A Study of Home Advantage in Football and Other Contributions to Sports Data Analysis

Author:
George Foulds

Supervisors:
Roger Brooks
Jonathan Tawn
Mike Wright

Submitted for the degree of Doctor of Philosophy at Lancaster University

November 28, 2016

Abstract

This thesis considers the application of statistical exploratory methods and modelling techniques to sports data. Key to this investigation is the analysis of home advantage and the factors which drive it. The review of literature has shown that much conjecture has been made about the cause of home advantage, but little statistical investigation has been pursued into this area.

Building on the model for association football goal counts discussed in Dixon and Coles (1997), reparameterisation to reflect time and team dependent home advantage was explored, alongside the effect of cards on home advantage. Covariate analysis was performed using parametric and semi-parametric models in an attempt to better interpret home advantage by analysing regularly hypothesised causal relationships.

Over and under dispersion in goal counts may be the result of variation in team skill or the lack thereof. Censoring and threshold mixture models were explored to try and capture any over or under dispersion, with the aim of creating a more flexible model. As an aside, weighted likelihood based changepoint methods were also explored as a method of considering the reduction in information about the threshold position carried by observations far from the threshold position.

Finally, a brief but insightful analysis of changes of performance in golf was carried out. The research contained within can be used to inform statistical models for sports results and impact betting strategies based upon such models.

Acknowledgements

I would like to acknowledge the support of my '11-'15 contemporaries within the STOR-i Centre for Doctoral Training for their guidance and support as well as their continued friendship.

Thank you to all of my supervisors, Jonathan Tawn, Roger Brooks and Mike Wright, for their time, encouragement and advice throughout my PhD. Special thanks to Jon for all of his support and patience throughout the conception and writing of this thesis. You are an inspiration to your students.

I am grateful to the EPSRC and ATASS Sports for their financial support. Within ATASS Sports, I would like to thank Jonathan Croft and Tim Paulden for their fresh perspectives on my work.

Finally, I would like to thank my fiancé and parents for their continued support throughout this experience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Sport has never been more prevelant in society. The international federation of association football (FIFA) claim viewing figures for the 2010 World Cup in South Africa reached 3.2 billion people, or 46.4% of the global population (FIFA , 2010). However, the current crop of statistical research into many sports is basic, often relying on raw outcomes with little further analysis than common sense. It is easy to argue that the general populous does not need to be able to accurately predict the outcomes of games. They may even solely watch sport to enjoy not knowing. Sports betting, through bookmakers or various online internet sites, allows an outlet for those who wish to gamble on these games. Betting on your favourite team is common place and can be done from anywhere with a mobile phone signal.

However, there are markets other than recreational betting. Many individuals and companies are beginning to accept that sports betting provides investment opportunities equal to those provided by the various stock exchanges around the world. The entry of this new strategy has had a difficult birth, with a large amount of media coverage given to the collapse of Centaur Galileo, the first sports betting hedge fund early in 2012 (Wager Minds, 2012).

To prevent such losses, the use of effective high level statistical modelling must be implemented. Before an investment strategy is designed, there must be a combination of statistical analysis with in-depth sports knowledge to formulate a better understanding of the sport in question. ATASS Sports provides this service, with one of the largest commercial statistical research teams in Europe (ATASS Sports, 2012). The combination of resources and funding they are prepared to share with the STOR-i doctoral training centre shows a commitment to ground breaking research and the development of key relationships in the academic sector.

It is not only the consumer who seeks to benefit from sports data analysis; successful coaching choices and in-game decisions rely on having the best information available, such as that resulting from high level statistical analysis of the available data.

Two possible routes of investigation, that are of value to both investors and sporting bodies, were identified as *home advantage, the distribution of goals scored in association football* and *changes of performance in sport*, due to alteration of the rules or evolving technologies. It became evident that home advantage and goals scored would be the main body of research as the PhD progressed, however, changes of performance in sport provided a short interlude and as such shall be supplied at the end of the research chapters.

The term home advantage came into existence in its most abstract form at the start of the 19th century, coinciding with the start of association football (Pollard, 2008). It has since been widely discussed in literature (Roth, 1957; Koppet, 1972; Lane, 1976; Morris, 1981; Dowie, 1982; Pollard, 1986). Home advantage is recognised as the positive effect experienced by a team playing at home and can also be attributed to the negative effects experienced by the away team (Pollard, 2008). There have been many theories hypothesised regarding the causes of home advantage; including crowd support, learning factors, travel by the away team and the bias of rules and their enforcement. Much of the work which has been done is of a non-statistical nature, including papers in psychological and social journals (Varca, 1980; Courneya and Carron, 1992). These do however, give direction to more mathematically orientated research, such as that carried out by Dixon and Coles (1997).

Golf provides an open and attractive research opportunity for changes in performance. Much data exists for the sport, with round by round scores available freely on the internet (ESPN, 2012). There are various aspects to the game which could allow performance enhancement through technological and coaching innovation. This has created a large amount of interest from both professional golf bodies and amateurs alike. Any improvement that can be made, in overall performance or player consistency, is highly regarded. Therefore, statistically modelling these improvements allows for better predictions of outcome.

Statistical modelling in both of these areas is important to allow the prediction of performance, which is important to the sports themselves and those who wish to make profitable betting strategies.

1.2 Content Summary

This thesis contains statistical analyses and literature reviews separated into the following Chapters:

- Hypothetical Covariates of Home Advantage (literature review)
- An Initial Analysis of Home Advantage
- Dixon and Coles: Model and Development
- Covariate Modelling of Home Advantage
- Overdispersion and Threshold Effects (in the assessment of the distribution of goals)
- Weighted Likelihood Based Changepoint Detection Methods
- Changes of Performance in Golf

It is clear from the review of literature in Chapter 2, that there are many conflicting ideas and opinions on the causes, effects and modelling approaches with regards to home advantage (Roth, 1957; Koppet, 1972; Lane, 1976; Morris, 1981; Dowie, 1982; Pollard, 1986). Much of this research has little supporting evidence and supplies only the views of the author. The objective of this PhD is to consolidate these ideas with statistical findings and hypothesise new directions.

Chapter 3 serves as an initial analysis of home advantage, carried out using a basic model for goal counts in association football designed by Clarke (1996). Data for many leagues were provided for over 105,000 matches by the industrial partner ATASS Sports. This analysis allows an insight into the nature of home advantage for each individual team in the top four English divisions between 2001/2002 and 2011/2012.

It becomes clear that a more complex generalised linear model describing the joint probabilities of home and away goal counts was needed than the Clarke (1996) approach. More stringent analyses are then carried out in Chapter 4 using a bivariate Poisson model for home and away goal counts based on a model proposed by Dixon and Coles (1997). Various additions to the model are considered including team dependent home advantage, changes in home advantage over time and the relation of home advantage to the number of cards given in each match. Various statistics used to analyse these models suggest that a number of additions are significant. However, the predictive power (which motivates this problem) is reduced in each case due to the reduction in available information to derive the parameter estimates.

Covariate analysis regarding home advantage in association football is then performed in

Chapter 5. This entails the use of semiparametric and parametric, linear and non-linear models, including changepoints, to effect a complete investigation into the presence and structure of any relationships between home advantage and distance, attendance, referee experience and pitch dimensions. This process results in a model for home advantage using multiple covariates and models, which shows a slight improvement in the predicitive capabilities of the base model considering constant home advantage.

Overdispersion is found to be present in goal count data for both home and away counts, leading to an investigation into possible models which consider this feature. Chapter 6.2 considers both negative binomial and censored Poisson models to allow more accurate prediction of the majority of goal counts. Following this, threshold mixture modelling is implemented, which allows the adjoining of two distributions, transformed to ensure that the second axiom of probablity is satisfied, i.e. $P(S) = 1$, where S represents the sample space. This model shows a reduction in the RMSE over the null Poisson model. However, it causes a reduction in the model's predicitive power, suggesting overfitting of the training data set.

The use of changepoints and threshold mixture modelling motivated the investigation of a model considering smooth changes. Chapter 7 discusses the use of a weighting function across the transition boundary (or changepoint), to mix either parameters within the probability distribution function (pdf) or two or more pdfs themselves. The later of these two models shows itself to be the most promising. An interesting result of this modelling process is the possiblity of smooth change prediction, if the change can be assumed to follow the form imposed by the weighting function. Therefore, the model can predict the timing of the change, something that is lacking from current changepoint methods.

Many sports use technology and coaching which can be optimised in an attempt to aid performance. However, the extent to which this may be done is often unclear. Chapter 8 attempts to quantify changes in player consistency (as a measure of performance) for the golf European Tour. Player consistency is particularly important in golf, due to the relatively narrow margins between the player that wins a tournament and the player which comes last. The consistency of some of the best professional players over the last 40 years are analysed, using their normalised round scores. A pertinent question which shall be addressed is: Does perceived change in consistency actually represent an increase in strength in depth (with reference to the increasing skill of the players in the tour)?

Chapter 2

Hypothetical Covariates of Home Advantage

There are many theoretical explanations for home advantage. These include biologically-founded theories of territoriality and circadian rhythm changes, social psychological theories such as social facilitation or perceived social support, and sociological-based theories such as ritual integration (Edwards, 1989). At present, there is little evidence in favour of one theory over another. Courneya and Carron (1992) proposed a framework for research into the effect of game location that allows the inclusion of many different constructs from varying theories. This framework is shown in Figure 2.1 below.

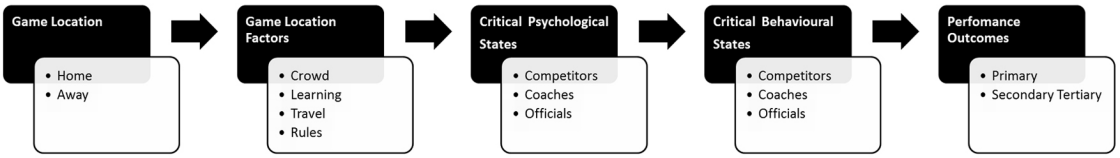


Figure 2.1: Game Location Research framework suggested by Courneya and Carron (1992).

Game location refers to the particular venue for the competition event. There are three possible alternatives, home, away and neutral ground. Neutral sites do not experience a home advantage and therefore are not included in Figure 2.1. The variable factors that occur due to the location, (*Game location factors*, Figure 2.1) can be grouped into four sectors: *crowd*, *learning*, *travel* and *rule factors*. The main theories are derived from these four categories.

Crowd effects may have many factors, for example, size, density and proximity. These effects (either a positive home reinforcement or a negative away effect) reflect the social support provided by the home crowd. Familiarity of the home ground facilities allows the team to experience learning factors, for example surface type. The distance between grounds prompts the possibility of travel effects introducing factors such as mental fatigue and routine disruption. In some sports, rules may influence the outcome of an event in

favour of the home team. For example, in baseball, the home team gets to bat last in each innings. It has been hypothesised that the effects of game location manifest themselves through the varying psychological states of the participants.

Critical psychological states may be influenced by the location of an event. These may vary at the same location for *competitors*, *coaches* and *officials*. Cognitive (e.g. anxiety, confidence, evaluation apprehension and outcome expectation) and affective states (e.g. excitement, anger, stress and pride) may lead to changes in behaviour of one or more of the groups involved with the match.

Critical behavioural states must be affected if a home advantage or game location is to influence any of the participants' performances, and thereby the match outcome. For competitors, these states may include the effort level used or their persistence. Coaches may implement strategic and tactical decisions and, finally, officials may be influenced to make subjective decisions.

The measure of performance outcomes can be split into three, or more, levels. A *primary* performance measure may be the first stage of outcome and would represent the fundamental quality of a team, for example batting average in cricket or free throw percentile in basketball. *Secondary*, or intermediate, performance measures would seek to reflect the necessary scoring system, goals allowed in association football for example. Finally, the *tertiary* measure would represent the overall outcome of a contest, for example win or loss or points differential (Edwards, 1989; Irving and Goldstein, 1990; Schwartz and Barsky, 1977).

2.1 Game Location Factors

2.1.1 Crowd Effects

The effect of the crowd is the most common factor attributed to home advantage, one which at least the fans believe is dominant (Wolfson et al., 2005). The extent and nuances of this effect are difficult to quantify (Morris, 1981; Pollard, 1986). The crowd size, density, support intensity and proximity to players all require detailed consideration (Nevill et al., 1996). However, the effect of each factor is unclear. For example, advantages have been shown with both small and large crowds, providing conflicting evidence that is difficult to interpret (Courneya and Carron, 1992; Pollard, 1986).

Dowie (1982), Pollard (1986) and Clarke and Norman (1995) all found that there was little variation of home advantage over the four divisions of the English Football League, despite the obvious differences in crowd size. When Nevill et al. (1996) added another lower English league and three Scottish divisions to the previously considered set, they

found a linear relationship between crowd size and home advantage. Unfortunately, the sample size used in this analysis was relatively small (only one season). They also showed little difference between the top three English divisions.

League	Level	Home advantage	Average Attendance
Premier	1	60.7%	31009
Division 1	2	61.2%	14160
Division 2	3	60.3%	6649
Division 3	4	61.9%	3757
Conference	5	56.7%	1484
Ryman Premier	6	56.7%	487
Ryman Division 1	7	54.1%	247
Ryman Division 2	8	53.3%	129
Ryman Division 3	9	55.1%	89

Table 2.1: Home advantage and average attendance of nine ranked levels of competition in English football for the seasons 1996/97 - 2001/02 (Pollard, 2006).

Whether the effect of crowd interaction is to give a positive advantage to the home team or a disadvantage to the away team is not known. It is also unclear whether the players themselves are directly affected or whether the referee conveys the crowd influence to them (Boyko et al., 2007). All-seater stadiums may also modify the crowd effect.

Table 2.1 shows the value of home advantage and average attendance for nine English leagues, ranked with respect to their playing level. The value of home advantage is given as the percentage of all games played that are won at home (Pollard, 1986). Although crowd size varies dramatically, very little difference in home advantage can be seen between the top four levels, which all experience a home advantage of just over 60%. There is a step change below this level to a home advantage of around 55%, with teams experiencing the effect even with crowds on average less than 100. This shows that a home advantage is experienced in all English football leagues, and that when crowd size increases above around 3,000 the extent to which home advantage is experienced increases by around 5%. Pollard (1986) also found that there was little relationship between crowd density and the level of home advantage experienced by a team.

2.1.2 Learning

Repeated training and playing at home grounds generally results in familiarity with the local playing conditions, which can be considered as a factor of home advantage. Barnett and Hilditch (1993) showed that English football teams playing on grounds which use artificial turf gain an increased home advantage over those which practise on traditional grass pitches. Clarke and Norman (1995) confirmed this finding, which resulted in a

ban being imposed on the use of artificial turf in the Football League. However, Pollard (1986) produced evidence that the use of abnormal pitch dimensions did not increase the home advantage experienced. Moving teams or players to new home grounds has been shown to decrease the home advantage, which can be accounted for by the reduction or loss of familiarity with playing conditions (Pollard, 2002).

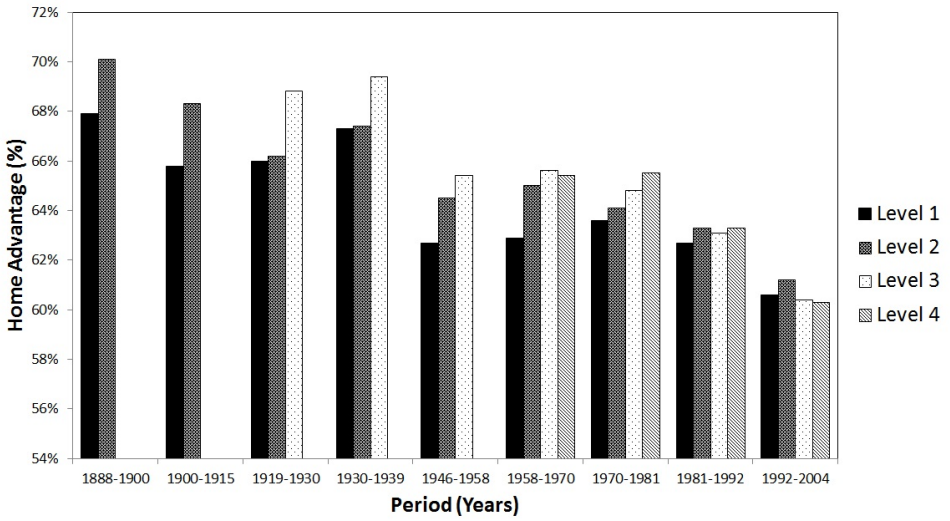
There are numerous factors which may contribute to the familiarity of home grounds. For example, the effects of prevailing winds and daylight depend on the alignment of a stadium and facility design or layout can provide visual cues to home players. Pollard and Pollard (2005) showed that there was a significant drop in home advantage in the Football League of England after the seven year suspension during the Second World War. Upon resuming League play in 1946, participating teams contained many new players, unfamiliar with their local playing environment. Therefore, this can be attributed to familiarity or learning if it is to be considered a contributing factor with regards to home advantage.

2.1.3 Travel

Fatigue experienced by away teams due to travel has been extensively analysed with respect to the extent to which it contributes to home advantage (Pollard, 2006). Reduced home advantage levels exhibited in local derbies could be attributed to the lack of travel for the players involved, as well as the local crowd support for the away team (Clarke and Norman, 1995). Further evidence was obtained by Pollard (2006) from European Cup and Champions League semi-finals between 1960/61 - 2003/04, where competing teams experienced an average home advantage of 71.7% [1]. This high level could be explained by the long distances travelled between matches. Also, Brown et al. (2002) and Clarke and Norman (1995) found that the home advantages of individual teams, in international competitions and English leagues respectively, increased as a function of distance between the teams.

Pollard (1986) produced a contradicting argument, showing no difference in home advantage for teams that were more than or less than 200 miles (320km) apart. Figure 2.2 is a bar plot of home advantage at different levels of English Football for the period 1888-2004. This shows that the home advantage slowly decreased over the twentieth century. This could be explained by the speed and comfort brought about with the development of modern transport (Pollard, 2006).

[1] Pollard (2006) defined home advantage as the number of points gained at home as a percentage of the number of points gained in all matches, with 50% indicating no home advantage.



Level 1: Div 1 (1888-1992), Premier (1992-2004).

Level 2: Div 2:(1892-1992), Div 1 (1992-2004).

Level 3: Div 3 (1920-1921), Div 3 North & South (1921-1958), Div 3 (1958-1992), Div 2 (1992-2004).

Level 4: Div 4 (1958-1992), Div 3 (1992-2004).

Figure 2.2: Home advantage for different levels of English football from 1888-2004. (Pollard, 2006).

2.1.4 Rules

The role that rules play in home advantage is not clear in many sports, as many sports use an unbiased rule book with respect to game location. Therefore, research on this factor is limited to certain sports. In baseball the opportunity of batting last may allow for distinct advantages with respect to game strategy if the game goes into extra innings. However, Courneya and Carron (1990) showed this not to be the case, using slo-pitch softball on neutral grounds, where there appeared to be no advantage provided by batting last. They postulated that any advantage may be eliminated as opportunties for offensive and defensive strategies may be equal.

During stoppages of play, ice hockey teams playing at home are allowed to substitute players after the away team allowing for better strategies. Also, during a faceoff between two players, the home player places his stick on the ice after the visitor, increasing the odds of him winning the faceoff. These rules may translate some home advantage to neutral games as a home designation is assigned to one team (IHHF, 2012).

2.1.5 Critical Psychological and Behavioural States

Competitors: The psychological state of a team or individual may be affected by game location, which may in turn affect the outcome of a competitive event. An examination of player perception of the home advantage was carried out by Jurkovac (1995). Basketball

players were surveyed with questions which could allow an insight into their attitudes when playing home and away. The results of the survey showed that 97% of the players asked felt they played better when backed by a loud and active crowd at home and 74% felt the same when playing at away grounds. Signs of support in the home arena also gave an apparent boost to morale and motivation, with an 89% agreement in the survey. Also, 47% believed that their personal statistics were improved when playing at home and 76% felt more confident at home.

Psychological effects are often the result of physiological reactions to environmental stimuli, such as the fight or flight mechanism (Gleitman et al., 2010). Hormones such as testosterone and adrenalin can act to alter the mental state of a player. Much research has been carried out into the effect of the steroid hormone testosterone. Studies into seasonal variations in male animal aggression have been shown to coincide with seasonal variations in testosterone (Lincoln et al., 1972). Monoghan and Glickman (1992) showed that artificially increasing testosterone increases aggression. Levels of testosterone have been shown by Mazur and Booth (1998) to increase competitiveness and dominance in humans, as testosterone levels rise when a challenge is faced. This may translate to the behavioural implications which cause enhanced status, although not all researchers agree with the conclusions proposed (Neave and Wolfson, 2003).

Varca (1980) measured aggressiveness to study home and away team behaviour in collegiate basketball games. Measures such as fouls, steals, rebounds and blocked shots were used to gauge the nature of play. These measures were then divided into subsets of functional (advantageous) and dysfunctional (disadvantageous) for the respective team. Varca hypothesised that home and away teams differed in type rather than level of aggression, a hypothesis that was supported by his findings. Teams playing at home showed more functional aggressive behaviour, blocking shots and achieving more rebounds, whilst away teams showed more dysfunctional behaviour, committing a higher number of fouls.

Officials: An official distributes rules from a position of authority, allowing gameplay to follow the set of predetermined rules. However, due to the nature of certain rules, subjective decisions have to be made. This provides a line of inquiry with respect to the effect of these choices on home advantage. Various studies have been documented that exhibit a leaning towards more decisions in favour of home teams or against away teams (Lefebvre and Passer, 1974; Varca, 1980; Sumner and Mobley, 1981; Greer, 1983; Glamser, 1990).

It must be considered, however, that a greater percentage of subjective decisions being made against the visiting team may not be a cause of home advantage, but rather a consequence (Sumner and Mobley, 1981). Teams playing away may spend more time in

defence and be more dsyfunctionally aggressive. This must be addressed by considering the team and individual player's skill status.

Lehman and Reifman (1987) hypothesised that officials who were present at professional basketball games could feel more pressured to behave less negatively towards the home team's star players. They reasoned that this pressure would originate from crowd influence and believed that investigation into this area could present evidence of an officiating bias. The data set used to test this hypothesis contained games involving the Los Angeles Lakers during the 1984/85 season. Lehman and Reifman concluded that their reasoning was correct, as fewer fouls were awarded to star players of the home team than the away team. No difference was found for players classified as non-stars.

Chapter 3

A First Analysis of Home Advantage

3.1 Introduction

Although widely accepted and discussed, there is often little attempt to quantify home advantage (Clarke, 1996). Hypothetical causes are well documented and an overview of the main considerations is discussed in Pollard (2008). The direction of research will be towards the cause of home advantage, rather than to solely document its existence. However, playing characteristics, stadium layout and thus crowd dynamics differ from club to club and match to match, requiring the estimation of home advantage for individual matches.

Pollard (1986) created a quantifiable definition of home advantage, for a balanced number of games. He defined home advantage as the number of home games won expressed as a percentage of total games played (where draws are omitted), with a value of 50% indicative of zero home advantage. However, this is an unsuitable method for modelling individual clubs, where strength of position in their league or division must be considered. This is often difficult to determine as the draw may not be balanced (e.g. the Australian Football League, AFL).

Snyder and Purdy (1985) analysed home advantage in collegiate basketball and found that the quality of opposition[1] had more of an effect on outcome than home advantage. Because of this, for balanced games such as English football, it makes sense to take into account team ability and to measure home advantage by comparing the home and away fixture results.

3.2 Modelling Home Advantage and Team Ability

Table 3.1 shows the final table for the 2011/12 Barclay's Premier League, a typical example for English soccer. It can be seen from Table 3.1 that the home and away scores are

[1] A team may win more (or less) than 50% at home because it is a relatively strong (or weak) team.

separated. This has become the standard layout due to the acceptance and recognition of home advantage. Without considering team ability, QPR could be perceived to have almost zero home advantage. They have won exactly 50% of their games at home, taking a draw as 0.5 of a win, and have scored 24 home goals and conceded 25. However, by taking away performance into account, team ability can be allowed for. Again, accounting for draws in a similar manner, QPR only won 4 out of the 19 away games played, with a goal difference of -22.

Team	Home					Away							
	Wins	Draws	Losses	Goals For	Goals Against	Wins	Draws	Losses	Goals For	Goals Against	Pts	h	u
Manchester C.	18	1	0	55	12	10	4	5	38	17	89	0.81	1.38
Manchester U.	15	2	2	52	19	13	3	3	37	14	89	0.14	1.52
Arsenal	12	4	3	39	17	9	3	7	35	32	70	0.64	0.49
Tottenham H.	13	3	3	39	17	7	6	6	27	24	69	0.64	0.49
Newcastle	11	5	3	29	17	8	3	8	27	34	65	0.64	-0.01
Chelsea	12	3	4	41	24	6	7	6	24	22	64	0.42	0.45
Everton	10	3	6	28	15	5	8	6	22	25	56	0.47	0.20
Fulham	10	5	4	36	26	4	5	10	12	25	52	0.86	-0.32
Liverpool	6	9	4	24	16	8	1	10	23	24	52	0.08	0.32
Norwich C.	7	6	6	28	30	5	5	9	24	36	47	0.14	-0.23
Swansea C.	8	7	4	27	18	4	4	11	17	33	47	0.97	-0.48
Westbrom. A.	6	3	10	21	22	7	5	7	24	30	47	-0.14	0.08
Stoke C.	7	8	4	25	20	4	4	11	11	33	45	1.08	-0.78
Sunderland	7	7	5	26	17	4	5	10	19	29	45	0.64	-0.16
Wigan A.	5	7	7	22	27	6	3	10	20	35	43	0.14	-0.38
Aston Villa	4	7	8	20	25	3	10	6	17	28	38	-0.08	-0.17
QPR	7	5	7	24	25	3	2	14	19	41	37	0.75	-0.76
Bolton W.	4	4	11	23	39	6	2	11	23	38	36	-0.47	-0.35
Blackburn R.	6	1	12	26	33	2	6	11	22	45	31	0.47	-0.80
Wolver. W.	3	3	13	19	43	2	7	10	21	39	25	-0.75	-0.49
Totals	171	93	116	604	462	116	93	171	462	604	1047	7.47	0

Table 3.1: 2011/12 Barclay's Premier League final table, including individual clubs' home advantage (h) and quality (u).

It is not widely appreciated that a team experiencing a true home advantage will create an apparent home advantage to each other team in their competition. Clarke and Norman (1995) provides an example of spurious and real home advantage.

Clarke and Norman (1995) considered three teams, A, B and C ranked by their skill in alphabetical order (i.e. A is better than B which is better than C), each with no home ground advantage. If it were to be supposed that the results were as in Table 3.2 and the final table took the form of Table 3.3, where draws are counted as 0.5 of a win in column ten, it is obvious that each team experiences the same home and away performance with regards to goals scored and wins.

However, if team C were to be given two additional goals when playing at home the final results would be as in Table 3.4, with Table 3.5 representing the final table. It can be seen from Table 3.5 that all teams now have better results in terms of both goal difference and matches won at home than away (as in columns ten and eleven of Table 3.5).

Home team	Away team		
	A	B	C
A	-	2-1	3-1
B	1-2	-	2-1
C	1-3	1-2	-

Table 3.2: Clarke and Norman (1995) final results.

Team	Home				Away				Home-away	
	Wins	Draws	Losses	Goals	Wins	Draws	Losses	Goals	Wins	Goals
A	2	0	0	5-2	2	0	0	5-4	0	0
B	1	0	1	3-3	1	0	1	3-5	0	0
C	0	0	2	6-5	0	0	2	2-5	0	0

Table 3.3: Clarke and Norman (1995) final ladder.

It is now easy to see how false conclusions can be drawn about the home advantage of individual teams. Even though only C experiences a real home advantage, teams A and B experience a spurious home advantage.

Home team	Away team		
	A	B	C
A	-	2-1	3-1
B	1-2	-	2-1
C	3-3	3-2	-

Table 3.4: Clarke and Norman (1995) final results when C is given a 2-goal advantage.

Team	Home				Away				Home-away	
	Wins	Draws	Losses	Goals	Wins	Draws	Losses	Goals	Wins	Goals
A	2	0	0	5-2	1	1	0	5-2	0.5	2
B	1	0	1	3-3	0	0	2	3-3	1	2
C	1	1	0	2-5	0	0	2	2-5	1.5	4

Table 3.5: Clarke and Norman (1995) final ladder when C is given a 2-goal home advantage.

Estimation of an individual team home advantage requires a model for the ability of a team (Clarke, 1993; Stefani, 1983; Stefani and Clarke, 1987). Clarke and Norman (1995) suggested a derivation of a formula for the calculation of home advantage and that of team performance by the use of least squares. Suppose the winning margin[2] for home team i against away team j is represented by w_{ij}, which is negative for the case of a loss.

[2]There are two possible definitions of w_{ij} which could be used. Firstly, it may be defined as 1 if the team won, 0 for a draw or -1 if they lost. Alternatively, and the preferred method for this prediction model, is to define it as the goals margin (or difference), as this allows for more sensitivity to home advantage. To illustrate this, imagine a particular team wins 4-0 at home and also wins 2-1 away. Using the win, lose or draw method would give zero home advantage, however, there would be a 3 goal home advantage using goal difference.

This would produce an $N \times N$ matrix for N teams, with no diagonals. Summing across the Ith row gives the value of home goal difference (HGD_I) and summing down the Ith column gives the negative away goal difference (AGD_I) for team I, i.e.

$$HGD_I = \sum_{j=1(j\neq I)}^{N} w_{Ij}, \quad AGD_I = -\sum_{i=1(i\neq I)}^{N} w_{iI}.$$

This can be modified as shown by

$$\sum_{i=1}^{N} HGD_i = -\sum_{i=1}^{N} AGD_i,$$

where it can be seen that the w_{ij} are simply being summed in a different order.

By defining h_i as a measure of the home ground advantage of an individual team i, u_i as a measure of their quality and ε_{ij} as a zero-mean random error, Clarke and Norman (1995) modelled w_{ij} by

$$w_{ij} = u_i - u_j + h_i + \varepsilon_{ij}. \tag{3.1}$$

It is assumed in this model that u_i and h_i are constant throughout the progression of the season. This is something that may need to be addressed in future models.

To make the model identifiable, an additional constraint is added, ensuring that $\sum_{i=1}^{N} u_i = 0$ for N teams. The model can then be fitted to the data using a simple regression package. Dummy variables would be used for u_i (1 for a home team, -1 for an away team and 0 for other teams in the league) and h_i (1 for the home team and 0 for other teams).

The values of u_i and h_i can also be found using a Lagrange Multiplier technique which seeks to minimise the sums of the squares of the errors. The derivation is shown in Appendix A. If data from complete balanced seasons are analysed, complicated regression methods do not need to be employed in favour of simple calculus using the final table. This is the method used in Table 3.1.

In a balanced league of N teams, each team plays the other $N-1$ teams once at home and once away. The total home advantage of the entire league is represented by H. This is calculated by

$$H = \sum_{i=1}^{N} h_i = \frac{\sum_{i=1}^{N} HGD_i}{N-1}.$$

Team i's home advantage, h_i, is given by the difference in their home and away goal differences, minus the total home advantage, divided by $N-2$, as shown by

$$h_i = \frac{HGD_i - AGD_i - H}{N-2}.$$

Finally, the individual team ability measure, u_i, is given by

$$u_i = \frac{HGD_i - (N-1)h_i}{N}.$$

3.3 Data and Results

Data were provided by the industrial partner (ATASS Sports) for over 105,000 association football matches, spread globally over 98 leagues with varying levels from 2001 to 2012. This initial analysis will restrict research to the top four English leagues (Premier League, Championship, League One and League Two).

The data were arranged into the final end-of-year tables for each league and each year (where data were available). These results were checked against the BBC sports website to ensure the tables were correct and altered until agreement was obtained (BBC, 2012).

For the 22206 games played over all four leagues examined, 9892 (44.6%) home wins were recorded, 6049 (27.2%) draws and 6262 (28.2%) losses. In terms of goals, home teams scored 32758 (56.8%) of the total 57682 goals scored. As found by Clarke (1996) this is just under the percentage of wins (1 for a win, 0.5 for a draw) $44.6+0.5\times27.2 = 58.2\%$.

It was also found that the average home advantage per team from seasons 2001/02 to 2010/11 (due to incomplete data for 2011/12 in the lower leagues) decreased as the league level decreased. This can be seen in Table 3.6.

League	**Average Home Advantage (Goals per Match, H)**										**Overall Average**
	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	
Premier League	0.30	0.37	0.35	0.43	0.44	0.46	0.42	0.32	0.62	0.45	0.43
Championship	0.42	0.32	0.36	0.33	0.33	0.39	0.34	0.33	0.43	0.33	0.36
League One	0.42	0.31	0.42	0.33	0.34	0.32	0.34	0.31	0.43	0.31	0.35
League Two	0.53	0.34	0.40	0.39	0.31	0.32	0.11	0.32	0.34	0.25	0.33

Table 3.6: Average home advantage (H) per team for the top 4 leagues in English football.

The average per team home advantage in terms of goals scored for each league differs from that found by Clarke and Norman (1995), who found the values to be 0.521, 0.529, 0.529

League	Home Advantage (Goals per Match)											Overall	Standard
	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	Average	Error
Sheffield U.	-	-	-	-	-	1.11	-	-	-	-	-	1.11	-
Stoke C.	-	-	-	-	-	-	-	1.37	0.42	1.06	1.08	0.98	0.17
Burnley	-	-	-	-	-	-	-	-	0.97	-	-	0.97	-
Swansea C.	-	-	-	-	-	-	-	-	-	-	0.97	0.97	-
Crystal Palace	-	-	-	0.91	-	-	-	-	-	-	-	0.91	-
Queens Park R.	-	-	-	-	-	-	-	-	-	-	0.75	0.75	-
Norwich City	-	-	-	1.13	-	-	-	-	-	-	0.14	0.63	0.35
Tottenham H.	0.78	0.31	0.84	0.74	0.35	0.66	0.59	0.87	0.97	0.22	0.64	0.63	0.07
Portsmouth	-	-	1.78	0.85	0.13	0.77	0.20	0.37	0.20	-	-	0.61	0.21
Newcastle U.	0.34	0.87	1.06	0.07	0.68	0.33	0.53	0.14	-	1.11	0.64	0.58	0.11
Fulham	0.67	0.98	0.17	0.30	1.18	0.72	-0.25	0.70	1.03	-0.06	0.86	0.57	0.13
Birmingham C.	-	0.26	0.11	0.85	0.63	-	1.20	-	0.47	0.33	-	0.55	0.13
Southampton	0.23	0.76	0.45	0.68	-	-	-	-	-	-	-	0.53	0.10
Ipswich T.	0.50	-	-	-	-	-	-	-	-	-	-	0.50	-
Watford	-	-	-	-	-	0.49	-	-	-	-	-	0.49	-
Liverpool	-0.27	0.04	0.17	0.68	0.40	1.38	0.70	-0.02	0.97	1.22	0.08	0.49	0.16
West Ham U.	2.17	0.20	-	-	0.24	0.61	-0.02	-0.08	0.47	0.22	-	0.48	0.24
Manchester C.	-	0.20	0.34	0.18	0.46	-0.34	0.87	1.64	0.08	0.44	0.81	0.47	0.16
Manchester U.	-0.11	0.70	0.45	-0.15	0.63	0.16	0.75	0.53	0.53	1.33	0.14	0.45	0.12
Middlesbrough	0.00	1.04	0.28	0.24	-0.15	0.55	0.53	0.92	-	-	-	0.43	0.14
Blackburn R.	0.89	0.09	-0.61	0.02	0.57	0.27	0.20	0.64	1.20	0.89	0.47	0.42	0.14
Everton	0.67	0.65	1.06	0.57	0.35	0.38	0.20	-0.13	0.25	0.06	0.47	0.41	0.09
Sunderland A.	1.11	-0.19	-	-	-0.87	-	1.03	0.31	1.20	-0.11	0.64	0.39	0.24
Chelsea	0.56	0.81	-0.11	-0.43	0.96	0.16	-0.08	-0.47	1.36	0.39	0.42	0.33	0.17
Bolton W.	-0.55	0.48	0.39	0.02	0.85	0.44	1.09	0.31	0.14	0.83	-0.47	0.32	0.15
Derby County	0.67	-	-	-	-	-	-0.08	-	-	-	-	0.30	0.27
Aston Villa	0.28	1.09	-0.05	0.91	0.24	0.05	-0.25	-0.02	0.03	0.89	-0.08	0.28	0.13
Arsenal	-0.83	0.20	-0.11	0.57	1.35	0.94	0.03	-0.41	0.64	-0.11	0.64	0.26	0.18
Reading	-	-	-	-	-	0.22	0.25	-	-	-	-	0.24	0.01
West Bromich A.	-	-0.30	-	0.13	0.68	-	-	0.59	-	0.33	-0.14	0.22	0.15
Leeds U.	-0.11	-0.58	1.11	-	-	-	-	-	-	-	-	0.14	0.41
Wigan A.	-	-	-	-	-0.32	-0.62	0.92	0.14	1.08	-0.67	0.14	0.10	0.24
Charlten A.	-0.50	-0.24	-0.39	0.41	0.40	0.83	-	-	-	-	-	0.09	0.20
Blackpool	-	-	-	-	-	-	-	-	-	0.00	-	0.00	-
Wolverhampton	-	-	0.45	-	-	-	-	-	-0.36	0.61	-0.75	-0.01	0.28
Hull	-	-	-	-	-	-	-	-0.97	0.81	-	-	-0.08	0.63
Leicester C.	-0.55	-	-0.44	-	-	-	-	-	-	-	-	-0.50	0.04
Average	0.30	0.37	0.35	0.43	0.44	0.46	0.42	0.32	0.62	0.45	0.37	0.43	0.03

Table 3.7: Individual team home advantage for all teams playing in the Premier League for the seasons starting in 2001/02 to 2011/12.

and 0.533 for divisions 1-4 respectively using similar calculations to those used here and relating to seasons from 1981/82 to 1990/91. This may suggest that there is a negative trend in home advantage over time.

To understand home advantage fully, a model may need to account for individual team or match pairing home advantage. Utilising the methods laid out in Section 3.2, the home advantages experienced by each team that played in the Premier League from 2001/02 to 2011/12 are shown in Table 3.7, which is arranged in decreasing order with regards to the mean home advantage.

Because of the inherent variable nature of football games, an average value of home advantage over a number of years is required to allow for reasonable conclusions. It is

obvious from Table 3.7 that some teams have rarely been involved in the Premier League, as such little significance can be attributed to their average home advantage. It should also be noted that certain teams exhibit negative home advantage. This is noteworthy, as a positive home advantage is thought to be present over all teams (Pollard, 1986).

One possible cause of home advantage that has been previously discussed is the team's familiarity with their home ground. However, this could also be due to the unfamiliarity of visiting teams. Therefore, teams that are new to the division, or who have re-entered after a period of absence, may be expected to have a high home advantage. Clarke and Norman (1995) found this to be the opposite of empirical evidence, with a small non-significant positive correlation between home advantage and years in the division. The data analysed from the top four English leagues seems to follow the same result.

However, there are a few teams that exhibit the hypothesised behaviour. For example, Sheffield United and Stoke City entered the Premiership in 2006 and 2008 respectively, both showing high home advantage. Also after a period of a one season absence, Birmingham City exhibited their highest home advantage over the period analysed. This can be countered with examples of Reading and Derby County, who exhibited low or negative home advantage upon joining the league, which could suggest that certain aspects of the grounds which follow this behaviour may help in causing unfamiliarity for visitors, whilst other grounds less so. It may also suggest that upon promotion from one league to another various factors such as increased crowd support and a better psychological condition may affect home advantage. If this is the case, by deduction the home advantage should be lower as teams go down by a division. This is shown for both Birmingham City and Sheffield United, where home advantage decreased to 0.045 and 0.13 respectively in the season following their demotion.

Many studies tend to look for seperate effects individually. Through a more sophisticated modelling approach, different factors may be assesed simultaneously. This could potentially reveal much more information. One such model is that proposed by Dixon and Coles (1997), which considers the goals counts as Poisson random variables. This will be covered in the following chapters.

3.4 Comparison of Clarke and Norman and Dixon and Coles Models

Clarke and Norman (1995) modelled the winning margin of the home team, which gives information at a match level about the relative strengths of teams. However, it is an unsuitable model for many betting strategies as bookmakers asign probabilities to results

rather than winning margins. More sophisticated models for the prediction of match outcome exist, including that proposed by Dixon and Coles (1997).

Dixon and Coles (1997) based their model on Maher (1982), allowing each team an attack and defence parameter and a home advantage parameter (γ) for the league. The model defined in Dixon and Coles (1997) is used as a basis of research in the following chapters, and as such is well defined there. However, using the simpler model defined in Maher (1982), which considers home and away goals to occur independently of each other, a comparison of home advantage was made as shown in Figure 3.1. It can be seen, that both follow similar trends, which supports conclusions made from the Clarke and Norman (1995) model about the nature of home advantage.

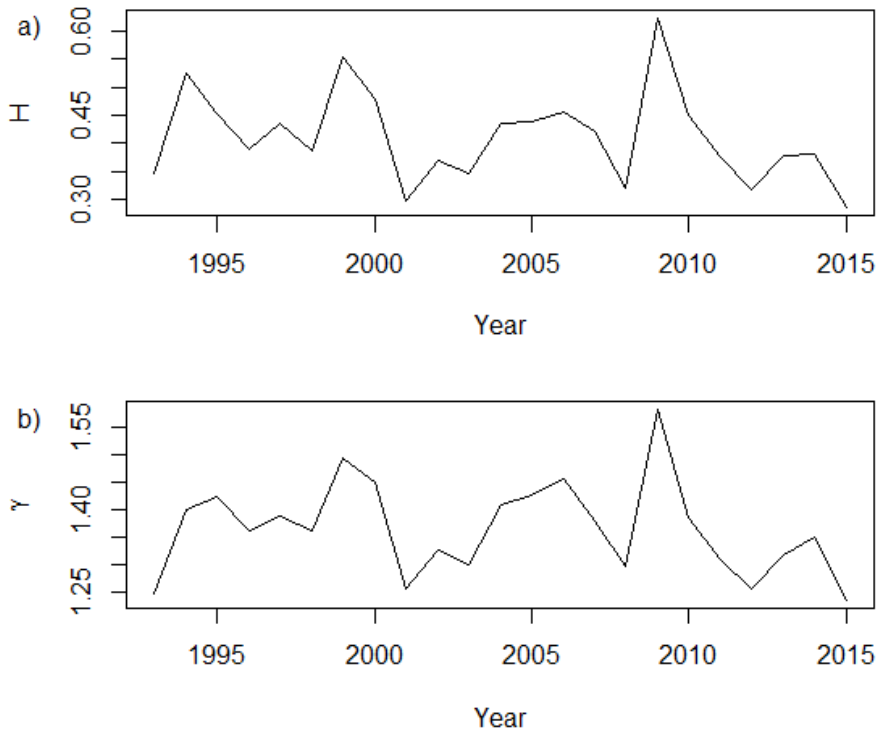


Figure 3.1: Comparison of additive and multiplicative home advantage parameters (a) H (Clarke and Norman, 1995) and (b) γ (Maher, 1982) respectively.

Table 3.8 shows the root mean squared error (RMSE) of the winning margin estimates (calculated as home goals minus away goals) from each model, using individual complete seasons to allow a fair comparison. As discussed, one of the benefits of the generalised linear models discussed in Maher (1982) and Dixon and Coles (1997) is that they allow home and away goal estimates to be made, however, these cannot be compared. Under this comparison, the RMSE of the Maher (1982) model is lower for most years, showing a better fit of the data.

Year	RMSEr of Winning Margin Estimate	
	Clarke and Norman	Maher
1993	1.48	1.48
1994	1.54	1.47
1995	1.51	1.48
1996	1.56	1.54
1997	1.65	1.65
1998	1.56	1.53
1999	1.65	1.59
2000	1.54	1.51
2001	1.52	1.56
2002	1.46	1.45
2003	1.49	1.50
2004	1.45	1.39
2005	1.50	1.47
2006	1.46	1.41
2007	1.52	1.49
2008	1.43	1.46
2009	1.63	1.53
2010	1.58	1.55
2011	1.64	1.62
2012	1.49	1.48
2013	1.60	1.59
2014	1.49	1.46
2015	1.51	1.52
All Years	1.53	1.51

Table 3.8: Using data from the English Premier League between 1993/1994 to 2015/2016, root mean squared error (RMSE) of the winning margin estimates under the models defined by Clarke and Norman (1995) and Maher (1982).

