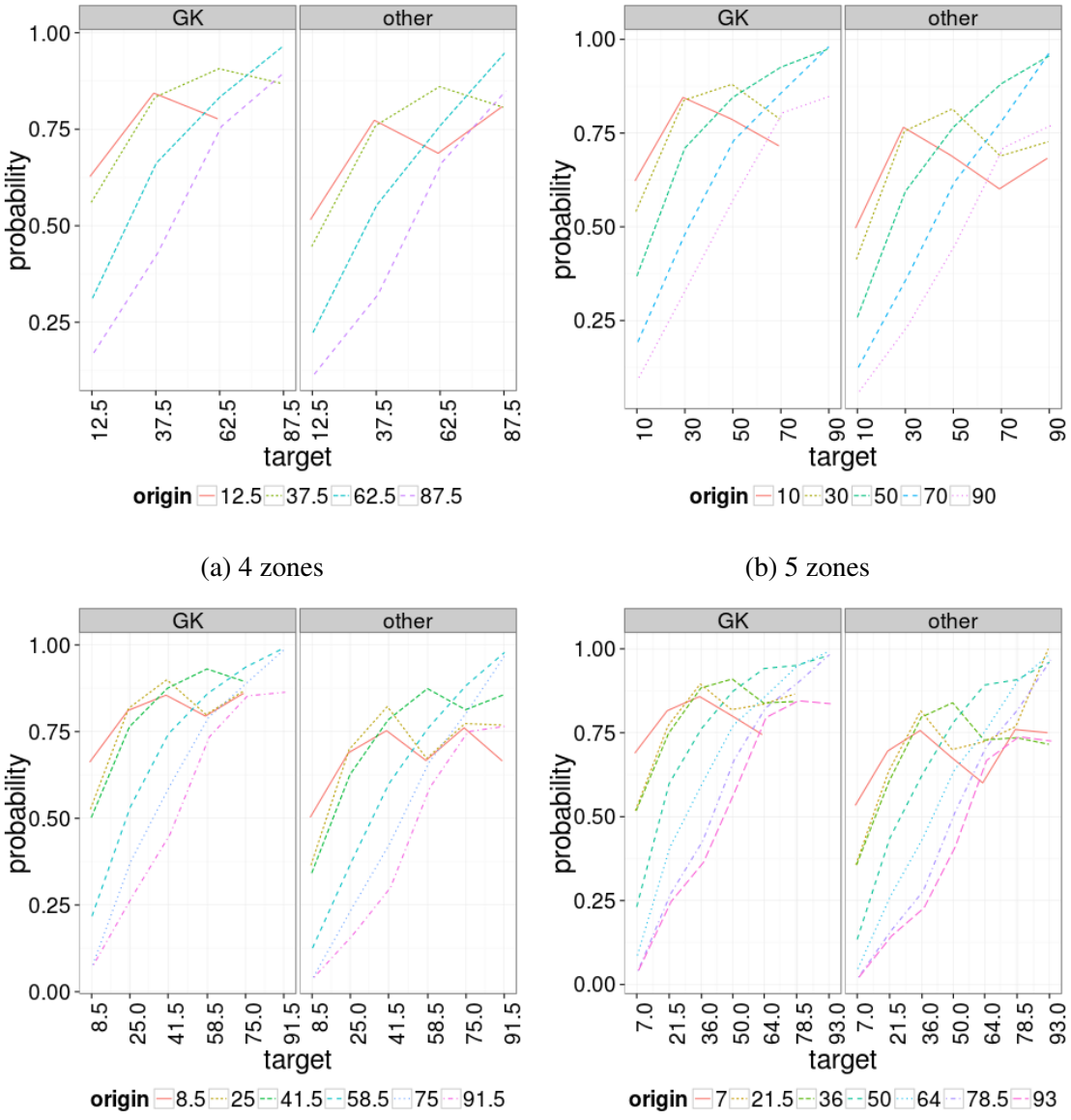


(c) 6 zones



(d) 7 zones

Figure C.3: Predicted probability of a pass being directed to a given zone depending on its origin and the nominal position of the executing player.