Chapter 4

Dixon and Coles: Model and Development

4.1 Introducing the Dixon and Coles Model

Dixon and Coles (1997) specified a feature set for the development of a statistical model to predict the outcome of association football matches. The intended use of the model was to create a profitable betting strategy. The specifications were:

- the quality of the two participating teams should both be taken into account;
- home advantage should be allowed for;
- it is likely that the most reasonable measure of a team's quality is to base it on a summary measure of recent performance;
- it is also reasonable to assume that due to the nature of football, team ability should be quantified in terms of their attacking and defending abilities;
- the ability of the opposing teams should be taken into account when calculating a specific team's performance.

Dixon and Coles (1997) speculated that obtaining empirical estimates of the probabilities of match outcomes which take into account all these parameters is impractical. Therefore, a statistical model which incorporates the various factors needs to be formulated. The basis of their model relied on that proposed by Maher (1982), which assumes the number of home and away goals scored, X and Y respectively, are both independent Poisson variables.

Let $X_{i,j}$ be the number of home goals scored in a match between home team i and away team j, and similarly let $Y_{i,j}$ be the number of goals scored by the away team in the same match. The model formulated by Maher (1982) is given in equation (4.2), where $X_{i,j}$ and $Y_{i,j}$ are independent and $\alpha_i, \beta_i > 0, \forall i$, where α_i represents a measure of the

attack rate for team i and β_i measures their defence rate. Maher (1982) originally used attack and defence parameters for home and away performance for each team, however, this is encapsulated by the home effect experienced, $\gamma > 0$:

$$X_{i,j} \sim \text{Poisson}(\alpha_i \beta_j \gamma), \tag{4.1}$$
$$Y_{i,j} \sim \text{Posson}(\alpha_j \beta_i). \tag{4.2}$$

In a competition with n teams, there are then $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_n\}$ attack parameters, $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_n\}$ defence parameters and the extent to which home advantage affects the competition, γ, to be estimated. The constraint $n^{-1} \sum_{i=1}^{n} \alpha_i = 1$ is introduced to ensure that the model does not become over-parameterised. Alternatively, α_n may be fixed, for example $\alpha_n = 1$. Dixon and Coles (1997) found that there is a departure from independence for low scoring games. They proposed a modification to the Maher (1982) model which included a dependence parameter ρ, given by

$$\Pr(X_{i,j} = x, Y_{i,j} = y) = \tau_{\lambda,\mu}(x, y) \frac{\lambda^x \exp(-\lambda)}{x!} \frac{\mu^y \exp(-\mu)}{y!}, \tag{4.3}$$

where

$$\lambda = \alpha_i \beta_j \gamma,$$
$$\mu = \alpha_j \beta_i,$$

and

$$\tau_{\lambda,\mu}(x, y) = \begin{cases} 1 - \lambda\mu\rho & \text{if } x = y = 0, \\ 1 + \lambda\rho & \text{if } x = 0, y = 1, \\ 1 + \mu\rho & \text{if } x = 1, y = 0, \\ 1 - \rho & \text{if } x = y = 1, \\ 1 & \text{otherwise,} \end{cases} \tag{4.4}$$

where

$$\max(-1/\lambda, -1/\mu) \leq \rho \leq \min(1/\lambda\mu, 1).$$

In this model, $\rho = 0$ corresponds to independence. However, for events which correspond to $x \leq 1$ and $y \leq 1$ the independence distribution, $\tau_{\lambda,\mu}(x, y)$, is perturbed when $\rho \neq 0$. The use of a dependence function was tested using data from the Premier League between seasons 1995/1996 and 2013/2014. A hypothesis test was performed with a null hypothesis that $\rho = 0$, and an alternative hypothesis that $\rho \neq 0$. In a chi-squared test with one degree of freedom, the null was rejected at a 0.05 significance level, with a p-value of 0.006 [1].

[1] For reference, estimates for α_i, β_i, γ and associated ρ parameters for a range of seasons are given in Tables B.1, B.2, B.3 and B.4 in Appendix B respectively

4.1.1 Model Inference

Dixon and Coles (1997) and Maher (1982) used the likelihood function as their main tool of inference for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, γ and ρ. For a total of N match pairings corresponding to team $i(k)$ at home and team $j(k)$ away, where $k = 1, \ldots, N$, with corresponding home and away scores (x_k, y_k) and including the independence distribution, the likelihood is given up to proportionality by

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \rho, \gamma; \boldsymbol{x}, \boldsymbol{y}) = \prod_{k=1}^{N} \tau_{\lambda_k, \mu_k}(x_k, y_k) \exp(-\lambda_k) \lambda_k^{x_k} \exp(-\mu_k) \mu_k^{y_k}, \tag{4.5}$$

where

$$\begin{aligned} \lambda_k &= \alpha_{i(k)} \beta_{j(k)} \gamma, \\ \mu_k &= \alpha_{j(k)} \beta_{i(k)}. \end{aligned} \tag{4.6}$$

Dixon and Coles (1997) restricted their inference to direct numerical maximisation of equation (4.5). Although this method has to handle the aspect of high dimensionality, many parameter combinations are near orthogonal, making the process relatively simple.

Dixon and Coles (1997) stated that the attack and defence parameters must reflect the relative quality of different divisions if more than one division is to be analysed. Unfortunately, reliable estimation of these parameters requires that teams from different divisions have engaged in competition. This is resolved by the fact that there is a degree of mobility between divisions due to relegation and promotion. Data from cup games may also be included, which includes inter-division play. As was expected, Dixon and Coles (1997) found that the average attack and defence ratings increased with higher league level. This corresponds to an increase in the value of α and a decrease in the value of β.

4.1.2 Derivation of the Closed Form Expression for Home Advantage, γ

The home advantage parameter, γ, can be estimated as with the other parameters via numerical maximisation of the likelihood. However, the closed form expression for home advantage allows the estimation of γ without the need for numerical maximisation. This can be derived using the model outlined by equation (4.3), provided data for complete seasons of the league in question are available. The log-likelihood function describing home and away goal counts for n teams in one season of a balanced league, is given by

$$\begin{aligned} \ell\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \rho, \gamma; \boldsymbol{x}, \boldsymbol{y}\right) = &\sum_{i=1}^{n} \sum_{j \neq i} \log\left[\tau_{\lambda_{i,j}, \mu_{i,j}}\left(x_{i,j}, y_{i,j}\right)\right] + x_{i,j} \log\left(\lambda_{i,j}\right) \\ &- \lambda_{i,j} - \log\left(x_{i,j}!\right) + y_{i,j} \log\left(\mu_{i,j}\right) - \mu_{i,j} - \log\left(y_{i,j}!\right), \end{aligned} \tag{4.7}$$

where

$$\lambda_{i,j} = \alpha_i \beta_j \gamma,$$
$$\mu_{i,j} = \alpha_j \beta_i,$$

and $\tau_{\lambda_{i,j},\mu_{i,j}}(x_{i,j}, y_{i,j})$ is given by equation (4.4).

Taking the first derivative of the log-likelihood function gives the Fisher's score function denoted by

$$\mathbf{u}(\boldsymbol{\theta}) = \frac{\partial \log \mathrm{L}(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}}.$$

Considering the log-likelihood is concave, maximum likelihood estimators (MLEs) can be calculated by setting the score function to zero and solving the system of equations

$$\mathbf{u}\left(\hat{\boldsymbol{\theta}}\right) = \mathbf{0}.$$

Taking the first derivative with respect to γ of the log-likelihood shown in equation (4.7) is then given by

$$\frac{\partial \log \mathrm{L}}{\partial \gamma} = \sum_{i=1}^{n} \sum_{j \neq i} \left(-\alpha_i \beta_j + \frac{x_{i,j}}{\gamma} \right).$$

Equating the score to zero and solving for γ gives the MLE

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} \sum_{j \neq i} x_{i,j}}{\sum_{i=1}^{n} \sum_{j \neq i} \hat{\alpha}_i \hat{\beta}_j}. \tag{4.8}$$

The first derivative with respect to α_i of the log-likelihood shown in equation (4.7) is then given by

$$\frac{\partial \log \mathrm{L}}{\partial \alpha_i} = \sum_{j \neq i} \left(-\beta_j \gamma + \frac{x_{i,j}}{\alpha_i} - \beta_j + \frac{y_{j,i}}{\alpha_i} \right).$$

Equating this to zero gives

$$\sum_{j \neq i} \left(-\hat{\beta}_j \hat{\gamma} + \frac{x_{i,j}}{\hat{\alpha}_i} - \hat{\beta}_j + \frac{y_{j,i}}{\hat{\alpha}_i} \right) = 0.$$

Rearranging gives

$$\sum_{j \neq i} \left(-\hat{\beta}_j \left(\hat{\gamma} + 1 \right) + \frac{x_{i,j} + y_{j,i}}{\hat{\alpha}_i} \right) = 0.$$

This is true for all $i = 1, \ldots, n$ so

$$\sum_{i=1}^{n} \sum_{j \neq i} \left(-\hat{\beta}_j \left(\hat{\gamma} + 1 \right) + \frac{x_{i,j} + y_{i,j}}{\hat{\alpha}_i} \right) = 0$$

and then

$$\sum_{i=1}^{n} \sum_{j \neq i} \hat{\alpha}_i \hat{\beta}_j = \sum_{i=1}^{n} \sum_{j \neq i} \frac{x_{i,j} + y_{i,j}}{\hat{\gamma} + 1}.$$

Substituting this expression into equation (4.8) gives

$$\hat{\gamma} = (1 + \hat{\gamma}) \frac{\sum_{i=1}^{n} \sum_{j \neq i} x_{i,j}}{\sum_{i=1}^{n} \sum_{j \neq i} (x_{i,j} + y_{i,j})}.$$

Rearranging gives

$$\hat{\gamma} \sum_{i=1}^{n} \sum_{j \neq i} (x_{i,j} + y_{i,j}) = (1 + \hat{\gamma}) \sum_{i=1}^{n} \sum_{j \neq i} x_{i,j}$$

and therefore

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} \sum_{j \neq i} x_{i,j}}{\sum_{i=1}^{n} \sum_{j \neq i} y_{i,j}}, \tag{4.9}$$

which, in words, equates home advantage to the sum of home goals over the sum of away goals in a balanced league. This expression allows the use of final league tables to calculate home advantage as only the total home and away goal counts are needed.

4.1.3 Dynamic Behaviour

Each team's performance in terms of both offence and defence tends to be dynamic not static. This means that it will vary from one time point to the next. Dixon and Coles (1997) also included this in their model. Due to the nature of football, team formations change almost every season as players are bought and sold or injured. Therefore, it is reasonable to assume that a team's performance is more likely to be related to their performance in recent matches. Therefore, it is assumed that parameters are locally constant in time and there is less value to older information. Parameter inference is then performed using information which occurred before a certain time, t. This leads to the 'pseudo-likelihood', which is given for each time point t by

$$L_t(\boldsymbol{\alpha}, \boldsymbol{\beta}, \rho, \gamma; \boldsymbol{x}, \boldsymbol{y}) = \prod_{k \in A_t} \{ \tau_{\lambda_k, \mu_k}(x_k, y_k) \exp(-\lambda_k) \lambda_k^{x_k} \exp(-\mu_k) \mu_k^{y_k} \}^{\phi(t - t_k)}, \tag{4.10}$$

where t_k is the time of match k and $A_t = \{k : t_k < t\}$ and ϕ is a positive non-increasing function of time.

The use of such ϕ allows the model a greater capacity to reflect current team performance, with smaller ϕ values giving less weight. However, this induces the issue of how much historical data are to be taken into account, i.e. how much downweighting should be applied. Dixon and Coles (1997) chose to use the function

$$\phi(t - t_k) = \exp\left[-\xi\left(t - t_k\right)\right], \tag{4.11}$$

where $\xi > 0$.

A value of $\xi = 0.0065$, with time units of half week, was chosen which optimised the prediction of match outcomes, although it was found that parameter estimates were robust when using a range of values for ξ. Due to the nature of the analysis being an investigation of home advantage, prior to any predictory analysis for betting strategies, it is prudent to use all of the available data not just data before time point t. This may be done by allowing parameters to remain static within seasons, i.e. α_i, β_i and γ change with seasons, s, for all i, and

$$\begin{aligned} \lambda_k &= \alpha_{i(k,s)}\beta_{j(k,s)}\gamma_{s(k)}, \\ \mu_k &= \alpha_{j(k,s)}\beta_{i(k,s)}, \end{aligned} \tag{4.12}$$

or by employing down weighting on matches that occured after t, as well as before t, where

$$\phi(t - t_k) = \begin{cases} \exp\left[-\xi\left(t - t_k\right)\right] & \text{if } t \leq t_k, \\ \exp\left[-\xi\left(t_k - t\right)\right] & \text{if } t > t_k. \end{cases} \tag{4.13}$$

To obtain a value of ξ, alternatively to optimising the prediction of match outcomes, cross validation was performed using data from the Premier League between seasons 2001/2002 and 2011/2012, where each team has the same identifier, i, over all seasons (Arlot and Celisse, 2009). A training set was created by removing every other 5 matches, without replacement, from the sample. The log-likelihood given by

$$\begin{aligned} \ell\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \rho, \gamma; \boldsymbol{x}, \boldsymbol{y}\right) = \sum_{k=1}^{N} \phi\left(t - t_k\right) \{ &\log\left[\tau_{\lambda_k,\mu_k}\left(x_k, y_k\right)\right] + x_k \log\left(\lambda_k\right) \\ &-\lambda_k - \log\left(x_k!\right) + y_k \log\left(\mu_k\right) - \mu_k - \log\left(y_k!\right)\}, \end{aligned} \tag{4.14}$$

was then maximised using the down-weighting function shown in equation (4.13), over a grid of values for ξ. The value of the log-likelihood was then evaluated for the validation

set using the MLEs, at each value of ξ as shown in Figure 4.1 (left). The value of ξ which maximises this curve is $\xi = 0.0041$.

As an alternative, a transformed normal pdf down-weighting function, as shown by

$$\phi(t - t_k) = \exp\left[-\frac{(t_k - t)^2}{2\sigma^2}\right], \tag{4.15}$$

was tested in the same way as above, where σ controls the extent of the down-weighting. Figure 4.1 (right) shows the log-likelihood evaluated on the validation set using the MLEs over a grid of values for σ, resulting in a value of σ which maximises the curve of $\sigma = 147$.

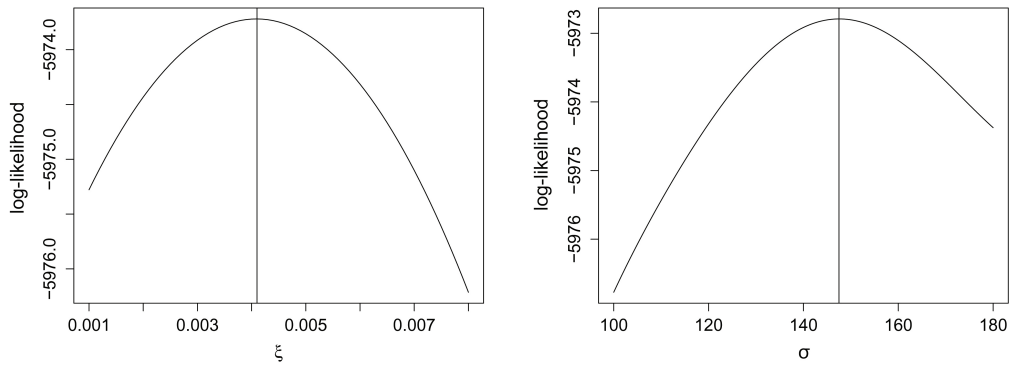


Figure 4.1: Log-likelihood evaluated using the validation set and parameter estimates for the training set for (left) the exponential weighting function shown in equation (4.13) and (right) the transformed normal pdf weighting function shown in equation (4.15), over a grid of values for ξ and σ respectively.

Finally, cross validation was performed using the down-weighting function given in equation (4.13). However, each team was given a different identifier for each season, $i(s)$, treating them as independent teams and home advantage was also treated seasonally, γ_s. For this analysis, the time scale was changed from half weeks, to every match, to give better fidelity. This was not done previously, as the computational time would have been too great. The value of ξ^* which maximises the log-likelihood is found to be $\xi^* = 0.005$, as shown in Figure 4.2. This value cannot be compared directly with the above value unless it is scaled, as t is incremented by matches and not half weeks, and so is denoted by $*$.

4.2 Home Advantage Over Time

Using a data set of final tables from the the English Division 1, now the Premier League, between seasons 1900/1901 and 2013/2014, the home advantage was estimated using $\hat{\gamma}$ of

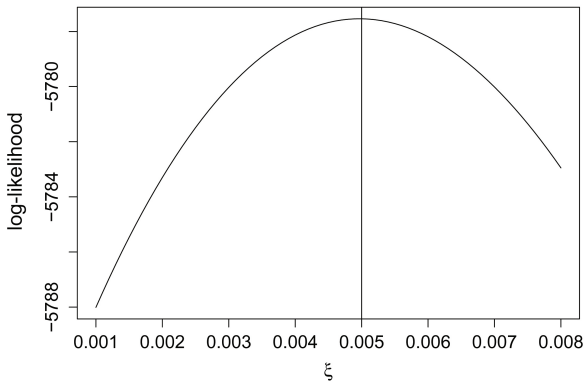


Figure 4.2: Seasonal log-likelihood evaluated using the validation set and parameter estimates for the training set for the exponential weighting function shown in equation (4.13) and over a grid of values for ξ.

equation (4.9) with estimates shown over seasons in Figure 4.3 (left). It can be seen that the home advantage decreases over time. As can be seen in Figure 4.4, the average total number of home and away goals scored per team varies over time due to rule changes and changes in attacking and defensive technique. For example, between 1920 and 1970 considerably more goals were scored, home and away, than during the period between 1970 and 2005. Figure 4.3 (right) shows a ratio of the home advantage and total goals scored per season, which takes into account changes in the number of teams in the league and any rule changes (e.g. 3 points for a win) which led to changes in the number of goals scored. Under this measure, the decrease in home advantage over time appears less significant.

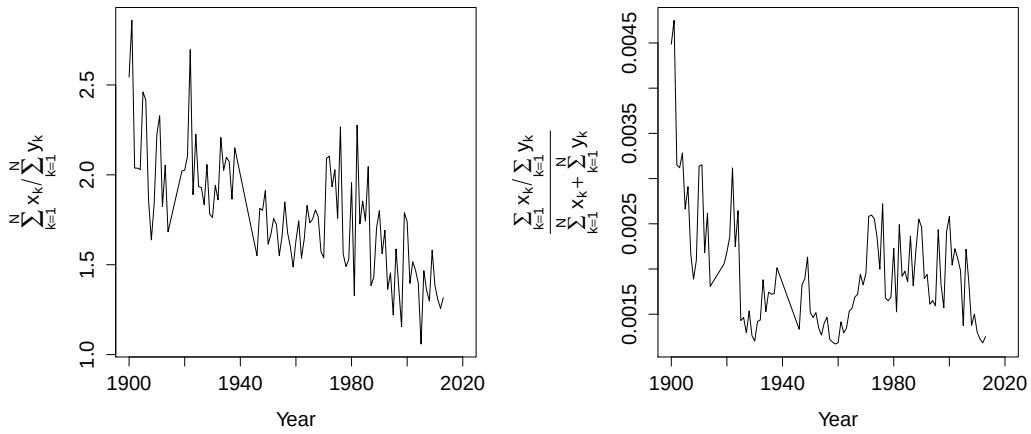


Figure 4.3: (Left) Seasonal home advantage and (right) seasonal home advantage over seasonal total goals scored for English Division 1 between seasons 1900/1901 and 2013/2014.

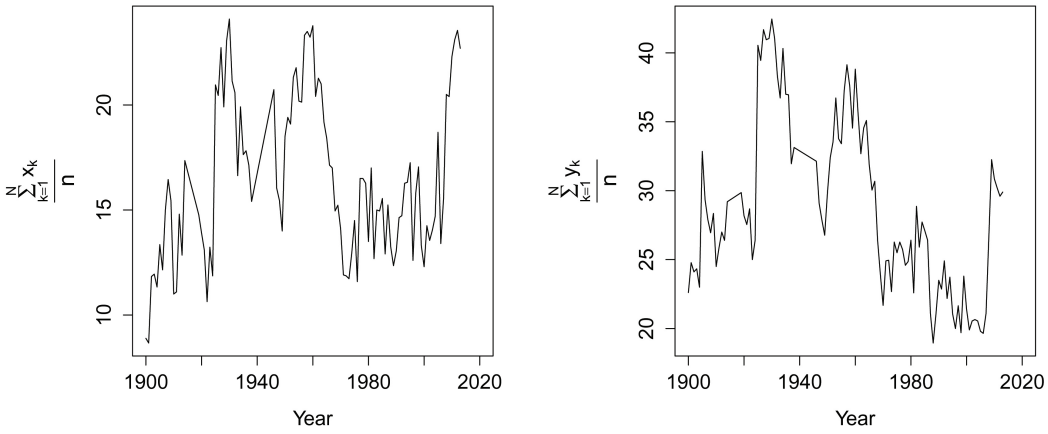


Figure 4.4: (Left) Average total home goals per team and (right) average total away goals per team each season for English Division 1 between seasons 1900/1901 and 2013/2014.

Using the log-likelihood given by equation (4.7) and data from the Premier League between 1995/1996 and 2013/2014, a hypothesis test was carried out between an alternative hypothesis that α, β and γ change with season, s, as given in equation (4.12), and a null hypothesis that only α and β change over time, whilst γ remains constant over seasons, i.e. $\lambda_k = \alpha_{i(k,s)}\beta_{j(k,s)}\gamma$. The deviance value obtained was 16.06, which proves insignificant in a chi-squared test with 18 degrees of freedom at a 0.05 significance level with a p-value for the alternative hypothesis of 0.65, therefore the null is not rejected. This provides some evidence that change in home advantage over time may not need to be taken into account in the modelling process for this recent period of data, despite the evidence of change from earlier seasons. Clearly, time is not a causal factor. It is however, the measure over which rule changes (such as three points for a win) and new technology and training methods enter the sport. This may lead to more significant results when analysing how the relationship of home advantage with various covariates changes over time.

Cross validation was also carried out under the above null and alternative hypotheses, to analyse each model's predictive power. Every other match in the data set was removed (starting with the second match in the 1995/1996 season) without replacement and the log-likelihood given by equation (4.7) was maximised to give MLEs which were used to calculate the root mean squared error (RMSE) for home and away goal estimates as given for individual parameters (represented by θ) by

$$\text{RMSE}(\theta) = \sqrt{\frac{\sum_{k \in C} 2\left(\hat{\theta}_k - \theta_k\right)^2}{N}}, \tag{4.16}$$

where C represents the validation set, which has length $N/2$.

This was repeated removing every other match in the data set starting with the first match in the 1995/1996 season, and the average of the two RMSEs were calculated for each model. Under the null model of constant home advantage, the average RMSE for home goal estimates ($\theta = \lambda$) was 1.300, whilst that for away goal estimates ($\theta = \mu$) was 1.110. Comparing this to that for the alternative model, that home advantage also changes over seasons, the average RMSE for home goal estimates increased to 1.301 and that for away goal estimates also increased to 1.111, which indicates a very slight loss in predictive power when including varying home advantage over seasons.

The reduction in predictive power may be due to the removal of large amounts of data, leaving small samples from which to estimate seasonal home advantage. Leave-one-out cross validation is an alternative method which removes only a single observation from the data, then maximises the log-likelihood for the remaining observations. This process is then repeated for each observation and the RMSE calculated using MLEs for each point when that observation is removed. Due to its exhaustive approach, using the entire data set would be computationally intensive. Therefore, a subset of seasons 2009/2010 to 2011/2012 was used in this analysis. Under the null model of constant home advantage, the RMSE for home goal estimates was 1.274, whilst that for away goal estimates was 1.1012. The RMSE for the alternative model of seasonally varying home advantage decreased slightly for home goal estimates by 3×10^{-4}, however, that for away goal estimates increased to 1.106, showing an overall decrease in predictive power, and agreeing with the intitial analysis.

To look in more detail at any trend in home advantage over time, $\hat{\gamma}$ can be evaluated using the down-weighting function given by equation (4.13), applied either over all seasons, with attack and defence parameters indexed by i and home advantage γ, or within seasons with attack and defence indexed by $i(s)$ and home advantage γ_s. For computational reasons, the data were reduced to seasons between 2001/2002 and 2011/2012. Under the non-seasonal parameter model and using time increments of half weeks, the log-likelihood was maxmised at each time step using values of $\xi = 0.004$ and 0.0065, as found in Section 4.1.3 and by Dixon and Coles (1997) respectively. The resulting values of $\hat{\gamma}$ are shown in Figure 4.5, alongside those estimated using a seasonal parameter model as discussed in Section 4.1.3, with time increments of each match and a value of $\xi^* = 0.005$.

It can be seen from Figure 4.5, that the estimates for γ follow a slightly positive trend over the eleven seasons of data analysed. However, when treated seasonally, the home advantage appears to increase over most seasons, with some large step changes between

seasons. This trend can be explored using a log-linear model such as

$$\log(\gamma_{s,t}) = \phi_s + \eta t, \tag{4.17}$$

where $\eta = 0$ represents constant home advantage with a value of ϕ_s within seasons. For this model, t is incremented by 5 matches. It should be noted at this juncture that confidence intervals were not used in Figure 4.5 to allow a visual comparison of estimates by overlaying them. However, comparing against them would aid the interpretation of the different estimates. Confidence intervals may be derived from the associated estimate and standard error (SE), where the SE is derived from the Hessian at the MLE. Then 95% confidence intervals are constructed by adding or subtracting $1.96 \times$ SE from the estimate.

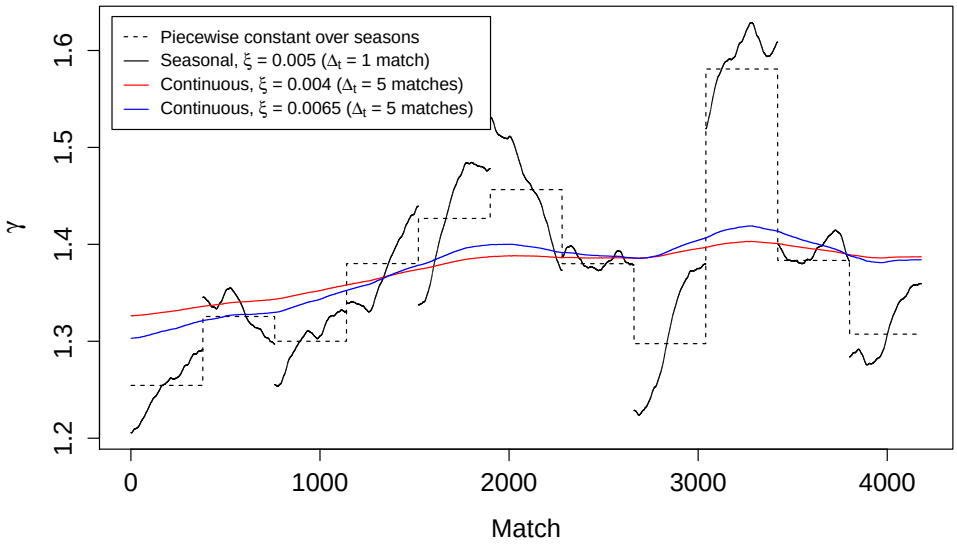


Figure 4.5: Values of $\hat{\gamma}$ evaluated at each time step (half week or per match) for the non-seasonal and seasonal team definitions. The dotted line shows the per season estimate of home advantage evaluated using the closed form expression. Note that confidence intervals were not plotted to allow visual comparison of estimates. However, 95% confidence intervals can be constructed by taking the associated estimate and adding or subtracting $1.96 \times$ SE. Comparing against these intervals would aid the interpretation of the different estimates.

A hypothesis test was performed on data from the Premier League between seasons 1995/1996 and 2013/2014 with a null hypothesis of constant home advantage within seasons, and an alternative hypothesis that there is a trend in home advantage within seasons. In a chi-squared test with 1 degree of freedom, the null was not rejected at a 0.05 significance level, with a deviance of 3.26 and a p-value of 0.071. This result tests significant at a level of 0.1, so it could be hypothesised that the within season trend of home advantage changes over seasons, which can be written as

$$\log(\gamma_{s,t}) = \phi_s + \eta_s t, \tag{4.18}$$

where t is incremented by each match. A second hypothesis test was performed with a null hypothesis of no change in linear trend describing home advantage within seasons, for all seasons, and an alternative hypothesis of varying trends in home advantage within seasons. In a chi-squared test with 19 degrees of freedom, the null was not rejected at a 0.05 significance level, with a deviance of 9.59 and a p-value of 0.962.

4.3 Team Dependent Home Advantage

As discussed in the previous section, it is prudent to assume that home advantage may not be a constant over time. However, it may also be necessary to consider the effect of team dependent home advantage. Dixon and Coles (1997) did not implement this in their model, as they took γ as constant over teams. The likelihood for this model is as given by equation (4.5). However

$$\lambda_k = \alpha_{i(k)}\beta_{j(k)}\gamma_{i(k)}, \tag{4.19}$$

where γ_i represents team specific home advantage and the likelihood is instead maximised over a vector of parameters describing home advantage, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$ as well as $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

4.3.1 Simulation Study

To illustrate the effect this may have, a simulation study was performed assuming team dependent home advantages. For this simulation it was assumed that the outcome of the home and away teams are independent, as in the model by Maher (1982), i.e. $\rho = 0$ in the independence function given by equation (4.4).

Individual team parameters for attack, defence and home advantage were assigned to a twenty team balanced league as shown in Tables 4.1 and 4.2. Note the value of $\alpha_1 = 1$, this is to ensure that the model is not over parameterised. Outcomes for a full season, in which every team played every other team once home and once away were simulated, and the process was repeated 1000 times, with no seasonal change in parameter values.

These simulated match results could then be used in an attempt to estimate the individual attack and defence parameters, α_i and β_i and either a constant home advantage over the league, as assumed by Dixon and Coles (1997), or a team dependent home advantage, using the definitions for λ_k and μ_k as given by equations (4.6) and (4.19)

respecitively. The log-likelihoods are then given by equation (4.7) and

$$\ell\left(\alpha_i, \beta_i, \rho, \gamma_i; i = 1, \ldots, n\right) = \sum_{k=1}^{N} \log\left[\tau_{\lambda_k,\mu_k}\left(x_k, y_k\right)\right] + x_k \log\left(\lambda_k\right) - \lambda_k - \log\left(x_k!\right) + y_k \log\left(\mu_k\right) - \mu_k - \log\left(y_k!\right), \quad (4.20)$$

respectively.

The log-likelihoods were maximised for both models over each of the 1000 seasons' final results. The average MLEs for α_i and β_i are shown in Table 4.1 when testing with both constant home advantage and team dependent home advantage models. Table 4.2 shows the MLEs for γ_i when testing with a team dependent home advantage model. These tables also show the RMSE and the absolute difference as given for a parameter θ_i by

$$\text{Abs. Diff.} = |\overline{\hat{\theta}_i} - \theta_i|,$$

where

$$\overline{\hat{\theta}_i} = \frac{1}{1000} \sum_{z=1}^{1000} \hat{\theta}_{i,z}$$

and z represents the simulation number.

When testing with a league constant home advantage model, the average MLE for home advantage was $\bar{\hat{\gamma}} = 1.32$, which is close to the average of the true individual team home advantages, $\sum_{i=1}^{n} \gamma_i / n = 1.3$. However, the other parameters seem to have suffered under this model with greater inaccuracies in many parameter estimates, the greatest of which being the maximum absolute difference between the true value and the MLE of 0.73, compared to 0.08 for the team dependent home advantage model.

Although much further investigation is needed, it may be concluded that ignoring team or match dependent home advantage could lead to inaccuracies in the estimates of model parameters. Including parameterisations such as team dependent home advantage greatly increases the dimensionality of the model, though, modern methods of numerical optimisation could handle the additional parameters and the information gained could potentially justify their use.

Parameter	True Value	League Home Advantage			Team Home Advantage		
		Mean MLE	Abs. Diff.	RMSE	Mean MLE	Abs. Diff.	RMSE
α_2	1.3	1.45	0.15	0.34	1.38	0.08	0.45
α_3	0.7	0.65	0.05	0.16	0.74	0.04	0.28
α_4	0.9	0.96	0.06	0.22	0.95	0.05	0.35
α_5	0.8	0.90	0.10	0.22	0.85	0.05	0.31
α_6	1.1	1.29	0.19	0.33	1.17	0.07	0.38
α_7	1.8	1.93	0.13	0.39	1.90	0.10	0.61
α_8	0.6	0.67	0.07	0.17	0.63	0.03	0.24
α_9	1.4	1.49	0.09	0.31	1.47	0.07	0.48
α_{10}	1.3	1.58	0.28	0.43	1.36	0.06	0.43
α_{11}	1.6	2.33	0.73	0.86	1.68	0.08	0.54
α_{12}	1.1	1.39	0.29	0.41	1.14	0.04	0.39
α_{13}	0.8	0.93	0.13	0.25	0.84	0.04	0.31
α_{14}	1.2	1.11	0.09	0.26	1.27	0.07	0.43
α_{15}	0.9	1.18	0.28	0.38	0.95	0.05	0.34
α_{16}	1.3	1.33	0.03	0.28	1.38	0.08	0.46
α_{17}	1.0	0.97	0.03	0.22	1.06	0.06	0.38
α_{18}	0.5	0.56	0.06	0.16	0.53	0.03	0.23
α_{19}	0.7	0.82	0.12	0.23	0.73	0.03	0.28
α_{20}	1.2	1.29	0.09	0.29	1.27	0.07	0.42
β_1	0.8	0.73	0.07	0.18	0.80	0.00	0.22
β_2	1.1	1.00	0.10	0.23	1.10	0.00	0.29
β_3	1.3	1.19	0.11	0.26	1.31	0.01	0.34
β_4	0.4	0.36	0.04	0.11	0.40	0.00	0.13
β_5	0.8	0.73	0.07	0.18	0.80	0.00	0.23
β_6	0.7	0.64	0.06	0.16	0.71	0.01	0.20
β_7	1.2	1.09	0.11	0.25	1.20	0.00	0.31
β_8	1.1	1.00	0.10	0.24	1.10	0.00	0.30
β_9	1.0	0.91	0.09	0.21	1.00	0.00	0.27
β_{10}	1.3	1.18	0.12	0.26	1.30	0.00	0.34
β_{11}	0.7	0.63	0.07	0.16	0.70	0.00	0.20
β_{12}	0.9	0.82	0.08	0.20	0.90	0.00	0.24
β_{13}	0.8	0.73	0.07	0.18	0.80	0.00	0.22
β_{14}	1.1	1.01	0.09	0.23	1.11	0.01	0.29
β_{15}	1.8	1.63	0.17	0.35	1.79	0.01	0.45
β_{16}	0.6	0.55	0.05	0.14	0.60	0.00	0.17
β_{17}	1.4	1.27	0.13	0.29	1.40	0.00	0.37
β_{18}	1.3	1.19	0.11	0.27	1.31	0.01	0.34
β_{19}	1.6	1.46	0.14	0.31	1.61	0.01	0.40
β_{20}	1.1	1.00	0.10	0.23	1.09	0.01	0.29

Table 4.1: Simulation results (mean MLEs, absolute difference and RMSEs for attack, α, and defence, β, parameters) using individual team home advantage, γ_i.

Parameter	True Value	Mean MLE	Abs. Diff.	RMSE
γ_1	1.1	1.17	0.07	0.39
γ_2	1.3	1.34	0.04	0.35
γ_3	0.9	0.97	0.07	0.42
γ_4	1.2	1.28	0.08	0.44
γ_5	1.3	1.39	0.09	0.51
γ_6	1.4	1.45	0.05	0.44
γ_7	1.2	1.24	0.04	0.30
γ_8	1.3	1.44	0.14	0.68
γ_9	1.2	1.24	0.04	0.35
γ_{10}	1.5	1.59	0.09	0.45
γ_{11}	2.0	2.09	0.09	0.48
γ_{12}	1.6	1.71	0.11	0.49
γ_{13}	1.4	1.50	0.10	0.55
γ_{14}	0.9	0.94	0.04	0.29
γ_{15}	1.7	1.79	0.09	0.60
γ_{16}	1.1	1.14	0.04	0.32
γ_{17}	1.0	1.05	0.05	0.38
γ_{18}	1.3	1.49	0.19	0.78
γ_{19}	1.4	1.55	0.15	0.72
γ_{20}	1.2	1.25	0.05	0.37

Table 4.2: Simulation results (mean MLEs, absolute difference and RMSEs for home advantage, γ) using individual team home advantage, γ_i.

4.3.2 Application to Premiership Data

The log-likelihood given by equation (4.20) was maximised over data from the Premier League between seasons 1995/1996 and 2013/2014 for three models, firstly considering that home advantage varies over teams, secondly that home advantage varies over teams and seasons, i.e. $\gamma_{i,s} = \gamma_i + \phi_s$, where ϕ_s represents a seasonal home advantage, and finally that team dependent home advantage varies over seasons, which can be written $\gamma_{i(s)}$ as for the attack and defence parameters, which are allowed to vary between seasons in all of these models.

A hypothesis test was performed between a null hypothesis that home advantage is constant over teams and seasons, $H_0 : \gamma_i = \gamma$, and an alternative hypothesis that the home advantage varies over teams, $H_1 : \gamma_i \neq \gamma$. In a chi-squared test with 43 degrees of freedom (the number of teams observed in the data minus one), the null is rejected at a 0.05 significance level, with a deviance of 64.53 and a p-value of 0.018.

The log-likelihood of the model and the number of free parameters, p, can be used to calculate the Akaike Information Criterion (AIC), which can be used as a measure of

statistical quality of a model and is given by

$$\text{AIC} = 2p - 2\log(L).$$

The model with the minimum AIC value is preferred. The AIC value is 41971.42 under the null model and 41992.89 for the alternative model, which opposes the p-value. Table B.5 in Appendix B shows the value of $\hat{\gamma}_i$ for all the teams that were present in the Premier League between seasons 1995/1996 and 2013/2014.

Secondly, a hypothesis test was performed between a null hypothesis that home advantage varies over teams, $\text{H}_0 : \gamma_{i,s} = \gamma_i$, and an alternative hypothesis that home advantage varies over teams and seasons, $\text{H}_1 : \gamma_{i,s} \neq \gamma_i$. In a chi-squared test with 19 degrees of freedom (the number of seasons observed in the data), the null is not rejected at a 0.05 significance level, with a deviance of 18.32 and a p-value of 0.501. The AIC supports this result with a value of 42012.56 for a model following the alternative hypothesis.

Finally, a hypothesis test was carried out between a null hypothesis that the home advantage varies over teams, $\text{H}_0 : \gamma_{i(s)} = \gamma_i$, and an alternative hypothesis that team dependent home advantage varies over seasons, $\text{H}_1 : \gamma_{i(s)} \neq \gamma_i$. In a chi-squared test with 336 degrees of freedom (the number of teams in a complete season times the number of seasons minus the number of teams obsevered in the data) the null is not rejected at a 0.05 significance level, with a deviance of 371.16 and a p-value of 0.091. The AIC supports this finding, with a value of 42293.71 for a model following the alternative hypothesis. Table B.6 in Appendix B shows the individual home advantage MLEs, $\hat{\gamma}_i(s)$, for each team, over each season, obtained from this maximisation. The standard errors for each team over the included seasons are relatively high. Performing a chi-squared hypothesis test, considering a null hypothesis of constant home advantage over all teams and seasons, $H_0 : \gamma_{i(s)} = \gamma$, against an alternative hypothesis of team dependent home advantage, varying over seasons, $\text{H}_0 : \gamma_{i(s)} \neq \gamma$, the null is rejected at a significance level of 0.05, with a p-value for the alternative hypothesis of 0.023.

The RMSE of the home and away goal counts achieved in the alternative and null models may be compared as a percentage improvement (or reduction), $v(\theta)$, as given by

$$v(\theta) = \frac{\text{Null RMSE}(\theta) - \text{Alternative RMSE}(\theta)}{\text{Null RMSE}(\theta)} \times 100, \tag{4.21}$$

where $\text{RMSE}(\theta)$ is given by equation (4.16) and a value of alternative $\text{RMSE}(\theta) = 0$ gives a value of $v(\theta) = 100$ and represents a 100% reduction in the RMSE of the Null, or zero error. It should be noted that this statistic requires null $\text{RMSE}(\theta) > 0$, which is empircally acceptable given a large number of observations and the nature of random

effects. Evaluating this function for home goal count, x, and away goal count, y, with a null hypothesis of constant home advantage and an alternative hypothesis of seasonally constant team home advantage, results in values of $v(x) = 0.21\%$ and $v(y) = 0.24\%$ respectively, whilst an alternative of seasonally varying team home advantage results in values of $v(x) = 1.48\%$ and $v(y) = 1.98\%$ respectively. These both indicate that the alternative models are fitting the data better than the null. However, this does not mean that they are more accurately estimating the true parameters.

Cross validation was carried out under the constant team home advantage and seasonally varying team home advantage hypotheses, to analyse each model's predictive power. Alternate matches in the data set were removed (starting with the second match in the 1995/1996 season) without replacement and the log-likelihood given by equation (4.7) was maximised to give MLEs which were used to calculate the RMSE for home and away goal estimates. This was repeated removing every other match in the data set starting with the first match in the 1995/1996 season, and the average of the two RMSEs were calculated for each model. Under the model of constant team home advantage, the average RMSE for home goal estimates was 1.313, whilst that for away goal estimates was 1.114. Under the model of seasonally varying team home advantage the average RMSE for home goal estimates was 1.301 and that for away goal estimates was 1.130. These values indicate a loss in predictive power in comparison to the model of constant home advantage over teams and seasons, as given in Section 4.2.

As discussed in Section 4.2, the removal of large amounts of information from the data set may impede the predictive capabilities of the models. Therefore, leave-one-out cross validation shall also be used here (as described in Section 4.2), using the subset of seasons 2009/2010 to 2011/2012. Under the model of constant team home advantage, the RMSE for home goal estimates was 1.280, whilst that for away goal estimates was 1.110. Under the model of seasonally varying team home advantage the RMSE for home goal estimates was 1.289 and that for away goal estimates was 1.120. These results agree with the above findings, as the RMSEs are greater than those values relating to the constant home advantage model.

It was hypothesised that home advantage may be directly linked to the skill of a team, more specifically, their attacking ability, α_i. To investigate this over multiple seasons for the full data set, a correction factor had to be applied to account for the fact that $\alpha_{1(s)} = 1$, normalising the values to each season seperately. Considering that the normalisation method prescribed by Dixon and Coles (1997) of $\sum_{i=1}^{n} \alpha_i/n = 1$ would allow comparable parameters over seasons, the correction factor, c_s, was derived as

$$c_s = \frac{n}{\sum \alpha_{i(s)}},$$

and $\alpha^*_{i(s)} = c_s\alpha_{i(s)}$ and $\beta^*_{i(s)} = \beta_{i(s)}/c_s$, are now comparable over seasons.

No significant relationship between either attack or defence and team dependent home advantage, varying over seasons, $\gamma_{i(s)}$ was found when comparing their estimates. Therefore, it was further hypothesised that $\gamma_{i(s)}$ may instead be dependent on the product of a team's attacking skill and the average defensive skill for the league. This can be represented by $T_{i(s)}$ as given by

$$\begin{aligned} T_{i(s)} &= c_s\hat{\alpha}_{i(s)} \sum_{j=1}^{n} \frac{\hat{\beta}_{j(s)}/c_s}{n} \\ &= \hat{\alpha}_{i(s)} \sum_{j=1}^{n} \frac{\hat{\beta}_{j(s)}}{n}. \end{aligned} \tag{4.22}$$

Figure 4.6 shows the seasonally varying team dependent home advantage MLEs, $\hat{\gamma}_{i(s)}$ with a confidence interval calculated using the Hessian, plotted against $T_{i(s)}$. This can be thought of as the expected goal count that team i would achieve on neutral ground, when playing a team with the average defensive ability of those teams in each season. Visually, Figure 4.6 shows a decreasing trend between $\gamma_{i(s)}$, and the skill of the home team measured through $T_{i(s)}$, which could possibly be used in the development of a model for home advantage.

One possible form for a team and season specific home advantage is

$$\log\left(\gamma_{i(k,s)}\right) = \phi + \eta_0 \exp\left(-\eta_1 T_{i(k,s)}\right), \tag{4.23}$$

for some constants ϕ, η_0 and η_1. This relates $\gamma_{i(s)}$ to $T_{i(s)}$ in a match k using an exponential curve model.

A hypothesis test was carried out with a null hypothesis of a constant home advantage over seasons, against an alternative hypothesis that home advantage follows the relationship with $\hat{\alpha}_{i(k,s)} \sum_{j=1}^{n} \frac{\hat{\beta}_{j(k,s)}}{n}$ given in equation (4.23). There was no improvement in the log-likelihood value between the null and alternative hypotheses. The fitted relationship under the alternative hypothesis is shown in Figure 4.6 as a blue dotted line, which appears flat, but has a slight negative trend. This would be interpreted as a more skillful team in attack and defence experiences a lower home advantage.

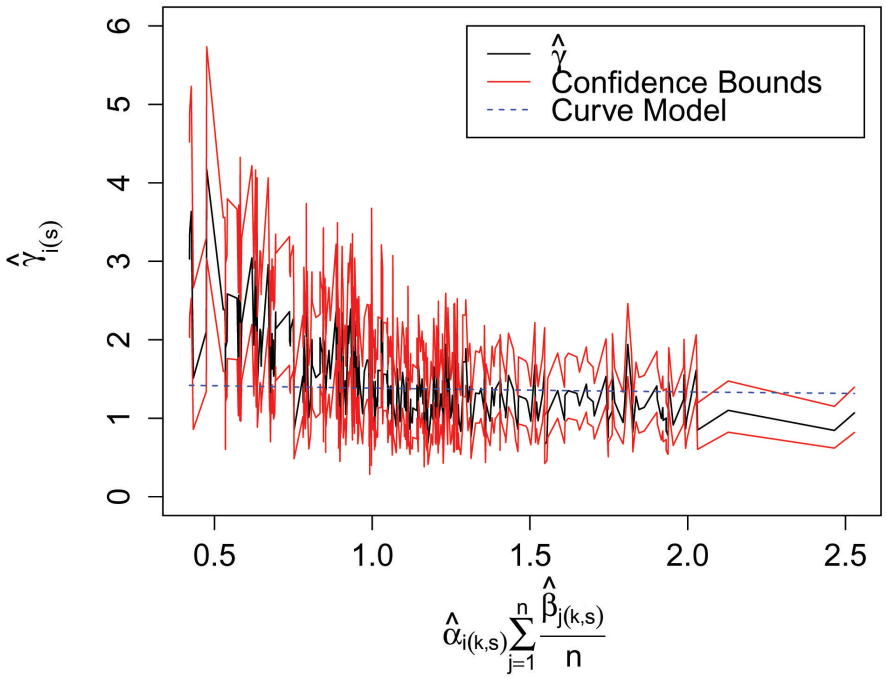


Figure 4.6: Individual team home advantage as a function of $\hat{\alpha}_{i(k,s)} \sum_{j=1}^{n} \frac{\hat{\beta}_{j(k,s)}}{n}$, using data from the Premiership between seasons 1995/1996 and 2013/2014.

4.4 The Effect of Cards on Goals

Referee bias is often seen as a leading cause of home advantage, due to pressures arising from crowd dynamics and other factors. Referees are susceptible to the same psychological influences as players when on the pitch, and could provide a conduit of crowd influence to the game (Boyko et al., 2007). In the Premiership between seasons 2001/2002 to 2011/2012, away teams received 1.49 times more red cards than home teams and 1.33 times more yellow cards, suggesting a very real home advantage in terms of cards given.

The effect of yellow cards on the actual gameplay is small in comparison to red cards, as the latter changes the number of players on a team. The difference between the number of players on each team at the end of the match, $R^h_{i,j}$, can be calculated by the number of red cards awarded to away team j minus those awarded to home team i, where a value of 1 indicates one additional player to the home team i.

The difference in cards can be employed in a log-linear model of home advantage as given by

$$\log(\gamma_{i,j}) = \phi + \eta R^h_{i,j},$$

where η describes a log-linear trend in γ with R^h. A hypothesis test was carried out using the goal count model given by equation (4.5), with a null model of constant home advantage, i.e. $\eta = 0$, and an alternative model where red cards are thought to be a covariate of home advantage as modelled by a log-linear relationship of γ and R^h, i.e. $\eta \neq 0$. Maximising the log-likelihood given by equation (4.7) under both the null and alternative models resulted in a p-value of 0.000 when comparing to a chi-squared distribution with one degree of freedom, which proves to be highly significant and in turn the null can be rejected. The MLEs take the value $\hat{\phi} = 0.30$ and $\hat{\eta} = 0.24$, which suggests a positive trend between home advantage and R^h, i.e. more players on the home team results in a higher home advantage.

It can be hypothesised that the away team would also benefit from an advantage if they had more players than the home teams. Then, μ_k can be written

$$\mu_k = \alpha_{j(k,s)}\beta_{i(k,s)} \exp\left(1 - \nu R^h_{i,j}\right),$$

where ν allows the consideration of some away effect relative to $-R^h$. A hypothesis test was then carried out between a null of home advantage with log-linear dependence on R^h, i.e. $\eta \neq 0$ and $\nu = 0$, and an alternative of both home and away effect equally dependent on R^h and $-R^h$ respectively, i.e. $\eta = \nu \neq 0$. As there are no additional free parameters, a positive value of deviance can be used to acertain the better model. The log-likelihood given by equation (4.7) was maximised under the two models, giving a deviance of 52.20, which indicates that the null can be rejected. Under this model, $\hat{\phi} = 0.30$ and $\hat{\eta} = \hat{\nu} = 0.26$. This model accounts for both home and away card effects; comparing $\exp(\hat{\phi}) = 1.35$ to the MLE for constant home advantage $\hat{\gamma} = 1.37$, the reduction in home advantage when considering games where equal amounts of players finish on each team is small (i.e. $R^h = 0$), suggesting that there are other important factors to consider.

4.5 Conclusion

Estimates of home advantage can be seen to reduce over time for a historical data set taken from the English Premier League. This reduction can be partially explained by the evolution of strategy to a lower average goal count per match. Rule changes such as the migration from two points for a win to three points may explain this change. However, home advantage estimates appear to be increasing in recent years, although no statistically significant change in home advantage with time was found for the period 2001 to 2012, suggesting that playing styles and any other driving factors have somewhat stabilised in recent years.

Home advantage was found to vary significantly over teams. However, performing a form of exhaustive cross validation under this model showed that the predictive capabilities were reduced from the null model which considers a constant home advantage over teams. This suggests that the model could be either overfitting the data or that there is too little data to accurately estimate these parameters. Modelling the heterogeneity of the home advantage parameters on some predictor variables would serve to mitigate this issue, but these are currently unknown and will require extensive hypothesis testing. This topic is covered in the next chapter.

It has often been hypothesised that home advantage results from referee bias, which could in turn be linked to other variables such as crowd dynamics. It was found that home and away team advantages are both heavily dependent on red cards, or rather the reduction in players and effects on team and crowd dynamics. However, modelling the evolution of cards in a match is another issue that is beyond the scope of this analysis. Titman et al. (2015) cover the joint modelling of goals and bookings in association football, using a Markov counting process to accrue goals and bookings.

It has been discussed that any relationship of home advantage with time is not a causal factor, but a measure over which changes happen to other possible covariates. Also, team dependent home advantage and cards could be thought of as simple model extensions. However, less obvious external covariates such as crowd dynamics and distance between teams may also be affecting the outcome. This will be discussed in the following chapter.

4.6 Future Work

Extending this study to a larger data set, with more leagues and seasons, would allow for a better insight into whether the provision of team dependent home advantage in the model can be used to better predict goal counts.

The assessment of different team sports would aid the conclusions of this chapter. Although different sports play to different rule sets and can be subcategorised in terms of their winning aims, they have many common factors, such as players and referees, travel and spectators.

Application of some joint models for home advantage and bookings may allow for considerably better predictions of goal counts. In addition, red and yellow card count estimates can be used in betting strategies to further profit from individual matches. Testing against the model defined by Titman et al. (2015) would allow an assessment of the accuracy of model estimates against current methods.

Chapter 5

Covariate Modelling of Home Advantage

5.1 Introduction

Extensive discussion as to what causes home advantage in team sports has been documented in both academic journals and the media. Suggested causes include but are not limited to: distance travelled, crowd size, stadium size, referee bias and pitch dimensions (Pollard, 1986). However, little statistical evidence has been produced in the public domain to support any of these arguments.

This chapter aims to investigate a number of possible causes, focusing primarily on distance travelled between grounds and crowd attendance, for association football. Various exploratory methods will be employed to examine the effects of the possible driving factors of home advantage, specifically changepoint methods and piecewise regression, which allow us to build a picture of any relationships to allow ease of fitting a more sensible regression model.

Non-parametric regression will also be explored, whereby the predictor derives its form according to information obtained from the data. Unlike parametric regression the functional relationships between response and explanatory variables can adapt better to non-linear relationships, allowing them to capture more features present in the data. It is also prudent to use non-parametric regression when the nature of the relationship is unknown, whilst recognising a potential reduction in power of the inference.

Various methods and models will be used to analyse the extent to which the possible driving factors relate to home advantage. Details of these will be given, including a description of how these models will be applied to the specific conditions of the analysis.

These models will follow a basis of the Dixon and Coles (1997) model, which is described

in Section 4.1. Equations (4.5) and (4.7) describe the likelihood and log-likelihood of this model respectively. The following analysis will use the same parameters and notation as given in Section 4.1 unless otherwise stated. Attack and defence parameters, $\alpha_{i(s)}$ and $\beta_{i(s)}$, will be treated as seasonally varying. The derivation of the closed form expression for home advantage, γ, under the basic model is outlined in Section 4.1.2. When maximising over parameters, γ in equation (4.7) can be replaced by the parameters describing the relation of a certain covariate to γ.

5.2 Covariate and Data Selection

The covariates under analysis, for a match pairing of home team i and away team j, are: distance (km), $d_{i,j}$, attendance (total home and away supporters at the match), $A_{i,j}$, referee experience (years), $R_{i,j}$, pitch length (m), $P^L_{i,j}$, and pitch width (m), $P^W_{i,j}$. It was noted during the analysis, that relative transformations of these covariates should be used in some cases as they provided more significant results, whilst remaining conceptually logical. For attendance $A_{i,j}/A_{j,i}$ was employed to represent the ratio between attendances at each of the two grounds where the pairing i, j played. For pitch length and width $P^D_{i,j}/P^D_{j,i}$ was used, where D indicates the dimension, width or length. This represents the ratio of each dimension of the pitch size between the two grounds where the pairing i, j played.

Data were provided by the industrial partner, ATASS Sports, for match results over many leagues worldwide and seasons between 2001 and 2012 (See Appendix C). Supplementary data from England's Premier League (2001/2002 to 2011/2012), France's Ligue 1 (2003/2004 to 2011/2012) and Italy's Serie A (2004/2005 to 2011/2012) were selected to provide a range of distances, over a suitably large number of seasons. The following sections will use these leagues to compare and contrast results ascertaining to the nature of any relationship between distance between teams and home advantage.

Further data were obtained from thefootballarchives.com (2014) regarding the combined home and away attendance, match referee, and home and away goal counts for the four highest English divisions. Data for Division 1 (currently the Premier League) were obtained between seasons 1995/1996 and 2013/2014, whilst that for the three lower leagues were obtained between seasons starting in 2004/2005 and 2013/2014. The use of four leagues within one specific country is more likely to ensure similar referee training, similar distances between teams and similar psychological attitudes of players. A separate breakdown of the crowd at a match into home and away team supporter attendances would be beneficial to this analysis, however, they were not readily available in the public domain. It is noteworthy that the level of attendance may imply a large team is within a higher wealth bracket, which will be considered in any conclusions.

Referee experience was determined by the length of time since they started professional refereeing as shown by internet records. A consensus was formed using various sites including: wikepedia.com, soccerway.com, soccerbase.com, premierleague.com and football-lineups.com. In a similar manner, pitch length and width were gathered for Premier League grounds, ascertaining to seasons 2001/2002 to 2011/2012.

The resulting relationships for the Premier League will be presented in the body of the text in each section, due to its status and presence in all analyses, whilst other leagues will be presented in appendices to allow a concise analysis.

5.3 Non-Linear Exploratory Methods

Linear models, generalised linear models (GLMs) and nonlinear models are all examples of parametric regression, as the function describing the relation of the response to the explanatory variable is known (Freund et al., 2006; Lee et al., 2006). In situations such as this, the relationship is not known. In these cases, semi-parametric regression, which employs parametric and non-parametric components, allows the incorporation of unknown, nonlinear relationships into regression analyses by capturing unusual or unexpected features of the data, and does not seek to employ a predetermined relationship between the dependent response and independent explanatory variables (Ruppert et al., 2003).

The following sections will detail the methods used to explore any relationships between home advantage and various hypothesised covariates. These methods include: changepoints, which refers to any abrupt changes in the statistical properties of the data when ordered by a covariate; piecewise constant regression, which can be treated as an extension of changepoint theory; and penalised spline smoothing, which is a method of fitting a smooth curve using a spline function. Piecewise constant regression can also be thought of as a reduction of penalised spline smoothing to the most basic form.

5.3.1 Changepoint Methods and Piecewise Constant Regression

Changepoint models may be used to decide whether a stochastic process is homogeneous or not with respect to some or all of its descriptive statistics (Eckley et al., 2011; Chen and Gupta, 2011; Muller, 1992). Within the context of this application detection models will base their decision on data ordered sequentially according to the covariate in question, distance for example.

Formally, in a discretised case, respective to the measure being used (time in the majority of changepoint applications), and considering a single changepoint $X_1, X_2, \ldots, X_n$ denote the sequence of n independent random variables. The elements $X_1, \ldots, X_\tau$ are identi-

cally distributed according to some density function f_0 and $X_{\tau+1}, \ldots, X_n$ are identically distributed according to some density function f_1. If the location of the changepoint position, τ, is unknown a hypothesis test would then seek to compare a null hypothesis of $f_0 = f_1$ against an alternative of $f_0 \neq f_1$ for each value $1 < \tau < n$ (Eckley et al., 2011).

By discretising the covariate to some resolution such models can be used to test for the existence of a changepoint within home advantage relative to the covariate in question and estimate its position and parameter set. Both Bayesian and non-Bayesian approaches exist (Atherton et al., 2009; Eckley et al., 2011). For this initial analysis we will use a classical maximised likelihood method, due to the nature of the base model for goal count prediction being employed.

Changepoint detection methods may be used as an exploratory tool to assess the relationship of a covariate, $\delta_{i,j}$, to home advantage. This can be done using a hypothesis test, as stated earlier in this section, which allows an assessment of the goodness of fit using the deviance statistic, however, this is a time consuming process. These methods may also allow for the use of the closed form expression for home advantage, which may create a better picture of the relationship by allowing many leagues to be quickly analysed separately or together. This is only possible if pairing i, j is in the same segment as j, i. Consider the single changepoint case, whereby

$$\gamma_{i,j} = \begin{cases} \gamma_1 & if \quad \delta_{i,j} \leq \delta^* \\ \gamma_2 & if \quad \delta_{i,j} > \delta^* \end{cases}, \tag{5.1}$$

where τ is relabelled as δ^* to signify the different measures of various covariates. This may be expanded to a multiple changepoint model describing γ_m, where $m = 1, \ldots, M+1$, which indicates M changepoints, as given by

$$\gamma_{i,j} = \begin{cases} \gamma_1 & if \quad \delta_0^* < \delta_{i,j} \leq \delta_1^* \\ \gamma_2 & if \quad \delta_1^* < \delta_{i,j} \leq \delta_2^* \\ \quad \vdots & \\ \gamma_{M+1} & if \quad \delta_M^* < \delta_{i,j} \leq \delta_{M+1}^* \end{cases}, \tag{5.2}$$

where δ_0^* represents the minimum value and δ_{M+1}^* represents the maximum value of the covariate's range. In the case of changepoint analysis, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{M+1})$.

A special case of this model is when $\delta_{i,j} = \delta_{j,i}, \forall i, j$, for example under the covariate of distance, whereby the closed form expression for the home advantage estimate in each

segment can be calculated, allowing the estimation of $\gamma_{i,j}$ without the use of numerical optimisation. Consider the single changepoint model, as given by equation (5.1), the derivation follows: the first derivative with respect to γ_1 of the log-likelihood shown in equation (4.7), under this model for home advantage, is given by

$$\frac{\partial \log \mathrm{L}}{\partial \gamma_1} = \sum_{i=1}^{n} \sum_{\{j \neq i : \delta_{i,j} \leq \delta^*\}} \left(-\alpha_i \beta_j + \frac{x_{i,j}}{\gamma_1} \right). \tag{5.3}$$

Indexing N pairings of i and j by $k = 1, \ldots, N$ and equating the score given in (5.3) to zero and solving for $\hat{\gamma}_1$ gives

$$\hat{\gamma}_1 = \frac{\sum_{\{k:\delta_{i,j} \leq \delta^*\}} x_{i,j}}{\sum_{\{k:\delta_{i,j} \leq \delta^*\}} \hat{\alpha}_i \hat{\beta}_j}. \tag{5.4}$$

Similarly the MLE $\hat{\gamma}_2$ is given by

$$\hat{\gamma}_2 = \frac{\sum_{\{k:\delta_{i,j} > \delta^*\}} x_{i,j}}{\sum_{\{k:\delta_{i,j} > \delta^*\}} \hat{\alpha}_i \hat{\beta}_j}.$$

The first derivative with respect to α_i of the log-likelihood is given by

$$\frac{\partial \log \mathrm{L}}{\partial \alpha_i} = \sum_{\{j \neq i : \delta_{i,j} \leq \delta^*\}} \left(-\beta_j \gamma_1 + \frac{x_{i,j}}{\alpha_i} - \beta_j + \frac{y_{j,i}}{\alpha_i} \right) + \sum_{\{j \neq i : \delta_{i,j} > \delta^*\}} \left(-\beta_j \gamma_2 + \frac{x_{i,j}}{\alpha_i} - \beta_j + \frac{y_{j,i}}{\alpha_i} \right).$$

Equating this to zero and rearranging gives

$$\sum_{\{j \neq i\}} (x_{i,j} + y_{j,i}) - (\gamma_1 + 1) \sum_{\{j \neq i : \delta_{i,j} \leq \delta^*\}} \hat{\alpha}_i \hat{\beta}_j - (\gamma_2 + 1) \sum_{\{j \neq i : \delta_{i,j} > \delta^*\}} \hat{\alpha}_i \hat{\beta}_j = 0.$$

Summing over $i : \delta_{ij} \leq \delta^*$ and representing the pairings of i and j by k, gives

$$\sum_{\{k:\delta_k \leq \delta^*\}} (x_{i,j} + y_{j,i}) - (\gamma_1 + 1) \sum_{\{k:\delta_k \leq \delta^*\}} \hat{\alpha}_i \hat{\beta}_j = 0.$$

Substituting the expression for $\hat{\gamma}_1$ given in equation (5.4) gives

$$\sum_{\{k:\delta_k \leq \delta^*\}} (x_{i,j} + y_{j,i}) = \sum_{\{k:\delta_k \leq \delta^*\}} (x_{i,j}) + \sum_{\{k:\delta_k \leq \delta^*\}} \hat{\alpha}_i \hat{\beta}_j.$$

Therefore,

$$\sum_{\{k:\delta_k \leq \delta^*\}} y_{j,i} = \sum_{\{k:\delta_k \leq \delta^*\}} \hat{\alpha}_i \hat{\beta}_j$$

and

$$\hat{\gamma}_1 = \frac{\sum_{\{k:\delta_k \leq \delta^*\}} x_{i,j}}{\sum_{\{k:\delta_k \leq \delta^*\}} y_{j,i}},$$

and similarly the closed form expression for $\hat{\gamma_2}$ may be given by

$$\hat{\gamma}_2 = \frac{\sum_{\{k:\delta_{i,j} > \delta^*\}} x_{i,j}}{\sum_{\{k:\delta_{i,j} > \delta^*\}} y_{j,i}}.$$

This derivation can be expanded simply to the multiple changepoint form, where

$$\hat{\gamma}_m = \frac{\sum_{\left\{k:\delta^*_{m-1} < \delta_k \leq \delta^*_m\right\}} x_{i,j}}{\sum_{\left\{k:\delta^*_{m-1} < \delta_k \leq \delta^*_m\right\}} y_{j,i}},$$

where $\hat{\gamma}_m$ is the ratio of the total home and away goal counts that occur in all matches where the home and away team grounds are a distance between δ^*_{m-1} and δ^*_m from each other.

If δ^*_m are determined by equally sized bins of match pairings ordered by the covariate, rather than being included as free parameters in the maximum likelihood estimates, this can be referred to as piecewise constant regression. This method allows a more complete picture of the relationship between the dependent variable and regressors, with respect to the data. However, it should be noted that as the bin size is decreased, by increasing segmentation, the approximation error increases. In the case of $\delta_{i,j} = \delta_{j,i}, \forall i, j$, using a closed form expression for the home advantage will allow a much faster computational time and thereby, the use of many matches from many leagues. In doing so, the approximation error of the parameter estimates will be decreased as overall sample size is made much larger. Even so, without some measure of fit, this method can only be used as an outline for other methods.

The sole covariate in this analysis which follows this prerequisite is distance, as teams do not regularly move grounds and always in a league play once at home and once away against each opposing team. Figure 5.1 shows a piecewise constant regression relating distance to home advantage, at a resolution of 50 bins of equal size, carried out on approximately 104,000 matches over all complete, balanced leagues in the data set provided by ATASS Sports (See Appendix C for full details). It can be seen that there is a positive relationship between increasing distance between teams and home advantage. As distance increases, the rate of increase of home advantage seems to decrease, although care has to be taken when interpreting this as there are few data at large distance values.

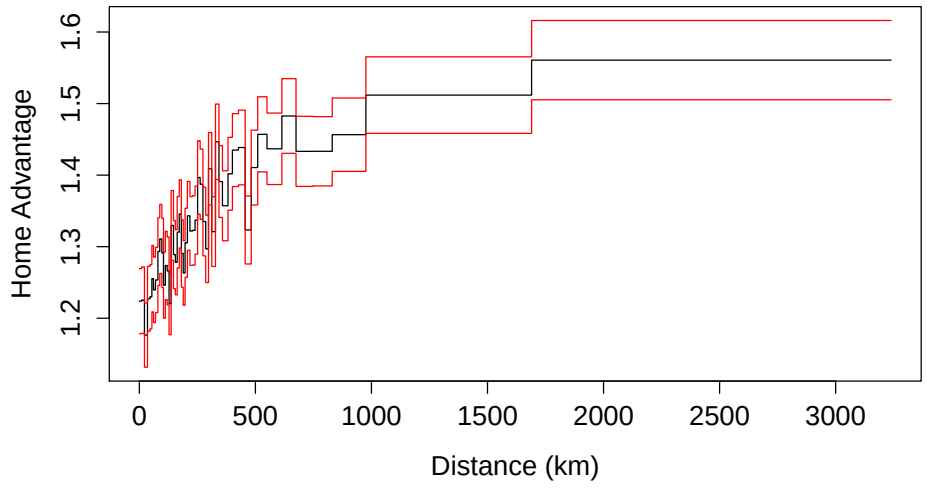


Figure 5.1: Piecewise constant regression performed at a resolution of 50 bins on 104000 matches between 2001/2002 and 2011/2012 seasons from the data set provided by ATASS Sports (see Appendix C for inclusive leagues and time periods) where 95% confidence intervals are shown in red.

To estimate the location of δ^* in a single changepoint model, as given by equation (5.1), the log-likelihood given by equation (4.7) is maximised over the leagues selected for the analysis (see Section 5.2), using incremental values of $\Delta_\delta = \frac{1}{q}\delta_{\max}$, where $\delta_{\max}$ is the maximum value of the covariate for the data set and q controls the number of increments. The value of $q = 100$ was used in the analysis. However, this leads to plateaus in the maximum log-likelihood; although the covariates are continuous, only certain values are observed and they are given to a certain degree of accuracy (for example years of referee experience). In this case, the smallest covariate value which maximises the log-likelihood is considered the changepoint, which follows the above definition.

The results of this analysis for the covariates of distance, relative attendance, referee experience, relative pitch length and relative pitch width, are shown in Tables 5.1 to 5.4 respectively. Hypothesis tests were carried out with a null hypothesis that $\gamma_1 = \gamma_2$ and an alternative hypothesis that $\gamma_1 \neq \gamma_2$ for each of the covariates. The resultant deviance statistics can be used to perform a chi-squared test with two degrees of freedom, considering the location parameter δ^* as an additional free parameter. The relevant p-values are shown along with the AIC in Tables 5.1 to 5.4. It should be noted that in the presence of a nuisance parameter, such as δ^*, which does not exist in the null hypothesis, traditional methods of hypothesis testing, such as a chi-squared test, come into question due to non-regularity (Davies, 1987). In such cases a sampling distribution could be sought numerically. However, changepoint detection methods often use a chi-squared test with 2 degrees of freedom per changepoint (Zhang et al., 2010; Hawkins, 2001).

Under the distance based changepoint model, as can be seen in Table 5.1, the null is rejected at a 0.05 significance level for both the Premier League and Ligue 1. However, this is not the case for Serie A. When all three leagues are combined the null is rejected and the value of δ^* is the same as for Ligue 1. It can be seen in Table 5.2 that the null is rejected under the attendance based changepoint model for both the Premier League and League 2, though it is not rejected for the Championship and League 1. None of the leagues display significant changepoints under the referee experience based changepoint model or those for relative pitch dimensions as can be seen in Tables 5.3 and 5.4.

League	$\hat{\gamma}_1$	$\hat{\gamma}_2$	p-value	AIC	$\hat{\delta}^*$ (km)	95% CI
Premier League (England)	1.33	1.43	0.038	24090.94	193	[164, 273]
Ligue 1 (France)	1.31	1.48	0.004	18777.53	290	[277, 302]
Serie A (Italy)	1.36	1.47	0.167	17258.46	710	NA
Combined	1.35	1.44	0.000	60126.96	290	[126, 302]

Table 5.1: Resultant parameter estimates, changepoint position and test statistics for distance based single changepoint model for home advantage using data from Premier League (England) 2001/2002-2011/2012, Ligue 1 (France) 2003/2004 - 2011/2012 and Serie A (Italy) 2004/2005 - 2011/2012

English League	$\hat{\gamma_1}$	$\hat{\gamma_2}$	p-value	AIC	$\hat{\delta}^*$	95% CI
Premier League	0.35	0.29	0.042	41969.09	1.0	[0.8,0.1]
Championship	1.15	0.26	0.239	31543.27	0.2	NA
League 1	0.26	0.20	0.118	31933.39	1.1	NA
League 2	0.26	0.18	0.017	31845.2	0.9	[0.5,1.1]

Table 5.2: Resultant parameter estimates, changepoint position and test statistics for attendance (A_{ij}/A_{ji}) based single changepoint model for home advantage using data from the Premier League (1995/1996 - 2013/2014), Championship, League 1 and League 2 (2004/2005 - 2013/2014).

5.3.2 Penalised Spline Smoothing

Parametric models used to describe the dependence of the expected value of some response variable on one or more covariates are often not flexible enough to describe the data (Dierckx, 1995; Ahlberg et al., 1967). Nonparametric regression models allow the response to be modelled using a smooth, unspecified function of covariates. The drawback of nonparametric regression is the possibility that such methods are too flexible, leading to overfitting of the data. Therefore, it still relies on 'professional' input to the model constraints.

English League	$\hat{\gamma}_1$	$\hat{\gamma}_2$	p-value	AIC	$\hat{\delta}^*$ (years)	95% CI
Premier League	1.33	1.39	0.320	26287.77	6	NA
Championship	1.41	1.31	0.598	37969.46	1	NA
League 1	1.33	1.29	0.364	38446.27	2	NA
League 2	1.33	1.26	0.076	38068.81	6	NA

Table 5.3: Resultant parameter estimates, changepoint position and test statistics for referee experience based single changepoint model for home advantage using data from the Premier League, Championship, League 1 and League 2 (2000/2001 - 2011/2012).

Dimension	$\hat{\gamma}_1$	$\hat{\gamma}_2$	p-value	AIC	$\hat{\delta}^*$ (years)	95% CI
Length	0.84	1.38	0.130	24089.40	0.94	NA
Width	1.75	1.36	0.147	24093.64	0.93	NA

Table 5.4: Resultant parameter estimates, changepoint position and test statistics for pitch length and width based single changepoint models for home advantage using data from the Premier League (2001/2002 - 2011/2012).

Penalised spline regression is one such nonparametric regression technique, which is defined by piecewise polynomial functions linked by smooth transitions at values called knots (Hall and Opsomer, 2005). Now consider a p-degree spline model between the log of home advantage, $\log(\gamma)$, and some covariate δ in match observation k, as given by

$$\log(\gamma_k) = a_0 + a_1\delta_{i,j(k)} + a_2\delta_{i,j(k)}^2 + \ldots + a_p\delta_{i,j(k)}^p + \sum_{q=1}^{Q} \phi_q \left(\delta_{i,j(k)} - t_q\right)_+^p,$$

where $a_0, \ldots, a_p$ are unkown parameters that need to be estimated, the ϕ_q denote the spline coefficients at knots $t_1, \ldots, t_Q$ and $(\delta - t)_+^p$ represents the truncated basis function and is given by

$$(\delta - t)_+^p = \begin{cases} (\delta - t)^p & \delta > t \\ \\ 0 & \text{otherwise.} \end{cases}$$

The smoothness of the estimated function can be controlled by limiting the number of basis functions. However, this method is discontinuous and cannot allow for local features. Therefore, the smoothness at each knot is controlled by a roughness penalty term which can be added to the log-likelihood to give a pseudo log-likelihood (Bell et al., 2012; Dierckx, 1995; Ahlberg et al., 1967). Let the basis function be given by $B_b(x)$ for $b = 0, \ldots, p+Q$, i.e. $1, \ldots, x^p, (x-t_1)_+^p, \ldots, (x-t_Q)_+^p$. Let $\mathbf{\Omega}$ be a $(p+Q+1)\times(p+Q+1)$ penalty matrix with elements Ω_{ij}, given by

$$\Omega_{ij} = \int_u^v B_i''(z) B_j''(z) dz,$$

where $u \leq t_1 \leq \ldots \leq t_K \leq v$. The penalty term is then written as $\Psi \boldsymbol{\Phi}' \boldsymbol{\Omega} \boldsymbol{\Phi}$, where $\boldsymbol{\Phi} = (\boldsymbol{a}, \boldsymbol{\phi})$, and Ψ is the smoothing parameter, which is typically (but not necessarily) in (0,1].

The pseudo log-likelihood under this model is then given by

$$\begin{aligned} \ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \rho, a_0, a_1, b; \boldsymbol{x}, \boldsymbol{y}) = & \sum_{k=1}^{N} \log\left[\tau_{\lambda_k, \mu_k}(x_k, y_k)\right] + x_k \log(\lambda_k) - \lambda_k - \log(x_k!) \\ & + y_k \log(\mu_k) - \mu_k - \log(y_k!) + \Psi \boldsymbol{\Phi}' \boldsymbol{\Omega} \boldsymbol{\Phi}, \end{aligned} \tag{5.5}$$

where

$$\begin{aligned} \lambda_k = & \alpha_{i(k)} \beta_{j(k)} \exp\left[a_0 + a_1 \delta_{i,j(k)} + a_2 \delta_{i,j(k)}^2 + \ldots + a_p \delta_{i,j(k)}^p + \sum_{s=1}^{Q} \phi_s \left(\delta_{i,j(k)} - t_s\right)_+^p\right], \\ \mu_k = & \alpha_{j(k)} \beta_{i(k)}. \end{aligned}$$

In this analysis, the number of knots, $Q = 40$. The value of Ψ may be best estimated using cross validation or similar methods, however, this would be computationally slow. Therefore, two values of Ψ were used to give a low penalty, $\Psi = 1$, and a high penalty, $\Psi = 10$, to allow a representation of any high and low smoothing features. The pseudo log-likelihood given in equation (5.5) is maximised for both values of Ψ, using each covariate and relevant data set. The resultant curves for home advantage regarding the Premier League and covariates of distance, relative attendance, referee experience, relative pitch length and relative pitch width are plotted in Figures 5.2 to 5.5 respectively.

For all other leagues in the analysis (as given in Section 5.2), the resulting high and low penalty linear-spline curves, alongside piecewise constant regression models for home advantage with respect to the various covariates are shown in Appendices D to F ; for distance see Figures D.1 and D.2, for relative attendance see Figures E.1 to E.3, and for referee experience see Figures F.1 to F.3.

Note that splines inherently may overfit extreme values in a sample, leading to poor extrapolation under the penalised spline regression model. This can be see, for example, in Figure 5.2, where the spline fits both deviate highly from the piecewise constant regression at high values.

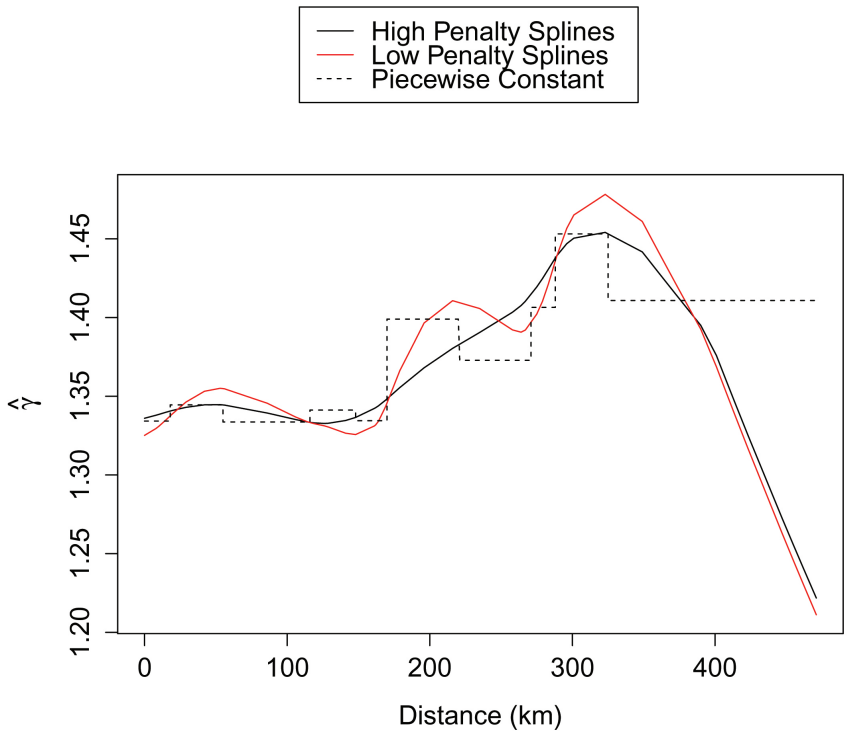


Figure 5.2: Premier League (England) 2001/2002 - 2011/2012: Penalised spline smooth curves describing the relationship of home advantage with distance (km), with high ($\Psi = 10$) and low ($\Psi = 1$) penalties, compared to a piecewise constant regression.

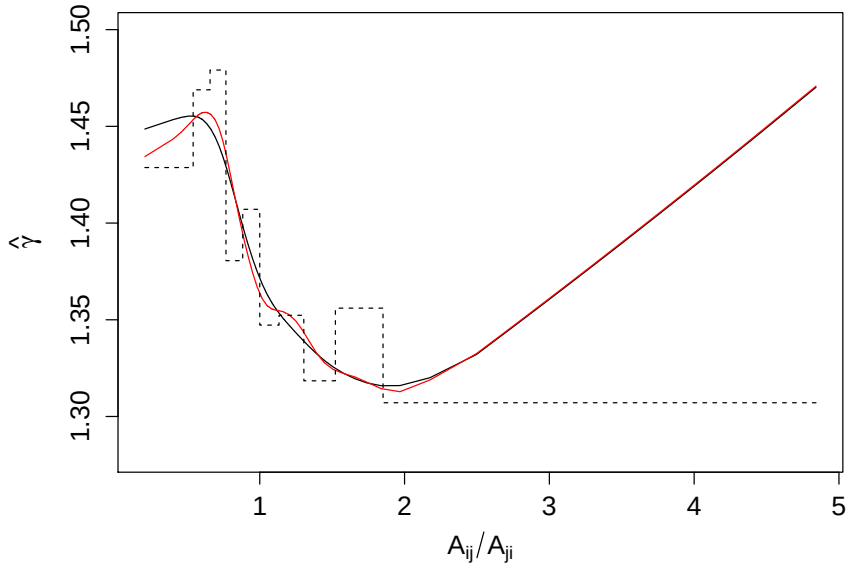


Figure 5.3: Premier League (England) 1995/1996 - 2013/2014: Penalised spline smooth curves describing the relationship of home advantage with relative attendance, with high ($\Psi = 10$) and low ($\Psi = 1$) penalties, compared to a piecewise constant regression.

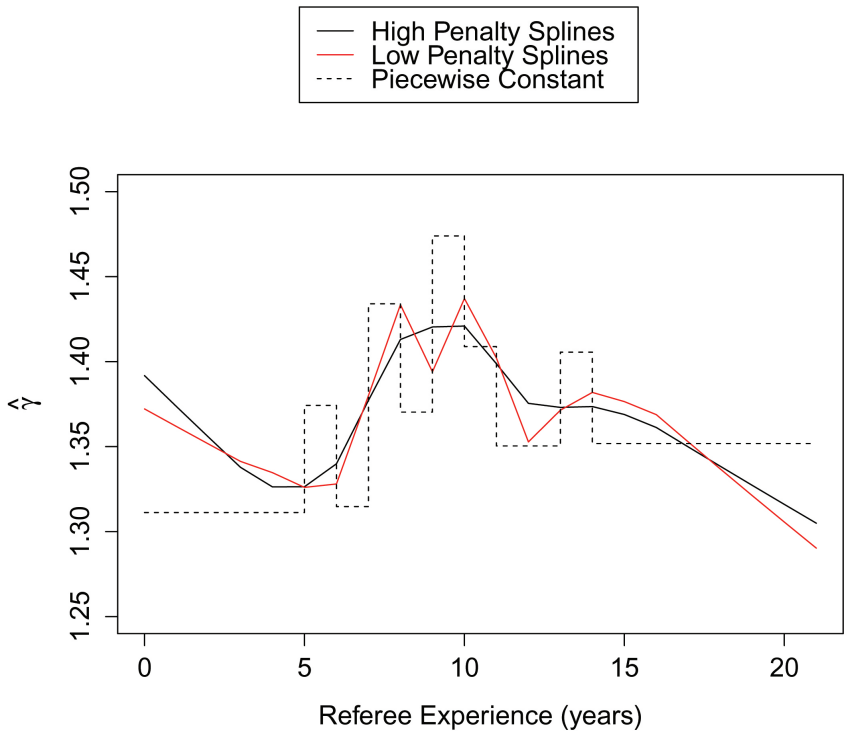


Figure 5.4: Premier League (England) 2000/2001 - 2011/2012: Penalised spline smooth curves describing the relationship of home advantage with referee experience, with high ($\Psi = 10$) and low ($\Psi = 1$) penalties, compared to a piecewise constant regression.

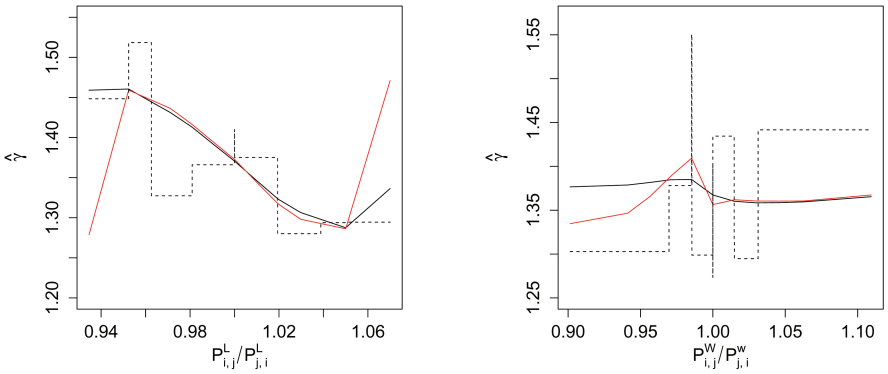


Figure 5.5: Premier League (England) 2001/2002 - 2011/2012: Penalised spline smooth curves describing the relationship of home advantage with (left) relative pitch length and (right) relative pitch width, with high ($\Psi = 10$) and low ($\Psi = 1$) penalties, compared to a piecewise constant regression.

5.4 Parametric Models

5.4.1 Log-Polynomial Regression Model

Consider the classical linear regression model

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ik}\beta_k + \varepsilon_i,$$

where the indexing $i = 1, \ldots, N$ describes N successive observations (Seber and Lee, 2012). This model assumes that the variable y_i is linearly dependent on the regressors x_i. An unobserved error variable ε_i adds noise to the relationship between y_i and x_i. Written in vector form this gives

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}$ are the regression coefficients and

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Simple linear regression describes the relationship between x and y for observations $i = 1, \ldots, n$ using the function

$$y_i = a_0 + a_1 x_i + \varepsilon_i,$$

where each ε_i is normally distributed with a zero mean conditioning on the regressor x_i and a_0 and a_1 are the regression coefficients. This may also be referred to as a first order polynomial.

The response is unbounded and therefore cannot be used directly to model non-negative goal counts without some modification. Taking the exponential allows the regression model to be applied to output quantities lying in the range 0 to ∞ and this is known as a log-linear regression model. Considering home advantage, γ, to be log-linearly dependent on some covariate, δ, then

$$\gamma_{i,j} = \exp\left(a_0 + a_1 \delta_{i,j}\right).$$

Polynomial models are linear when considering estimation, though they allow a non-linear relationship between the independent variable and the conditional mean of the dependent variable (Fan and Gijbels, 1996). Log-polynomial regression employs an exponentiated p-th order polynomial, and can be used to describe the relationship of γ and δ, as given by

$$\gamma_{i,j} = \exp\left(a_0 + a_1 \delta_{i,j} + a_2 \delta_{i,j}^2 + a_3 \delta_{i,j}^3 + \cdots + a_p \delta_{i,j}^p\right) \tag{5.6}$$

Caution must be taken to prevent overfitting. For example a high degree log-polynomial function may pass through each data point in a series, however, a first order log-polynomial or log-linear function may be a better model of the true relationship. If the regression curve were to be used to extrapolate the findings, a model which overfits to the data may produce considerably worse estimates of future observations.

Using the data specified in Section 5.2 first ($p = 1$) to third ($p = 3$) order log-polynomials were used in an attempt to model any relationship between γ and the covariates of distance, relative attendance, referee experience, relative pitch length and relative pitch width. The log-likelihood given by equation (4.7) was maximised relating γ to the covariate using equation (5.6) and the data selected for each analysis. The resultant parameters relating to the Premier League are given in Tables 5.5 to 5.9 and the parametric relationships are shown visually in Figures 5.6 to 5.9. Note that in Figure 5.9 for relative pitch length and relative pitch width, the curves under the first, second and third order log-polynomial models are similar.

For all other leagues in the analysis (as given in Section 5.2), the resulting polynomial regression models for home advantage with respect to the various covariates are shown in Appendices D to F ; for distance see Tables D.1 and D.2 and Figures D.3 and D.4, for relative attendance see Tables E.1 to E.3 and Figures E.4 to E.6, and for referee experience see Tables F.1 to F.3 and Figures F.4 to F.6.

The appropriateness of each model, in terms of significance and predictive capability, will be discussed in a later model comparison in Section 5.5. Visually, under the first order polynomial model, home advantage appears to have a relatively highly positive relationship with distance, and a negative relationship with relative attendance and relative pitch length, whilst referee experience shows only a slight positive relationship and relative pitch width shows a slight negative relationship. At the highest values of distance and relative attendance, the second and third order polynomials appear to provide evidence against the overall positive and negative trends respectively. It should be noted that there are few observations of high values in both cases.