(a) 4 zones

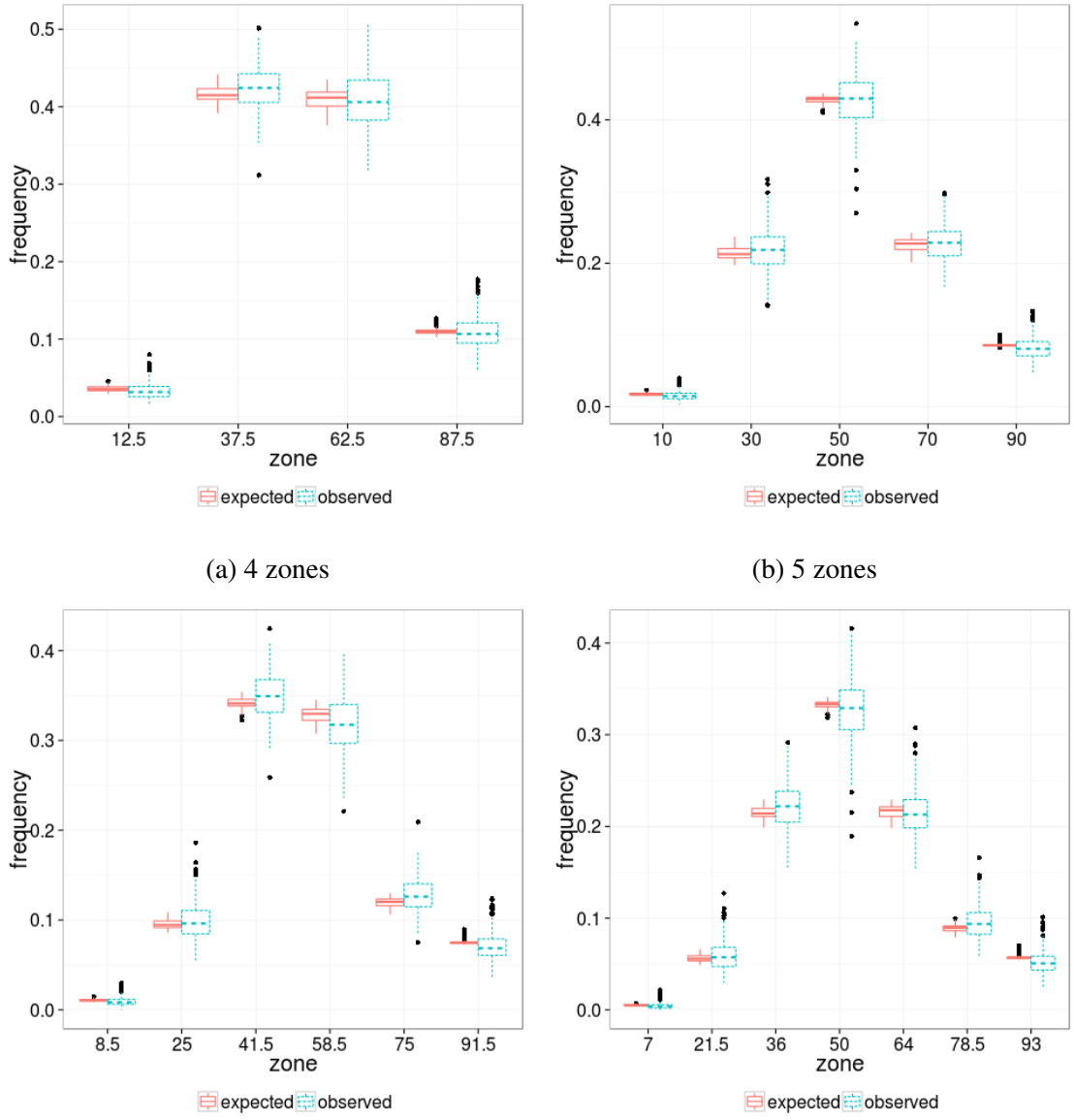
(b) 5 zones

(c) 6 zones

(d) 7 zones

Figure C.4: Predicted pass completion rate of an average player (a goalkeeper or not) depending on the zone of origin and the targeted zone.

Figures C.5 and C.6 compare the observed and the model implied distribution of the location attribute of pass and shot events respectively. All the models fit the data quite well.



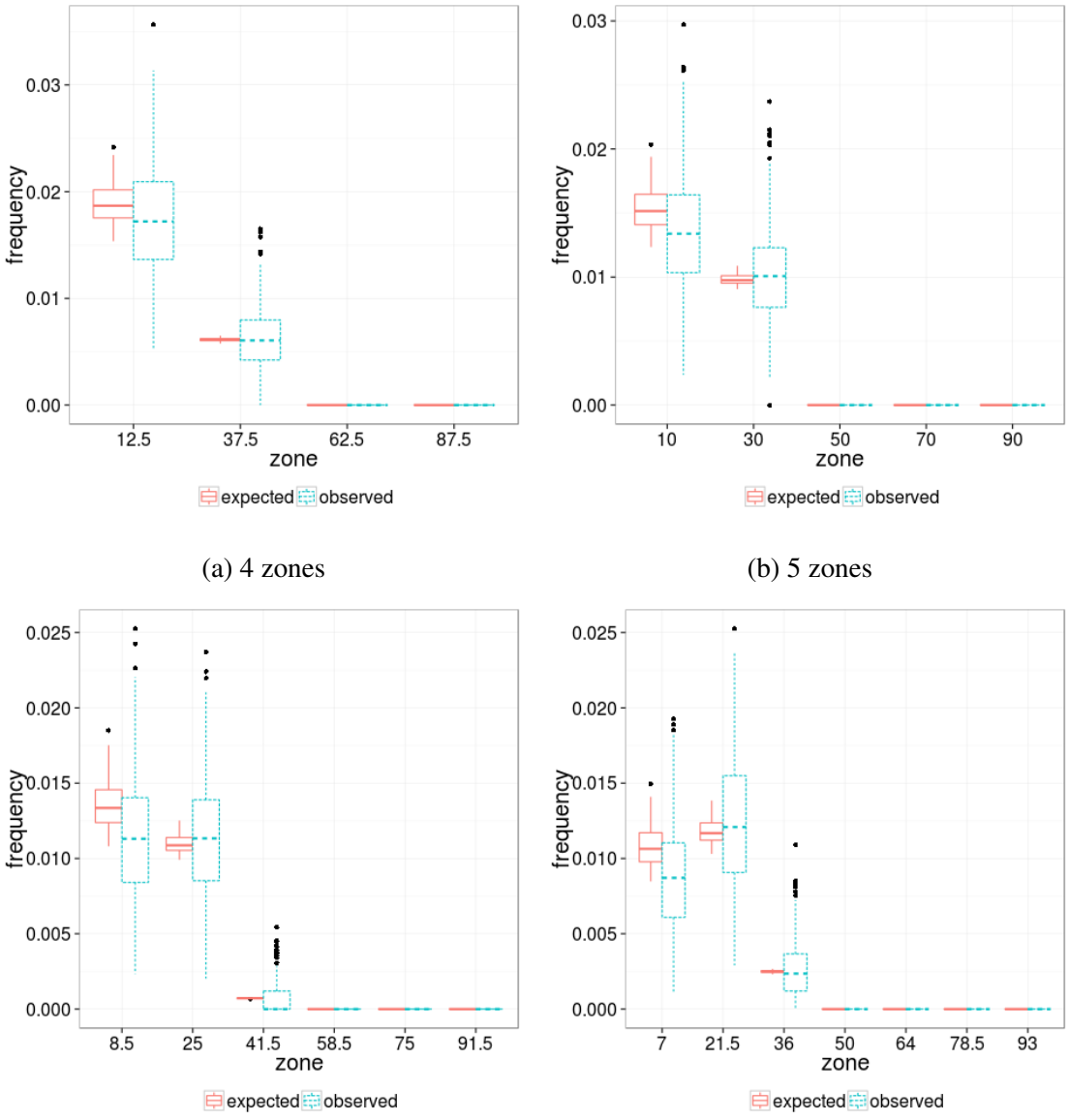
(a) 4 zones

(b) 5 zones

(c) 6 zones

(d) 7 zones

Figure C.5: Expected (solid red) and observed (dashed blue) frequency of passes per game in a given zone.