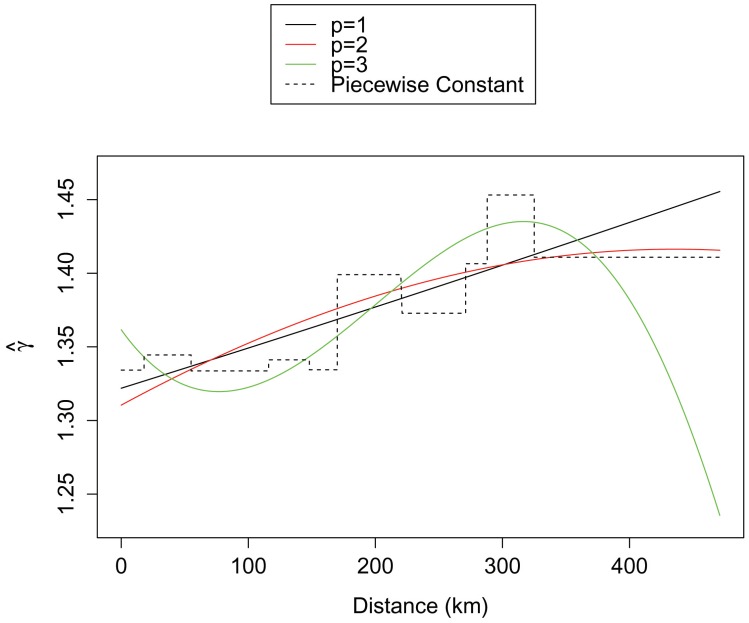


Figure 5.6: Premier League (England) 2001/2002 - 2011/2012: Log-polynomial models describing the relationship of home advantage with distance (km), compared to a piecewise constant regression.

	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
$p = 1$	2.78×10^{-1}	2.10×10^{-4}	NA	NA
$p = 2$	2.69×10^{-1}	3.61×10^{-4}	-4.15×10^{-7}	NA
$p = 3$	2.94×10^{-1}	-4.69×10^{-4}	4.74×10^{-6}	-8.34×10^{-9}

Table 5.5: Premier League (England) 2001/2002 - 2011/2012: Parameter values for first, second and third order polynomial regressions relating a regressor of distance between teams to home advantage.

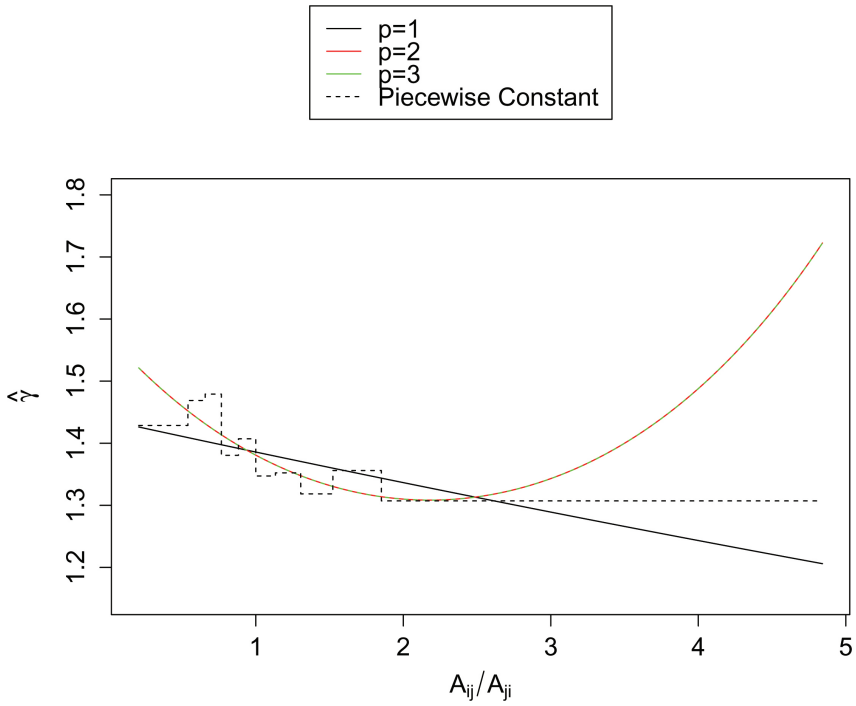


Figure 5.7: Premier League (England) 1995/1996 - 2013/2014: Log-polynomial models describing the relationship of home advantage with relative attendance, compared to a piecewise constant regression.

	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
$p = 1$	3.62×10^{-1}	-3.62×10^{-2}	NA	NA
$p = 2$	4.53×10^{-1}	-1.69×10^{-1}	3.88×10^{-2}	NA
$p = 3$	4.53×10^{-1}	-1.69×10^{-1}	3.88×10^{-2}	1.27×10^{-8}

Table 5.6: Premier League (England) 1995/1996 - 2013/2014: Parameter values for first, second and third order polynomial regressions relating a regressor of relative attendance to home advantage.

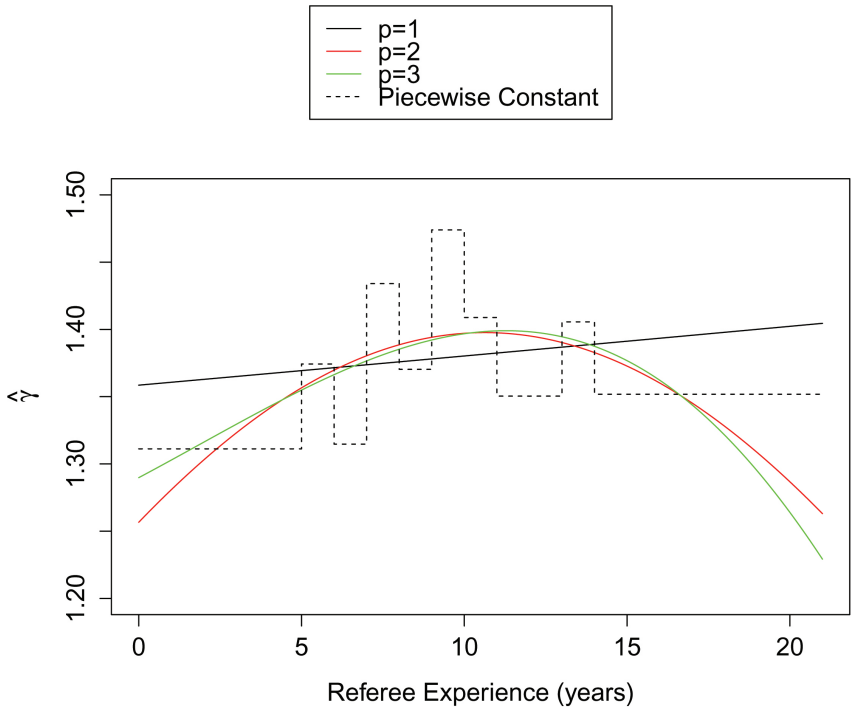


Figure 5.8: Premier League (England) 2000/2001 - 2011/2012: Log-polynomial models describing the relationship of home advantage with referee experience, compared to a piecewise constant regression.

	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
$p = 1$	3.06×10^{-1}	1.58×10^{-3}	NA	NA
$p = 2$	2.28×10^{-1}	2.00×10^{-2}	-9.40×10^{-4}	NA
$p = 3$	2.54×10^{-1}	9.99×10^{-3}	1.46×10^{-4}	-3.48×10^{-5}

Table 5.7: Premier League (England) 2000/2001 - 2011/2012: Parameter values for first, second and third order polynomial regressions relating a regressor of referee experience to home advantage.

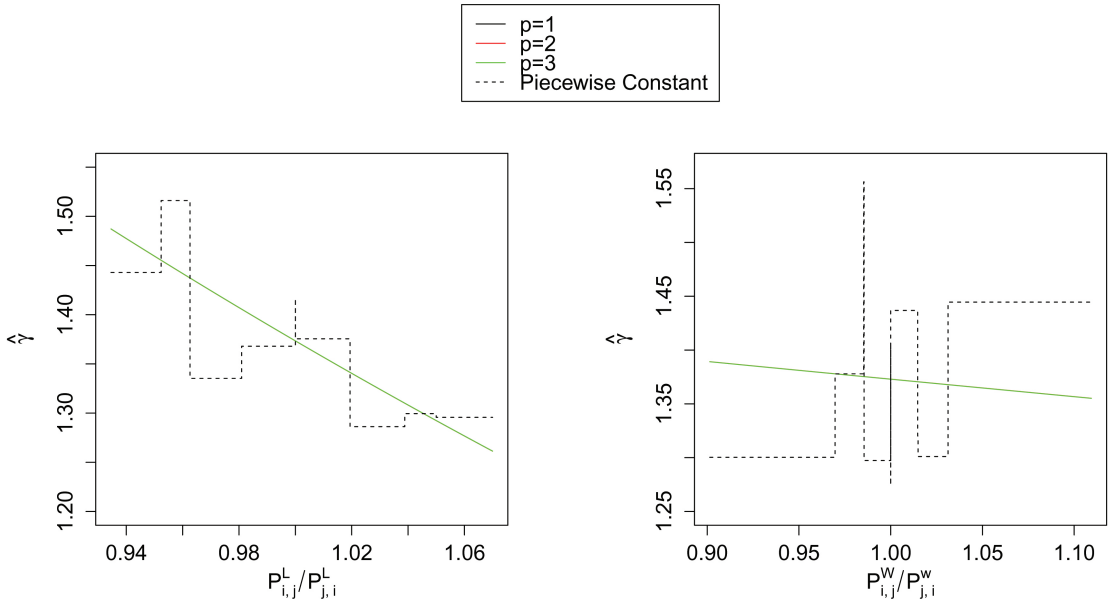


Figure 5.9: Premier League (England) 2001/2002 - 2011/2012: Log-polynomial model curves describing the relationship of home advantage with (left) relative pitch length and (right) relative pitch width, compared to a piecewise constant regression.

	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
$p = 1$	1.53	-1.22	NA	NA
$p = 2$	1.53	-1.22	-1.90×10^{-5}	NA
$p = 3$	1.53	-1.22	-1.95×10^{-5}	-1.50×10^{-5}

Table 5.8: Premier League (England) 2001/2002 - 2011/2012: Parameter values for first, second and third order polynomial regressions relating a regressor of relative pitch length to home advantage.

	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
$p = 1$	4.36×10^{-1}	-1.19×10^{-1}	NA	NA
$p = 2$	4.36×10^{-1}	-1.19×10^{-1}	-6.93×10^{-6}	NA
$p = 3$	4.36×10^{-1}	-1.19×10^{-1}	-1.7×10^{-5}	-9.44×10^{-6}

Table 5.9: Premier League (England) 1995/1996 - 2013/2014: Parameter values for first, second and third order polynomial regressions relating a regressor of relative pitch width to home advantage

5.4.2 Exponential Curve Model

Figures 5.1 and 5.3 suggest that a more subtle exponential curve may describe the relationships of distance and relative attendance with home advantage more accurately, i.e.

$$\gamma = \exp\left[c_0 + c_1 \exp\left(c_2 \theta\right)\right]. \tag{5.7}$$

As with all regressions there are issues if the sub-sample that is being analysed does not display the overall characteristics of the sample, i.e. the range of values for the regressor is too narrow to define the curve. Also, the data may be highly noisy. Therefore it is easier in this situation to choose initial starting values that are not too far from their correct values when numerically maximising the likelihood. This may be overcome to some extent by using starting values similar to those achieved under, for example, the log-linear or log-polynomial models.

The log-likelihood given by equation (4.7), under the model for home advantage given by equation (5.7), was maximised for the covariates of distance and relative attendance, using the data selected for each analysis as given in Section 5.2. The resultant parameters are given in Tables 5.10 and 5.11 respectively, and the parametric relationships are shown visually in Figures 5.10 and 5.11 respectively, where home advantage has a positive relationship with distance and a negative relationship with relative attendance. These findings are consistent with the log-polynomial and piecewise constant models.

League	$\hat{c}_0$	$\hat{c}_1$	$\hat{c}_2$
Premier League	0.28	0.10	0.89
Ligue 1	0.09	0.33	0.17
Serie A	0.18	0.18	0.20
Combined	0.25	0.15	0.40

Table 5.10: Exponential curve parameters using individual and combined data from Premier League (England) 2001/2002-2011/2012, Ligue 1 (France) 2003/2004 - 2011/2012 and Serie A (Italy) 2004/2005 - 2011/2012.

5.5 Model Comparisons

A number of models have been investigated, relating various covariates to home advantage using the model outlined by Dixon and Coles (1997). Three statistics shall be used to compare and contrast the models; the p-value, the AIC and the RMSE. It should be noted that the penalised spline model cannot be compared using the p-value or AIC statistics due to its pseudo-likelihood.

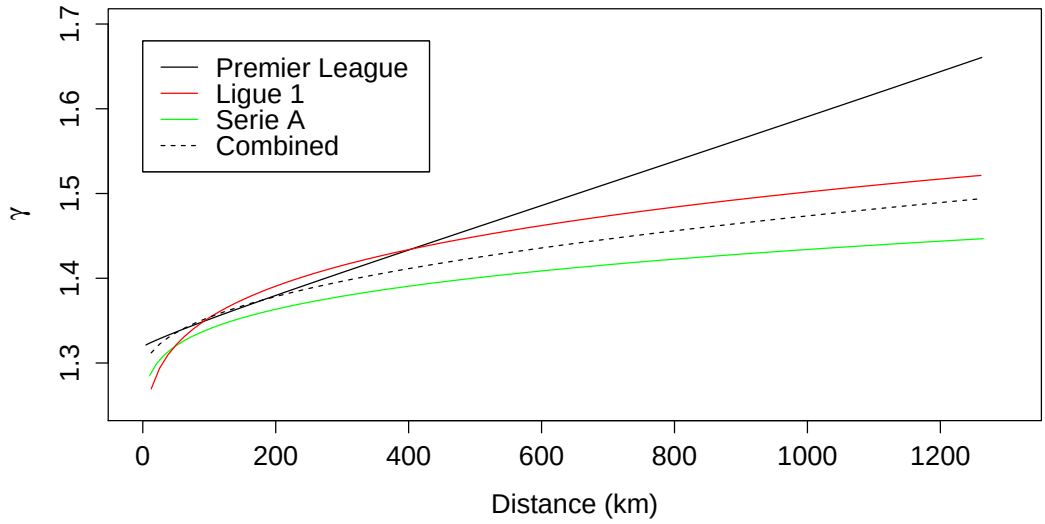


Figure 5.10: Comparison of exponential curves using individual and combined data from Premier League (England) 2001/2002-2011/2012, Ligue 1 (France) 2003/2004 - 2011/2012 and Serie A (Italy) 2004/2005 - 2011/2012.

League	$\hat{c}_0$	$\hat{c}_1$	$\hat{c}_2$
Premier League	0.16	0.16	-0.42
Championship	0.13	0.13	-0.01
League 1	-0.14	0.38	-0.12
League 2	0.13	0.08	-0.72

Table 5.11: Parameters relating to exponential curve model for home advantage as a function of $A_{i,j}/Aj, i$ using data from the Premier League (1995/1996 - 2013/2014), Championship, League 1 and League 2 (2004/2005 - 2013/2014).

Deviance cannot be used to compare all models as they are not necessarily nested. However, it may be used in conjunction with the additional number of degrees of freedom over the null model of constant home advantage (shown in Table 5.12) to perform a chi squared test between a null of constant home advantage and an alternative of each model. This will allow an assessment of the statistical significance of each variable as a covariate for home advantage under each model.

5.5.1 Distance

Table 5.13 shows the p-value for each parametric model for each league used in the analysis of distance as a covariate of home advantage (see Section 5.2), and all leagues combined, with the cases where the null is rejected at a 0.05 significance level highlighted.

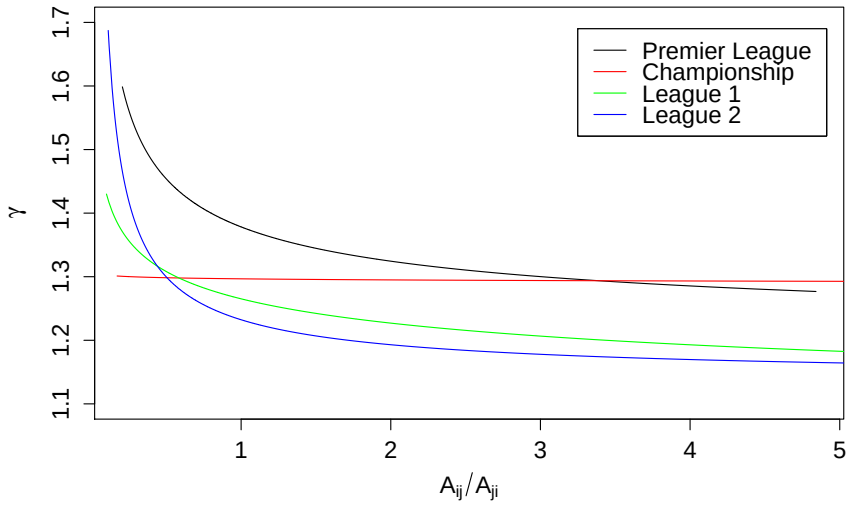


Figure 5.11: Comparison of exponential curve models relating A_{ij}/A_{ji} to home advantage using data from the Premier League (1995/1996 - 2013/2014), Championship, League 1 and League 2 (2004/2005 - 2013/2014).

Model	+D.F.	$\chi^2_{.050}$
Changepoint	2	5.99
Log-Linear	1	3.84
Log-Quadratic	2	5.99
Log-Cubic	3	7.82
Exponential Curve	2	5.99
Piecewise Constant	9	16.92

Table 5.12: Additional degrees of freedom (compared to null hypothesis of constant home advantage) for each model in the comparison and the associated 0.05 significance level in a chi squared test.

When considering individual leagues, the changepoint model may be considered statistically significant at a level of 0.05 in a chi squared test for Ligue 1 and the Premier League. All models are significant at this level for the combined analysis of all leagues, suggesting that there is a statistically significant relationship between distance and home advantage.

The AIC was calculated for each parametric model in each league in the distance study and all leagues combined, and is presented in Table 5.14 as the AIC value for each model in the comparison minus the AIC of the model with the lowest (or best) value. In this case, zero represents the best model. For both the Premier League and Ligue 1, the changepoint model relating distance to home advantage is the best model according to this statistic. For Serie A of Italy, the constant home advantage model is the best model

Model	Ligue 1(France)	Serie A (Italy)	Premier League (England)	Combined
Changepoint	0.004	0.167	0.038	0.000
Log-Linear	0.129	0.192	0.083	0.007
Log-Quadratic	0.101	0.391	0.204	0.005
Log-Cubic	0.205	0.532	0.165	0.011
Exponential Curve	0.208	0.375	0.219	0.001
Piecewise Constant	0.167	0.251	0.862	0.014

Table 5.13: p-values values comparing alternate hypotheses of each model to a null hypothesis of constant home advantage for each model in the comparison relating to distance between teams as a covariate of home advantage.

according to the AIC. Finally, when combining data from all leagues, the changepoint model for home advantage with respect to distance is the best model under the AIC.

Model	Ligue 1 (France)	Serie A (Italy)	Premier League (England)	Combined
Constant (Null)	6.86	0	2.54	12.71
Changepoint	0	0.42	0	0
Log-Linear	6.55	0.3	1.54	7.44
Log-Quadratic	6.27	2.12	3.36	6.1
Log-Cubic	8.28	3.8	3.44	7.62
Exponential Curve	7.72	2.03	3.5	2.64
Piecewise Constant	11.95	6.63	15.86	9.96

Table 5.14: AIC values with division minimum AIC value subtracted for each model and division in the comparison relating to distance between teams as a covariate of home advantage. A zero value represents the best model under this test statistic.

The p-value and AIC both give a representation of statistical significance, though they do not give an impression of the practical benefits. The RMSE may be used to compare the actual goal counts to the expected values under each model. It can also be used to compare the spline model fit to the other models, which could not be done with the p-value or AIC. To allow the impression of any improvement, this will be presented as a percentage improvement in RMSE of goal estimates over the model for constant home advantage, as given by equation (4.21), for home and away goals, $v(x)$ and $v(y)$ respectively,

Table 5.15 shows the values of $v(x)$ and $v(y)$ for all models tested for the Premier League. It can be seen that the linear spline fit shows the greatest overall improvement, which is not surprising as it allows the largest amount of flexibility. However, the amount of improvement is extremely low in all cases. It can therefore be concluded, that the inclusion of a model solely relating distance to home advantage and not allowing for other factors,

may not be of practical benefit in a betting strategy over a short time horizon.

Model	$v(x)$	$v(y)$
Log-Linear	0.036	0.003
Log-Quadratic	0.044	0.001
Log-Cubic	0.078	-0.004
Exponential Curve	0.035	0.006
Piecewise Constant	0.087	0.002
Changepoint	0.061	0.015
Linear Splines ($\Psi = 1$)	0.129	0.007
Linear Splines ($\Psi = 10$)	0.102	0.001

Table 5.15: Values of $v(x)$ and $v(y)$ given for each model relating distance to home advantage, to allow a comparison of the RMSE values to a constant home advantage model for the Premier League (England) 2001/2002 - 2011/2012

5.5.2 Relative Attendance

Comparing a null hypothesis of a constant home advantage to alternatives of each other parametric model, a chi-squared test was performed to allow the assessment of significant deviation of the log-likelihood values for each league in the analysis of relative attendance as a covariate of home advantage (see Section 5.2). Table 5.16 shows the p-values under an alternative hypothesis of each parametric model against a null of constant home advantage, with the values which test significant at a 0.05 significance level highlighted (see Table 5.12 for significance levels).

Model	Premier League	Championship	League 1	League 2
Changepoint	0.042	0.239	0.118	0.017
Log-Linear	0.111	0.920	0.126	0.036
Log-Quadratic	0.024	0.568	0.097	0.079
Log-Cubic	0.059	0.009	0.198	0.065
Exponential Curve	0.077	1.000	0.177	0.014
Piecewise Constant	0.310	0.580	0.284	0.023

Table 5.16: p-values values comparing alternate hypotheses of each model to a null hypothesis of constant home advantage for each model in the comparison relating to relative attendance as a covariate of home advantage.

Table 5.17 shows the AIC values of each parametric model minus the lowest AIC values in the analysis of relative attendance as a covariate of home advantage. Again, in this analysis, the best model then has a value of zero. The log-quadratic and log-cubic models

are the best models under the AIC for the highest three divisions, whilst the exponential curve model fits the data from League 2 better according to this statistic.

	Premier League	Championship	League 1	League 2
Constant (Null)	3.43	5.68	0.67	4.54
Log-Linear	2.89	7.67	0.33	2.12
Log-Quadratic	0.00	8.55	0.00	3.46
Log-Cubic	2.00	0.00	2.00	3.32
Exponential Curve	2.30	9.67	1.22	0.00
Changepoint	1.10	6.81	0.40	0.39
Piecewise Constant	10.91	16.12	7.79	3.21

Table 5.17: AIC values with division minimum AIC value subtracted for each model and division in the comparison relating to relative attendance as a covariate of home advantage. A zero value represents the best model under this test statistic.

As discussed in Section 5.5.1, the RMSE allows a view of the practical benefits, and also permits a comparison of the semi-parametric spline model. Table 5.18 shows the value of $v(x)$ and $v(y)$, as defined in equation (4.21), for each model in the comparison, including the linear spline models, using data from the Premier League. Under this statistic, the piecewise constant model displays the greatest percentage improvement in RMSE. Again, little practical impact to the values of RMSE can be seen when implementing relative attendance as a covariate for home advantage in this way.

Model	$v(x)$	$v(y)$
Log-Linear	0.007	0.015
Log-Quadratic	0.061	0.011
Log-Cubic	0.061	0.011
Exponential Curve	0.033	0.013
Piecewise Constant	0.072	0.012
Changepoint	0.041	0.015
Linear Splines ($\Psi = 1$)	0.064	0.013
Linear Splines ($\Psi = 10$)	0.060	0.014

Table 5.18: Values of $v(x)$ and $v(y)$ given for each model relating attendance to home advantage, to allow a comparison of the RMSE values to a constant home advantage model for the Premier League (England) 2001/2002 - 2011/2012.

5.5.3 Referee Experience

Similarly to the previous sections, chi-squared tests were performed considering a null hypothesis of constant home advantage and alternative hypotheses of each parametric model tested, relating home advantage and a covariate of referee experience for all leagues chosen for this analysis as discussed in Section 5.2. Table 5.19 shows the p-values from these tests. None of the models relating home advantage and referee experience tested significant at a 0.05 level (see Table 5.12 for significance levels).

Model	Premiership	Championship	League 1	League 2
Changepoint	0.321	0.596	0.363	0.076
Log-Linear	0.635	0.164	0.103	0.066
Log-Quadratic	0.380	0.179	0.167	0.128
Log-Cubic	0.569	0.119	0.203	0.111
Piecewise Constant	0.538	0.084	0.080	0.175

Table 5.19: p-values values comparing alternate hypotheses of each model to a null hypothesis of constant home advantage for each model in the comparison relating to referee experience as a covariate of home advantage.

Table 5.20 shows the AIC values of each parametric model minus the lowest AIC values in the analysis of referee experience as a covariate of home advantage. Interestingly, the changepoint model proved to have the lowest AIC value for the Premiership and League 2, whilst the log-linear model had the lowest AIC value for League 1. These results are contradictory to the p-values given in Table 5.19.

Model	Premiership	Championship	League 1	League 2
Constant (Null)	0.28	0.00	0.66	3.16
Changepoint	0.00	0.97	0.63	0.00
Log-Linear	2.05	0.06	0.00	1.79
Log-Quadratic	2.34	0.56	1.08	3.04
Log-Cubic	4.26	0.15	2.05	3.14
Piecewise Constant	10.32	2.74	3.23	8.41

Table 5.20: AIC values with division minimum AIC value subtracted for each model and division in the comparison relating to referee experience as a covariate of home advantage. A zero value represents the best model under this test statistic.

Again, little practical impact to the values of RMSE can be seen when implementing referee experience as a covariate for home advantage in this way. It shall not be shown in a tabular format as in previous sections, as little benefit can be perceived due to the

lack of signficance for all models.

5.5.4 Relative Pitch Dimensions (Length and Width)

Finally, chi-squared tests were performed comparing a null hypothesis of constant home advantage against alternatives of each other parametric model, for each league in the analysis of relative pitch length and width as covariates of home advantage as discussed in Section 5.2. Table 5.21 shows the respective p-values, with the values which test significant at a 0.05 significance level highlighted (see Table 5.12 for significance levels). Only the log-linear model caused the null hypothesis to be rejected, suggesting a significant negative log-linear trend between home advantage and relative pitch length.

Model	$P^L_{i,j}/P^L_{j,i}$	$P^W_{i,j}/P^W_{j,i}$
Changepoint	0.131	0.147
Log-Linear	0.034	0.825
Log-Quadratic	0.107	0.976
Log-Cubic	0.214	0.997
Piecewise Constant	0.572	0.063

Table 5.21: p-values values comparing alternate hypotheses of each model to a null hypothesis of constant home advantage for each model in the comparison relating to relative pitch length and relative pitch width as covariates of home advantage.

Table 5.22 shows the AIC values of each parametric model minus the lowest AIC values in the analysis of relative pitch width as a covariate of home advantage. Again, in this analysis, the best model then has a value of zero. The log-linear model proved to be the best model under the AIC for relative pitch length in the Premier league, whilst no model relating home advantage to relative pitch width could be found that had a lower AIC than the null model of constant home advantage.

The values of $v(x)$ and $v(y)$ are given in Table 5.23. It can be seen that there is some improvement in the RMSE over the null model for all tested models, though it is only slight.

5.6 Combining Models

To ascertain the best combination of covariates to employ when modelling home advantage, data from the Premiership between seasons 2001/2002 and 2011/2012 were analysed for each of the best covariate models. Table 5.24 shows the resulting AIC and p-values

Model	$P^L_{i,j}/P^L_{j,i}$	$P^W_{i,j}/P^W_{j,i}$
Constant (Null)	2.48	0.00
Changepoint	2.41	0.16
Log-Linear	0.00	1.95
Log-Quadratic	2.00	3.95
Log-Cubic	4.00	5.95
Piecewise Constant	12.85	1.78

Table 5.22: AIC values with division minimum AIC value subtracted for each model and division in the comparison relating to relative pitch length and relative pitch width as covariates of home advantage. A zero value represents the best model in under this test statistic.

Model	$v(x)$	$v(y)$
Log-Linear	0.04	0.05
Log-Quadratic	0.04	0.05
Log-Cubic	0.04	0.05
Changepoint	0.04	0.07
Piecewise Constant	0.05	0.01
Linear Splines ($\Psi = 1$)	0.08	0.05
Linear Splines ($\Psi = 10$)	0.06	0.05

Table 5.23: Values of $v(x)$ and $v(y)$ given for each model relating relative pitch length to home advantage, to allow a comparison of the RMSE values to a constant home advantage model for the Premier League (England) 2001/2002 - 2011/2012

from a hypothesis test between a null of constant home advantage and an alternative of each individual model under test. Note, referee experience showed no significant relationship with home advantage in any of the models tested, so no model for it is tested here.

In this time range, the changepoint model relating distance and home advantage and the log-linear model relating relative pitch length and home advantage caused the null to be rejected at a 0.05 significance level in a chi-squared test. However, the log-linear model for the covariate of relative attendance did not prove significant at this level for the shorter time period used here.

Covariate	Model	p-value	AIC
Distance	Changepoint	0.038	24090.94
Relative Attendance	Log-quadratic	0.376	24095.52
Relative Pitch Length	Log-linear	0.034	24091.00

Table 5.24: AIC and p-value results for each best covariate model.

All combinations of these models (both for two or three covariates) were tested, combining parameters where appropriate. Hypothesis tests were carried out with null hypotheses of each individual component covariate model for home advantage and an alternative of a model formed from the combination of those components, e.g. an alternative of a combined model for home advantage of a changepoint with distance and a log-quadratic relationship with relative attendance was tested against a null model of a changepoint with distance and a null model relating relative attendance by a log-quadratic curve. Under this cross testing, the only combination which rejected all null models at a 0.05 significance level, was that relating home advantage to a changepoint over distance (at 193 km) and a log-linear trend with pitch length (with p-values of 0.035 and 0.039 respective to the null component models). This can be written

$$\gamma_{i,j} = \begin{cases} \gamma_1 + a_1 P^L_{i,j}/P^L_{j,i} & if \quad D_{i,j} \leq D^* \\ \gamma_2 + a_1 P^L_{i,j}/P^L_{j,i} & if \quad D_{i,j} > D^* \end{cases} .$$

This model also had the lowest value of AIC of any of the models under testing for this data set, suggesting that it is the best model with respect to this statistic. To test the impact on the predictive power of the model, a leave one out cross validation study was carried out in a similar fashion as described in Section 4.2. The subset of seasons 2009/2010 to 2011/2012 was used, siimilarly to the study performed in Section 4.2. Under the model relating home advantage to a changepoint over distance and a log-linear trend with pitch length, the RMSE for home goals was 1.270, whilst that for away goals was 1.100. These result in values of $v(x) = 0.31\%$ and $v(y) = 0.11\%$, when compared to the model of constant home advantage, suggesting that the predictive power has increased when considering parametric models for home advantage relating to covariates of distance and pitch length.

Table 5.25 shows the ten highest and ten lowest home advantage estimates for the dataset. From these data, it can be seen that a common theme of Manchester City having the lowest home advantage, and playing as away team in the matches which experience the highest home advantage.

5.7 Conclusion

Various linear and non-linear models have been used to explore the relationship of home advantage with predictor variables, such as distance between teams, match attendance, pitch dimensions and referee bias. Exploratory methods such as piecewise regression and penalised spline smoothing provided an initial picture of any trends.

Using the closed form expression for piecewise constant home advantage, regression could

Home Team	**Away Team**	x	y	**Distance**	**Pitch Length**	$\hat{\gamma}$
Manchester City	Bolton Wanderers	2	0	17	1.07	1.22
Manchester City	Bolton Wanderers	6	2	17	1.07	1.22
Manchester City	Everton	3	1	53	1.07	1.22
Manchester City	Liverpool	0	3	53	1.07	1.22
Manchester City	Liverpool	2	2	53	1.07	1.22
Manchester City	Everton	5	1	53	1.07	1.22
Manchester City	Wolverhampton Wanderers	3	3	102	1.07	1.22
Manchester City	Birmingham City	1	0	117	1.07	1.22
Manchester City	Birmingham City	0	0	117	1.07	1.22
Aston Villa	Birmingham City	0	2	4	1.05	1.25
...	...	...	...	...	...	...
Fulham	Newcastle United	1	0	402	0.95	1.51
Fulham	Newcastle United	5	2	402	0.95	1.51
Crystal Palace	Newcastle United	0	2	414	0.95	1.51
Tottenham Hotspur	Manchester City	0	2	258	0.93	1.54
Tottenham Hotspur	Manchester City	1	1	258	0.93	1.54
Arsenal	Manchester City	2	1	264	0.93	1.54
Fulham	Manchester City	0	1	264	0.93	1.54
Fulham	Manchester City	2	2	264	0.93	1.54
Arsenal	Manchester City	2	1	264	0.93	1.54
West Ham United	Manchester City	0	0	266	0.93	1.54

Table 5.25: English Premier League 2001/2002 - 2012/2013: Home advantage estimates, distance between teams and pitch lengths for the ten highest and ten lowest estimates.

be quickly carried out over many leagues simultaneously. It became clear that home advantage has a positive relationship with distance. This hypothesis was confirmed with statistical hypothesis testing and under the AIC. However, it appears that a changepoint model better modelled the hidden relationship than a smooth curve as was seen in the exploratory analysis. This could be because individual leagues with varying changepoints combine to form the curve seen under the piecewise constant regression, or that the leagues tested did not have a wide enough range of distance values.

Crowd dynamics have often been discussed as a driving factor in home advantage. One possible measure of this is the crowd size or match attendance, which in this case was treated as the relative attendance. No obvious trend was found under exploratory analysis. However, under hypothesis testing between various alternative models and the null of a standard Poisson model, a significant negative trend was found between home advantage and relative attendance for three of the four leagues tested. No consistent regression model was found across all leagues tested, however.

No significant relationship between referee experience and home advantage was found. This might suggest that although experience is thought to bring consistency, some referees may still suffer a predisposition to awarding home advantage far into their careers.

Pitch dimensions were tested finally (although only for one league, due to a lack of data), with home advantage experiencing a significant negative log-linear relationship with relative pitch length. No significant relationship was found between home advantage and relative pitch width. This suggests that home teams playing on a short pitch as opposed to the away team who usually plays on a longer pitch, will experience a higher home advantage. This could be because the team does not have as much room to play to their usual style.

In combining these models, it was found that the most statistically significant combination was that of a changepoint with distance and a log-linear relationship with pitch length. Table 5.25 shows the matches displaying the highest and lowest home advantages under this model, whereby Manchester City appears to display the most disadvantageous characteristics, as opposed to West Ham United, Arsenal, Fulham and Tottenham Hotspur.

5.8 Future Work

Alongside further testing of the models discussed in this chapter, on both larger data sets and different leagues, the evolution of models through innovative regression characteristics and higher order changepoint models may lead to a better fit and higher predictive power.

Crowd dynamics may have shown some significant relationship with home advantage, however, it was not conclusive. The lack of available home and away crowd split has limited the analysis. If these data were attained, it may show a much more significant relationship between home advantage and attendance.

Chapter 6

Overdispersion and Threshold Effects

6.1 Introduction

Previous attempts to increase the goodness of fit and predictive capabilities of the model defined in Dixon and Coles (1997) have related to solely parametric additions to allow the model to consider various variables which goal count may depend on (see Chapters 4 and 5). This chapter aims to challenge the class of distribution used in the modelling process, and to ascertain whether or not a mixture distribution of some sort would better fit the data.

Right censoring of the Poisson model defined by equation (4.3), hereafter referred to as Dixon and Coles model (4.3), will be used to illustrate the effect that poor fitting in the extreme right tail has on the fit of the body. Under this censored model the extent to which the Poisson distribution underperforms in modelling the home and away goal counts will be discussed.

Finally, threshold mixture modelling will be introduced as a method of increasing the goodness of fit in the right tail, with a selection of distributions. The use of threshold mixture modelling opens up the ability to implement a high scoring home advantage parameter, which may help to reduce the probabilistic error and better model home advantage when teams of very high and low abilities play each other.

It should be noted that more traditional mixture models were tested, though there were none which were found to increase the goodness of fit, as all regressed to a sole Poisson distribution. This is not to say that some other reparameterisation would not allow such models. This could be an area of further work.

6.2 Overdispersion

Goal count data in association football is thought to follow a Poisson distribution (Dixon and Coles, 1997). One of the main features of the Poisson distribution is the equality of the mean and variance. A sensible measure of dispersion in home and away goal counts can therefore be calculated using the variance-to-mean ratio, i.e.,

$$D = \frac{\sigma^2}{\mu}. \tag{6.1}$$

Under the assumption that goal counts are identically distributed within teams and using a data set of over 100,000 professional level association football matches[1], $D = 1.07$ for both home and away goals. This may be interpreted as a slight overdispersion in goal count. However, as we knew teams differ in ability this could equally be due to that reason.

Table 6.1 shows the quantiles for home and away goal count in steps of 10%. In fact, less than 1% of all goal counts are greater than 5. These few high goal count events have a greater effect on the variance than on the mean.

Quantile	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Home	0	0	0	1	1	1	2	2	2	3	10
Away	0	0	0	0	1	1	1	2	2	3	11

Table 6.1: 10% quantiles of home and away goal counts

Figure 6.1 shows a plot comparing the empirical and estimated Poisson cumulative probabilities, assuming identically distributed goals over home teams and identically distributed goals over away teams, given a single parameter describing home goal count ($\lambda = \mathrm{E}(X) = 1.51$) and a single parameter describing away goal count ($\mu = \mathrm{E}(Y) = 1.34$). This shows that the probability estimates reflect the empirical probabilites with little discrepancy. Figure 6.2 shows the ratio between the emprical probabilities and the estimated Poisson probabilities, which indicates that the estimated Poisson probabilities depart from the empirical probabilities as counts increase, placing relatively less probability mass on high goal count events.

However, as can be seen in Figure 6.3, most of this error occurs below 5 counts; for home and away goal counts, 85% and 90% of the error in the probability mass occurs

[1]Dataset provided by ATASS sports, multiple leagues from 2001-2012, see Table 6.2 and Appendix C for further details.

below 5 counts, respectively.

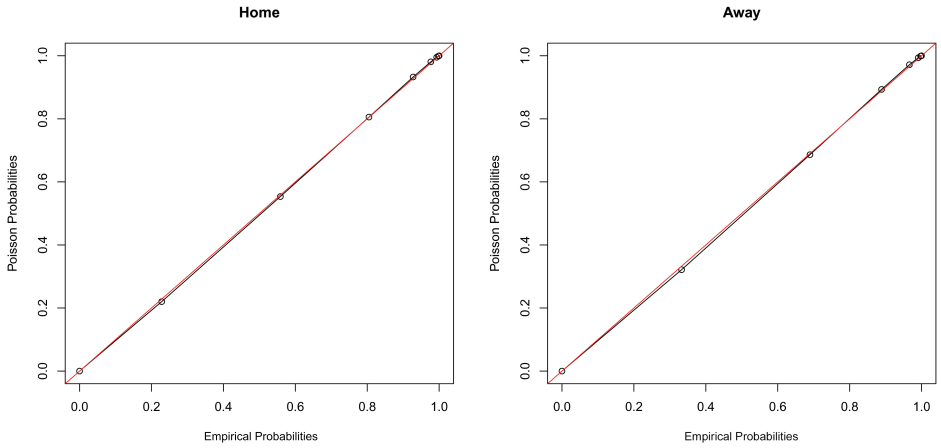


Figure 6.1: Comparison plot of empirical and Poisson cumulative probabilities for (left) home goal count and (right) away goal count, using a single parameter model for each distribution.

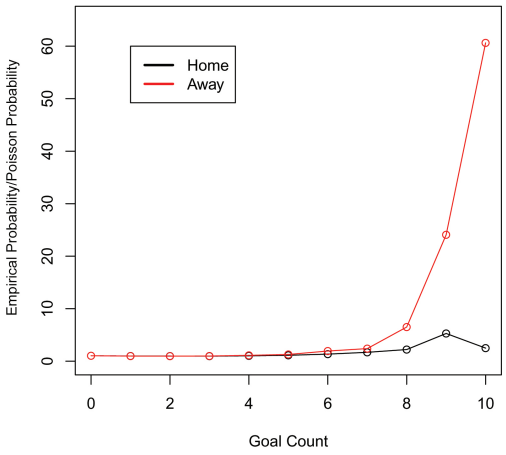


Figure 6.2: Ratio of empirical over estimated Poisson probabilities for home and away goal count, using a single parameter model for each distribution.

The apparent over-dispersed nature of goal count data, when considering goal counts for each team to be identically distributed, can be accounted for in a number of ways. Another approach could be used, for example the negative binomial distribution which considers any unobserved heterogeneity. Alternatively, data above some level, c, could be censored to account for any doubt in the model above a certain threshold value.

However, the assumption of identically distributed goal counts for all teams or games is a naive one, and the model can be altered to capture the varying mean goal counts for each team or each game. The following sections will cover the approaches that have been

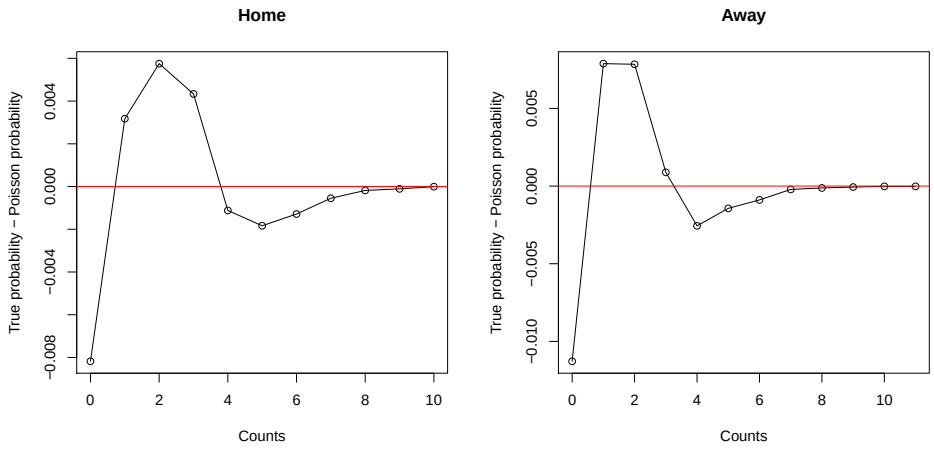


Figure 6.3: Difference between emprical and estimated Poisson probabilities for home and away goal count, using a single parameter model for each distribution.

considered and how these affect the predictive nature of the model.

6.3 Increased Poisson Variation

An important feature of the Poisson distribution is that the mean and variance are equal. Empirical count data, such as goal counts, often exhibit overdispersion, whereby the variance is greater than the mean. A common alternative to the Poisson distribution when encountering overdispersion is to add a multiplicative gamma random effect, which can be used to represent any unobserved heterogeneity (Hilbe, 2011). Negative binomial regression is an example of this process and is commonly used in cases of overdispersion in count data.

Consider the conditional distribution of X given an unobserved variable θ to be Poisson with mean and variance $\theta\mu$, so

$$X|\theta \sim \text{Poisson}\left(\theta\mu\right).$$

This model still assumes Poisson distributed goal counts, but recognises that the mean goal counts of all teams are not the same, and θ captures the variability in performance of each team over games.

Consider θ to follow a gamma distribution with a shape parameter α and scale parameter β. The mean (α/β) and variance (α/β^2) of the θ distribution are taken to be 1 and σ^2 respectively, therefore $\alpha = \beta = 1/\sigma^2$. Using this assumption θ may be integrated out allowing the effective computation of the unconditional distribution of X:

$$\Pr(X=x)=\int_0^\infty \Pr(X=x|\theta)\, f(\theta)\, d\theta=\frac{\Gamma(\nu+x)}{x!\Gamma(\nu)}\frac{\nu^\nu\mu^x}{(\mu+\nu)^{\nu+x}}, \qquad x=0,1,\ldots, \quad (6.2)$$

where $\nu=1/\sigma^2$ is referred to as the dispersion parameter or 'size'. The negative binomial distribution which uses this definition of α and β has mean $E(X)=\mu$ and variance $\mathrm{Var}(X)=\mu(1+\sigma^2\mu)$. and $D=1+\sigma^2\mu$. If $\sigma^2=0$ there is no unobserved heterogeneity and the variance is equal to the mean, i.e. $D=1$. If $\sigma^2>0$ the variance will be larger than the mean and $D>1$, therefore the distribution will be over-dispered relative to the Poisson distribution.

6.3.1 Negative Binomial vs Poisson Regression Models

This is a simple analysis, whereby all match pairings are assumed to follow identical distributions, with a single set of parameters (one in the Poisson case and two in the negative binomial case) estimated over all the available match results in each league (from the data set provided by ATASS Sports - seasons starting in years 2001 - 2011 in most cases, only those leagues with over 500 matches were used).

To analyse the extent of dispersion in goal data the variance of goal counts may be divided by the mean, to gain the variance-to-mean ratio as defined in equation (6.1). This will show whether the variance is greater than, less than or similar to the mean, allowing the identification of any cases of over or under dispersion.

Two tests will be carried out to analyse the most appropriate model to use, the Poisson or negative binomial model. To test whether the sample goal counts are consistent with the hypothesised distribution, a goodness of fit test was carried out. To test whether the sample goal counts are more likely to follow the negative binomial model, a hypothesis test was carried out between a null hypothesis that $\sigma=0$ and an alternative hypothesis that $\sigma>0$ for both home and away goals.

The overall goodness of fit may be analysed using the right tailed p-values, calculated using the Pearson chi-squared statistic, which compares the estimated count frequency to the empirical count frequency. When the right tailed p-value is larger than 0.1 it can be said that the result provides a strong presumption that the data follows the fitted distribution. The p-values for both Poisson and negative binomial regression models have been collated, along with the measure of dispersion mentioned in the previous paragraph in Table 6.2.

It can be seen from Table 6.2 that in many cases overdispersion occurs. On average the variance is approximately 7% larger than the mean goal count. In these cases the negative binomial model, which can account for such over-dispersed behaviour, may seem

to produce a better fit.

Further evidence may be seen in Table 6.2 in the form of the resultant p-values in hypothesis tests which compare an alternative hypothesis that the data follow a negative binomial distribution to a null hypothesis that they are Poisson distributed. In those leagues highlighted in grey, the null hypothesis may be rejected in a chi-squared test at a significance level of 0.05. In numerous cases the null is not rejected and in some cases under dispersion is present, which raises the question whether the variability comes from any unseen heterogeneity or whether it is coming from another factor. The use of a more complex model is necessary to find out what is the cause of these features.

It is noteworthy that the Australian, German, Scottish and Swedish leagues all reject the null, suggesting an overdispersed nature in goal count, whilst the Brazilian, French, Italian and Mexican leagues are better fit by either a Poisson distribution or an underdispersed distribution. The English leagues are mixed in terms of over- and underdispersion, as are the Spanish leagues. Overdispersion in this case suggests leagues which have teams of more varying ability, leading to higher variability in the goal counts, and vice versa for underdispersion.

6.3.2 Goodness of Fit in the Right Tail

The initial analysis, performed in Section 6.3.1, compared the overall fit of the Poisson distribution to that of the negative binomial using the naive assumption of identically distributed data. The aim of the following process is to ascertain whether the goodness of fit is better throughout the range of the distribution using a more complex model allowing for different means for each match pairing. The right-tail fit, describing high scoring matches, is the current area of interest due to the apparent over-dispersed nature of the goal count data.

Regression is initially performed on goal count data from the English Premier League between seasons 2001/2002 - 2011/2012, using both Poisson and negative binomial models. The regression uses a multiplicative mean in both cases, as in the model proposed by Dixon and Coles (1997), as discussed in Section 4.1, and the likelihood of the Poisson model is given by equation (4.5).

The negative binomial is described by the mean, μ, and dispersion parameter, ν, as in Section 6.3. In both cases the home and away goal counts are analysed seperately and there is no dependence between home and away goals in the initial fit. Equation (6.2) can be rewritten, substituting ν^x into the numerator and denominator for numerical stability, as given by

Division	Var/Mean		$P(>\chi^2)$ Poisson		$P(>\chi^2)$ Neg. Binomial		p-value	
	Home	Away	Home	Away	Home	Away	Home	Away
aus0	1.15	1.17	0.00	0.00	0.20	0.17	0.000	0.000
aus1	1.03	1.16	0.73	0.00	0.68	0.30	0.493	0.000
bel0	1.12	1.09	0.00	0.00	0.31	0.10	0.000	0.002
bra0	0.98	1.03	0.17	0.03	0.05	0.02	1.000	0.225
bra1	0.92	0.97	0.03	0.75	0.01	0.50	1.000	1.000
brassp	0.91	1.01	0.01	0.79	0.01	0.66	1.000	0.888
eng0	1.10	1.11	0.00	0.00	0.42	0.15	0.000	0.000
eng1	0.99	1.01	0.35	0.16	0.13	0.12	1.000	0.488
eng2	0.99	1.01	0.33	0.72	0.16	0.61	1.000	0.777
eng3	1.03	0.98	0.17	0.07	0.27	0.02	0.099	1.000
eng4	1.06	1.08	0.00	0.00	0.03	0.01	0.002	0.000
engfac	1.15	1.13	0.00	0.00	0.32	0.08	0.000	0.001
englcup	1.16	1.18	0.06	0.01	0.90	0.36	0.001	0.000
fre0	0.99	1.06	0.34	0.03	0.18	0.38	1.000	0.004
fre1	0.93	0.98	0.00	0.46	0.00	0.22	1.000	1.000
ger1	1.06	1.11	0.27	0.00	0.86	0.06	0.013	0.000
ger2	1.09	1.10	0.03	0.00	0.60	0.13	0.001	0.000
ger3	1.09	1.11	0.16	0.05	0.60	0.48	0.018	0.005
ger4	1.10	1.12	0.00	0.00	0.50	0.46	0.000	0.000
ger4n	1.31	1.11	0.00	0.02	0.75	0.08	0.000	0.028
ger4s	1.15	1.14	0.00	0.03	0.00	0.26	0.004	0.006
ger4w	1.15	1.10	0.02	0.01	0.65	0.04	0.001	0.000
ire0	1.18	1.07	0.01	0.01	0.57	0.01	0.000	0.130
ita0	0.97	1.01	0.21	0.29	0.06	0.21	1.000	0.632
ita1	0.90	0.99	0.00	0.76	0.00	0.54	1.000	1.000
mex0	1.01	1.01	0.37	0.38	0.28	0.29	0.718	0.632
sco0	1.14	1.10	0.00	0.00	0.51	0.08	0.000	0.000
sco1	1.07	1.13	0.24	0.00	0.68	0.31	0.028	0.000
sco2	1.12	1.06	0.00	0.60	0.12	0.87	0.000	0.084
sco3	1.24	1.19	0.00	0.00	0.17	0.53	0.000	0.000
spa1	1.06	1.11	0.20	0.00	0.74	0.63	0.012	0.000
spa2	0.97	1.00	0.48	0.25	0.20	0.13	1.000	1.000
swe0	1.13	1.05	0.01	0.65	0.78	0.88	0.000	0.121
swe1	1.18	0.99	0.00	0.74	0.09	0.58	0.001	1.000

Table 6.2: Representation of dispersion and p-values for Poisson and Negative Binomial Regressions for Home and Away Goals

$$\Pr\{X = x\} = \frac{\Gamma(\nu + x)}{\nu^x x! \Gamma(\nu)} \frac{\nu^{\nu+x} \mu^x}{(\mu + \nu)^{\nu+x}}, \tag{6.3}$$

where

$$\frac{\Gamma(\nu + x)}{\nu^x \Gamma(\nu)} = \left(1 + \frac{x-1}{\nu}\right) \ldots \left(1 + \frac{1}{\nu}\right),$$

and

$$\frac{\nu^{\nu+x}}{(\mu+\nu)^{\nu+x}} = \left(1+\frac{\mu}{\nu}\right)^{-(\nu+x)}.$$

This numerical stability ensures that larger dispersion parameters can be used in the optimisation process.

The likelihood for the negative binomial model may be defined by

$$L(\boldsymbol{\alpha},\boldsymbol{\beta},\rho,\nu_x,\nu_y,\gamma;\boldsymbol{x},\boldsymbol{y}) = \prod_{k=1}^{N} \frac{\Gamma(\nu_x+x_k)}{\nu_x^{x_k}x_k!\Gamma(\nu_x)} \frac{\nu_x^{\nu_x+x_k}\lambda_k^{x_k}}{(\lambda_k+\nu_x)^{\nu_x+x_k}} \frac{\Gamma(\nu_y+y_k)}{\nu_y^{y_k}y_k!\Gamma(\nu_y)} \frac{\nu_y^{\nu_y+y_k}\mu_k^{y_k}}{(\mu_k+\nu_y)^{\nu_y+y_k}},$$

where ν_x and ν_y represent the respective home and away dispersion parameters.

A hypothesis test may be performed to compare an alternative hypothesis of the negative binomial model to a null of the Poisson model. However, the fitted negative binomial model is identical to an equivalent Poisson model as $\hat{\sigma} = 0$. Therefore, the negative binomial model provides no improvement in terms of goodness of fit and shall not be further compared.

To analyse the goodness of fit of the Poisson distribution in the right tail, the probability that, above a certain threshold, the observed home and away goals resulted from a Poisson distribution can be estimated. This may be done using the following process.

Both the expected and observed values are reordered to reflect the expected values in ascending order. A value c above which to analyse can now be chosen, discarding all expected values, and their observed counterparts, with values less than or equal to c. The expected values and observations are then pooled according to the ordered expected values to ensure that each pool is greater than 10, this accounts for the low nature of the counts. The Pearson's residuals are then evaluated and summed for all pools $1,\ldots,n$, as shown by

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{6.4}$$

to obtain a test statistic, χ^2.

An empirical sampling distribution of goal count (home or away) was evaluated under the null model using the pooled expected values, with a sample size of 200 for each pooled expectation. The observed test statistic was then compared to the sampling distribution

to obtain the right tailed p-value. This process was repeated for a range of values for c.

Figures 6.4 and 6.5 show the test statistic and right tailed p-values as functions of c for Poisson regression performed on the English Premiership inclusive of seasons 2001/2002 to 2011/2012 comparing a model using seperate home advantage parameters γ_s for each season s, where $\mathrm{E}(X_{i,j}) = \alpha_i \beta_j \gamma_s$ and a model using constant home advantage over all seasons, for home and away goals respectively.

It can be seen from Figures 6.4 and 6.5 that, although erratic, the p-value is relatively high over most of the tested values of c, suggesting that the null Poisson model should not be rejected, for both home and away goals. It should also be noted, that adding a seasonally varying home advantage does not improve the fit in the right tail. Further investigation is therefore required in an attempt to better explain and model the difference between the mean and variance of goal counts.

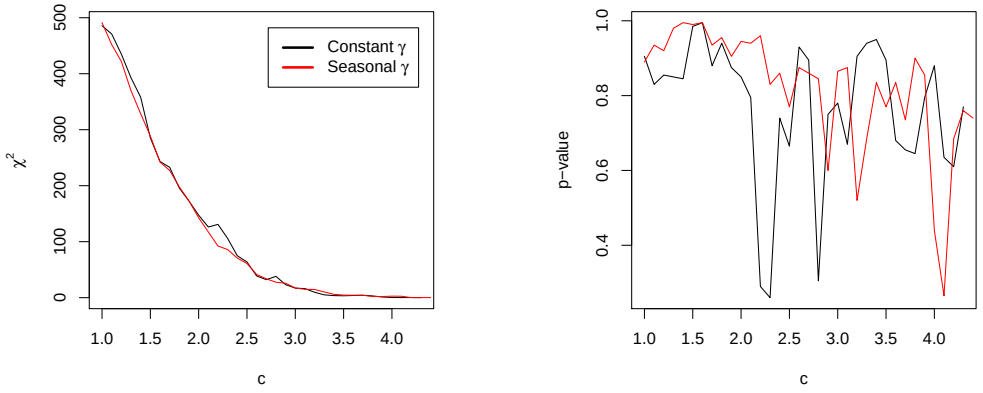


Figure 6.4: Constant and seaonal home advantage parameter models of home goals (Left) χ^2 test statistic and (Right) right-tailed p-value as functions of c.

6.4 Censored Likelihood

Further investigation is required to account for discrepancies in the model distribution: overdispersion and error in the fitted Poisson probability mass function (pmf) relative to the empirical probabilities are present. However, the negative binomial regression model regressed to a homogeneous Poisson model, where $\sigma = 0$, so a more complex approach is required. Censoring may be employed to account for the low confidence in the model above a certain level.

If we have a model and choose to censor aspects of the model it is helpful to know

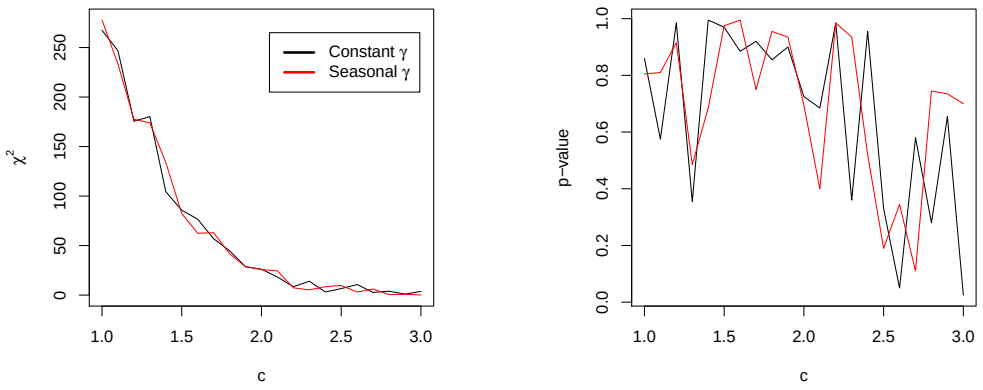


Figure 6.5: Constant and seaonal home advantage parameter models of away goals (Left) χ^2 test statistic and (Right) right-tailed p-value as functions of c

what information is lost if in fact the model was correct. To estimate this loss of information the variance of parameter estimates can be compared with those of an uncensored model. The following sections give the derivations of the variance of a single mean estimate under the uncensored and censored Poisson models, to allow their comparison for different levels of censoring.

6.4.1 Uncensored Poisson Likelihood

The likelihood for a Poisson (λ) distributed independent and identically distributed variable X is defined by

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}.$$

Then the log likelihood is given by

$$\ell(\lambda) = \sum_{i=1}^{n} x_i \log(\lambda) - n\lambda - \sum_{i=1}^{n} \log(x_i!).$$

The first and second derivatives are then

$$\ell'(\lambda) = \sum_{i=1}^{n} \frac{x_i}{\lambda} - n,$$
$$\ell''(\lambda) = -\sum_{i=1}^{n} \frac{x_i}{\lambda^2}.$$

The expected information can be calculated by

$$\begin{aligned}\mathrm{I_E}(\lambda) &= \mathrm{E}\left[-\ell''(\lambda)\right] \\ &= \frac{1}{\lambda^2}\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{n}{\lambda^2}\mathrm{E}(X) \\ &= \frac{n}{\lambda},\end{aligned}$$

which can be used to calculate the variance

$$\begin{aligned}\mathrm{var}\left(\hat{\lambda}\right) &= \mathrm{I_E}(\lambda)^{-1}. \\ &= \frac{\lambda}{n}\end{aligned}$$

6.4.2 Independent Poisson Likelihood, Censoring Above c

The likelihood function for a Poisson distributed variable X_c with censored observations greater than a known value c is defined by

$$L(\lambda) = \prod_{i=1}^{m} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \prod_{i=m+1}^{n} \left(1 - \sum_{u=0}^{c} \frac{\lambda^u e^{-\lambda}}{u!}\right),$$

when $x_1, \ldots, x_m$ are all less than or equal to c, and $x_{m+1}, \ldots, x_n$ are all greater than c.

Taking the logarithm of the above gives the log-likelihood

$$\ell(\lambda) = \sum_{i=1}^{m} x_i \log(\lambda) - m\lambda - \sum_{i=1}^{m} \log(x_i!) + (n-m)\log\left(1 - \sum_{u=0}^{c} \frac{\lambda^u e^{-\lambda}}{u!}\right).$$

The first and second derivatives of this log-likelihood are calculated as

$$\begin{aligned}\ell'(\lambda) &= \sum_{i=1}^{m} \frac{x_i}{\lambda} - m + (n-m)\left(\frac{\lambda^c e^{-\lambda}/c!}{1 - \sum_{u=0}^{c} \lambda^u e^{-\lambda}/u!}\right), \\ \ell''(\lambda) &= \sum_{i=1}^{m} -\frac{x_i}{\lambda^2} + (n-m)\frac{d}{d\lambda}\left(\frac{\lambda^c e^{-\lambda}/c!}{1 - \sum_{u=0}^{c} \lambda^u e^{-\lambda}/u!}\right).\end{aligned}$$

and

$$\frac{d}{d\lambda}\left(\frac{\lambda^c e^{-\lambda}/c!}{1-\sum_{u=0}^{c}\lambda^u e^{-\lambda}/u!}\right) = \frac{\frac{e^{-\lambda}(c-\lambda)\lambda^{c-1}}{c!}\left(1-\sum_{u=0}^{c}\frac{\lambda^u e^{-\lambda}}{u!}\right)-\left(\frac{\lambda^c e^{-\lambda}}{c!}\right)^2}{\left(1-\sum_{u=0}^{c}\frac{\lambda^u e-\lambda}{u!}\right)^2}$$
$$= \frac{e^{-\lambda}(c-\lambda)\lambda^{c-1}}{c!\left(1-\sum_{u=0}^{c}\frac{\lambda^u e^{-\lambda}}{u!}\right)} - \left(\frac{\lambda^c e^{-\lambda}}{c!\left(1-\sum_{u=o}^{c}\frac{\lambda^u e^{-\lambda}}{u!}\right)}\right)^2.$$

The expected information is then given by

$$\mathrm{I_E}(\lambda_0) = \mathrm{E}\left[-\ell''(\lambda_0)\right]$$
$$= \mathrm{E}\left[\sum_{i=1}^{m}\frac{x_i}{\lambda_0^2}\right] + \mathrm{E}\left\{(m-n)\left[\frac{e^{-\lambda}(c-\lambda)\lambda^{c-1}}{c!\left(1-\sum_{u=0}^{c}\frac{\lambda^u e^{-\lambda}}{u!}\right)} - \left(\frac{\lambda^c e^{-\lambda}}{c!\left(1-\sum_{u=0}^{c}\frac{\lambda^u e^{-\lambda}}{u!}\right)}\right)^2\right]\right\}.$$

As m is the number of complete observations out of a total n observations and so $m \sim \text{Binomial}(n,p)$, where

$$p = \mathrm{P}(X_c \leq c) = \sum_{u=0}^{c}\frac{\lambda^u e^{-\lambda}}{u!}$$

and

$$\mathrm{E}(m) = np = n\sum_{u=0}^{c}\frac{\lambda^u e^{-\lambda}}{u!}.$$

This property can be used to derive an expression for $\mathrm{E}\left(\sum_{i=1}^{m}x_i\right)$, as given by

$$\mathrm{E}\left(\sum_{i=1}^{m}x_i\right) = \mathrm{E}(m)\,\mathrm{E}(X_c),$$

where

$$\mathrm{E}(X_c) = \sum_{i=0}^{c} i\mathrm{P}(X=i|X\leq c)$$
$$= \frac{\sum_{i=1}^{c} i\lambda^i e^{-\lambda}/i!}{\sum_{j=0}^{c}\lambda^j e^{-\lambda}/j!}.$$

The expected information is then given by

$$
\begin{aligned}
\mathrm{I_E}(\lambda) =& \frac{n}{\lambda^2}\left(\sum_{u=0}^{c}\frac{\lambda^u e^{-\lambda}}{u!}\right)\frac{\sum_{i=1}^{c} i\lambda^i e^{-\lambda}/i!}{\sum_{j=0}^{c}\lambda^j e^{-\lambda}/j!} \\
&+\left(n\sum_{u=0}^{c}\frac{\lambda^u e^{-\lambda}}{u!}-n\right)\left[\frac{e^{-\lambda}(c-\lambda)\lambda^{c-1}}{c!\left(1-\sum_{u=0}^{c}\frac{\lambda^u e^{-\lambda}}{u!}\right)}-\left(\frac{\lambda^c e^{-\lambda}}{c!\left(1-\sum_{u=0}^{c}\frac{\lambda^u e^{-\lambda}}{u!}\right)}\right)^2\right] \\
&=\frac{n}{\lambda^2}\sum_{i=1}^{c}\frac{i\lambda^i e^{-\lambda}}{i!} \\
&+\left(n\sum_{u=0}^{c}\frac{\lambda^u e^{-\lambda}}{u!}-n\right)\left[\frac{e^{-\lambda}(c-\lambda)\lambda^{c-1}}{c!\left(1-\sum_{u=0}^{c}\frac{\lambda^u e^{-\lambda}}{u!}\right)}-\left(\frac{\lambda^c e^{-\lambda}}{c!\left(1-\sum_{u=0}^{c}\frac{\lambda^u e^{-\lambda}}{u!}\right)}\right)^2\right].
\end{aligned}
$$

The expected information can then be inverted to derive the variance of $\hat{\lambda}_c$ the mle of λ under censoring at c.

Figures 6.6 and 6.7 show the variance calculated with uncensored and censored likelihoods and a comparative ratio of the two for different levels of censoring and different values of λ. It can be seen that the ratio of the variances rapidly decreases after the point of censoring. As the value of c decreases the information lost at higher values of λ become increasingly great, suggesting that caution should be used when interpreting estimates resulting from low values of c.

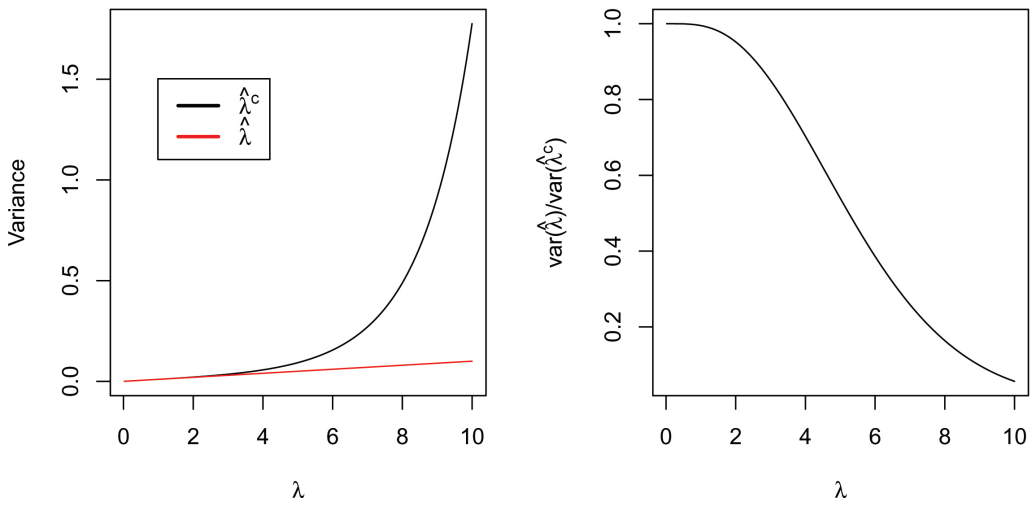


Figure 6.6: (Left) Variance of Poisson parameter estimated from an uncensored likelihood, giving $\hat{\lambda}$, and a likelihood censored above $c = 3$, giving $\hat{\lambda}^c$. (Right) Ratio of variances, $\mathrm{var}(\hat{\lambda})/\mathrm{var}(\hat{\lambda}^c)$. In each subplot, values are plotted against the true value, λ.

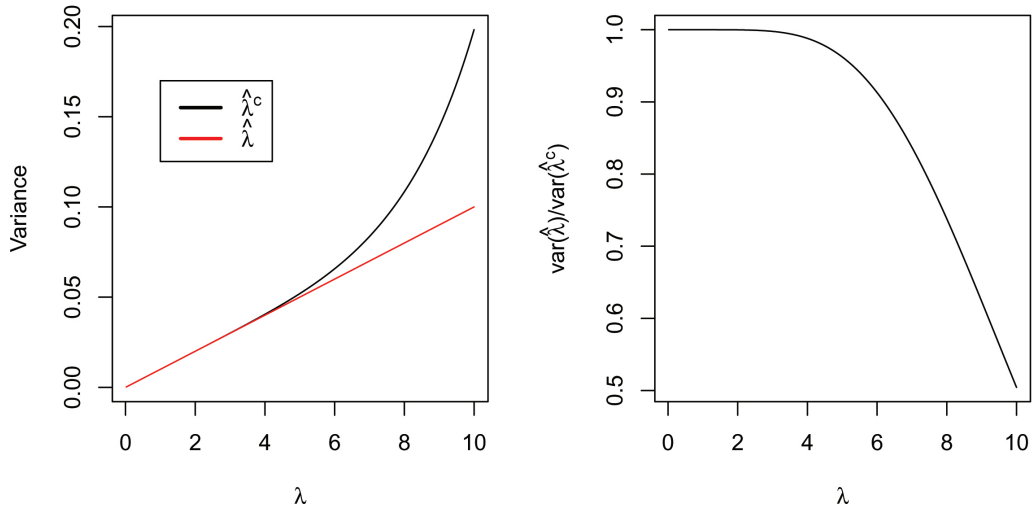


Figure 6.7: (Left) Variance of Poisson parameter estimated from an uncensored likelihood, giving $\hat{\lambda}$, and a likelihood censored above $c = 7$, giving $\hat{\lambda}^c$. (Right) Ratio of variances, $\text{var}(\hat{\lambda})/\text{var}(\hat{\lambda}^c)$. In each subplot, values are plotted against the true value, λ.

6.4.3 Dependent Joint Poisson Likelihood, Censoring Above c

We now consider a bivariate Poisson distribution (X_k, Y_k) for home and away goals in a match k, with means λ_k and μ_k respectively and some dependence as in the Dixon and Coles model (4.3). To assess the effect of censoring goal counts over c for either team on the variance of estimators in censored and uncensored cases a numerical approach must be employed. The resultant characteristics of the loss of efficiency may then be compared to the relationships found in Sections 6.4.2 for the tested values of c.

By simulating goal count data under the Dixon and Coles model (4.3) and fitting with the correct model and a censored model, we can estimate the variance of parameter estimates in each case. This is shown for 1000 observations ($n = 1000$) in Figures 6.8 and 6.9 for parameter values between 0.5 and 10 at a resolution of 0.5 and values of c equal to 3 and 7 respectively. Each plot includes the equivalent analytically obtained expression for the independent case, as in Section 6.4.2, displayed as a dotted line. It should be noted that 5 is approximately the limit of the Poisson mean parameter values encountered when regressing real data.

It can be seen from Figures 6.8 and 6.9 that the efficiency of censoring the Dixon and Coles model (4.3) follows similar features of increasing variance under the censored model as that seen for the independent goals case as in Section 6.4.2. This is consistent for both values of c tested, suggesting that the inclusion of an independence function and a joint likelihood model describing home and away goals, as in the Dixon and Coles model (4.3),

does not affect the increase in variance (interpreted as a loss of information) when censoring.

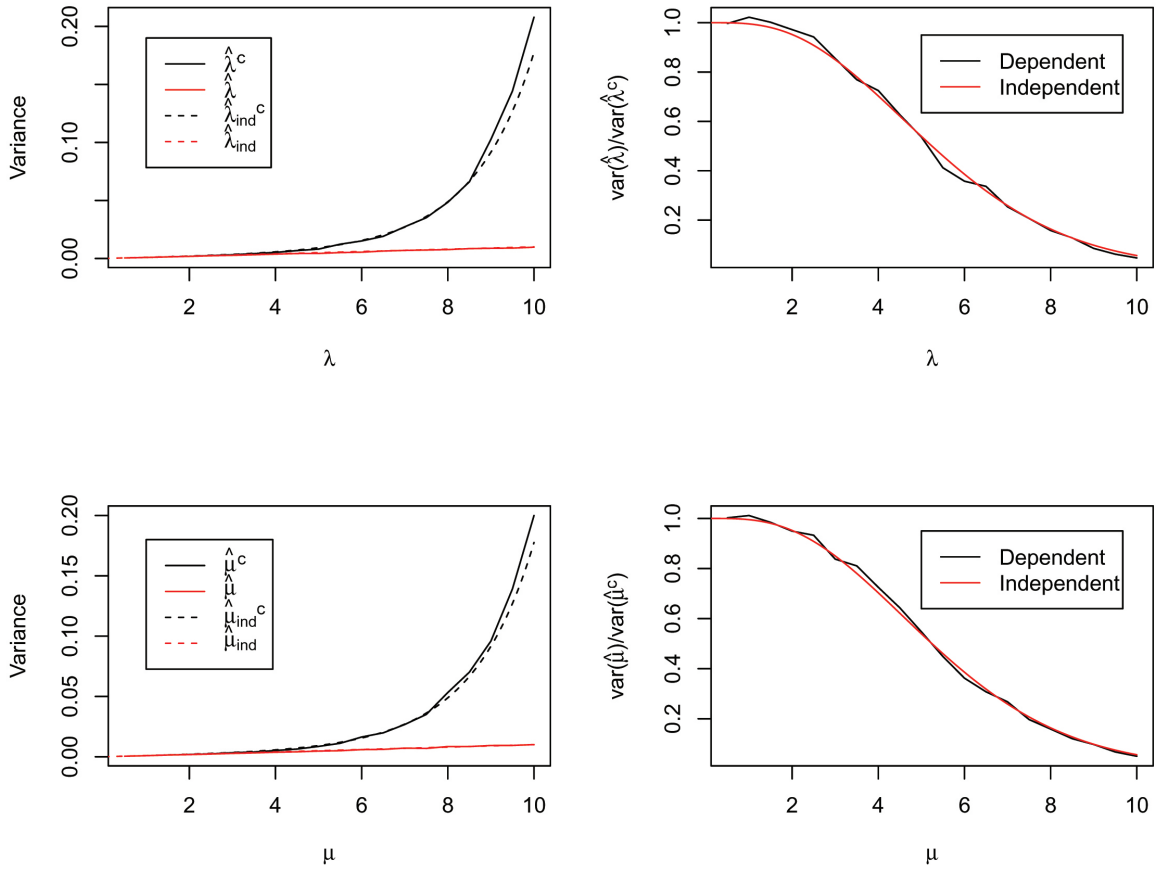


Figure 6.8: (Left) Variance of Poisson parameter estimates calculated from an uncensored likelihood which gives (top) $\hat{\lambda}$ describing home goals and (bottom) $\hat{\mu}$ describing away goals, and a likelihood censored above $c = 3$ and $n = 1000$, which gives (top) $\hat{\lambda}^c$ and (bottom) $\hat{\mu}^c$. Dotted lines describe the equivalent independent models. (Right) Ratio of variances for the dependent and independent models.

6.4.4 Goodness of Fit

It is prudent to check the goodness of fit of the censored model at varying levels of censoring for both home and away goals, based on the Dixon and Coles model (4.3), compared to the null uncensored model. This will allow the analysis of whether either or both home and away goals would benefit from the right tail beyond a certain value, c_x for home goals and c_y for away goals, being modelled in some alternative fashion.

The test statistic prescribed in Section 6.3.2, as given by equation (6.4), may be calculated for the censored model and compared to that obtained under the uncensored

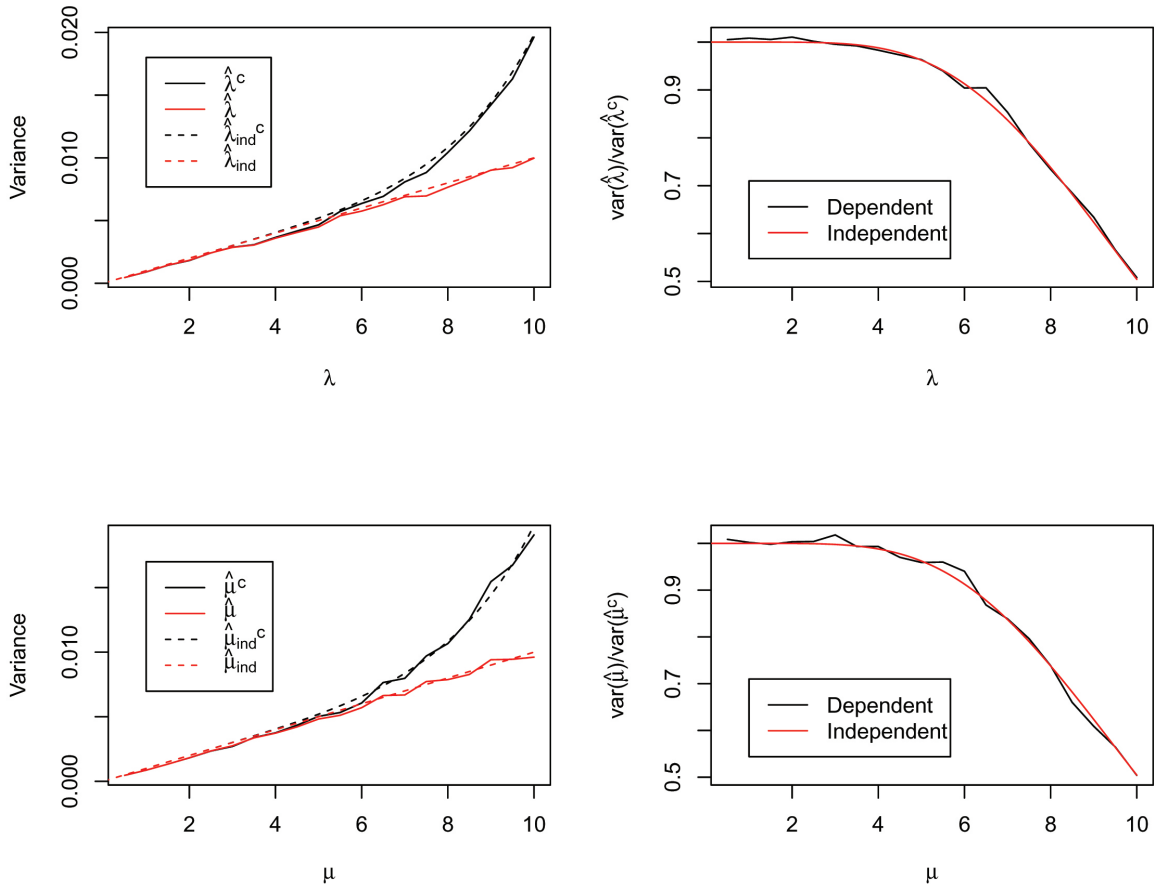


Figure 6.9: (Left) Variance of Poisson parameter estimates calculated from an uncensored likelihood which gives (top) $\hat{\lambda}$ describing home goals and (bottom) $\hat{\mu}$ describing away goals, and a likelihood censored above $c = 7$ and $n = 1000$, which gives (top) $\hat{\lambda}^c$ and (bottom) $\hat{\mu}^c$. Dotted lines describe the equivalent independent models. (Right) Ratio of variances for the dependent and independent models.

Dixon and Coles model (4.3). Larger pools of 100 counts shall be used in this case as only the overall goodness of fit is to be calculated.

A way of censoring the right tail is to shift the probability mass from counts greater than c onto $c+1$, to represent the portion of counts which are greater than the threshold value c, as given for home goals by

$$X_c = \min\left(X, c_x + 1\right).$$

Censoring $X > c_x$ and $Y > c_y$ in this way, it follows that X_c has a probability mass function (and similarly for Y) defined by

$$P_{X_c}(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & \text{if } x = 0, \ldots, c_x, \\ 1 - \sum_{u=0}^{c_x} \frac{\lambda^u e^{-\lambda}}{u!}, & x = c_x + 1. \end{cases} \tag{6.5}$$

The censored likelihood based on the Dixon and Coles model (4.3) is then given by

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \rho, \gamma, c_x, c_y; \boldsymbol{x}, \boldsymbol{y}) = \prod_{k=1}^{N} & \tau_{\lambda_k, \mu_k}(x_k, y_k) \\ & \left(\frac{\lambda_k^{x_k} e^{-\lambda_k}}{x_k!} \right)^{\mathbb{1}\{x_k \leq c_x\}} \left(1 - \sum_{u=0}^{c_x} \frac{\lambda_k^u e^{-\lambda_k}}{u!} \right)^{\mathbb{1}\{x_k > c_x\}} \\ & \left(\frac{\mu_k^{y_k} e^{-\mu_k}}{y_k!} \right)^{\mathbb{1}\{y_k \leq c_y\}} \left(1 - \sum_{u=0}^{c_y} \frac{\mu_k^u e^{-\mu_k}}{u!} \right)^{\mathbb{1}\{y_k > c_y\}}, \end{aligned}$$

where k denotes the match pairing of home team i and away team j.

To calculate the goodness of fit statistic, a well defined expected value must first be derived for the censored model, as given by

$$\mathbb{E}(X_c) = \sum_{x=0}^{c_x} \frac{x \lambda^x e^{-\lambda}}{x!} + (c_x + 1) \left(1 - \sum_{u=0}^{c_x} \frac{\lambda^u e^{-\lambda}}{u!} \right).$$

As the censored probability mass is not removed but placed on index $c+1$, the test statistic is obtained for the censored model by comparing the expected values to observations which are adjusted to reflect the censoring. Specifically, the observations which contain greater count values than c are given the value of $c + 1$, as the model is not attempting to be exact above this censoring value.

Figure 6.10 shows the percentage increase in the χ^2 test statistic, defined in equation (6.4), between that calculated under the censored model, χ_c^2, and that calculated under the uncensored model, χ_u^2, for home and away goal counts combined, as given by

$$\frac{\chi_c^2 - \chi_u^2}{\chi_u^2} \times 100,$$

where a negative value indicates a better fit. Negative values are shown in green and positive in red, with a gradient of colour in between. The values of $c_x = 1$ and $c_y = 6$ give the best overall improvement in goodness of fit under the conditions of the test. This suggests that the censoring of home goals causes an improvement on the fit of data at lower values, though censoring the away goals has little effect.

Figure 6.11 shows this measure broken down for home and away counts seperately. Regarding home goals, the censoring levels which show the greatest improvement in the test statistic, $c_x = 1$ and $c_y = 4$. Censoring values must be considered together in this way due to the low scoring dependence considered in the Dixon and Coles model (4.3). For away goals, the censoring levels which give the greatest improvement are the same as those for combined, however, there is an improvement for almost all combinations of c_x and c_y. Considered together, this may suggest that the distribution describing home goals would benefit most from a change in modelling technique to aid the fit in the right tail.

		c_y					
		1	2	3	4	5	6
c_x	1	-9.47%	-2.12%	-9.59%	-6.34%	-8.46%	-11.09%
	2	-4.87%	-5.34%	-3.86%	-3.30%	-3.16%	-3.31%
	3	-3.14%	-2.60%	-1.73%	-1.25%	-1.86%	-1.64%
	4	-1.41%	-0.90%	-0.75%	0.03%	-0.25%	-0.19%
	5	-0.72%	-0.55%	-0.09%	0.11%	-0.17%	0.01%
	6	-1.37%	-0.63%	-0.09%	0.06%	0.07%	0.03%
	7	-1.96%	-0.38%	0.03%	0.15%	-0.03%	0.06%
	8	-2.22%	-0.32%	-0.17%	0.13%	0.11%	0.02%
	9	-1.53%	-0.32%	-0.21%	0.11%	0.08%	0.00%

min max

Figure 6.10: Percentage difference in χ^2 test statistic between the censored model defined in equation (6.5) and the standard Poisson model (including dependency function) over home and away goal counts

6.5 Threshold Mixture Regression

In Section 6.4.4 a goodness of fit test was used to determine whether censoring home and away goals to some threshold would increase the goodness of fit below the threshold level. It was determined that censoring did increase the goodness of fit when comparing to the null Dixon and Coles model (4.3). Following this result, it could be hypothesised that some non-Poisson class of left truncated count distribution could be used in conjunction with a right truncated Poisson distribution to create a more effective model.

Threshold mixture regression refers to the use of different distributions seperated by a threshold level, c. In the simplest two distribution case, the portion to the right of the

Home

c_x \ c_y	1	2	3	4	5	6
1	-17.19%	-18.02%	-18.70%	-21.83%	-20.04%	-20.56%
2	-6.15%	-7.00%	-7.88%	-7.39%	-8.03%	-8.15%
3	-3.67%	-2.60%	-2.28%	-2.48%	-2.61%	-2.55%
4	-1.77%	0.26%	0.19%	0.22%	0.07%	0.01%
5	-0.80%	-0.01%	0.61%	0.37%	0.20%	0.15%
6	-1.29%	0.40%	0.59%	0.06%	0.12%	0.04%
7	-1.67%	-0.23%	0.49%	0.46%	0.04%	0.07%
8	-1.41%	-0.17%	0.56%	0.33%	0.27%	0.07%
9	-1.54%	-0.17%	0.15%	0.26%	0.24%	0.00%

Away

c_x \ c_y	1	2	3	4	5	6
1	-3.03%	-7.58%	-12.51%	-9.83%	-12.49%	-17.60%
2	-5.29%	-3.62%	-3.71%	-4.68%	-4.65%	-4.63%
3	-10.06%	-2.85%	-2.11%	-0.67%	-2.45%	-2.37%
4	-8.95%	-2.46%	-1.17%	-0.45%	-0.87%	0.01%
5	-9.25%	-2.08%	-1.16%	-0.12%	-0.52%	-0.30%
6	-9.43%	-1.94%	-1.39%	-0.37%	-0.20%	-0.34%
7	-9.94%	-2.10%	-1.03%	-0.61%	-0.28%	-0.12%
8	-9.98%	-2.20%	-1.20%	-0.35%	-0.46%	-0.22%
9	-9.72%	-2.23%	-1.06%	-0.33%	-0.41%	0.00%

Figure 6.11: Percentage difference in χ^2 test statistic between the censored model defined in equation (6.5) and the standard Poisson model (including dependency function) (top) home goal count, X, (bottom) away goal count, Y.

threshold is adjusted to match the missing probability mass of the initial distribution.

Modelling the goal counts as a threshold mixture of a Poisson and another discrete distribution which allows for overdispersion may increase the accuracy of predictions by increasing the goodness of fit in the right tail. Two such right tail models which can be used in conjunction with a Poisson body are the negative binomial distribution and the geometric distribution. The following sections will derive the threshold mixture models for these cases.

6.5.1 Poisson-Geometric

Section 6.4.4 showed that the upper tail of the emprical goal count distribution is not well captured by a Poisson distribution, therefore, it is prudent to seek a mixture model that better fits the right tail, one example being the Poisson-Geometric threshold mixture model. The Dixon and Coles model (4.3) uses the Poisson rate parameter to determine the mean. To allow ease of comparison a similar parameterisation describing the mean through attack, defence and home advantage, a model considering the mean must be saught. The pmf of the geometric distribution is typically written as

$$p\left(n\right) = p(1-p)^n, \qquad n = 0, 1, \ldots$$

However, it can be reparameterised so that the mean, λ, is the parameter (as with the Poisson distribution) as given by

$$\lambda = \frac{1-p}{p}.$$

Therefore,

$$p = \frac{1}{\lambda + 1},$$

so

$$p\left(n\right) = \frac{1}{\lambda+1}\left(\frac{\lambda}{\lambda+1}\right)^n, \qquad n = 0, 1, \ldots$$

Using this definition of a geometric pmf, a threshold-mixed Poisson-Geometric model for goal counts may be implemented, where counts in the range $0, \ldots, c$ follow a Poisson distribution and from $c+1, \ldots, \infty$ follow a geometric distribution, adjusted to match the remaining probability mass. The condition $c \geq 1$ is introduced to ensure a mixture with a Poisson model describing the body. This also allows the independence function for low scoring games as given in Dixon and Coles (1997) to be used. To ensure that the probabilities sum to 1, the geometric portion is adjusted by

$$h\left(x\right) = \frac{g\left(x\right)\left[1 - \sum_{u=0}^{c} f\left(u\right)\right]}{\sum_{u=c+1}^{\infty} g\left(u\right)}, \qquad \text{for} \quad x > c$$

where

$$f\left(x\right) = \frac{\lambda^x e^{-\lambda}}{x!}$$

and

$$g(x) = \left(\frac{1}{\lambda+1}\right)\left(\frac{\lambda}{\lambda+1}\right)^{x},$$

so

$$\sum_{u=c+1}^{\infty} g(u) = \left(\frac{\lambda}{\lambda+1}\right)^{c+1},$$

and

$$h(x) = \frac{\left(\frac{\lambda}{\lambda+1}\right)^{x-c-1}\left(1-\sum_{u=0}^{c}\frac{\lambda^{u}e^{-\lambda}}{u!}\right)}{\lambda+1}, \quad x > c+1.$$

Now, c_x and c_y are parameters which are used to describe the threshold level for home and away goals, i.e. they are considered as unknown and treated as parameters in the likelihood inference. The pmf describing home goal count probabilities may then be given by

$$P_{X_c}(x) = \begin{cases} \frac{\lambda^{x}e^{-\lambda}}{x!}, & \text{if } x = 0,\ldots,c_x, \\ \left[\frac{\left(\frac{\lambda}{\lambda+1}\right)^{x-c-1}\left(1-\sum_{u=0}^{c}\frac{\lambda^{u}e^{-\lambda}}{u!}\right)}{\lambda+1}\right], & \text{if } x > c_x. \end{cases}$$

This may be regarded as a threshold mixture of a Poisson and a Geometric variable, i.e.,

$$P_{X_c}(x) = \left[\frac{\lambda^{x}e^{-\lambda}}{x!}\right]^{\mathbb{1}_{\{x\le c\}}}\left[\frac{\left(\frac{\lambda}{\lambda+1}\right)^{x-c-1}\left(1-\sum_{u=0}^{c}\frac{\lambda^{u}e^{-\lambda}}{u!}\right)}{\lambda+1}\right]^{\mathbb{1}_{\{x>c\}}},$$

and the expected value of X is given by $\mathrm{E}(X) = \sum_{x=0}^{x=\infty} xP_{X_c}(x)$.