(a) 4 zones

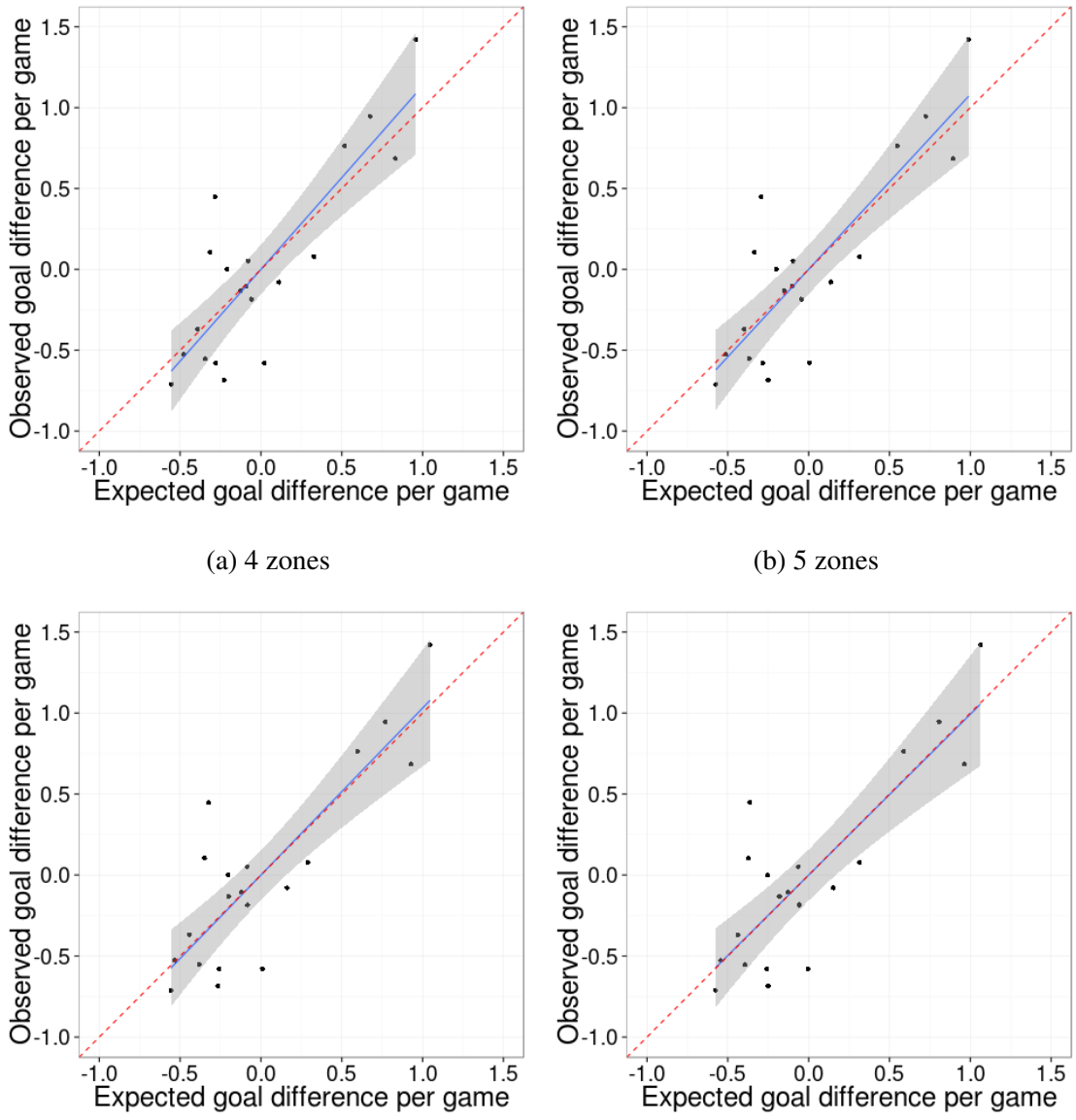
(b) 5 zones

(c) 6 zones

(d) 7 zones

Figure C.6: Expected (solid red) and observed (dashed blue) frequency of shots per game in a given zone.

Figure C.7 presents the relationship between the average expected and the observed goal supremacy per team per game. All the models perform similarly.