The likelihood is then defined by

$$L\left(\boldsymbol{\alpha},\boldsymbol{\beta},\rho,\gamma,c_x,c_y;\boldsymbol{x},\boldsymbol{y}\right)=\prod_{k=1}^{N}\tau_{\lambda_k,\mu_k}(x_k,y_k)$$
$$\left[\frac{\lambda_k^{x_k}e^{-\lambda_k}}{x_k!}\right]^{\mathbb{1}_{\{x_k\leq c_x\}}}\left[\frac{\left(\frac{\lambda_k}{\lambda_k+1}\right)^{x_k-c_x-1}\left(1-\sum_{u=0}^{c_x}\frac{\lambda_k^u e^{-\lambda_k}}{u!}\right)}{\lambda_k+1}\right]^{\mathbb{1}_{\{x_k>c_x\}}}$$
$$\left[\frac{\mu_k^{y_k}e^{-\mu_k}}{y_k!}\right]^{\mathbb{1}_{\{y_k\leq c_y\}}}\left[\frac{\left(\frac{\mu_k}{\mu_k+1}\right)^{y_k-c_y-1}\left(1-\sum_{u=0}^{c_y}\frac{\mu_k^u e^{-\mu_k}}{u!}\right)}{\mu_k+1}\right]^{\mathbb{1}_{\{y_k>c_y\}}}.$$

Under testing, the mixture regressed to a Poisson variable, with a value of $\hat{c}_x$ and $\hat{c}_y$ equal to the highest respective home and away observations, which shows that the mixture does not increase the accuracy of estimates, for all leagues tested (see Section 5.2 for details).

6.5.2 Poisson-Negative Binomial

The same procedure as the previous section may be used to produced a Poisson-negative binomial mixture model, the pmf of this model may be defined as

$$P_{X_c}\left(x\right)=\begin{cases}\frac{\lambda^x e^{-\lambda}}{x!}, & \text{if } x=0,\ldots,c_x,\\ \frac{\Gamma(\nu_x+x)}{\nu_x^x x!\Gamma(\nu_x)}\frac{\nu_x^{\nu_x+x}\lambda^x}{(\lambda+\nu_x)^{\nu_x+x}}\left(1-\sum_{u=0}^{c_x}\frac{\lambda^u e^{-\lambda}}{u!}\right) & \\ \left(1-\sum_{u=0}^{c_x}\frac{\Gamma(\nu_x+u)}{\nu_x^u u!\Gamma(\nu_x)}\frac{\nu_x^{\nu_x+u}\lambda^u}{(\lambda+\nu_x)^{\nu_x+u}}\right)^{-1}, & \text{if } x>c_x.\end{cases}$$

This variable may be regarded as a mixture of a Poisson and a Negative Binomial variable, i.e.

$$P_{X_c}\left(x\right)=\left[\frac{\lambda^x e^{-\lambda}}{x!}\right]^{\mathbb{1}_{\{x\leq c\}}}\left[\frac{\Gamma\left(\nu_x+x\right)}{\nu_x^x x!\Gamma\left(\nu_x\right)}\frac{\nu_x^{\nu_x+x}\lambda^x}{\left(\lambda+\nu_x\right)^{\nu_x+x}}\left(1-\sum_{u=0}^{c_x}\frac{\lambda^u e^{-\lambda}}{u!}\right)\right.$$
$$\left.\left(1-\sum_{u=0}^{c_x}\frac{\Gamma\left(\nu_x+u\right)}{\nu_x^u u!\Gamma\left(\nu_x\right)}\frac{\nu_x^{\nu_x+u}\lambda^u}{\left(\lambda+\nu_x\right)^{\nu_x+u}}\right)^{-1}\right]^{\mathbb{1}_{\{x>c\}}},$$

and the expected value of X is again given by $\mathrm{E}(X)=\sum_{x=0}^{x=\infty}xP_{X_c}(x)$.

The likelihood may then be defined by

$$L\left(\boldsymbol{\alpha},\boldsymbol{\beta},\rho,\gamma,c_x,c_y,\nu_x,\nu_y;\boldsymbol{x},\boldsymbol{y}\right)=\prod_{k=1}^{N}\tau_{\lambda_k,\mu_k}(x_k,y_k)$$
$$\left[\frac{\lambda_k^{x_k}e^{-\lambda_k}}{x_k!}\right]^{\mathbb{1}\{x_k\leq c_x\}}\left[\frac{\Gamma\left(\nu_x+x_k\right)}{\nu_x^{x_k}x_k!\Gamma\left(\nu_x\right)}\frac{\nu_x^{\nu_x+x_k}\nu_x^{x_k}}{\left(\lambda_k+\nu_x\right)^{\nu_x+x_k}}\left(1-\sum_{u=0}^{c_x}\frac{\lambda_k^u e^{-\lambda_k}}{u!}\right)\right.$$
$$\left.\left(1-\sum_{u=0}^{c_x}\frac{\Gamma\left(\nu_x+u\right)}{\nu_x^u u!\Gamma\left(\nu_x\right)}\frac{\nu_x^{\nu_x+u}\lambda_k^u}{\left(\lambda_k+\nu_x\right)^{\nu_x+u}}\right)^{-1}\right]^{\mathbb{1}\{x_k>c_x\}}$$
$$\left[\frac{\mu_k^{x_k}e^{-\mu_k}}{y_k!}\right]^{\mathbb{1}\{y_k\leq c_y\}}\left[\frac{\Gamma\left(\nu_y+y_k\right)}{\nu_y^{y_k}y_k!\Gamma\left(\nu_y\right)}\frac{\nu_y^{\nu_y+y_k}\mu^{y_k}}{\left(\mu_k+\nu_y\right)^{\nu_y+y_k}}\left(1-\sum_{u=0}^{c_y}\frac{\mu_k^u e^{-\mu_k}}{u!}\right)\right.$$
$$\left.\left(1-\sum_{u=0}^{c_y}\frac{\Gamma\left(\nu_y+u\right)}{\nu_y^u u!\Gamma\left(\nu_y\right)}\frac{\nu_y^{\nu_y+u}\mu_k^u}{\left(\mu_k+\nu_y\right)^{\nu_y+u}}\right)^{-1}\right]^{\mathbb{1}\{y_k>c_x\}},$$

Similar to the Poisson-geometric threshold mixture model, the MLEs $\hat{c}_x$ and $\hat{c}_y$ were equal to the highest value of goal count and the model regressed to a standard Poisson.

6.6 Poisson-Poisson Threshold Mixture Regression

To address the issue of error in low and high goal count probabilities as discussed in Section 6.4, a more complex model must be implemented. Consider the expected home and away goals counts as given by equation (4.2), under this definition the expected away goal count in a match between home team i and away team j would be adjusted by a multiple γ to give the expected home goal count when played at team j's home ground. A simple two team league will be used to illustrate the fact that this current specification of home advantage parameter may be leading to a reduction in goodness of fit in the right tail. This will be followed by the introduction of a Poisson-Poisson threshold mixture model which adds an additional parameter to describe the home advantage in high scoring games. A further parameter may be added to aid the modelling of high away goal counts. The model will be verified using hypothesis testing and a goodness of fit test.

The previous threshold mixture models had an expected value which remained unchanged either side of the threshold. To create a Poisson-Poisson threshold mixture model the expected value of the distribution contributing to the probability mass function when $x > c_x$ or $y > c_y$ may be adjusted by a multiplicative factor, m_x and m_y respectively, allowing the use of a bivariate Poisson threshold mixture model, which is defined here.

The pmf of the Poisson-Poisson model is defined as

$$P_{X_c}(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & \text{if } x = 0, \ldots, c_x, \\ \left[\frac{(m_x\lambda)^x e^{-m_x\lambda}}{x!}\left(1 - \sum_{u=0}^{c_x}\frac{\lambda^u e^{-\lambda}}{u!}\right)\right. & \\ \left.\left(1 - \sum_{u=0}^{c_x}\frac{(m_x\lambda)^u e^{-m_x\lambda}}{u!}\right)^{-1}\right], & \text{if } x > c_x. \end{cases} \tag{6.6}$$

The resulting expected value of X_c is

$$\begin{aligned} \mathbb{E}(X) =& \sum_{x=0}^{c_x} xP(X = x) + \sum_{x=c_x+1}^{\infty} xP(X = x) \\ =& \sum_{x=0}^{c_x} \frac{x\lambda^x e^{-\lambda}}{x!} + \sum_{x=c_x+1}^{\infty} \left[\frac{x(m_x\lambda)^x e^{-m_x\lambda}}{x!}\right. \\ & \left.\left(1 - \sum_{u=0}^{c_x}\frac{\lambda^u e^{\lambda}}{u!}\right)\left(1 - \sum_{u=0}^{c_x}\frac{(m_x\lambda)^u e^{-m_x\lambda}}{u!}\right)^{-1}\right] \end{aligned}$$

and a joint likelihood describing home and away goal counts is given by

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \rho, \gamma, c_x, c_y, m_x, m_y; \boldsymbol{x}, \boldsymbol{y}) =& \prod_{k=1}^{N} \tau_{\lambda_k,\mu_k}(x_k, y_k) \\ & \left(\frac{\lambda_k^{x_k} e^{-\lambda_k}}{x_k!}\right)^{\mathbb{1}_{\{x_k \le c_x\}}} \left(\frac{\mu_k^{y_k} e^{-\mu_k}}{y_k!}\right)^{\mathbb{1}_{\{y_k \le c_y\}}} \\ & \left[\frac{(m_x\lambda_k)^{x_k} e^{-m_x\lambda_k}}{x_k!}\left(1 - \sum_{u=0}^{c_x}\frac{\lambda_k^u e^{-\lambda_k}}{u!}\right)\left(1 - \sum_{u=0}^{c_x}\frac{(m_x\lambda_k)^u e^{-m_x\lambda_k}}{u!}\right)^{-1}\right]^{\mathbb{1}_{\{x_k > c_x\}}} \\ & \left[\frac{(m_y\mu_k)^{y_k} e^{-m_y\mu_k}}{y_k!}\left(1 - \sum_{u=0}^{c_y}\frac{\mu_k^u e^{-\mu_k}}{u!}\right)\left(1 - \sum_{u=0}^{c_y}\frac{(m_y\mu_k)^u e^{-m_y\mu_k}}{u!}\right)^{-1}\right]^{\mathbb{1}_{\{y_k > c_y\}}}. \end{aligned}$$

This section will not include the specifications of the Poisson-geometric or Poisson-negative binomial bivariate threshold mixture models as it follows on from the definitions given in Sections 6.5.1 and 6.5.2.

6.6.1 Model Comparison

Three bivariate threshold mixture models were investigated, Poisson-geometric, Poisson-negative binomial and Poisson-Poisson. The deviance values from a hypothesis test comparing these models to a null Dixon and Coles model (4.3) using the English Premier

League data set described in Section 6.3.2 are shown in Table 6.3. It can be seen that the Poisson-Poisson model, with fewer degrees of freedom and a deviance approximately equal to that achieved under the Poisson-negative binomial model is the most statistically significant out of the three and in a chi-sqaured test at a 0.05 significance level the null hypothesis of a single Poisson model may be rejected in favour of this alternative. Maximum likelihood estimates for the threshold levels under this model took the values $\hat{c}_x = 1$ and $\hat{c}_y = 5$, whilst those for the multiplicative parameters were $\hat{m}_x = 0.85$ and $\hat{m}_y = 0.09$. This is contrary to the idea that the Poisson model is under-representing the tails, which could suggest that an underdispersed distribution should be investigated.

Model	Deviance	df	p-value
Poisson-Geometric	8.08	4	0.089
Poisson-Neg. Binomial	35.77	6	0.000
Poisson-Poisson	35.77	4	0.000

Table 6.3: Deviance values of the three threshold mixture models tested

Table 6.4 shows the deviance values for each possible threshold level of $\hat{c}_x$ and $\hat{c}_y$. The green cell shading represents a drop of 5.99 which is equivalent to a 0.05 significance level of a χ^2 test on 2 degrees of freedom. This value represents the range of values that $\hat{c}_y$ could take and still be a significant addition to the model. The maximum deviance is obtained at $c_x = 1$ and $c_y = 5$.

Figure 6.12 shows the percentage difference in the χ^2 test statistic defined by equation (6.4) between the bivariate Poisson threshold mixture model defined in equation (6.6) and the uncensored Dixon and Coles (1997) model. As previously stated in Section 6.4.4, an average was taken from measures created using pooling ordered by the expected values of the standard model and the censored model. Figure 6.13 shows this measure for home and away counts seperately. It is interesting to note that the greatest percentage difference in the goodness of fit statistics when combining home and away goals is achieved at $\hat{c}_x = 1$ and $\hat{c}_y = 5$, the same values which give the highest deviance. However, when analysing home and away goals seperately this is not the case.

Under seperate analysis, the home and away goals experience the lowest RMSE at values of $c_x = 1$ and $c_y = 5$, and $c_x = 1$ and $c_x = 4$ respectively. This finding, alongside the fact that the maximum reduction of RMSE is greater for home goals than away goals (and the maximum increase is also lower), suggests that this model is more important for home goals than away goals.

		$\hat{c}_y$					
		1	2	3	4	5	6
$\hat{c}_x$	1	32.43	31.24	31.64	30.52	35.77	29.20
	2	12.77	11.74	12.17	11.13	16.46	9.91
	3	5.02	4.05	4.52	3.53	8.89	2.38
	4	2.51	1.61	2.04	1.12	6.36	0.04
	5	2.58	1.69	2.12	1.15	6.53	0.08
	6	2.67	1.76	2.21	1.23	6.54	0.15
	7	2.82	1.86	2.35	1.42	6.68	0.29
	8	2.81	1.90	2.31	1.41	6.60	0.29
	9	2.50	1.61	2.06	1.12	6.34	0.00

Table 6.4: Deviance values for the Poisson-Poisson threshold mixture model at all possible threshold levels. Green cells indicate values contained within a drop 5.99, representing the equivalent χ^2 statistic at a 0.05 significance level for 2 degrees of freedom.

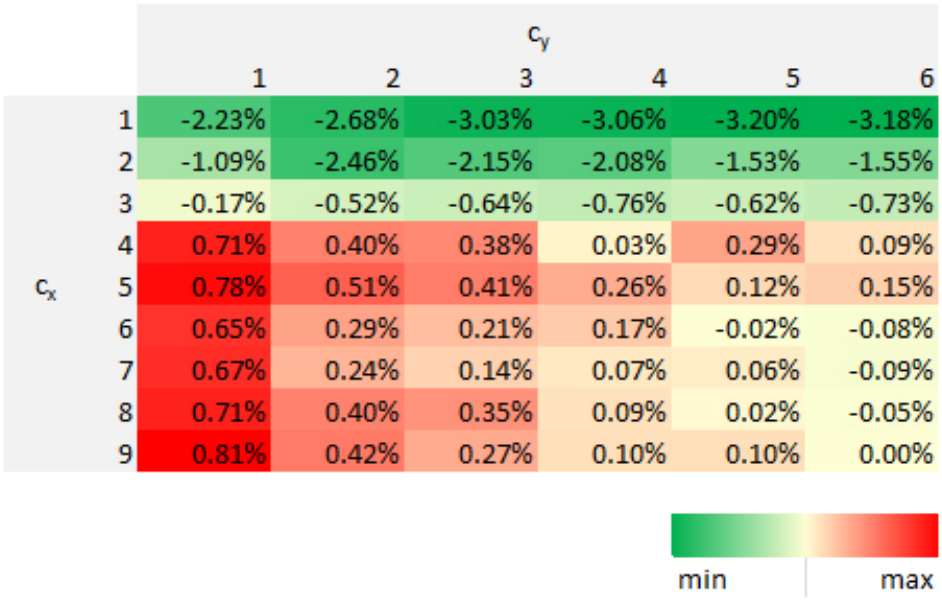


Figure 6.12: Percentage difference in χ^2 test statistic between the threshold mixture model defined in equation (6.6) and the standard Poisson Dixon and Coles model (4.3) for both home and away goal counts

To correctly understand how the threshold model affects probabilities, Table 6.5 shows the ratio of Poisson and Poisson-Poisson model probabilities for $\lambda = 1, \ldots, 10$ and $x = 1, \ldots, 10$, with $c_x = 4$ and $m_x = 1.3$. Table 6.6 shows the same ratio. However, now $m_x = 0.8$. Values of $m_x < 1$ result in an initial increase in probability, which

Home

c_x \ c_y	1	2	3	4	5	6
1	-3.75%	-4.03%	-3.22%	-5.83%	-3.76%	-5.37%
2	-1.52%	-1.78%	-2.06%	-2.07%	-2.23%	-2.54%
3	-0.35%	-0.89%	-0.85%	-1.17%	-0.94%	-1.24%
4	1.05%	0.75%	0.57%	0.33%	0.72%	-0.10%
5	1.21%	1.06%	0.85%	0.33%	0.48%	0.28%
6	0.91%	0.40%	0.47%	0.26%	0.11%	-0.12%
7	1.11%	0.47%	0.31%	0.38%	0.56%	-0.15%
8	1.16%	0.89%	0.45%	0.40%	0.37%	-0.05%
9	1.25%	0.75%	0.48%	0.36%	0.52%	0.00%

Away

c_x \ c_y	1	2	3	4	5	6
1	-2.77%	-3.44%	-3.40%	-3.10%	-3.45%	-3.33%
2	-1.30%	-1.41%	-1.40%	-1.95%	-1.53%	-1.87%
3	0.99%	0.56%	0.08%	-0.33%	-0.56%	-0.62%
4	2.18%	1.28%	1.42%	0.34%	0.25%	0.25%
5	2.10%	1.34%	0.96%	1.00%	0.51%	-0.13%
6	2.17%	1.51%	1.13%	0.51%	0.41%	-0.04%
7	1.77%	1.12%	1.06%	0.14%	-0.21%	-0.04%
8	1.89%	0.89%	1.28%	0.23%	0.00%	-0.06%
9	2.20%	1.25%	1.18%	0.48%	-0.05%	0.00%

min max

Figure 6.13: Percentage difference in χ^2 test statistic between the censored model defined in equation (6.6) and the standard Poisson Dixon and Coles model (4.3) (top) home goal count, X, (bottom) away goal count, Y.

also increased with λ, over the null Poisson model, followed by a decreasing ratio of probability as x increases at points above c_x, whilst values of $m_x > 1$ result in the opposite. This shows that if m is greater than 1, there is a lower probability of higher goal counts, however there is a sharp inorganic step between c_x and $c_x + 1$, and again vice versa for values of $m_x < 1$.

To assess the predictive power of the model a leave one out cross validation study was carried out as described in Section 4.2, using the same data set from the Premier League between seasons 2009/2010 - 2011/2012. Under the Poisson-Poisson threshold mixture

model, the RMSE for home goals was 1.293 and that for away goals was 1.106. Both of these values are greater than those for the standard Dixon and Coles (1997) model as given in Section 4.2, suggesting a loss in predictive power when employing this model.

		x									
		1	2	3	4	5	6	7	8	9	10
λ	1	1.00	1.00	1.00	1.00	1.07	0.83	0.63	0.49	0.38	0.29
	2	1.00	1.00	1.00	1.00	1.18	0.91	0.70	0.54	0.41	0.32
	3	1.00	1.00	1.00	1.00	1.33	1.03	0.79	0.61	0.47	0.36
	4	1.00	1.00	1.00	1.00	1.56	1.20	0.93	0.71	0.55	0.42
	5	1.00	1.00	1.00	1.00	1.90	1.46	1.12	0.86	0.66	0.51
	6	1.00	1.00	1.00	1.00	2.37	1.83	1.41	1.08	0.83	0.64
	7	1.00	1.00	1.00	1.00	3.05	2.35	1.81	1.39	1.07	0.82
	8	1.00	1.00	1.00	1.00	4.00	3.08	2.37	1.82	1.40	1.08
	9	1.00	1.00	1.00	1.00	5.31	4.08	3.14	2.42	1.86	1.43
	10	1.00	1.00	1.00	1.00	7.10	5.46	4.20	3.23	2.49	1.91

Table 6.5: Ratio of Poisson to Poisson-Poisson probabilities for a range of parameter values, $c_x = 4$ and $m_x = 1.3$.

		x									
		1	2	3	4	5	6	7	8	9	10
λ	1	1.00	1.00	1.00	1.00	0.96	1.19	1.49	1.87	2.33	2.92
	2	1.00	1.00	1.00	1.00	0.90	1.13	1.41	1.76	2.20	2.75
	3	1.00	1.00	1.00	1.00	0.84	1.05	1.31	1.64	2.05	2.56
	4	1.00	1.00	1.00	1.00	0.77	0.96	1.20	1.50	1.88	2.35
	5	1.00	1.00	1.00	1.00	0.69	0.87	1.08	1.35	1.69	2.11
	6	1.00	1.00	1.00	1.00	0.61	0.76	0.96	1.19	1.49	1.87
	7	1.00	1.00	1.00	1.00	0.53	0.66	0.83	1.04	1.30	1.62
	8	1.00	1.00	1.00	1.00	0.45	0.57	0.71	0.89	1.11	1.38
	9	1.00	1.00	1.00	1.00	0.38	0.48	0.60	0.75	0.93	1.17
	10	1.00	1.00	1.00	1.00	0.32	0.40	0.50	0.62	0.78	0.98

Table 6.6: Ratio of Poisson to Poisson-Poisson probabilities for a range of parameter values, $c_x = 4$ and $m_x = 0.8$.

6.7 Conclusion

The Poisson model for goal counts works under the assumption that the mean and variance are equal. However, it is clear that this is not always the case. Although dispersion in goal counts can be quickly analysed when assuming they are identically distributed, this ignores any hidden heterogeneity. Therefore, the only way to create better models is to hypothesise novel models and statistically test their fit.

Over or under dispersion in the overall mean of goal counts represents either more similarly or dissimilarly skilled teams respectively. However, an over or under dispersed nature when considering heterogeneity (in the form of the Dixon and Coles model (4.3)) represents any increased or decreased variability in goals scored, respectively. Under testing, evolving the Dixon and Coles model (4.3) to a negative binomial model did not more effectively model the data.

To allow a deeper investigation, the body and tail of the model were considered separately using censoring and threshold mixture modelling. Right censoring above some value led to a better fit of values below the censoring value and increased variability for higher estimates, due to the reduction in information given to the model. Although the body below the censoring value may have been better fit, this is not a total representation of the probability space.

Various threshold mixture models were tested, with the aim of better modelling the right tail and informing the interpretation of any over or under dispersion. From the models tested, a Poisson-Poisson threshold mixture model, with the right tail portion modified by a multiplicative parameter, proved the best fit, with threshold estimates of $\hat{c}_x = 1$ and $\hat{c}_y = 5$ and multiplicative parameter estimates of $\hat{m}_x = 0.85$ and $\hat{m}_y = 0.09$. These parameter estimates indicate that the right tail is under dispersed for both home and away goal counts. However, exhaustive cross validation indicated that the predictive power was lower than the null Dixon and Coles model (4.3).

Comparing the probabilities of goal counts at different parameter values for the Poisson-Poisson model and the Dixon and Coles model (4.3), showed that the Poisson-Poisson threshold mixture distribution does not have a smooth kernel at values where $m \neq 1$. Instead it has a sharp step change in the probability of observations above the threshold value. This could be the source of the reduction in predictive power and goal counts might be better modelled using some smooth transition.

6.8 Future Work

Modelling home and away goals using an under dispersed model, based on the binomial distribution, would capture any under dispersion. However, there is a difficulty in estimating the number of trials, which would logically be the number of times a goal was attempted. This could possibly be done using total of the goals scored, goals saved and corners given. However, the number of goals saved was unavailable for the data set used. This would, however, entail predicting the total number of shots and the conversion rate to forecast the estimated shots on target. It may also benefit the method to jointly model corners, goals and goals saved as this increases the range of possible bets an investor could make.

Chapter 7

Weighted Likelihood Based Changepoint Detection Methods

7.1 Introduction

In Chapter 6 a threshold mixture model between two Poisson distributions was formulated in an attempt to better model the right tail or extreme values in the distribution of goals in association football. This model can lead to a step change in the pdf as discussed in Section 6.6.1. Some form of smoothing across the transition could lead to a more natural pdf. Such a method could not only be used for smoothing the transition of probabilities, but also for modelling smooth changepoints in any ordered series.

The term changepoint refers to any abrupt change in the structure of a time series, dividing the data into distinct homogeneous sections (Eckley et al., 2011). Assuming an ordered sequence of data $x_{1:n} = (x_1, \ldots, x_n)$, the typical definition of a discrete single changepoint can be interpreted in two ways:

1. A changepoint can be thought to occur when there exists a time or index position, $\tau \in \{1, \ldots, n-1\}$, such that the statistical properties of $\{x_1, \ldots, x_\tau\}$ and $\{x_{\tau+1}, \ldots, x_n\}$ are different in some way.
2. A changepoint can be thought to occur when there exists a time or index position, $\tau \in \{2, \ldots, n\}$, such that the statistical properties of $\{x_1, \ldots, x_{\tau-1}\}$ and $\{x_\tau, \ldots, x_n\}$ are different in some way.

From this point on definition 1 will be used unless otherwise stated and will be referred to as a 'generic' discrete changepoint model. These definitions can be extended to the multiple changepoint case (see Section 7.5), though the focus here is primarily on the single changepoint case.

Figure 7.1 shows simulated Poisson distributed data exhibiting a changepoint from a mean of $\lambda_1 = 5$ to $\lambda_2 = 7$ at index position τ. Accurate changepoint detection methods

are of high importance to sports data analysis due to the occurrence of many time series and indexed data sets and the expectation of some change in the observed processes due to changes in strategies, training or rules. Often these data sets are small (under 100 data points) and are limited by both the age of the sport on a professional level and the amount of freely available information. Due to the small size of such data sets, likelihood based inference for the generic discrete changepoint model can produce a lot of noise in the likelihood surface.

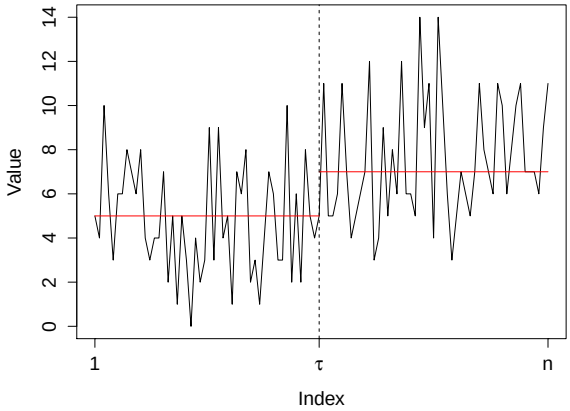


Figure 7.1: An example of a changepoint: Data from index position $1, \ldots, \tau$ follows a Poisson distribution with mean of $\lambda_1 = 5$ and that from $\tau + 1, \ldots, n$ follows a Poisson distribution with mean of $\lambda_2 = 7$.

The detection of a changepoint can be achieved using a hypothesis test, whereby the null hypothesis, H_0, considers the state of no changepoint and the alternative hypothesis, H_1, represents the existence of a changepoint (Eckley et al., 2011), where independent observations $x_1, \ldots, x_\tau$ are considered to be distributed according to $f(x_t, \boldsymbol{\lambda}_1)$ and independent observations $x_{\tau+1}, \ldots, x_n$ are considered to be distributed according to $f(x_t, \boldsymbol{\lambda}_2)$, where $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ refer to the vector of model parameters for the relevant segment. Then the log-likelihood is given by

$$\ell(\tau, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = \sum_{i=1}^{\tau} \log\left[f(x_t, \boldsymbol{\lambda}_1)\right] + \sum_{i=\tau+1}^{n} \log\left[f(x_t, \boldsymbol{\lambda}_2)\right]. \tag{7.1}$$

When testing for no changepoint in this model under H_0: $\boldsymbol{\lambda} = \boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2$, the deviance surface can be defined for an alternative hypothesis H_1: $\boldsymbol{\lambda}_1 \neq \boldsymbol{\lambda}_2$, by taking the deviance shown in equation (7.2) for each possible value of τ, the hypothesised changepoint position, from index position 1 to n.

$$D(\tau) = 2\{\ell(\tau, \hat{\boldsymbol{\lambda}}_1, \hat{\boldsymbol{\lambda}}_2) - \ell(0, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\lambda}})\}, \tag{7.2}$$

where $\hat{\boldsymbol{\lambda}}_1$ and $\hat{\boldsymbol{\lambda}}_2$ refer to the vectors of maximum likelihood estimates (MLEs) for the alternative hypothesis when the tested changepoint parameter is equal to τ (i.e. $\ell(\tau, \hat{\boldsymbol{\lambda}}_{\boldsymbol{1}}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{2}})$ is the profile log-likelihood for τ) and $\hat{\boldsymbol{\lambda}}$ is the vector of MLEs for the common value of $\boldsymbol{\lambda} = \boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2$ under the null hypothesis. Note here $\hat{\boldsymbol{\lambda}}_1$ and $\hat{\boldsymbol{\lambda}}_2$ are functions of τ but the notational dependence is suppressed.

Changepoints should not only be identified by their point estimates, but also by the confidence interval. In the single changepoint case for Poisson distributed data, as in Figure 7.1, two MLEs are to be calculated: $\hat{\lambda}_1$ using data between 1 and the index before the suspected changepoint parameter, τ, (i.e. $x_1, \ldots, x_\tau$), and $\hat{\lambda}_2$ using $x_{\tau+1}, \ldots, x_n$, where n is the total number of indices. As τ is moved an index value passes from one set to another altering the MLE estimates.

If the set from which an MLE of λ_1 or λ_2 is to be derived is small, the introduction of another data value can drastically change the estimate (and vice versa for the removal of the same value from the set which is used to calculate the MLE for the other side of the changepoint). This causes noise in the deviance surface, as described earlier, which may lead to the production of a broken confidence interval or confidence 'set' (Siegmund, 1988).

In the generic discrete changepoint model, each observation to one side of τ is considered to carry the same amount of information regarding τ and the associated parameters, $\boldsymbol{\lambda}_j$, in that segment. However, the level of information carried by observations is not constant. The observations far from τ, carry less information about τ and more information about $\boldsymbol{\lambda}_j$ than observations near to τ. The technique that is developed here to account for this relies on the use of a weighting function, $\phi_h(t, \boldsymbol{c})$, which uses information about how far an observation at index t is from the centre of a smooth change represented by the position of equal weighting, c, where h controls the smoothness of the weighting function, specifically $\phi_h(t, c) \to 0$ as $t \to -\infty$ and $\phi_h(t, c) \to 1$ as $t \to \infty$.

An example deviance surface, relating to a generic discrete changepoint model applied to the data shown in Figure 7.1 and a smooth surface resulting from a weighted likelihood method (that is developed in Section 7.2) are shown in Figure 7.2. It should be noted that the generic discrete method is joined linearly between integer values of the index in Figure 7.2 to look continuous and allow visual comparison. The 95% confidence set relating τ is $\{42, 44, \ldots, 60, 62, \ldots, 66\}$, whereas the 95% confidence interval for c is $[37.35, 75.01]$.

Traditional changepoint methods rely on a bounded discrete definition of changepoint position, i.e $\tau = \{1, \ldots, n\}$. However, due to the nature of the weighted method, the po-

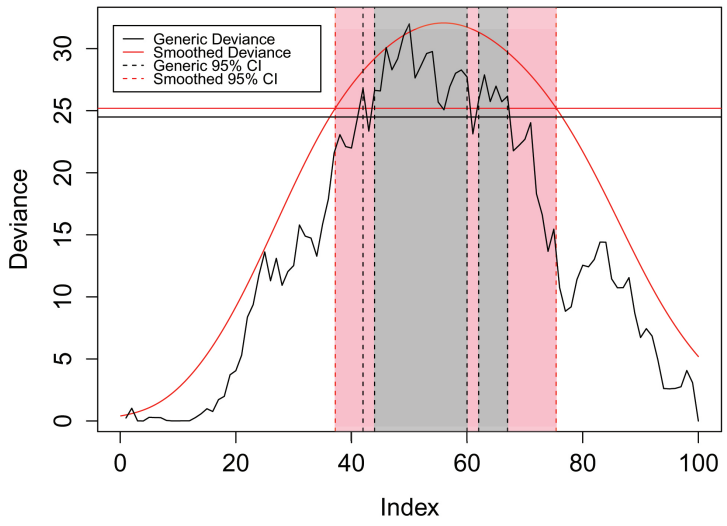


Figure 7.2: Deviance functions comparing alternative hypotheses of one generic discrete changepoint (black) or one smooth change between two distributions (red) to a null hypothesis of zero changepoints, simulating from a generic discrete changepoint model. Here $n = 100$ and the true changepoint position $\tau = 50$. In the case of the generic discrete changepoint the integer values are joined linearly.

sition of equal weighting between segments may be used as the changepoint position and this is a well defined unbound continuous definition of a changepoint, i.e. $c \in (-\infty, \infty)$. This permits the method to learn about c occurring outside of the sample unlike with existing changepoint methods. Of particular importance is the case when c occurs after the end of the sample, allowing the forecasting of a changepoint position to be explored using data only in the sample.

7.2 Defining the Method

The variation of confidence in information obtained from observations about changepoint position and parameter estimates close to and distant from the changepoint position may be accounted for using a smooth weighting function across the changepoint boundary. This weighting function is employed to mix parameter estimates and thereby aid the description of uncertainty over to which segment of observations a certain index should belong. Such a method inherently allows the modelling of both sudden (discrete) and smooth (continuous) changes.

A naive approach to account for the discontinuity of information about changepoint position, τ, and parameters, $\boldsymbol{\lambda}_j$, carried by observations is to mix model parameters within a distribution using a suitable smooth weighting function. This method is detailed in Section 7.2.1.

However, a more subtle approach, whereby differing distributions are mixed using a smooth weighting function, may result in better forecasting ability. This method, covered in Section 7.2.2, will be the focus of discussion, using the smoothly weighted mixed parameters method comparatively.

7.2.1 Smoothly Weighted Mixed Distribution Parameters

Consider the single changepoint case for n observations $x_1, \ldots, x_n$ of a random variable X, described by two vectors of distribution parameters, $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, mixed with the probability $\phi_h(t, c)$ to describe a smooth change. The continuous-index position which experiences equal weighting between $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ is defined as c, where $-\infty < c < \infty$, unlike the generic changepoint $\tau \in \{1, \ldots, n-1\}$. The weighting function is given by $\phi_h(t, c)$, where t describes the index of observations and h controls the amount of smoothing.

The likelihood is then given by

$$\mathcal{L}(c, h, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = \prod_{t=1}^{n} f\left\{x_t; \phi_h(t, c)\boldsymbol{\lambda}_1 + [1 - \phi_h(t, c)]\,\boldsymbol{\lambda}_2\right\}. \tag{7.3}$$

Taking the logarithm of this likelihood gives

$$\begin{aligned} \ell(c, h, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) &= \log\left\{\prod_{t=1}^{n} f\left\{x_t; \phi_h(t, c)\boldsymbol{\lambda}_1 + [1 - \phi_h(t, c)]\,\boldsymbol{\lambda}_2\right\}\right\} \\ &= \sum_{t=1}^{n} \log\left\{f\left\{x_t; \phi_h(t, c)\boldsymbol{\lambda}_1 + [1 - \phi_h(t, c)]\,\boldsymbol{\lambda}_2\right\}\right\}. \end{aligned}$$

To illustrate the approaches, all initial model comparisons will be carried out using a change in the Poisson mean, λ. Assuming that $X_t \sim \text{Poisson}\left(\phi_h(t, c)\lambda_1 + [1 - \phi_h(t, c)]\,\lambda_2\right)$ and substituting in a Poisson probability mass function to describe the data results in the log-likelihood

$$\ell(c, h, \lambda_1, \lambda_2) = \sum_{t=1}^{n} \log\left\{\frac{\left\{\phi_h(t, c)\,\lambda_1 + [1 - \phi_h(t, c)]\,\lambda_2\right\}^{x_t} e^{-\{\phi_h(t,c)\lambda_1 + [1-\phi_h(t,c)]\lambda_2\}}}{x_t!}\right\}. \tag{7.4}$$

Under this model, when $t \ll c$, $X_t \sim \text{Poisson}(\lambda_1)$ and when $t \gg c$, $X_t \sim \text{Poisson}(\lambda_2)$.

7.2.2 Smoothly Weighted Mixed Distributions

Consider the single changepoint case for n observations $x_1, \ldots, x_n$ of a random variable X, where $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ correspond to the vectors of parameters of two distributions which are

mixed with the probability $\phi_h(t,c)$ to describe a smooth change. The continuous-index position which experiences equal weighting between distributions $f(x,\boldsymbol{\lambda}_1)$ and $g(x,\boldsymbol{\lambda}_2)$ is defined, similarly to Section 7.2.1, as c.

The likelihood is then given by

$$\mathcal{L}(c,h,\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2) = \prod_{t=1}^{n} \left\{\phi_h(t,c)\, f(x_t;\boldsymbol{\lambda}_1) + [1-\phi_h(t,c)]\, g(x_t;\boldsymbol{\lambda}_2)\right\}. \tag{7.5}$$

Taking the logarithm of this likelihood gives

$$\begin{aligned} \ell(c,h,\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2) &= \log\left\{\prod_{t=1}^{n} \left\{\phi_h(t,c)\, f(x_t;\boldsymbol{\lambda}_1) + [1-\phi_h(t,c)]\, g(x_t;\boldsymbol{\lambda}_2)\right\}\right\} \\ &= \sum_{t=1}^{n} \log\left\{\phi_h(t,c)\, f(x_t;\boldsymbol{\lambda}_1) + [1-\phi_h(t,c)]\, g(x_t;\boldsymbol{\lambda}_2)\right\}. \end{aligned} \tag{7.6}$$

Assuming f an g are both Poisson, then

$$X_t \sim \begin{cases} \text{Poisson}(\lambda_1) & \text{w.p. } \phi_h(t,c), \\ \text{Poisson}(\lambda_2) & \text{w.p. } [1-\phi_h(t,c)]. \end{cases}$$

Substituting a Poisson probability mass function into equation (7.6) results in the log-likelihood given by

$$\ell(c,h,\lambda_1,\lambda_2) = \sum_{t=1}^{n} \log\left[\phi_h(t,c)\,\frac{\lambda_1^{x_t} e^{-\lambda_1}}{x_t!} + [1-\phi_h(t,c)]\,\frac{\lambda_2^{x_t} e^{-\lambda_2}}{x_t!}\right]. \tag{7.7}$$

Similarly to the case of mixed distribution parameters described in Section 7.2.1, under this model, when $t \ll c$, $X_t \sim \text{Poisson}(\lambda_1)$ and when $t \gg c$, $X_t \sim \text{Poisson}(\lambda_2)$.

7.2.3 Weighting Function

The weighting function considers the variation of confidence in information obtained from observations about the changepoint position and parameter estimates close to and distant from the centre of a smooth change or point of equal weighting between segments, c. Various weighting functions were tested. The most consistently performing method is based on the Gaussian cumulative distribution function (cdf), Φ, and is given by

$$\phi_h(t,c) = 1 - \Phi\left(\frac{t-c}{h}\right), \tag{7.8}$$

where t is the index value or time step and $h > 0$ controls the smoothing level. This may

be interpreted visually as shown in Figure 7.3.

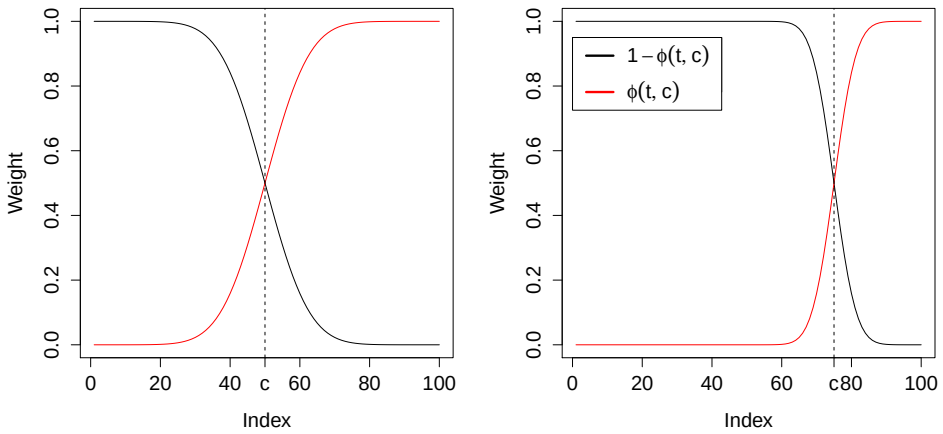


Figure 7.3: (Left) Gaussian weighting function for a single changepoint where $c = 50$ and $h = 10$. (Right) Gaussian weighting function for a single changepoint where $c = 75$ and $h = 5$.

As $h \to 0$ under the Gaussian weighting function, the likelihood shown by equation (7.5) reduces to the case of a single generic discrete changepoint. This can be shown by

$$\phi_h(t, c) = 1 - \Phi\left(\frac{t-c}{h}\right) \xrightarrow{h \to 0} \begin{cases} 0, & \text{if } t > c, \\ \frac{1}{2} & \text{if } t = c, \\ 1, & \text{if } t < c. \end{cases} \tag{7.9}$$

In the case where $t = c$ the weighting function $\phi_h(t, c) = \frac{1}{2}$ and, under the mixed distribution model described in Section 7.2.2, the log-likelihood component when $t = c$ would take the form

$$\ell(c, h, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = \log\left\{\frac{1}{2} f(x_c; \boldsymbol{\lambda}_1) + \frac{1}{2} g(x_c; \boldsymbol{\lambda}_2)\right\}.$$

However, c is continuous and $c \neq t$ for all values of t, so this term does not arise in practice. The log-likelihood in the case of $h \to 0$ is therefore given by

$$\ell(c, h, \boldsymbol{\lambda}_1, \boldsymbol{\lambda_2}) = \sum_{t=1}^{\lfloor c \rfloor} \log\left[f(x_t, \boldsymbol{\lambda}_1)\right] + \sum_{t=\lfloor c \rfloor + 1}^{n} \log\left[f(x_t, \boldsymbol{\lambda}_2)\right]. \tag{7.10}$$

A similar derivation can be followed for the mixed parameter model described in Section 7.2.1.

Figure 7.4 shows how the deviance surface varies as h is increased for both the mixed

parameter and mixed distribution models. It can be seen that as h increases the smoothness of the surface increases in both cases, resulting in the loss of information about the changepoint position and respective parameter values as $h \to \infty$.

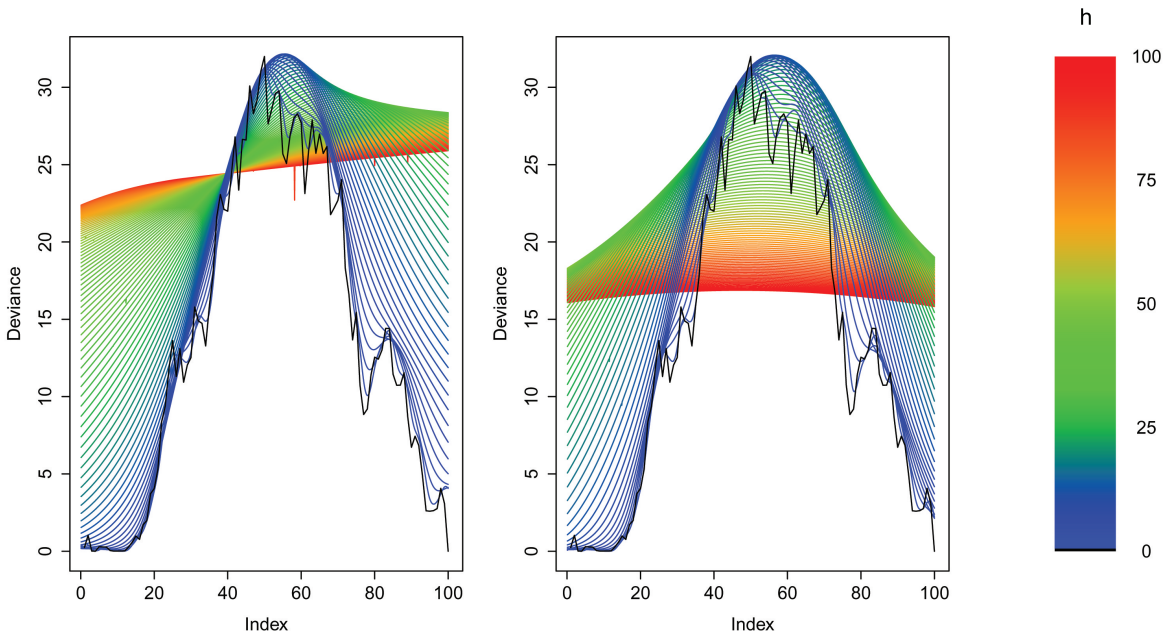


Figure 7.4: Deviance surfaces relating to varying h (as given by coloured legend) for (left) mixed parameter model and (right) mixed distributions model. Note, black curves relate to discrete generic changepoint case or $h = 0$ as shown by equation (7.10). Sample simulated using a generic discrete change (or $h = 0$), from a Poisson parameter of $\lambda_1 = 5$ to one of $\lambda_2 = 7$, where $n = 100$ and $c = 50$.

Cross-validation techniques have been explored to allow auto-tuning of h to the dataset (see Appendix G.1), however, these methods prove to be slow in practice. However, unlike in usual smoothing methods, h can be treated as a parameter and estimated via likelihod methods as shown in Sections 7.2.1 and 7.2.2. Under this definition of h, when $h = 0$ the model is the generic discrete changepoint method as shown by equation (7.10).

7.3 Overcoming Local Maxima in the Deviance Surface

Figure 7.5 shows a deviance surface for various h and c relating to a single smooth change between two Poisson distributions, simulated using $n = 200$, $c_0 = 100$, $h_0 = 50$ and parameter values $\lambda_1 = 5$ and $\lambda_2 = 7$. It can be seen that at low values of h local maxima occur in the deviance when varying c. This may cause generic optimisation methods to become trapped. This is also true for the mixed means model described in Section 7.2.1. The difficulties encountered due to the existence of local features in the deviance surface at low values of h may be overcome by using heuristic methods which seek to find the global maximum.

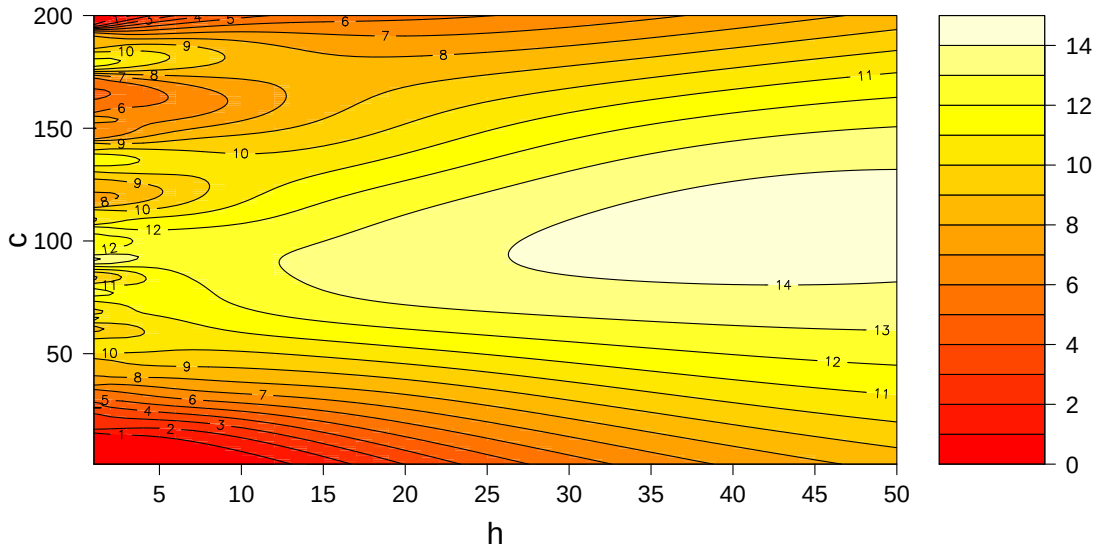


Figure 7.5: Contour plot showing the deviance surface of a simulated smoothly weighted mixed distribution change over a grid of c and h.

Figure 7.5 provides evidence that the parameters c and h are approximately orthogonal to one another under the smoothly weighted mixed distribution model when $0 < c \leq n$ and may be treated as independent from one another in this range. Therefore, fixing h or c at a sensible value (not too small in the case of h) and maximising over the other parameters may give a fast initial estimate. Discretising the parameter(s) may also aid in both overall speed and reducing the possibility of becoming trapped. Following this process with a constrained search around these intial estimates will provide the necessary accuracy for achieving a global maximum, i.e. the maximum likelihood estimates.

Section 7.3.2 will discuss how estimates of parameters c and h are no longer orthogonal when c occurs close to and outside of the bounds of the sample space. Finally, a method will be detailed, which seeks to account for this by quantifying correlation between the two parameters.

However, in Figure 7.4 the mixed parameter model can be seen to lack orthogonality within the bounds of the sample, $0 < c \leq n$. Therefore, the more complex method defined in Section 7.3.2 must be used for this method at all times.

7.3.1 Discretisation of c and Piecewise Linearisation of Parameter Estimates within $0 < c \leq n$ Under the Smoothly Weighted Mixed Distribution Model

To reduce the possibility of an optimisation routine becoming trapped by one of the local maxima when mixing two Poisson distributions in this way, in the first instance, the log-likelihood may be maximised using a reasonable value of h and discretising $c \in C = \{\Delta_c, 2\Delta_c, \ldots, n\}$ where Δ_c represents a sensible increment of c. The log-likelihood is maximised to give MLEs by

$$\hat{\boldsymbol{\theta}}_{|h} \subseteq \underset{\boldsymbol{\theta}_{|h} \in \Theta_{|h}}{\arg\max}\ \ell\left(c, h, \lambda_1, \lambda_2\right), \tag{7.11}$$

where $\boldsymbol{\theta}_{|h} = (c, \lambda_1, \lambda_2)$ and $\Theta_{|h} = C \times [0, \infty)^2$, where the Poisson parameters $\lambda_1, \lambda_2 > 0$. To clarify that the maximisation is over the grid for c we denote the value of c at $\hat{\boldsymbol{\theta}}_{|h}$ as $\hat{c}_\Delta$.

Define $\boldsymbol{\theta}_{|c,h} = (\lambda_1, \lambda_2)$ as the vector of parameters when fixed values of c and h are given. The log-likelihood is maximised under these conditions by

$$\hat{\boldsymbol{\theta}}_{|c,h} \subseteq \underset{\boldsymbol{\theta}_{|c,h} \in [0,\infty)^2}{\arg\max}\ \ell\left(c, h, \lambda_1, \lambda_2\right). \tag{7.12}$$

Figure 7.6 shows how $\hat{\lambda}_{1|c,h}$ and $\hat{\lambda}_{2|c,h}$ might vary with c. It can be seen that incremental linearisation of $\hat{\lambda}_{1|c,h}$ and $\hat{\lambda}_{2|c,h}$ gives a good approximation to the estimates. In this case, the value of $\Delta_c = 10$, so the range between piecewise segments is given by $2\Delta_c = 20$. To achieve linearisation between segments adjacent to $\hat{c}_\Delta$ parameter estimates $\hat{\lambda}_{i|c,h}$ for $i = 1, 2$ are obtained by evaluating $\hat{\boldsymbol{\theta}}_{|c,h}$ at values of $c = \hat{c}_\Delta - \Delta_c$ and c=$\hat{c}_\Delta + \Delta_c$, and are denoted $\hat{\lambda}_i^-$ and $\hat{\lambda}_i^+$ respectively.

The linearised estimates $\tilde{\lambda}_{1|c,h}$ and $\tilde{\lambda}_{2|c,h}$ corresponding to $\hat{\lambda}_{1|c,h}$ and $\hat{\lambda}_{2|c,h}$ respectively are calculated for a given value of c as

$$\tilde{\lambda}_{i|c,h} = \hat{\lambda}_i^- + \left(\frac{\hat{\lambda}_i^+ - \hat{\lambda}_i^-}{2\Delta_c}\right)\left[c - (\hat{c}_\Delta - \Delta_c)\right] \tag{7.13}$$

for $i = 1, 2$ and $c \in [\hat{c}_\Delta - \Delta_c, \hat{c}_\Delta + \Delta_c]$.

Penultimately, maximisation is performed over parameters h and c, using $\tilde{\lambda}_{1|c,h}$ and $\tilde{\lambda}_{2|c,h}$ in place of λ_1 and λ_2, as given by

$$\hat{\boldsymbol{\theta}}_{|\boldsymbol{\lambda}} \subseteq \underset{\boldsymbol{\theta}_{|\boldsymbol{\lambda}}}{\arg\max}\ \ell\left[c, h, \tilde{\lambda}_{1|c,h}, \tilde{\lambda}_{2|c,h}\right],$$

where $\boldsymbol{\theta}_{|\boldsymbol{\lambda}} = (c, h)$, $h \in \mathbb{R}^+$ and $c \in [\hat{c}_\Delta - \Delta_c, \hat{c}_\Delta + \Delta_c]$. The parameter estimates $\hat{c}_{|\boldsymbol{\lambda}}$ and $\hat{h}_{|\boldsymbol{\lambda}}$ and the corresponding estimates $\tilde{\lambda}_{1|c,h}$ and $\tilde{\lambda}_{2|c,h}$ gained from equation (7.13) may then

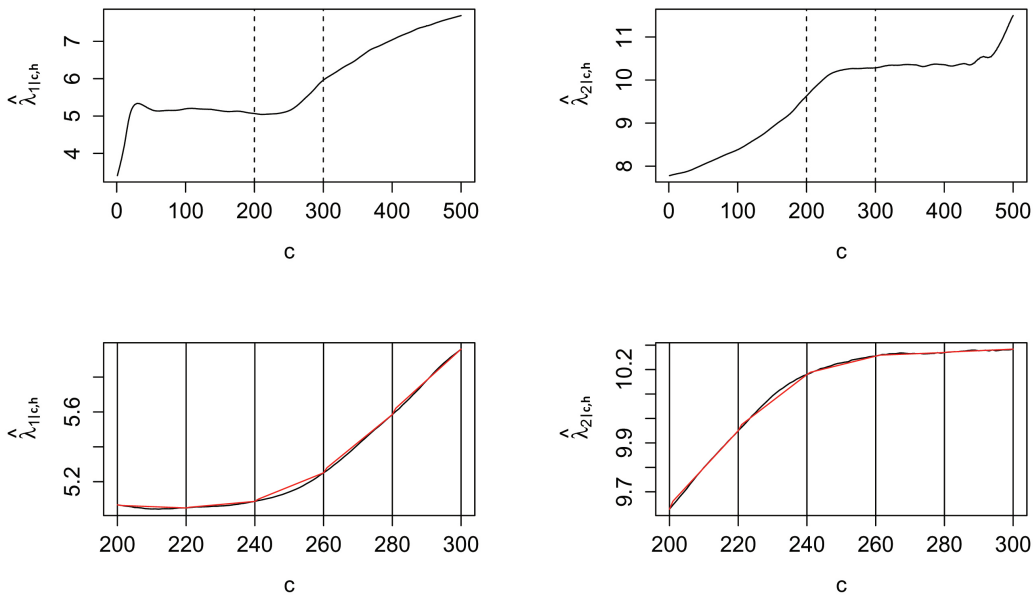


Figure 7.6: Using a simulated data set with $h = 10$, $c = 250$, $n = 500$, $\lambda_1 = 5$ and $\lambda_2 = 10$: (Top left) Maximum likelihood estimate $\hat{\lambda}_{1|c,h}$ for $c = 1, \ldots, n$. (Top right) Maximum likelihood estimate $\hat{\lambda}_{2|c,h}$ for varying values of $c = 1, \ldots, n$. (Bottom left) Piecewise linearisation of $\hat{\lambda}_{1|c,h}$ for $c = 200, \ldots, 300$ by increments of 20. (Bottom left) Piecewise linearisation of $\hat{\lambda}_{2|c,h}$ for $c = 200, \ldots, 300$ by increments of 20.

be used to give a starting value to maximise over all parameters, $\boldsymbol{\theta} = (c, h, \lambda_1, \lambda_2)$, allowing the optimiser to more effectively find the global maximum in the range of $c \in [0, n]$.

7.3.2 Quantifying the Correlation Between c and h

Contrary to the evidence provided by Figure 7.5, if c occurs close to or outside of the bounds of the sample under the smoothly weighted mixed distribution model, c and h are no longer orthogonal and cannot be treated as independent from one another. This is logical as the smoothness, h, of the weighting function controls the range of significant mixing between distributions, low smoothly weighted mixing centred distant from the bounds of the sample cannot be perceived.

It can also be seen in Figure 7.4 that the more naive method of smoothly weighted mixed means within a Poisson distribution lacks orthogonality whether or not c occurs inside or outside of the bounds of the sample space.

Therefore, the assumption of independence between c and h due to orthogonality when maximising the likelihood is not valid unless $0 < c \leq n$ under the model considering smoothly weighted mixed distributions. A new method seeks to linearise a vague esti-

mate of c with respect to the value of h and maximise over h to gain a starting position near to the global maximum parameter estimates.

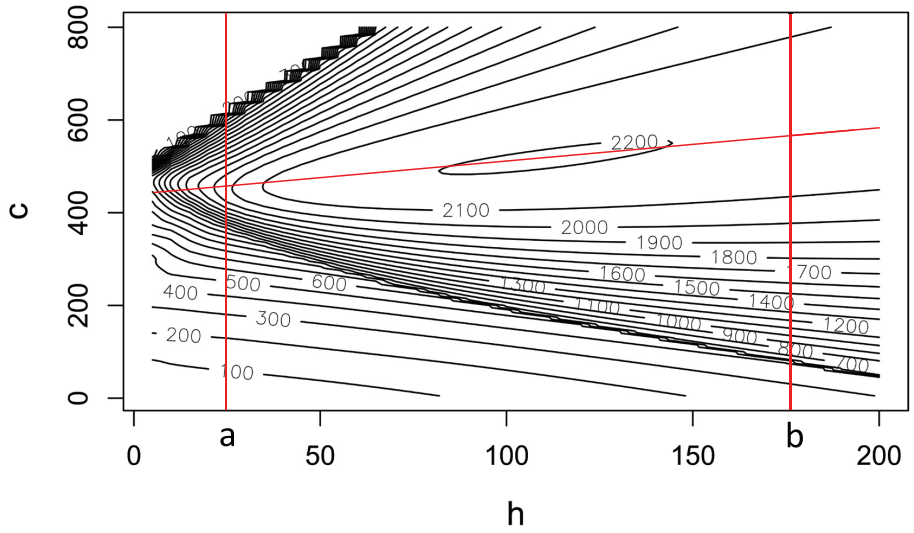


Figure 7.7: Contour plot showing the deviance surface of a simulated weighted changepoint over c and h, where $n = 500$ and $c = 510$, outside of the bounds of the sample space. The values of h which are fixed as to maximise over discretised c are represented by a and b.

Figure 7.7 relates to such a method which takes into account any correlation between c and h. Again, considering a pair of Poisson distributions as a weighted mixture according to equation (7.7), the method is as follows: Maximising the log likelihood, as shown by equation (7.11), over discretised $c \in C = \{\Delta_c, 2\Delta_c, \ldots, m_c\Delta_c\}$, where $m_c\Delta_c$ represents the maximum value of c to be analysed, at fixed low and high values of h, a and b respectively, gives a pair of estimates for c: $\hat{c}_{\Delta,a}$ and $\hat{c}_{\Delta,b}$ respectively. These estimates can be used to form a linear equation which allows the calculation of c at a given value of h, i.e. $f_c(h)$, as given by

$$f_c(h) = \hat{c}_{\Delta,a} + \left(\frac{\hat{c}_{\Delta,b} - \hat{c}_{\Delta,a}}{b - a}\right)(h - a)\,. \tag{7.14}$$

This function may then be used to maximise over discretised $h \in H = \{\Delta_h, 2\Delta_h, \ldots, m_h\Delta_h\}$, where $m_h\Delta_h$ represents the maximum value of h to be analysed.

$$\hat{\boldsymbol{\theta}}_{|c} \subseteq \underset{\boldsymbol{\theta}_{|c} \in \boldsymbol{\Psi}}{\arg\max}\ \ell\left[f_c(h), h, \lambda_1, \lambda_2\right],$$

where $\boldsymbol{\theta}_{|c} = (h, \lambda_1, \lambda_2)$ and $\boldsymbol{\Psi} = H \times [0, \infty)^2$. To clarify that the maximisation is over the grid for h we denote the value of h at $\hat{\boldsymbol{\theta}}_{|c}$ as $\hat{h}_\Delta$.

The parameter estimate $\hat{h}_{\Delta}$ and the corresponding value of $c_{|h}$ gained from equation (7.14) may then be used to give a starting value to maximise over $\boldsymbol{\theta}$, allowing the optimiser to more effectively find the global maximum in the range of $c \in (-\infty, \infty)$.

7.4 Model Comparisons

To analyse the usability of the models discussed in Section 7.2 they must be cross-tested to ensure that they work effectively. This testing will be achieved by simulation under each model and testing for changepoints within and outside of the bounds of the sample, effectively allowing an investigation of modelling and forecasting ability.

7.4.1 Within the Bounds of the Sample

To ensure a concise and meaningful comparison, the models under experimentation will be renamed and all models will use the Poisson distribution as their basis. Model 0 refers to the generic discrete changepoint model, which is equivalent to either alternative model when $h = 0$. The log-likelihood corresponding to Model 0 is given by equation (7.10). Model 1 will refer to the main model of interest, that of the smoothly weighted mixed distribution model described in Section 7.2.2. The log-likelihood corresponding to Model 1 is given by equation (7.7). Finally, Model 2 will refer to the more naive alternative of smoothly weighted mixed means within similar distributions, described in Section 7.2.1, where the log-likelihood is given by equation (7.4).

Models 1 and 2 separately nest Model 0, and therefore should never perform worse than Model 0 in terms of maximised likelihood. However, when fitting Models 1 and 2 it is assumed that $h > 0$ as the nature of the log-likelihood surface when h = 0 causes difficulty when optimising.

A series of experiments were performed to ascertain the ability of each model to fit observations simulated from each model according to two test statistics:

1. The summed root mean squared error (SRMSE), T, of MLEs $\hat{\lambda}_{i,k}$ to the true value λ_i, as given for Model k by

$$T_k = \sum_{i=1}^{2} \sqrt{\frac{\sum_{j=1}^{m} \left(\hat{\lambda}_{i,k,j} - \lambda_i\right)^2}{m}}, \tag{7.15}$$

   where i indicates parameter 1 or 2, j gives the simulation number and k refers to Model number 1 or 2. To allow comparison, the percentage relative improvement in, T, under Models 1 and 2 relative to Model 0 will be presented as a percentage relative difference, d, given for Model k by

$$d_k = \frac{(T_0 - T_k)}{T_0} \times 100, \tag{7.16}$$

where $d_k = 0$ indicates no improvement in T over Model 0, $d_k > 0$ indicates an improvement in T over Model 0 and $d_k < 0$ indicates a worse T than Model 0.

2. The log-likelihood, $\ln(\mathcal{L})$, of the model and the number of free parameters, p, are used to calculate the Akaike Information Criterion (AIC), which can be used as a measure of statistical quality of a model and is given by

$$\text{AIC} = 2p - 2\ln(\mathcal{L}).$$

The model with the minimum AIC value is preferred.

In each case the true statistics were evaluated over m repetitions of the data. This process of cross-testing allows the comparison of these statistics over a grid of h and c, or solely c in the case of Model 0. In this section c will be limited to the bounds of the sample, i.e. $c \in (0, n]$.

The smoothing parameter h in the Gaussian weighting function, $\phi_h(t, c)$, is constrained by $h \geq 0$. Therefore, when comparing to a null model where $h = 0$, if a 0.05 significance level is desired, a 0.025 significance test should be performed using the standard test. The reasons for this are explained in Appendix G.2.

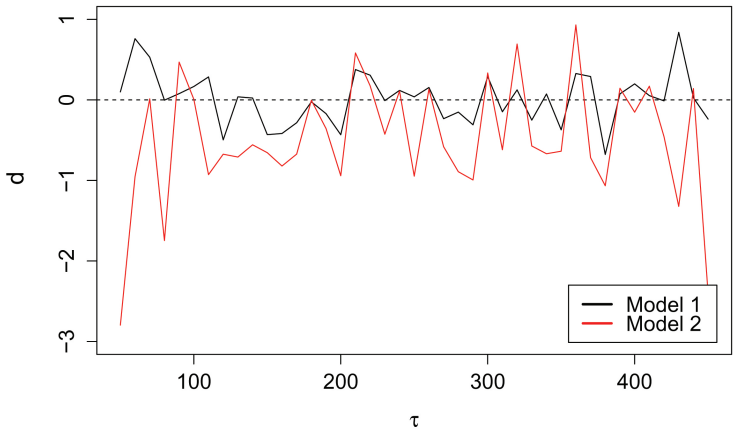


Figure 7.8: Percentage relative improvement in T over Model 0 for Models 1 and 2, d_1 and d_2 respectively, when simulating from Model 0 with $m = 200$, $n = 500$, $\lambda_1 = 5$ and $\lambda_2 = 10$ and changepoint τ at values of $\tau = 50, 60, \ldots, 450$.

Figure 7.8 shows d_1 and d_2 when simulating data from Model 0 with $m = 200$, $n = 500$, $\lambda_1 = 5$ and $\lambda_2 = 10$ and changepoint τ at values of $\tau = 50, 60, \ldots, 450$. It can be seen that

Model 1 performs similarly to Model 0 across the range of tested τ. However, at values near the bounds of the sample, Model 2 performs relatively worse by approximately 3% in terms of T.

When simulating under Model 0 and cross-testing alternative hypotheses of Models 1 and 2 against a null hypothesis of Model 0, Model 0 was rarely (less than 1%) rejected in favour of Models 1 or 2 under each simulation at all values of τ tested.

Figure 7.9 shows d_1 and d_2 over $m = 200$ repeated simulations under Model 1, where $n = 500$, $\lambda_1 = 5$, $\lambda_2 = 10$ over a discretised grid of $c = 50, 100, \ldots, 450$ and $h = 10, 20, \ldots, 100$. It can be seen that both Models 1 and 2 perform considerably better than Model 0 at high values of h. However, Model 1 exceeds Model 2 according to this statistic when simulating from Model 1. As $h \to 0$ both Models 1 and 2 collapse to Model 0 and their improvement decreases, which is expected given the results shown in Figure 7.8.

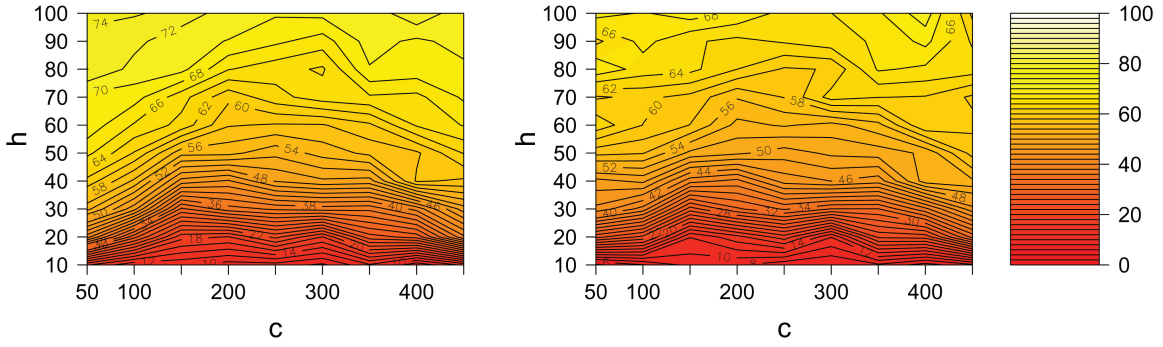


Figure 7.9: Contour plots (units of percentile) showing (left) d_1 and (right) d_2, when simulating from Model 1 using $n = 500$, $\lambda_1 = 5$, $\lambda_2 = 10$ over a discrete grid of $c = 50, 100, \ldots, 450$ and $h = 10, 20, \ldots, 100$.

Figure 7.10 shows the percentage of replicated cases, m, when either Model 0, 1 or 2 was the best fitting model according to the AIC when simulating under Model 1 over the same grid as above. This shows that Model 1 performs better than Models 0 or 2 at high values of h when simulating from Model 1. However, as $h \to 0$, the percentage of m where Model 0 was the best fitting model increases. Model 2 is rarely the best fitting model in this test.

Figure 7.11 shows d_1 and d_2 over $m = 200$ repeated simulations under Model 2, where $n = 500$, $\lambda_1 = 5$, $\lambda_2 = 10$ over a discretised grid of $c = 50, 100, \ldots, 450$ and $h = 10, 20, \ldots, 100$. It can be seen that both methods perform considerably better than Model 0 at high values of h when c is close to the centre of the sample, Model 2 in this sense being significantly better than Model 1 when simulating from Model 2 according to this statistic. In the same manner as before, as $h \to 0$ both Models 1 and 2 collapse to Model 0 and their im-

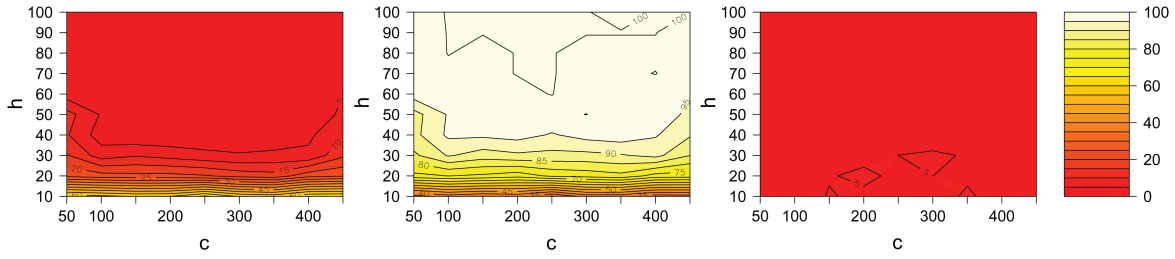


Figure 7.10: Contour plots (units of percentile) showing the percentage of occurrences when simulating from Model 1 using $n = 500$, $\lambda_1 = 5$, $\lambda_2 = 10$, where Model 0 (left), Model 1 (middle) or Model 2 (right) was the best fitting model according to the AIC, over a discrete grid of $c = 50, 100, \ldots, 450$ and $h = 10, 20, \ldots, 100$.

provement decreases. However, when cross-testing under Model 2, parameter estimates obtained under both Model 1 and 2 decrease in accuracy as c approaches the bounds of the sample space (i.e. $c \to 0$ or $c \to n$).

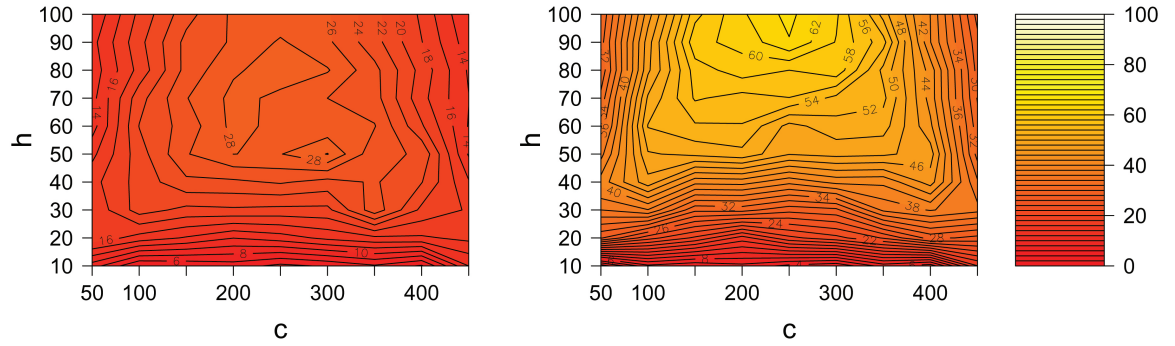


Figure 7.11: Contour plots (units of percentile) showing (left) d_1 and (right) d_2, when simulating from Model 2 using $n = 500$, $\lambda_1 = 5$, $\lambda_2 = 10$ over a discrete grid of $c = 50, 100, \ldots, 450$ and $h = 10, 20, \ldots, 100$

Figure 7.12 shows the percentage of m when either Model 0, 1 or 2 was the best fitting model according to the AIC when simulating under Model 2 over the same grid as above. This shows that Model 2 performs better than Models 0 and 1 at high values of h when simulating from Model 2. However, again, as $h \to 0$, the percentage of m where Model 0 was the best fitting model increases. Model 1 is rarely the best fitting model in this test. As $c \to 0$ or $c \to n$, Model 1 begins to lose out to Model 0. Due to these results it can be confirmed that data following Model 2, with c occuring outside of the sample space, would be difficult to model (e.g. when forecasting a change) using either Model 1 or 2.

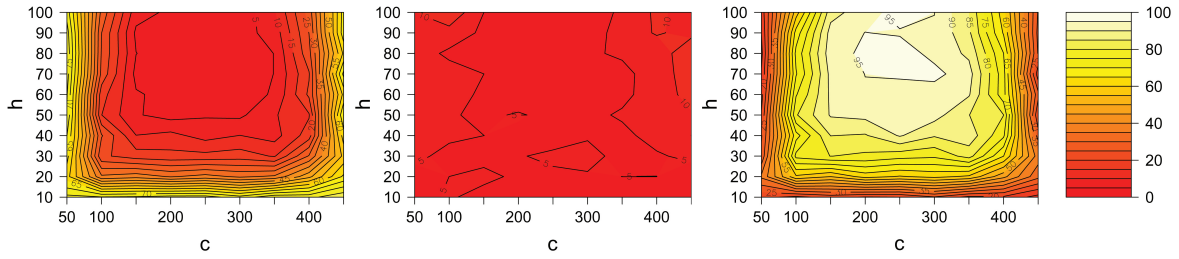


Figure 7.12: Contour plots (units of percentile) showing the percentage of occurrences when simulating from Model 2 using $n = 500$, $\lambda_1 = 5$, $\lambda_2 = 10$, where Model 0 (left), Model 1 (middle) or Model 2 (right) was the best fitting model according to the AIC, over a discrete grid of $c = 50, 100, \ldots, 450$ and $h = 10, 20, \ldots, 100$.

7.4.2 Forecasting Outside of the Bounds of the Sample

Time series forecasting is the process of utilising current observations to predict future ones. Forecasting the occurence of a step change, as used in the generic discrete change-point model, is impossible from information contained in a single series, due to the nature of the change. However, gradual changes often occur in practice and the prediction of a change and the related parameters is possible using the models discussed in Sections 7.2.1 and 7.2.2.

In Section 7.4 the log-likelihood was maximised over all data points in each sample. However, data streams can produce large volumes of data which require online processing, where memory is often limited. In these cases analysing historical data may be impractical or not useful. To account for this a rolling window can be used to check for structural changes in the time-series, whereby, as new data enter the sample, old data leave the sample. This addresses both speed and scalability.

In practical use on real data, the length of the window can be related to the timescale of the system. To allow the model to capture changes in the structure of the time series a smaller window should be used for a fast timescale and a longer window for a slow timescale. The evolution of parameter estimates from a rolling window method can also help to assess any statistical noise present in the data.

For the following examples a rolling window shall be used, with a length that represents the initial sample size over which the data is thought to follow one distribution. The window will move incrementally and remain constant in size, introducing new data points to allow a comparison of the forecasting ability of Models 0, 1 and 2 (see Section 7.4.1 for model nomenclature), whilst removing the oldest or first data points. The positioning of this window relative to the data index is given by the first point, w_1 and the end

point w_2 in the window, where $w_2 > w_1$. The window size is then $w_2 - w_1$. In practice, this increases the efficiency of analysing the data, rather than using fixed $w_1 = 0$, whilst ensuring a constant sample size.

Under Model 0, $m = 200$ data sets were simulated, where $n = 1100$, $\tau = 1000$, $\lambda_1 = 5$ and $\lambda_2 = 10$. A window of size $w_2 - w_1 = 900$ was then moved incrementally over $w_2 - \tau = -100, -90, \ldots, 100$, representing the point at which the window reaches the changepoint τ by $w_2 - \tau = 0$ and maximising the log-likelihood of Models 1 and 2 at each position. Using the MLEs obtained through these maximisations, two test statistics were applied, the individual RMSE for parameter λ_i, RMSE^{λ_i}, and the SMRSE, T. These are presented as percentage relative improvements over Model 0. The improvement in T, d, given by equation (7.16), is here relabelled as d^T whilst the improvment in RMSE for parameter λ_i, d_i^{RMSE}, is given for Model k by

$$d_{i,k}^{RMSE} = \frac{\text{RMSE}_0^{\lambda_i} - \text{RMSE}_k^{\lambda_i}}{\text{RMSE}_0^{\lambda_i}} \times 100. \tag{7.17}$$

A null model of no changepoint will also be introduced for comparative purposes. As model index 0 is already defined, this model will be referred to as null and given index $k = 3$ to ensure completeness in definition of k.

Figure 7.13 shows d_1^{RMSE}, d_2^{RMSE} and d^T for Model 1 and 2 and a null model of no changepoint for the tested sample subsets. It can be seen from Figure 7.13 that Models 0, 1, 2 and the null model all perform similarly well over the period of no change, with respect to the test statistics used. After $w_2 - \tau = 0$ it can be seen that Model 1 performs better than Model 2, but takes around 100 post τ observations to regain similar performance to Model 0.

This process was repeated simulating from Model 1 with $n = 1100$, $c = 1000$, $h = 100$, $\lambda_1 = 5$ and $\lambda_2 = 10$. Figure 7.14 shows d_1^{RMSE}, d_2^{RMSE} and d^T for Model 1 and 2 and a null model of no changepoint for the tested sample subsets. However the scale for the window is now $w_2 - c$ to represent the correct model parameter c, although the same window positions, $w_2 - c = -100, -90, \ldots, 100$, were analysed.

It can be seen from Figure 7.14 that Model 1 achieves the greatest improvement over Model 0 in the accuracy of both parameter estimates $\hat{\lambda}_1$ and $\hat{\lambda}_2$ prior to the position of equal weighting, c, when simulating from model 1. After c has passed, Model 2 steadily reaches a similar level of improvement as model 1 by approximately 50 post c observations.

Finally, simulating from Model 2 with $n = 1100$, $c = 1000$, $h = 100$, $\lambda_1 = 5$ and $\lambda_2 = 10$, Figure 7.15 shows d_1^{RMSE}, d_2^{RMSE} and d^T for Model 1 and 2 and a null model of no change-

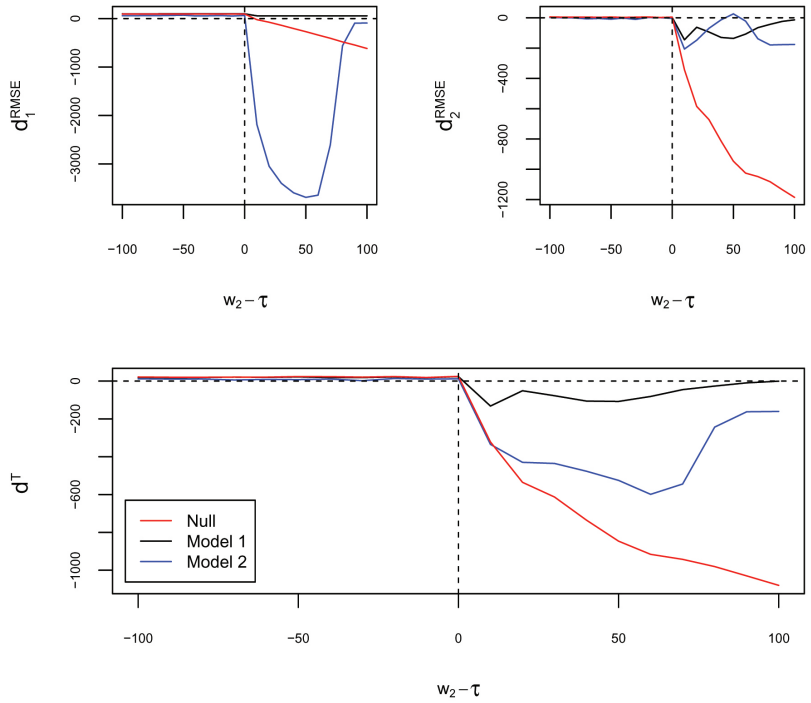


Figure 7.13: Simulating from Model 0 using $\tau = 1000$, $w_2 - w_1 = 900$, $\lambda_1 = 5$ and $\lambda_2 = 10$. (Top left) d_1^{RMSE}, (top right) d_2^{RMSE} and (bottom) d, for Models 1 and 2 and a null model of no changepoint.

point for the tested sample subsets using window positions $w_2 - c = -100, -90, \ldots, 100$. Interestingly, it can be seen that Model 1 achieves the greatest improvement over Model 0 in T prior to the position of c, when simulating from Model 2. This is in contrast to simulation from Models 0 and 1, where the correct model performs best out of those tested. After c has passed, Model 1 remains as the greatest improvement in T over Model 0 for approximately 20 post c observations and is consistently better than Model 2 in RMSE^{λ_1} over the tested range. After 20 post c observations, Model 2 experiences the most improved T with respect to Model 0.

7.4.3 Comparison Discussion

In this model comparison, the discrepancy between parameter estimates and their true values has been used as a measure of model quality. This is due to the range of possible applications which aim to predict or model the portion of the sample before and after the change. The changepoint position is not under scrutiny due to its ambiguity under a smooth change model when comparing to the generic discrete changepoint model. Models 1 and 2 may be referred to as smooth change models rather than changepoint models.

In a bounded scenario, when forecasting is not necessary, each model performs well when tested on simulations using that model. The situations which cause exception apply to

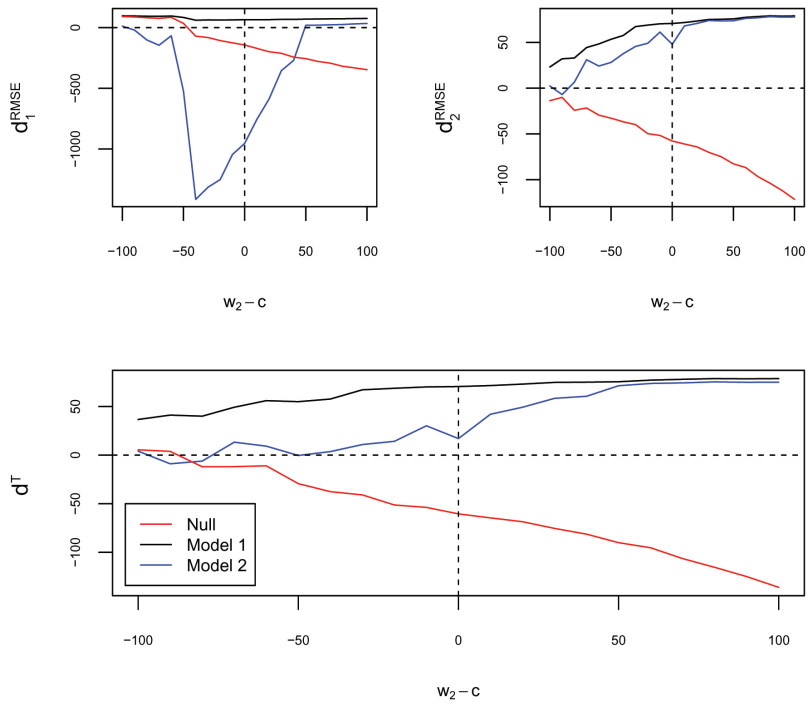


Figure 7.14: Simulating from Model 1 using $c = 1000$, $h = 100$, $w_2 - w_1 = 900$, $\lambda_1 = 5$ and $\lambda_2 = 10$. (Top left) d_1^{RMSE}, (top right) d_2^{RMSE} and (bottom) d, for Models 1 and 2 and a null model of no changepoint.

Model 2, which struggles to give a better estimate of parameter values when c (or τ) is close to the bounds of the sample. In these cases, Model 0 shows the highest quality when consulting the AIC, which takes into account the number of model parameters. This fact leads to the hypothesis that Model 2 may not be effective in a forecasting situation when the position of c may occur outside of the bounds of the sample.

Under forecasting conditions (i.e. $0 < c < \infty$) Model 1 continues to perform well. Under the measure of percentage improvement of T, d^T, it seems to be performing much worse than Model 0 when simulating from Model 0, but the actual difference has a relatively low order of magnitude.

Figure 7.15 shows that Model 1 performs better than Model 2 until approximately $w_2 - c = 20$ when simulating from Model 2. It can be deduced that simulated samples under Model 2 cause difficulties when forecasting parameters due to the fact that there is no visible end point to the smooth change. Whereas Model 1 uses the available data points in a more effective way, building a picture of future changes more slowly, but less chaotically. This is the reason why testing under Model 2 sees more noise in the parameter describing the data after the change.

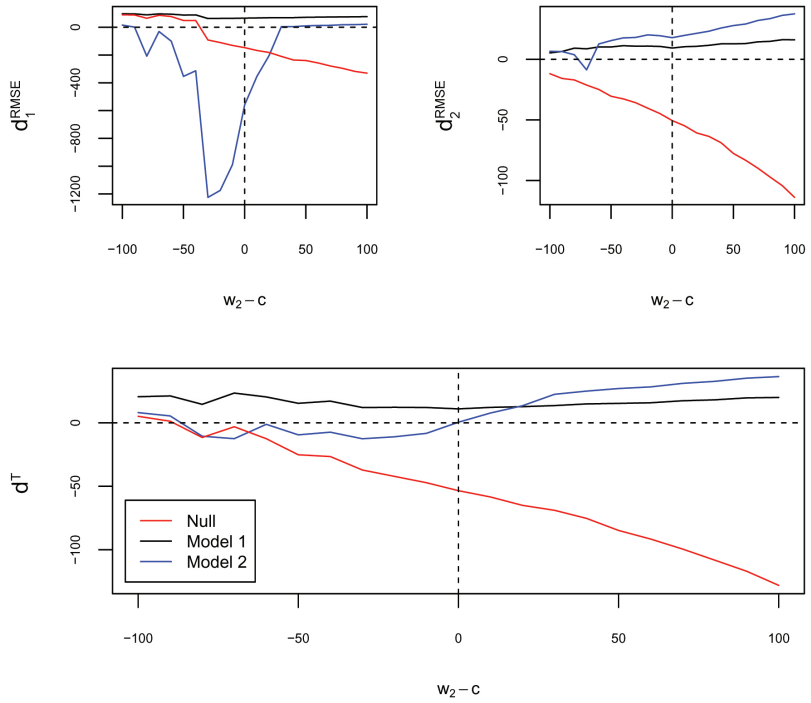


Figure 7.15: Simulating from Model 2 using $c = 1000$, $h = 100$, $w_2 - w_1 = 900$, $\lambda_1 = 5$ and $\lambda_2 = 10$. (Top left) d_1^{RMSE}, (top right) d_2^{RMSE} and (bottom) d, for Models 1 and 2 and a null model of no changepoint.