(a) 4 zones

(b) 5 zones

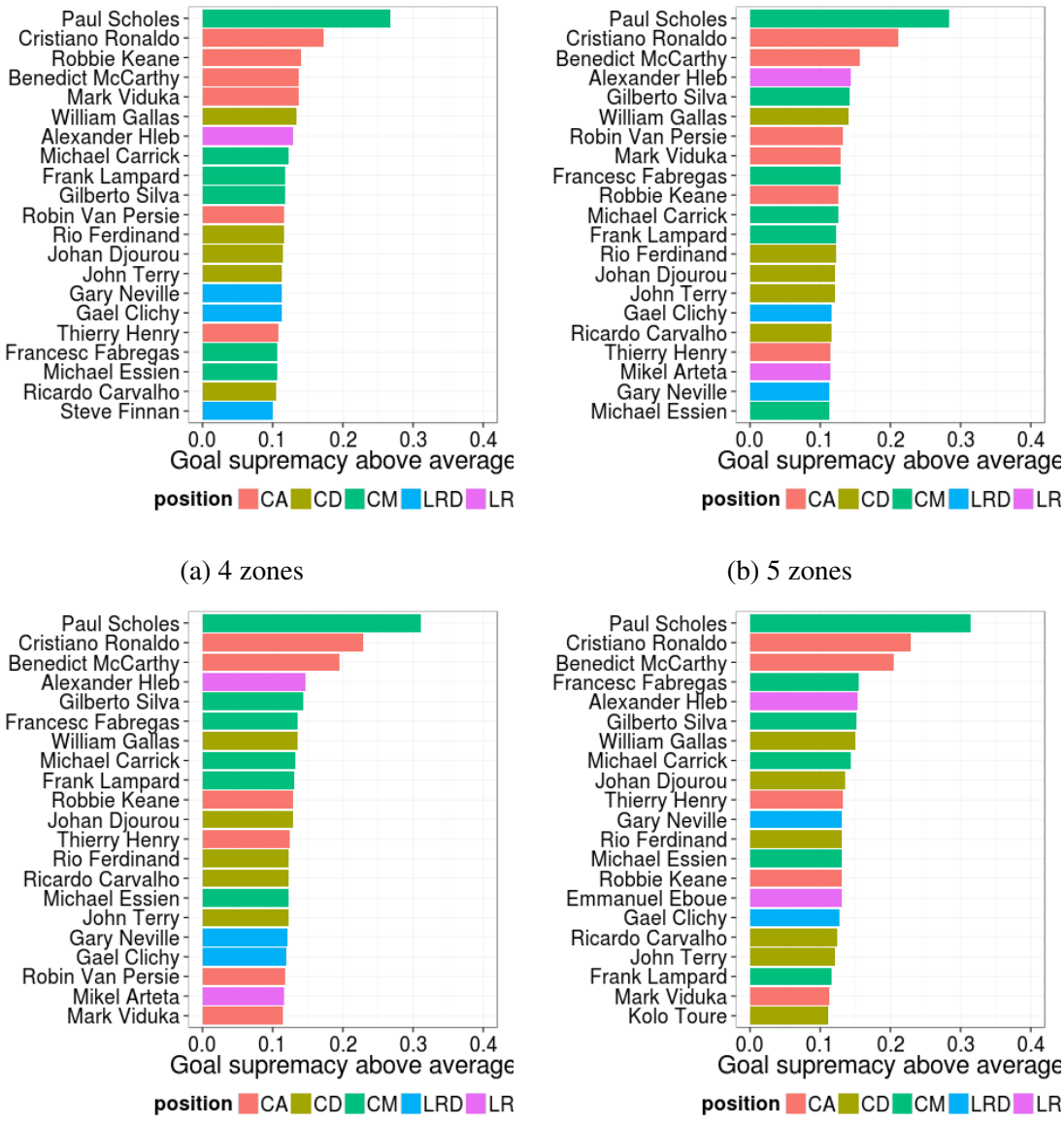
(c) 6 zones

(d) 7 zones

Figure C.7: Expected and observed goal supremacy per game by team. The dashed line is the identity function and the solid one is the linear model fit.

Figure C.8 presents the lists of the top 20 players according to the expected goals supremacy above average statistic. There is a good consistency between these lists with similar names appearing repeatedly on all of them.

Finally, the level of the expected goal supremacy contribution among a team's players has a similar predictive utility of its future performance irrespective of the number



(a) 4 zones

(b) 5 zones

(c) 6 zones

(d) 7 zones

Figure C.8: Top 20 players by the expected goal supremacy above an average player in their position in season 2006/07.

of the location zones in the model. The Pearson correlation coefficient for:

1. the difference in the average value of expected goal supremacy contribution among players of a given team compared to their opposition, and
2. the goal supremacy in a game between these two teams,

(see sections 6.4.4 and 7.4.3 for details) is:

- 0.383 with 90% confidence interval of $(0.308, 0.454)$ for 4 zones;

- 0.388 with 90% confidence interval of $(0.312, 0.459)$ for 5 zones;
- 0.381 with 90% confidence interval of $(0.305, 0.452)$ for 6 zones;
- 0.377 with 90% confidence interval of $(0.301, 0.449)$ for 7 zones.

Appendix D

Code

This appendix presents some of the code used to implement the ideas presented in this thesis. The `R` code to fit the shot count and shot conversion models of section 5.3.2 and 5.3.3 using the `lme4` package (Bates and Maechler, 2010) is presented in listings D.1 and D.2 respectively. Listing D.3 presents the algorithm to calculate goals predictions given shot predictions (obtained from the model in listing D.1) and a conversion rate model fit (from listing D.2).

```
if(modelName == "extended"){
  m.shot <- glmer(n ~ # number of shots
       host_ind + # home team player indicator
       (1|player_id) + # player random effect
       factor(opp_team_id) + # opponent fixed effect
       log(time_played/100) + # by the given player
       factor(player_position), # in the tactical formation
       family="poisson", data=dataset.fit)
} else if(modelName == "basic"){
  m.shot <- glmer(n ~
       host_ind +
       (1|player_id) +
       factor(opp_team_id)
       family="poisson", data=dataset.fit)
}
```

Listing D.1: Code to fit the shot count model.

```
if(modelName == "extended"){
  m.conv <- glmer(goals / shots ~
       host_ind +
       (1|player_id) +
       log(shots),
       family="binomial",
       weights=shots,
       data=dataset.fit)
} else if(modelName == "basic"){
  m.conv <- glmer(goals / shots ~
       host_ind +
       (1|player_id),
       family="binomial",
       weights=shots,
       data=dataset.fit)
}
```

Listing D.2: Code to fit the shot conversion model.

```
library(lme4) # for fixef function
library(boot) # for inv.logit function

# Function to predict conversion rate
# dataset contains:
# - home team indicator: host_ind;
# - player conversion rate random effects: conv_re
convPredict <- function(dataset, shots, model.fit, model.type){
  if(model.type == "basic"){
    predictions  <- with(dataset, {
      inv.logit(fixef(model.fit)[1] +
                  fixef(model.fit)[2]*host_ind +
                  conv_re)
    })
  } else if(model.type == "extended"){
    predictions  <- with(dataset, {
      inv.logit(fixef(model.fit)[1] +
                  fixef(model.fit)[2]*host_ind +
                  fixef(model.fit)[3]*log(shots) +
                  conv_re)
    })
  }
  return(predictions)
}
```

```
# Calculate goals predicitions
# convmodel.fit is the lme4 model fit for conversion rates
# dataset.fit$shots.pred is the shooting rate prediction
goals.pred <- rep(0, nrow(dataset.fit))
for(nShots in 1:20){
  goals.pred <- goals.pred +
    dpois(nShots, dataset.fit$shots.pred) *
    nShots *
    convPredict(dataset.fit, nShots, convmodel.fit, "extended")
}
```

Listing D.3: Code to calculate goal predictions.

The R code to fit the pass completion model of section 6.3 using the mgcv package (Wood, 2006) is presented in listing D.4. Listing D.5 presents the code to make various pass completion rate predictions defined in section 6.3.4.

```
m.pass <- bam(pass_successful ~
      host_ind + # home team player indicator
      pass_head + # headed pass
      pass_head_t_1 + # previous pass was headed
      after_duel_won_aerial + # following an aerial duel
      after_duel_won_tackle + # following a tackle
      after_duel_won_same_pl + # previous duel involved the passer
      s(player_id, bs="re", by=dumPl) + # player random effect
      # dumPl is a dummy variable set to 1 for all observations
      # for fitting but can be set to 0 in prediction to "switch off"
      # the player effect, i.e. assume an average player is passing
      s(team_id, bs="re", by=dumT) + # team random effect
      # dumT is a team dummy variable; see dumPl above
      s(opp_team_id, bs="re", by=dumO) + # opponent random effect
      # dumO is an opponent dummy variable; see dumPl above
      s(minute, k=6) + # time of the game
      s(pmin(15, possession_event), k=9) +
      # number of pass in possession
      s(timedelta1, k=8) + # time since the previous pass
      s(timedelta2, k=4) +
      # time between the previous pass and the one before
      te(posx, posy, k=4) +
      # anticipated average coordinates of the passer
      te(x, endx, abs(y - 0.5), abs(endy - 0.5), k=5) +
      te(x, endx, I((y - 0.5) * (end_y - 0.5)), k=5),
      # coordinates of the pass origin and target
```

```
      cluster=cl, # cluster of processors to use
      family="binomial", method="fREML",
      data=dataset.fit)
```

Listing D.4: Code to fit the passing model.

```
library(mgcv) # for predict function for gam object
library(plyr) # for ddply function
library(boot) # for inv.logit function

# Get player random effects
coefs <- coef(m.pass)
player_ids <- levels(dataset.fit$player_id)
player_ids.fit <- player_ids[player_ids %in% dataset.fit$player_id]
re <- data.frame(id = player_ids.fit,
                 re = coefs[grep("player_id", names(coefs))])

# Merge random effects to a data.frame with player names and ids
player.map <- merge(player.map, re)

# Completion rate predictions
dataset.fit <- within(dataset.fit, {
    dumPl <- 1
    dumT <- 1
    dumO <- 1
  })
dataset.fit <- within(dataset.fit, {
  pred <- predict(m.pass, newdata=dataset.fit, type="response")
}

# Pass easiness calculations
dataset.fit <- within(dataset.fit, {
  dumPl <- 0 # make predictions for an average player
  dumT <- 1
  dumO <- 1
})
dataset.fit <- within(dataset.fit, {
  lpred0 <- predict(m.pass, newdata=dataset.fit, type="link")
})
```

```
# Average predictions by player
completion.player <-
  ddply(dataset.fit, .variables="player_id",
        .fun=function(df)
          with(df, data.frame(easiness=mean(inv.logit(lpred0)),
                              pred.full=mean(pred)))
        )

# Global average linear predictor for an average player
completion.lpred0.av <- mean(dataset.fit$lpred0)

# Merge to a data.frame containing player names and ids
player.map <- merge(player.map, completion.player,
                    by.x="id", by.y="player_id")
player.map <-
  within(player.map,
         player_in_av_sit <- inv.logit(completion.lpred0.av + re))
```

Listing D.5: Code to calculate various passing predictions.

Figure D.1 presents an overview of the structure of the Markov chain model project of chapter 7. It consists of the following stages:

1. Prepare data:

   (a) Pre-process data. Opta event definitions are applied to the raw events data. Passes and shots data is selected.

   (b) Assign states. The pass and shot events are translated into states of the Markovian model like in tables 7.1 and 7.2. Consecutive states are joined to represent chain transitions.

   (c) Present matrix. Figures 7.4 and C.2 as well as C.1 are produced.

2. Fit models:

   - Action choice. The model from equations (7.3.21)-(7.3.22) is fitted.
   - Pass direction. The model from equations (7.3.13)-(7.3.14) is fitted.
   - Pass completion. The model from equations (7.3.18)-(7.3.20) is fitted.
   - Shot conversion. The model from equations (7.3.10)-(7.3.12) is fitted.
   - Next player. The distribution of a given player being in possession of the ball is worked out based on the logic from equation (7.3.23).

3. Create submatrices. Submatrices of the transition matrix from figure 7.3 are created based on the fitted models and the logic of appendix B.

4. Combine submatrices. The submatrices are combined into the game specific transition matrices from figure 7.3.

5. Postprocessing:

   (a) Unconditional distribution. Uses the transition matrices to calculate the unconditional distributions of equation (7.3.25) according to equation (7.3.26).

   (b) Compare model and empirical distributions. Produces figures 7.12, C.5 and C.6 as well as 7.13 and C.7.

   (c) Compare two model distributions. Compares the unconditional distributions with a given player in the team and substituted by an average player of his position. Produces figures 7.15 and 7.16. Calculates the expected contribution to the team supremacy above an average player and produces figures 7.17 and C.8.

   (d) Evaluate predictive utility. Compares the predictive utility of the above metric with that of simpler indices as described in section 7.4.3.

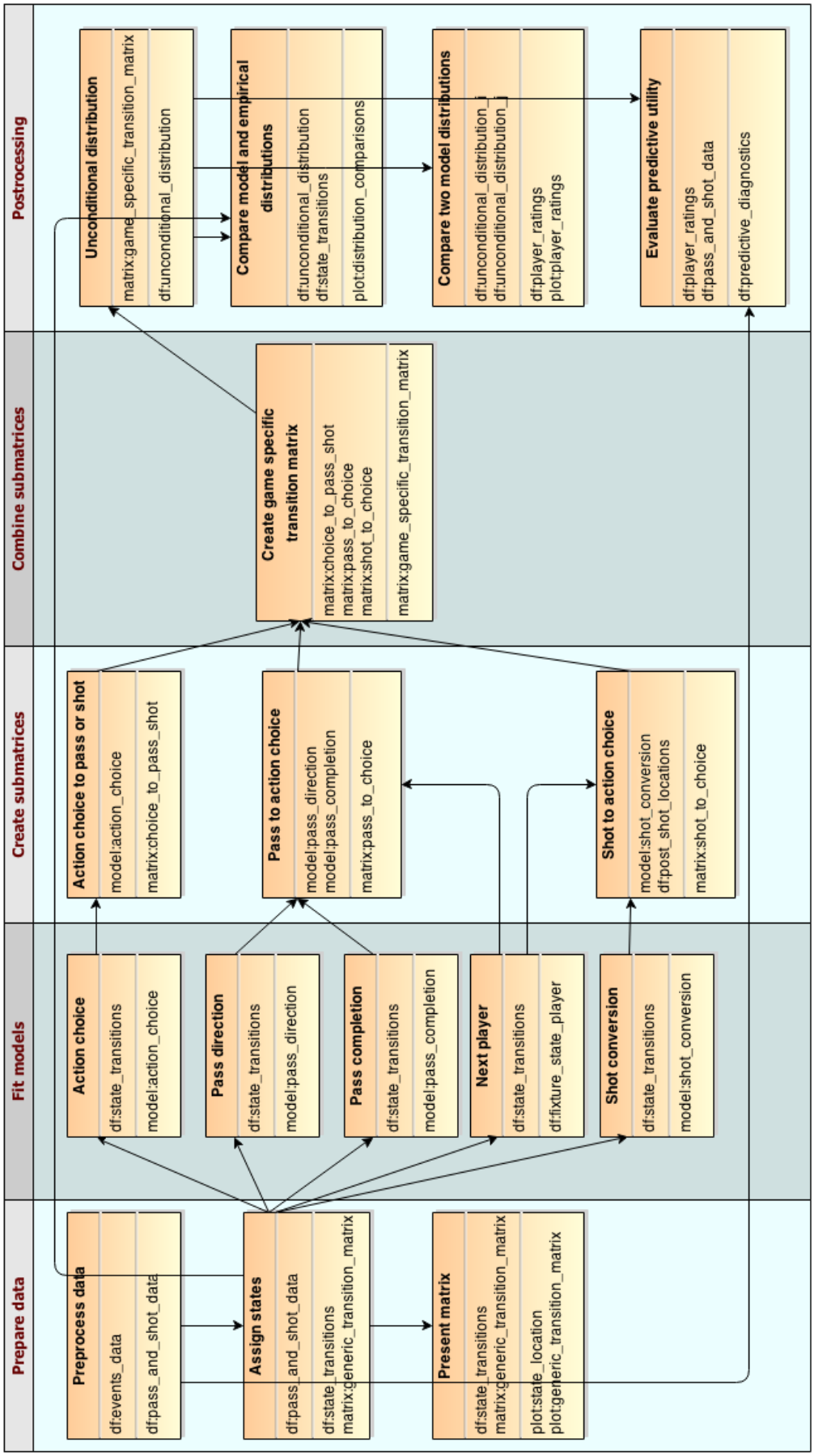


Figure D.1: Outline of the code structure for the Markov chain model of a football game. Each box represents one executable with the name (in bold), inputs and outputs separated by a horizontal line. *df* stands for *data.frame*.

Bibliography

Albert, J. (1992). A Bayesian analysis of a Poisson random effects model for home run hitters. *American Statistician*, 46(4):246–253.

Albert, J. (2006). Pitching Statistics, Talent and Luck, and the Best Strikeout Seasons of All-Time. *Journal of Quantitative Analysis in Sports*, 2(1).

Albert, J. (2007). Hitting in the pinch. In Albert, J. and Koning, R. H., editors, *Statistical thinking in sports*, pages 111–134. Chapman & Hall/CRC.

Baio, G. and Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37(2):253–264.

Baker, R. D. and McHale, I. G. (2014). Deterministic Evolution of Strength in Multiple Comparisons Models: Who is the Greatest Golfer? *Scandinavian Journal of Statistics (to appear)*.

Baseball-Reference.com (2013). WAR Comparison Chart. `http://www.baseball-reference.com/about/war_explained_comparison.shtml`. [Online; accessed 24-July-2013].

Bate, A. (2012). Scouting revolution. `http://www1.skysports.com/football/news/11096/8293811/scouting-revolution`. [Online; accessed 02-August-2013].

Bates, D. and Maechler, M. (2010). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-37.

Bennett, J. M. and Flueck, J. A. (1983). An evaluation of Major League Baseball offensive performance models. *The American Statistician*, 37(1):76–82.

Bennett, J. M. and Flueck, J. A. (1984). Player Game Percentage. In *Proceedings of the Social Statistics Section, American Statistical Association*, pages 378–380.

Berri, D. J. (1999). Who is 'most valuable'? Measuring the player's production of wins in the National Basketball Association. *Managerial and Decision Economics*, 20(8):411–427.

Branch, R. (1954). Some Old Baseball Ideas. *LIFE*, 37(5):79–82.

Bryant, A. (2013). In Head-Hunting, Big Data May Not Be Such a Big Deal. `http://www.nytimes.com/2013/06/20/business/in-head-hunting-big-data-may-not-be-such-a-big-deal.html`. [Online; accessed 02-August-2013].

Bukiet, B., Harold, E. R., Palacios, J. L., and Palacios, J. L. (1997). A Markov Chain Approach to Baseball. *Operations Research*, 45(1):14–23.

Cameron, D. (2008). Win Values Explained. `http://www.fangraphs.com/blogs/win-values-explained-part-one/`. [Online; accessed 24-July-2013].

Cameron, D. (2009). Pitcher Win Values Explained. `http://www.fangraphs.com/blogs/pitcher-win-values-explained-part-one/`. [Online; accessed 24-July-2013].

Carling, C., Williams, M., and Reilly, T. (2006). *Handbook of Soccer Match Analysis: A Systematic Approach to Improving Performance*. Routledge, New York.

Chang, J. (2012). Everton at forefront of performance analytics in Premier League. `http://sportsillustrated.cnn.com/2012/writers/jen_chang/04/29/performance.analytics/index.html?utm_source=SRC+Newsletter+List&utm_campaign=06b720f4ba-5_02_2012`. [Online; accessed 02-August-2013].

Connolly, R. A. and Rendleman, R. J. (2008). Skill, Luck, and Streaky Play on the PGA Tour. *Journal of the American Statistical Association*, 103(481):74–88.

Cover, T. M. and Keilers, C. W. (1977). An offensive earned-run average for baseball. *Operations Research*, 25(5):729–740.

Dawson, P., Dobson, S., and Gerrard, B. (2000). Estimating Coaching Efficiency in Professional Team Sports: Evidence from English Association Football. *Scottish Journal of Political Economy*, 47(4):399–421.

Deloitte (2013). Deloitte Annual Review of Football Finance 2013. Technical Report June.

Diez-Roux, A. V., Link, B. G., and Northridge, M. E. (2000). A multilevel analysis of income inequality and cardiovascular disease risk factors. *Social science & medicine*, 50(5):673–687.

Dixon, M. and Coles, S. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280.

Duch, J., Waitzman, J. S., and Amaral, L. A. N. (2010). Quantifying the performance of individual players in a team activity. *PloS one*, 5(6):e10937.

DuPaul, G. (2012). What is WAR good for? `http://www.hardballtimes.com/main/article/what-is-war-good-for/`. [Online; accessed 24-July-2013].

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319.

Ensum, J., Pollard, R., and Taylor, S. (2005). Applications of Logistic Regression to Shots at Goal in Association Football. In Reilly, T., Cabri, J., and Araujo, D., editors, *Science and Football V: The proceedings of the Fifth World Congress on Science and Football*, pages 211–218, New York. Routledge.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 50(2):201–220.

Fearnhead, P. and Taylor, B. M. (2011). On Estimating the Ability of NBA Players. *Journal of Quantitative Analysis in Sports*, 7(3).

Fowler, J. H., Baker, L. A., and Dawes, C. T. (2008). Genetic Variation in Political Participation. *American Political Science Review*, 102(02):233–248.

Frees, E. W. and Valdez, E. A. (2008). Hierarchical Insurance Claims Modeling. *Journal of the American Statistical Association*, 103(484):1457–1469.

Garcia-del Barrio, P. and Pujol, F. (2009). The Rationality of Under-employing the Best-performing Soccer Players. *Labour*, 23(3):397–419.

Gerrard, B. (2001). A new approach to measuring player and team quality in professional team sports. *European Sport Management Quarterly*, 1(3):219–234.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):377–394.

Gramacy, R. B., Jensen, S. T., and Taddy, M. (2013). Estimating player contribution in hockey with regularized logistic regression. *Journal of Quantitative Analysis in Sports*, 9(1):97–111.

Gulbrandsen, A. and Gulbrandsen, C. (2011). *Valuation of Football Players*. Master's thesis, Norges Handelshoyskole.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical science*, 1(3):297–318.

Hirotsu, N. and Wright, M. (2002). Using a Markov process model of an association football match to determine the optimal timing of substitution and tactical decisions. *Journal of the Operational Research Society*, 53(1):88–96.

Hirotsu, N. and Wright, M. (2003). An evaluation of characteristics of teams in association football by using a Markov process model. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(4):591–602.

Ilardi, S. and Barziali, A. (2008). Adjusted Plus-Minus Ratings: New and Improved for 2007-2008. `http://www.82games.com/ilardi2.htm`. [Online; accessed 26-July-2013].

Jarvandi, A., Sarkani, S., and Mazzuchi, T. (2013). Modeling team compatibility factors using a semi-Markov decision process: a data-driven approach to player selection in soccer. *Journal of Quantitative Analysis in Sports*, 9(4).

Jensen, S. T., Shirley, K. E., and Wyner, A. J. (2009). Bayesball: A Bayesian hierarchical model for evaluating fielding in major league baseball. *The Annals of Applied Statistics*, 3(2):491–520.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer Verlag.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin.

Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393.

Koopman, S. and Lit, R. (2014). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society: Series A (Statistics in Society) (to appear)*.

Kubatko, J., Oliver, D., Pelton, K., and Rosenbaum, D. T. (2007). A Starting Point for Analyzing Basketball Statistics. *Journal of Quantitative Analysis in Sports*, 3(3).

Kuper, S. (2011). A football revolution. `http://www.ft.com/cms/s/2/9471db52-97bb-11e0-9c37-00144feab49a.html`. [Online; accessed 02-August-2013].

Kuper, S. (2013). Everton: how the blues made good. `http://www.ft.com/cms/s/2/2fa7ef1e-b2c0-11e2-8540-00144feabdc0.html`. [Online; accessed 02-August-2013].

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis Via H-likelihood*. Chapman and Hall/CRC; Har/Cdr edition.

Lewis, A. (2004a). Towards fairer measures of player performance in one-day cricket. *Journal of the Operational Research Society*, 56(7):804–815.

Lewis, M. (2004b). *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton & Company.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B*, 61(2):381–400.

Lindsey, G. (1963). An investigation of strategies in baseball. *Operations Research*, 11(4):477–501.

Liu, Q. and Pierce, D. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, 81(3):624–629.

Loughin, T. M. and Bargen, J. L. (2008). Assessing pitcher and catcher influences on base stealing in Major League Baseball. *Journal of sports sciences*, 26(1):15–20.

Macdonald, B. (2011). A Regression-Based Adjusted Plus-Minus Statistic for NHL Players. *Journal of Quantitative Analysis in Sports*, pages 1–39.

Macdonald, B. (2012). Adjusted Plus-Minus for NHL Players using Ridge Regression with Goals, Shots, Fenwick, and Corsi. *Journal of Quantitative Analysis in Sports*, 8(3).

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118.

McCracken, V. (2001). Pitching and Defense. `http://baseballprospectus.com/article.php?articleid=878`. [Online; accessed 23-July-2013].

McHale, I. and Morton, A. (2011). A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting*, 27(2):619–630.

McHale, I. and Scarf, P. (2007). Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica*, 61(4):432–445.

McHale, I. G. and Scarf, P. (2005). Ranking football players. *Significance*, 2(2):54–57.

McHale, I. G., Scarf, P., and Folker, D. (2012). On the development of a soccer player performance rating system for the English Premier League. *Interfaces*, 42(4):339–351.

McHale, I. G. and Szczepański, Ł. (2014). A mixed effects model for identifying goal scoring ability of footballers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177(2):397–417.

Miller, B. (2011). Moneyball.

Moore, A. and Gould, R. (2005). Longitudinal patterns and predictors of alcohol consumption in the United States. *American Journal of Public Health*, 95(3):458–464.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370.

Null, B. (2009). Modeling Baseball Player Ability with a Nested Dirichlet Distribution. *Journal of Quantitative Analysis in Sports*, 5(2).

Oberstone, J. (2011). Evaluating English Premier League Player Performance Using the MAP Model. In Percy, D., Reade, J., and Scarf, P., editors, *3rd International Conference on Mathematics in Sport: Proceedings Papers*, pages 153–159. Institute of Mathematics And Its Applications.

Owen, A. (2011). Dynamic bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics*, 22:99–113.

Patton, G., Coffey, C., and Carlin, J. B. (2002). Cannabis use and mental health in young people: cohort study. *Bmj*, 325(November):1195–1198.

Pena, J. (2014). A Markovian model for association football possession and its outcomes. *arXiv preprint arXiv:1403.7993*.

Pinheiro, J. and Bates, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):12–35.

Pinquet, J. (1997). Allowance for cost of claims in bonus-malus systems. *Insurance: Mathematics and Economics*, 27(I):33–57.

Pollard, R., Ensum, J., and Taylor, S. (2004). Estimating the probability of a shot resulting in a goal: The effects of distance, angle and space. *International Journal of Soccer and Science*, 2(1):50–55.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raudenbush, S. W., Yang, M.-L., and Yosef, M. (2000). Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation. *Journal of Computational and Graphical Statistics*, 9(1):141–157.

Réale, D., McAdam, A. G., Boutin, S., and Berteaux, D. (2003). Genetic and plastic responses of a northern mammal to climate change. *Proceedings. Biological sciences / The Royal Society*, 270(1515):591–596.

Reep, C. and Benjamin, B. (1968). Skill and chance in association football. *Journal of the Royal Statistical Society. Series A (General)*, 131(4):581–585.

Reep, C., Pollard, R., and Benjamin, B. (1971). Skill and chance in ball games. *Journal of the Royal Statistical Society. Series A (General)*, 134(4):623–629.

Rigotti, N. and DiFranza, J. (1997). The effect of enforcing tobacco-sales laws on adolescents' access to tobacco and smoking behavior. *The New England Journal of Medicine*, 337(15):1044–1051.

Rosenbaum, D. T. (2004). Picking the Difference Makers for the All-NBA Teams. `http://www.82games.com/comm30.htm`. [Online; accessed 26-July-2013].

Rue, H. v., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Scarf, P., Akhtar, S., and Rasool, Z. (2011). Rating players in test match cricket. In Percy, D., Reade, J., and Scarf, P., editors, *3rd International Conference on Mathematics in Sport: Proceedings Papers*, pages 209–214. Institute of Mathematics And Its Applications.

Schuckers, M. and Curro, J. (2013). Total Hockey Rating (THoR): A comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events. `http://statsportsconsulting.com/main/wp-content/uploads/Schuckers_Curro_MIT_Sloan_THoR.pdf`. [Online; accessed 02-August-2013].

Schuckers, M., Lock, D., and Wells, C. (2011). National Hockey League Skater Ratings Based upon All On-Ice Events: An Adjusted Minus/Plus Probability (AMPP) Approach. `http://myslu.stlawu.edu/~msch/sports/LockSchuckersProbPlusMinus113010.pdf`. [Online; accessed 02-August-2013].

Silver, N. (2006). Lies, Damned Lies: Playoff Hurlers. `http://www.baseballprospectus.com/article.php?articleid=5560`. [Online; accessed 23-July-2013].

Slaton, Z. (2012). The Analyst Behind Manchester City's Rapid Rise. `http://www.forbes.com/sites/zachslaton/2012/08/16/the-analyst-behind-manchester-citys-player-investments-part-1/`. [Online; accessed 02-August-2013].

Slowinski, S. (2011). The Projection Rundown: The Basics on Marcels, ZiPS, CAIRO, Oliver, and the Rest. http://www.fangraphs.com/library/the-projection-rundown-the-basics-on- marcels-zips-cairo-oliver-and-the-rest/. [Online; accessed 24-July-2013].

Sueyoshi, T., Ohnishi, K., and Kinase, Y. (1999). A benchmark approach for baseball evaluation. *European Journal of Operational Research*, 115:429–448.

Swartz, M. and Seldman, E. (2010). Introducing SIERA. `http://www.baseballprospectus.com/article.php?articleid=10027`. [Online; accessed 23-July-2013].

Szczepański, Ł. (2008). Measuring the effectiveness of strategies and quantifying players' performance in football. *International Journal of Performance Analysis in Sport*, 8(2):55–66.

Szczepański, Ł. and McHale, I. G. (2015). Beyond completion rate: evaluating the passing ability of footballers. *Journal of the Royal Statistical Society. Series A (Statistics in Society) (to appear)*.

Tango, T., Lichtman, M., Dolphin, A., and Palmer, P. (2008). *The Book: Playing the Percentages in Baseball*. Potomac Books Inc.

Tekwe, C., Carter, R., and Ma, C. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, 29(1):11–35.

Thomas, A., Ventura, S., Jensen, S., and Ma, S. (2013). Competing Process Hazard Function Models for Player Ratings in Ice Hockey. *arXiv preprint arXiv:1208.0799*.

Tiedemann, T., Francksen, T., and Latacz-Lohmann, U. (2010). Assessing the performance of German Bundesliga football players: a non-parametric metafrontier approach. *Central European Journal of Operations Research*, 19(4):571–587.

Tunaru, R. S. and Viney, H. P. (2010). Valuations of Soccer Players from Statistical Performance Data. *Journal of Quantitative Analysis in Sports*, 6(2).

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Zak, T. A., Huang, C. J., and Siegfried, J. J. (1979). Production efficiency: the case of professional basketball. *Journal of Business*, 52(3):379–392.