7.5 Extension to Multiple Changepoints

The generic discrete changepoint model can be extended to the multiple changepoint case quite simply. Building on the current definition of the single changepoint (definition 1, Section 7.1) a series of J changepoints, $\tau_1, \ldots \tau_J$ can be thought to occur when the statistical properties of neighbouring segments $\{x_{\tau_{j-1}}, \ldots, x_{\tau_j}\}$ and $\{x_{\tau_{j+1}}, \ldots, x_{\tau_{j+2}}\}$, where τ_0 represents the first index and τ_{J+1} represents the last index of the sample, are different in some way.

To be used as an effective alternative or inline model to the generic discrete changepoint model the smoothly weighted mixed distribution model defined in Section 7.2.2 needs to account for the multiple changepoint scenario. The smoothly weighted mixed parameter model described in Section 7.2.1 will not be considered here, however, similar derivations can be applied. To mark its completeness in describing the multiple changepoint scenario, and to ensure concise and legible analysis, the smoothly weighted multiple mixed distribution model shall now be referred to as Smooth Distribution Transitions (SDT).

7.5.1 Definition of Smooth Distribution Transitions (SDT)

Consider J multiple smoothly weighted mixed distribution changes (as described in Section 7.2.2) indexed by j, $\boldsymbol{c} = (c_1, \ldots, c_J)$, between $J+1$ segments, indexed by i, following probabililty distribution functions given by $f(x_t, \boldsymbol{\lambda_i})$. The likelihood under this model is given by

$$\mathcal{L}(\boldsymbol{c}, \boldsymbol{h}, \boldsymbol{\lambda_1}, \ldots, \boldsymbol{\lambda_{J+1}}) = \prod_{t=1}^{n} \left[\sum_{i=1}^{J+1} \phi_{\boldsymbol{h},i}(t, \boldsymbol{c}) f(x_t, \boldsymbol{\lambda_i}) \right],$$

where

$$\sum_{i=1}^{J+1} \phi_{\boldsymbol{h},i}(t, \boldsymbol{c}) = 1,$$

and $\boldsymbol{h} = (h_1, \ldots, h_J)$ is the vector of smoothing parameters respective to $\boldsymbol{c}$.

To illustrate the model it shall be assumed that

$$X_t \sim \begin{cases} \text{Poisson}(\lambda_1) & \text{w.p. } \phi_{h,1}(t, c), \\ \text{Poisson}(\lambda_2) & \text{w.p. } \phi_{h,2}(t, c), \\ \vdots & \\ \text{Poisson}(\lambda_{J+1}) & \text{w.p. } \phi_{h,J+1}(t, c). \end{cases}$$

The likelihood is then given by

$$\mathcal{L}(\boldsymbol{c}, \boldsymbol{h}, \lambda_1, \ldots, \lambda_{J+1}) = \prod_{t=1}^{n} \left[\sum_{i=1}^{J+1} \phi_{\boldsymbol{h},i}(t, \boldsymbol{c}) \frac{\lambda_i^{x_t} e^{-\lambda_i}}{x_t!} \right].$$

Taking the log of the right hand side gives the log-likelihood as shown by

$$\begin{aligned} \ell(\boldsymbol{c}, \boldsymbol{h}, \lambda_1, \ldots, \lambda_{J+1}) &= \log \left\{ \prod_{t=1}^{n} \left[\sum_{i=1}^{J+1} \phi_{\boldsymbol{h},i}(t, \boldsymbol{c}) \frac{\lambda_i^{x_t} e^{-\lambda_i}}{x_t!} \right] \right\} \\ &= \sum_{t=1}^{n} \log \left\{ \sum_{i=1}^{J+1} \left[\phi_{\boldsymbol{h},i}(t, \boldsymbol{c}) \frac{\lambda_i^{x_t} e^{-\lambda_i}}{x_t!} \right] \right\}. \end{aligned}$$

7.5.2 SDT Weighting Functions

Similarly to that described in Section 7.2.3, a weighting function based on a Gaussian cumulative distribution function may be employed in SDT to mix the probability density functions used to describe each segment. In the two changepoint case, where $J = 2$, the

weighting function can be defined as shown in equation (7.18):

$$\begin{aligned}
\phi_{\boldsymbol{h},1}(t,\boldsymbol{c}) &= 1-\Phi\left(\frac{t-c_1}{h_1}\right),\\
\phi_{\boldsymbol{h},3}(t,\boldsymbol{c}) &= \Phi\left(\frac{t-c_2}{h_2}\right),\\
\phi_{\boldsymbol{h},2}(t,\boldsymbol{c}) &= \phi_{\boldsymbol{h},1}(t,\boldsymbol{c})-\phi_{\boldsymbol{h},3}(t,\boldsymbol{c}).
\end{aligned} \tag{7.18}$$

However, this cannot be translated into the case where the number of changepoints is greater than two. To account for this, we can set:

$$\phi_{\boldsymbol{h},j}\left(t,\boldsymbol{c}\right)=\frac{\alpha_{\boldsymbol{h},j}\left(t\right)}{\sum_{k=1}^{J+1}\alpha_{\boldsymbol{h},k}\left(t\right)}. \tag{7.19}$$

The two changepoint case can then be described by

$$\begin{aligned}
\alpha_{\boldsymbol{h},1}(t,\boldsymbol{c}) &= 1-\Phi\left(\frac{t-c_1}{h_1}\right),\\
\alpha_{\boldsymbol{h},2}(t,\boldsymbol{c}) &= \min\left\{\Phi\left(\frac{t-c_1}{h_1}\right),\left[1-\Phi\left(\frac{t-c_2}{h_2}\right)\right]\right\},\\
\alpha_{\boldsymbol{h},3}(t,\boldsymbol{c}) &= \Phi\left(\frac{t-c_2}{h_2}\right).
\end{aligned} \tag{7.20}$$

This translates into the J changepoint case, as shown in equation (7.21):

$$\begin{aligned}
\alpha_{\boldsymbol{h},1}(t,\boldsymbol{c}) &= 1-\Phi\left(\frac{t-c_1}{h_1}\right),\\
\alpha_{\boldsymbol{h},j}(t,\boldsymbol{c}) &= \min\left\{\Phi\left(\frac{t-c_{j-1}}{h_{j-1}}\right),\left[1-\Phi\left(\frac{t-c_j}{h_j}\right)\right]\right\} \quad \text{for} \quad j=2,\ldots,J,\\
\alpha_{\boldsymbol{h},J+1}(t,\boldsymbol{c}) &= \Phi\left(\frac{t-c_J}{h_J}\right).
\end{aligned} \tag{7.21}$$

Figure 7.16 shows two example weightings for two smooth changes at $c_1 = 250$ and $c_2 = 750$ for values of $h_1 = h_2$ equal to 10 and 100.

It is noteworthy that multiple (more than two) segments may be mixed due to large amounts of smoothing in the weighting function. This necessitates an addendum to the definition of c that c is the the point of equal weighting between *neighbouring* segments, which is not necessarily equal to 0.5. Figure 7.17 illustrates this.

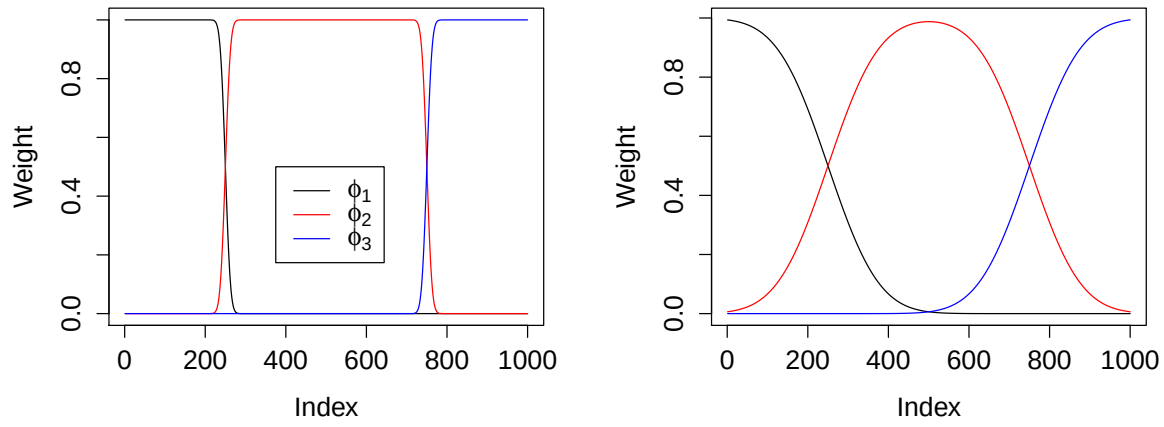


Figure 7.16: Weights for two smooth changes at $c_1 = 250$ and $c_2 = 750$ for (left) $h_1 = h_2 = 10$ and (right) $h_1 = h_2 = 100$.

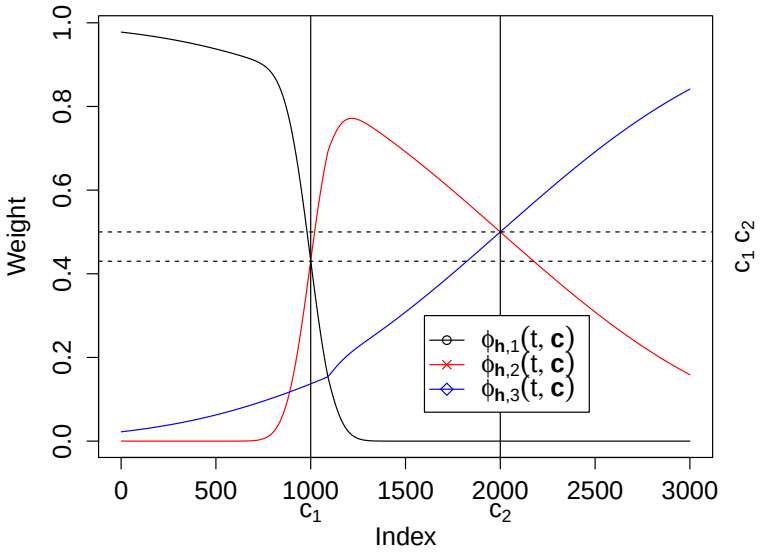


Figure 7.17: Weights for two smooth changes where $c_1 = 1000$, $c_2 = 2000$, $h_1 = 100$ and $h_2 = 1000$.

7.5.3 Illustration of Improvement of SDT over the Generic Discrete Changepoint Model

An illustration of the improvement of SDT over the generic discrete changepoint model can be achieved using the percentage relative improvement in RMSE over the generic discrete changepoint model, d_i^{RMSE} as given in equation (7.17). This will be done for each parameter, λ_i, where each segment follows a Poisson distribution.

Data were simulated using $m = 200$ repetitions of Poisson SDT with $n = 1000$, $c_1 = 250$ and $c_2 = 750$, $\lambda_1 = 5$, $\lambda_2 = 10$ and $\lambda_3 = 5$, whilst varying a common smoothing parameter h over discrete values $h = 0, 10, 20, \ldots, 150$. Example simulated samples using $h = 0$ and $h = 150$ are shown in Figure 7.18, where $h = 0$ is the generic discrete changepoint case. It can be seen that the changes are much less identifiable by eye when $h = 150$.

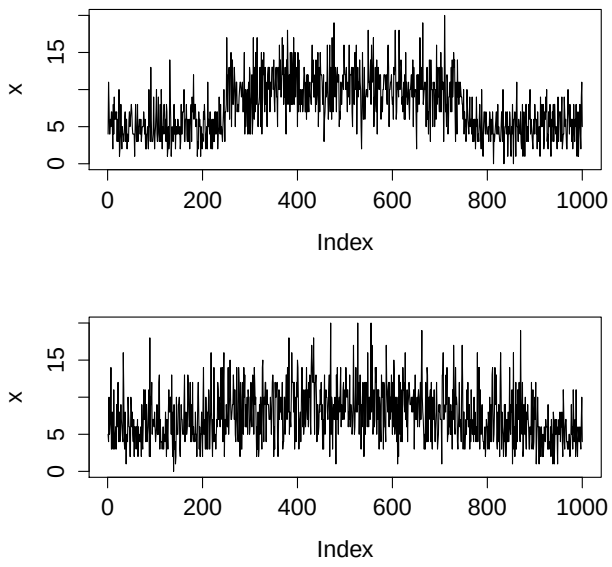


Figure 7.18: (Top) Simulated data set using two changepoints under the generic discrete Poisson changepoint model at $\tau_1 = 250$, $\tau_2 = 750$, $\lambda_1 = 5$, $\lambda_2 = 10$ and $\lambda_3 = 5$. (Bottom) Simulated data set using two smooth changes under Poisson SDT with $h = 150$, $c_1 = 250$, $c_2 = 750$, $\lambda_1 = 5$, $\lambda_2 = 10$ and $\lambda_3 = 5$.

Figure 7.19 shows d_i^{RMSE} testing Poisson SDT for each parameter λ_i, where in this case $i = 1, 2, 3$. Note $\lambda_1 = \lambda_3 = 5$ and therefore the percentage improvement relative to the generic discrete changepoint model is similar. It can be seen that d_i^{RMSE} increases to some limiting factor for all i, dependent on the true value of λ_i, as h increases. It should also be noted that SDT performs similarly to the generic discrete changepoint model in this test as $h \to 0$.

Figure 7.20 shows the average position of τ_1 and τ_2 for the generic discrete Poisson changepoint model and c_1 and c_2 under Poisson SDT at the tested values of h. It is noteworthy that the point estimate for τ when tested using the generic discrete changepoint model moves away from the segment simulated from a larger parameter value in both cases as h increases. This is also true in the case of $\lambda_1, \lambda_3 > \lambda_2$. Therefore, it cannot be equated to the start or end of the smooth change.

Figure 7.21 shows the percentage rejection in a hypothesis test using a null hypothesis of a generic discrete Poisson changepoint model compared to an alternative hypothesis

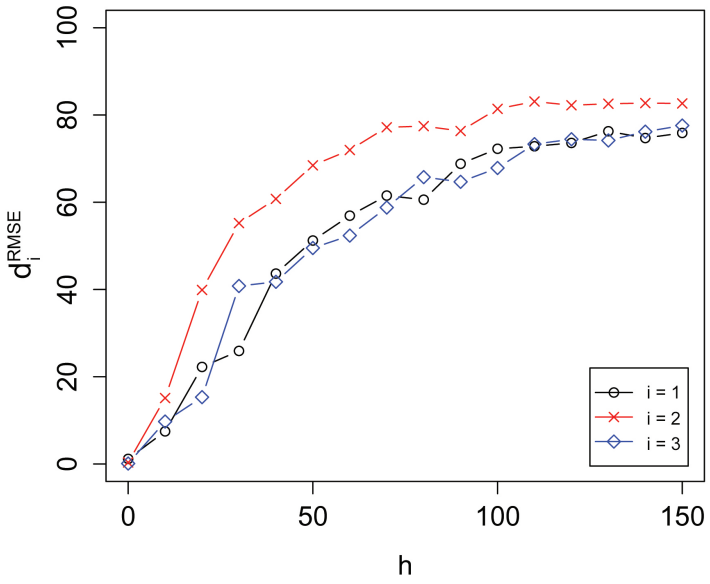


Figure 7.19: Values of d_i^{RMSE}, $i = 1, 2, 3$, testing $m = 200$ repetitions of Poisson SDT with $c_1 = 250$, $c_2 = 750$, $\lambda_1 = 5$, $\lambda_2 = 10$ and $\lambda_3 = 5$, and varying h over discrete values $h = 0, 10, 20, \ldots, 150$ using Poisson SDT.

of Poisson SDT, testing at a significance level of 0.05. It can be seen that the null is rejected 100% of the time by approximately $h = 40$, which indicates that the model with the highest quality in this situation is considered to be SDT.

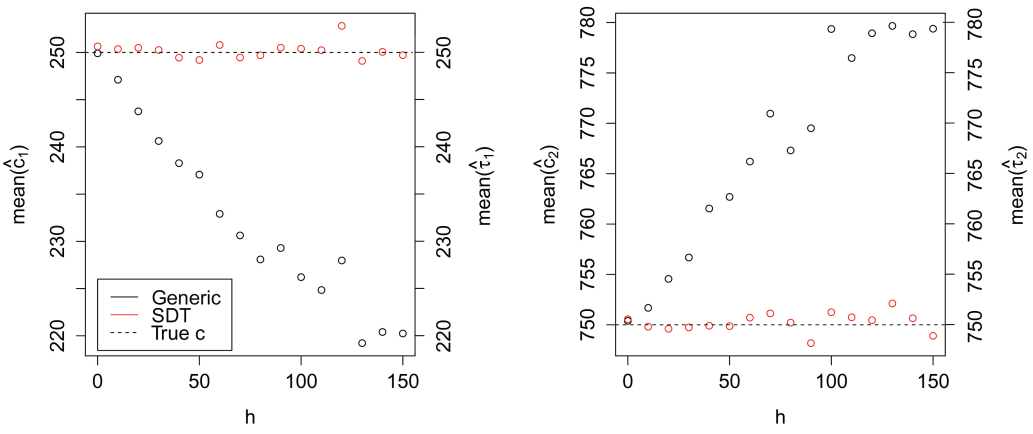


Figure 7.20: Average estimates of τ_1 and τ_2 for the generic discrete multiple changepoint model (black) and c_1 and c_2 for Poisson SDT (red) when simulating using Poisson SDT with $c_1 = 250$, $c_2 = 750$, $\lambda_1 = 5$, $\lambda_2 = 10$ and $\lambda_3 = 5$, and varying h over discrete values $h = 0, 10, 20, \ldots, 150$.

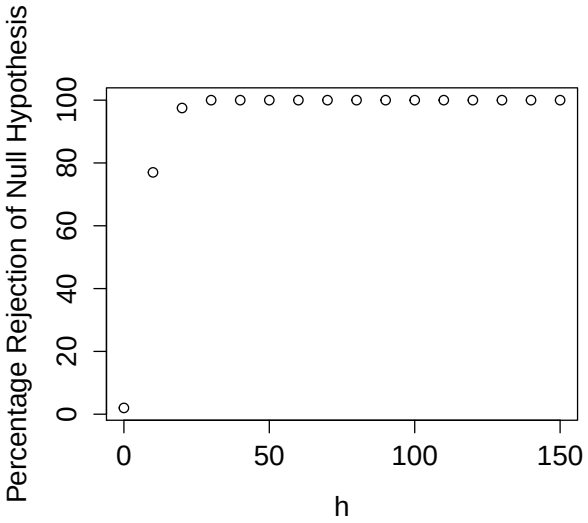


Figure 7.21: Percentage rejection of null hypothesis of a generic discrete Poisson changepoint model when compared to an alternative hypothesis of Poisson SDT when simulating using Poisson SDT with $c_1 = 250$, $c_2 = 750$, $\lambda_1 = 5$, $\lambda_2 = 10$ and $\lambda_3 = 5$, and varying h over discrete values $h = 0, 10, 20, \ldots, 150$.

7.5.4 Detecting Number of Changepoints and Comparison to PELT

Automated detection of the number of changepoints or smooth changes to be analysed is a prerequisite of any changepoint model to be used on non-simulated data where this number is not known. Using the AIC, or some penalty function, the number of parameters in the changepoints and smoothing parameters, $2 \times J$, can be estimated.

To further illustrate the model normal SDT data with a constant known standard deviation $\sigma = 1$ and varying means μ_i, will be used to test the choice of J. Assuming

$$X_t \sim \begin{cases} \text{Normal}\left(\mu_1, \sigma\right) & \text{w.p. } \phi_{h,1}\left(t, c\right), \\ \text{Normal}\left(\mu_2, \sigma\right) & \text{w.p. } \phi_{h,2}\left(t, c\right), \\ \quad \vdots & \\ \text{Normal}\left(\mu_{J+1}, \sigma\right) & \text{w.p. } \phi_{h,J+1}\left(t, c\right), \end{cases}$$

the likelihood is given by

$$\mathcal{L}\left(\boldsymbol{c}, \boldsymbol{h}, \mu_1, \ldots, \mu_{J+1}\right) = \prod_{t=1}^{n} \left[\sum_{i=1}^{J+1} \phi_{\boldsymbol{h},i}\left(t, \boldsymbol{c}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(x - \mu_i\right)^2}{2} \right) \right].$$

Taking the log of the right hand side gives the log-likelihood as shown by

$$\begin{aligned}\ell\left(\boldsymbol{c},\boldsymbol{h},\mu_1,\ldots,\mu_{J+1}\right) &= \log\left\{\prod_{t=1}^{n}\left[\sum_{i=1}^{J+1}\phi_{\boldsymbol{h},i}\left(t,\boldsymbol{c}\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\left(x-\mu_i\right)^2}{2}\right)\right]\right\}\\ &= \sum_{t=1}^{n}\log\left\{\sum_{i=1}^{J+1}\left[\phi_{\boldsymbol{h},i}\left(t,\boldsymbol{c}\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\left(x-\mu_i\right)^2}{2}\right)\right]\right\}.\end{aligned}$$

Figure 7.22 shows a sample of normal SDT data with $n = 3000$, $c_1 = 1000$, $c_2 = 2000$, $h_1 = 100$, $h_2 = 1000$, $\mu_1 = 5$, $\mu_2 = 10$, $\mu_3 = 5$ and $\sigma = 1$ for all segments. When tested using the AIC to ascertain the value of J in $m = 200$ simulations the results were highly dependent on starting value, most likely due to the optimisation routine used. However, if reasonable starting values were supplied SDT performed well under testing using the AIC to choose the correct J in all simulations. Note, this does not mean that it will be correct 100% of the time. However, under a limited number of simulations it performed admirably well.

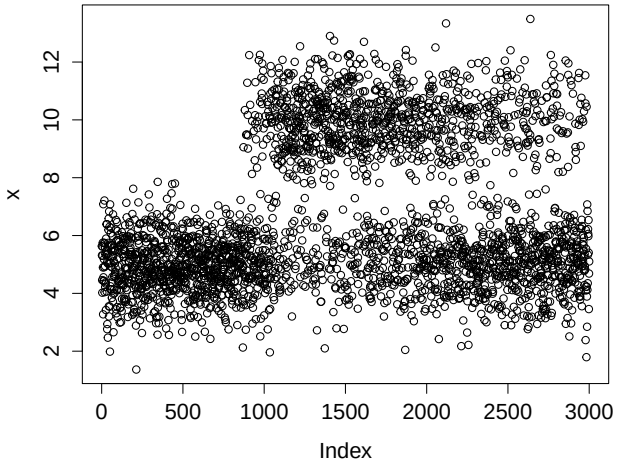


Figure 7.22: Example simulated data following normal SDT with $n = 3000$, $c_1 = 1000$, $c_2 = 2000$, $h_1 = 100$, $h_2 = 1000$, $\mu_1 = 5$, $\mu_2 = 10$, $\mu_3 = 5$ and $\sigma = 1$ for all segments. Note the weighting pattern used is the same as shown in Figure 7.17.

The introduction of differing values of h_j allows the consideration that some changes are more abrupt than others. It also incurs the benefit of understanding that multiple changes can happen at the same time to one data set, as can be seen in Figure 7.22, where regions experiencing smooth changes between distributions visibly overlap.

The RMSE of the various free parameters' estimates for the simulations are given in Table 7.1. Note the disparity in RMSE between h_1 and h_2. This does not translate into inaccuracy in estimates of the mean parameters, μ_i. If the RMSE values are considered

as percentage errors to the true parameter value, they are infact quite similar.

Parameter	RMSE
c_1	11.56
c_2	12.86
h_1	12.14
h_2	73.50
m_1	0.04
m_2	0.03
m_3	0.03

Table 7.1: RMSE values for $m = 200$ simulations of normal SDT data with $n = 3000$, $c_1 = 1000$, $c_2 = 2000$, $h_1 = 100$, $h_2 = 1000$, $\mu_1 = 5$, $\mu_2 = 10$, $\mu_3 = 5$ and $\sigma = 1$ for all segments.

Interestingly, using a readily available Pruned Exact Linear Time (PELT) method to estimate changes in mean (Eckley et al., 2011), under both penalty regimes which count the changepoints as parameters and those that do not, between 100-700 changepoints were found in the simulated data sets. The inability of the PELT method to identify smooth changes and only identify discrete steps meant a gross overestimation of the number of changepoints. Due to this fact, direct comparison of parameter estimates is unhelpful.

This also shows that SDT has evolved to a much larger application base as its similarities with the generic discrete changepoint model become blurred. This evolution brings into question the basis of the model on segmentation, as it was first conceived. The initial concept sought to account for uncertainty in parameter estimates and changepoint position by mixing segments, however, as can be seen in Figure 7.22 segments may not be distinguishable in the typical sense. Therefore, the definition of c should solely be the point of equal weighting between probability distribution functions of neighbouring indices.

7.6 Conclusion

Changepoint analysis is a useful tool for detecting structural changes in the distribution of a time series. They can model changes in mean, variance, correlation and spectral density. In this way they can be used to model smooth changes disjointly, though this does not allow them to smoothly model the transition of the change over time.

The methods laid out in this chapter give a framework for building changepoint models

that consider such smooth transitions, using a weighting function to continuously model the change in confidence of information from observations surrounding the changepoint(s). The model can fit sudden changes and smooth changes alike, allowing it to model changes which display either of these characteristics.

Further to modelling changepoints retrospectively, the method also allows for the estimation of upcoming changes and the period of transition. Obviously, this relies on having some prior data to test models for both the distribution of the data and the weighting function. A Gaussian weighting function was discussed in detail, though a linear or other curve weighting function could be used if it proves to fit the data better.

Performing a simulation study showed that the SDT method was highly advantageous over current methods (PELT) in identifying (at least a low number) of changepoints of this nature. The specific advantages of the model were in the estimation of number of changepoints, where PELT estimated a much higher number of changes. When interpreting such results, this would be misleading and could incorrectly identify causal events.

7.7 Future Work

Much further simulation analysis, including the testing of several transition weighting models, should be carried out to identify the most efficient models. The application to a real data set would allow much clearer conclusions about the properties and benefits of this model, possibly in which the changepoints are relatable to a known event.

The automation of multiple changepoint detection under the SDT model needs to be perfected before it could be released as a package. This is difficult due to the nature of the log-likelihood surface, in that it displays multiple local maxima. A different optimisation algorithm may allow a better result. Also, changes to the optimisation method could significantly decrease the time to convergence, which is an area where current methods are much more capable.

Chapter 8

Changes of Performance in Golf

8.1 Introduction

Due to the nature of the game, golf provides an interesting and diverse set of statistical problems, not least of which is the effect of technology and coaching on outcome. Prize money and a popular basis for betting ensure that there is a constant interest in factors which may be used to predict player performance. Modern players have the ability to choose equipment that is customised to their tastes. As technology becomes more advanced, with innovative designs and new combinations of material, the first to adopt can often achieve an advantage over their competitors. This can be seen through the uptake of new solid golf balls after Tiger Woods achieved great success subsequent to his replacement of wound balls with a multi-core, solid design in 2001 (Masataka, 2008). Appendix H.1 shows gives some examples of where technology might affect player performance.

Tiger's success throughout the season was not solely a case of absolute performance, but also one of consistency. It has been speculated that the choice of golf ball may affect a player's consistency throughout a tour (Johnson, 2001; Masataka, 2008). This exploratory analysis will outline the techniques that may be used to investigate player consistency, allowing the identification of periods of higher consistency with the aim of identifying the cause.

Perceived changes in consistency and relative performance may also be *strength in depth*, which refers to the existence of more good players in the competition. This could therefore provide spurious results and should be considered in any conclusions.

8.2 Background

Lili (2005) analysed various statistical aspects of golf play over the last 40 years. Although some of the methods contained in the thesis are basic, the approach was a good starting point. Lili (2005) suggested the use of round by round scores as a fair comparison of play

on that particular day. However, altering weather conditions within a tournament may mean that difficulty changes from day to day, requiring a normalisation of each player's score to that of the field. This is further compounded throughout the tour, as each course holds particular challenges that are difficult to rate with regards to their required level of play.

A standard score (often referred to as a z-score) signifies by how many standard deviations a single observation at time point i, x_i, in a data set is above or below the mean value. This is done in such situations where raw score cannot be affectively utilised. Normalisation is achieved by

$$z_i = \frac{x_i - \mu}{\sigma}, \tag{8.1}$$

where μ is the mean of the population and σ is the standard deviation of the population. A positive z-score represents the number of standard deviations above the mean and negative corresponds to below (Larsen and Marx, 2012).

Relatively little separates the scores of the highest ranked players in a golf tour. Because of this, in addition to overall performance, consistency allows players to dominate the season. Consistent players will achieve their average handicap more often and therefore maintain their performance and position in tournaments throughout a tour. This is increasingly important for top players, as it is common for many different players to win each tournament. One measure of consistency is the variance of the normalised score, where a lower variance corresponds to a more consistent performance and vice versa.

Agreement between initial analyses and the research by Lili (2005) could not be acheived, possibly due error in the research or the abscence of the original data set that was used. Therefore, only the consistency measure was brought forward into this analysis.

8.3 Player Consistency

Data containing player round score were initially collected for the PGA Tour, with the aim of calculating a consistency measure based on the average variance of a player's normalised round score. The normalised round score, or 'z-score', may be calculated using equation 8.1.

Normalising to each round of play for each tournament allows for course difficulty, variability of day to day weather conditions and any other factors. The limited availability of historical data only allowed the capture of 10 years for the PGA Tour, which gave little insight into change in player performance.

The European Tour provided a more comprehensive data set for the years 1975 onwards. Data has been collected from 1975-2010 for all tournaments following standard stroke play (discounting match play tournaments etc...). After considerable reformatting of the data using a 'pretty print' type program written in R, the z-scores were calculated for all rounds over the 36 year data set. The average consistency (variance of z-scores) of the top 50 players who competed in 5 tournaments or more can be seen in Figure 8.1.

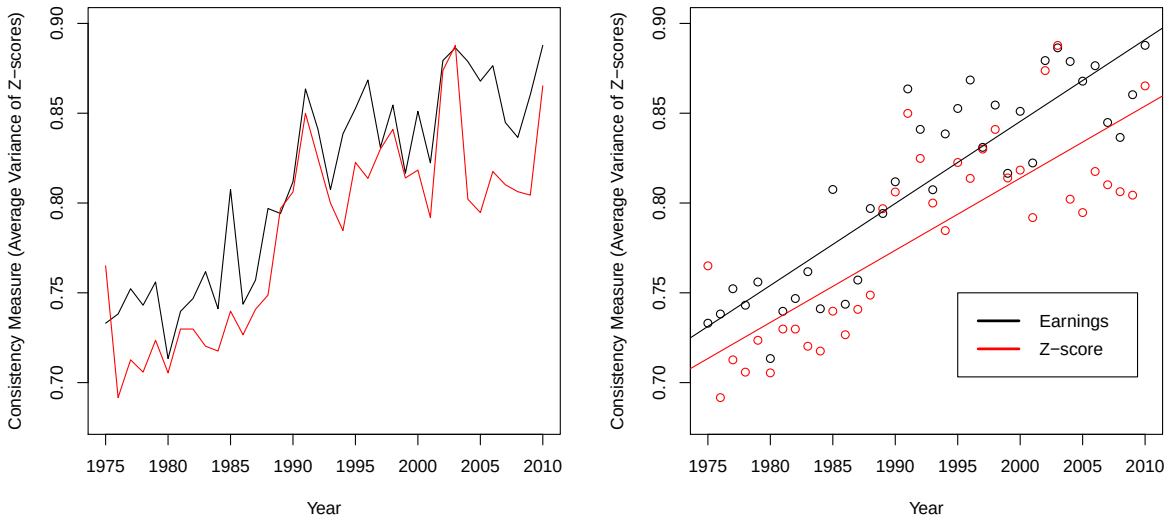


Figure 8.1: (left) Consistency measure of average variance of z-scores over time for both the top 50 players according to earnings and to z-scores (averaged over tournaments played by each player). (right) Linear regressions.

The consistency of the top 50 players were analysed using the variance of their z-scores over all rounds (data has also been captured for the individual round variance of z-scores for these players). The definition of the top 50 players was approached from two directions, the top 50 by earnings on tour (does not include advertising etc... or other tournament contributions) and the top 50 by z-score. Calculation of the top 50 by z-score takes into account not all players play every tournament by using X_i, as given by

$$X_i = \frac{\sum_{t \in T_i} Z_{t,i}}{|T_i|}, \tag{8.2}$$

where $t \in T_i$ represents the tournaments played by player i, $Z_{t,i}$ is the score in each tournament t, played by player i and $|T_i|$ is the number of tournaments played by player i, represented as the length of the set T_i. The results can be seen in Figure 8.1.

It was clear from initial analysis of this data set that there is a significant upward trend,

which can also be seen in Figure 8.1. Table 8.1 shows the results of linear regressions carried out on the overall consistency measure (average variance) over time for both ranking techniques (earnings and z-score). It can be seen that the p-values for both are much less than 0.05 and therefore, can be identified as significant trends.

Ranked By	Intercept	Gradient	p-value
Earnings	-8.28	0.00456	0.000
Z-Score	-7.22	0.00402	0.000

Table 8.1: Linear regression coefficients and p-values for both European Tour average top 50 consistency measures.

To visually enhance the relationship between average player consistency and time a 5 point average can be applied. This is shown in Figure 8.2, and reveals a possible change-point around 1988.

Figure 8.2: Consistency measure of average variance of z-scores over time for both the top 50 players according to earnings and to z-scores (averaged over tournaments played by each player), with 5 point average overlaid.

An important consideration was the exclusion of cut players in the calculation of z-scores. Figure 8.3 shows the consistency measure over time when all players, the top 100 players and the top 50 players in the tournament are used to calculate the z-score. It should be noted that the gradient component of linear regressions carried out upon these three results are decreasing as the number of players used to calculate the z-scores decreases (All: 0.0046, top 100: 0.0034, top 50: 0.0018).

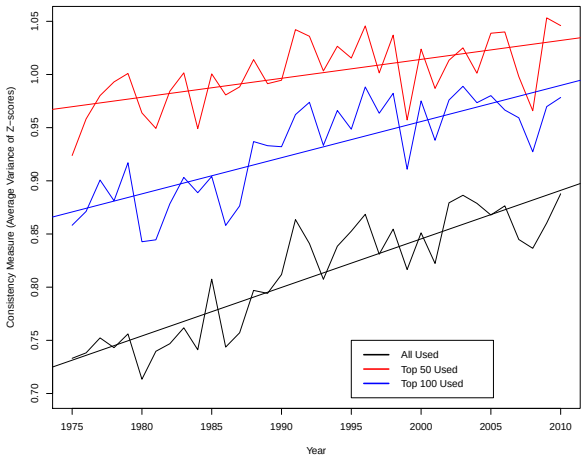


Figure 8.3: Consistency measure of average variance of Z-scores over time for both the top 50 players according to earnings and to z-scores, using all players in the tournament to calculate z-scores for each round (black), the top 50 (red) and top 100 (blue).

Another important decision relates to the variability measure of the z-score calculation. Usually this is left as the standard deviation, however, it could be replaced with the absolute difference between the median and either the upper or lower quartiles as can be seen in Figure 8.4. This would allow the effect of the upper and lower tails of the density distribution to be seen. Figure 8.5 shows a 100pt moving average of the mean and

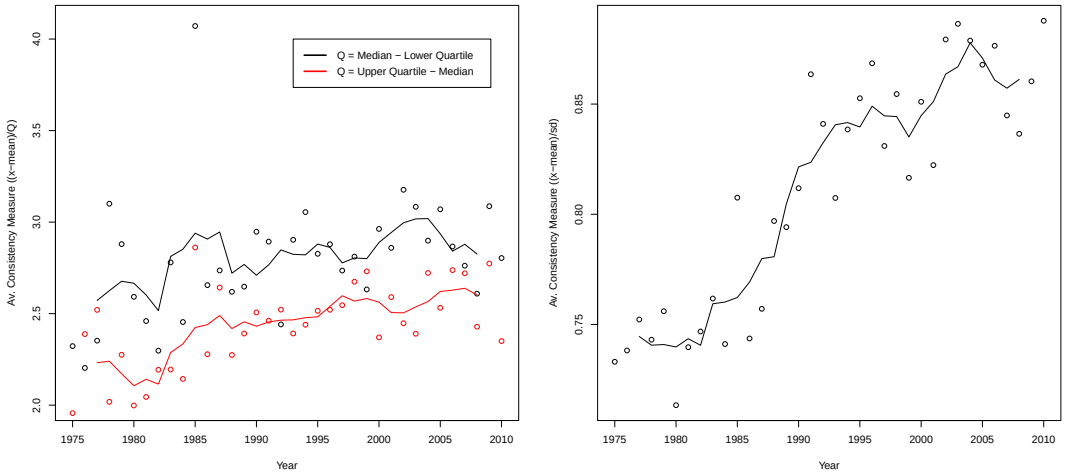


Figure 8.4: Consistency of average variance of z-scores over time for both the top 50 players according to earnings and to z-scores using different variability measures on the denominator of the z-score calculation (labelled as Q on the left plot). (right) standard deviation used (as before) for the denominator term.

standard deviation of the round score between 1975 and 2010. Both of these statistical measures can be seen to be decreasing. The decrease in standard deviation may point to an increase of strength in depth of the field as more players enter tournaments. This may thereby affect the perception of results achieved using the standard z-score method

as shown in equation (8.1).

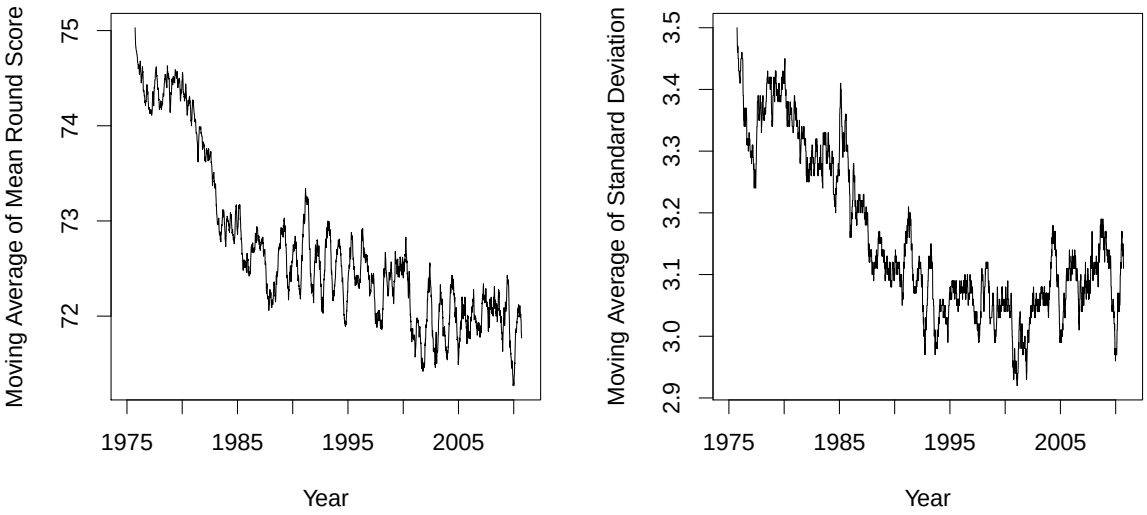


Figure 8.5: (Left) 100pt moving average mean round score for all rounds from 1975 to 2010. (Right) 100pt moving average standard deviation of round score for 1975 to 2010.

To address this issue, the variability measure, standard deviation in equation (8.1), may be removed from the denominator and a normalisation method of $z = x - \mu$ may be employed. Figure 8.6 shows the average consistency, measured as the variance of z-score of each player, for the top 50 players ranked by earnings using this normalisation method. In contrast to the results shown in Figure 8.1, the average consistency of the top 50 players is now relatively constant, with a slight increase in consistency between 1975 to 2000. This supports the theory that the decrease in consistency described by Figure 8.1 is related to strength in depth.

8.4 Discussion

Strength in depth is a term that is used widely in the golf community (Golfshake, 2015; Lindsay, 2010), and can be seen as an apparent cause of the perceived decrease in consistency shown in Figure 8.1. However, it should be considered that the reduction in standard deviation of round score between 1975 and 2010 shown in Figure 8.5 begins to plateau in the second half. This could give support to a counter arguement that as players become more consistent and hit the ball futher (due to improved technology and coaching methods), the courses are made more difficult by increasing their length (Diaze, 2011).

As the player is playing against a course and not an opponent, as in Tennis for example, if the course difficulty increases, the player's relative skill decreases. This could

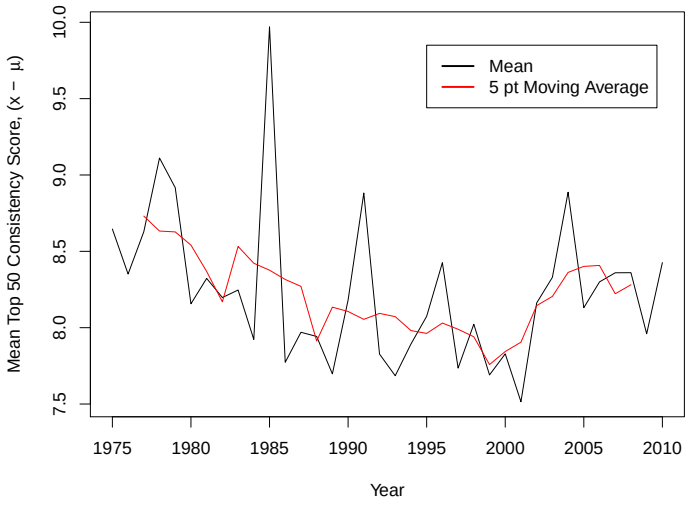


Figure 8.6: Mean consistency of the top 50 players from 1975-2010 using the consistency measure var $(\boldsymbol{x} - \mu)$.

be an explanation of the plateau of both the mean and standard deviation of round scores as seen in Figure 8.5, rather than their continued reduction - leading to the conjecture that this effect may have been engineered by the Tour organisers and course designers.

8.5 Future Work

As discussed in Section 8.3, the availability of free historical data from the PGA Tour was low. Future work would need to compare the results found in this analysis with those of the PGA Tour, over a similar time period.

ShotLink allows the capture of real-time statistical data regarding the trajectory and distance travelled by the ball for every player in the PGA Tour (ShotLink, 2015). Alongside other data, this would allow an investigation of the hypothesis that player form and technique have changed to reflect the increasing course length.

Chapter 9

Conclusion

This thesis has covered an analysis of home advantage and other sports data analysis: From a basic additive model for home and away goals in association football (Clarke, 1996) to a more complex model which considered the distribution of data (Dixon and Coles, 1997), various methods have been explored to better model the effect of home advantage. These include considering the effect of time, within game effects (such as red and yellow cards) and external covariates (such as distance between teams and match attendance). Modelling extensions which sought to account for extreme goal counts were performed, resulting in the evolution of a transition mixture model. Methods for considering smooth transitions were considered, and a changepoint model was discovered which allows the consideration of information close to and distant from the transition. Finally, a brief foray was made into changes in performance in professional golf.

Chapter 3 analysed home advantage at both the individual team and overall league levels. This analysis used the additive model for goal counts in association football defined in Clarke (1996). It was found that, under the definition of this model, home advantage tended to increase with increasing league level. Differing from the home advantage found in Clarke (1996), this also suggested a possible decrease in home advantage over time. Some teams were seen to have negative home advantages in some years. This contrasts with Pollard (1986), who suggested that all teams have a positive home advantage. There was little conclusive evidence found to suggest that a team newly entering a league has a higher home advantage, although some teams such as Sheffield United in 2006 and Stoke City in 2008 did exhibit this behaviour. This could suggest that some teams have grounds which contribute to home advantage more strongly than others.

It was concluded that a more complex modelling approach was required. That of Dixon and Coles (1997) was chosen as a base model to work from. Under comparison with the Clarke (1996) model, the Dixon and Coles (1997) model produced a better fit of winning margin and provided the added benefit of modelling individual home and away team goal counts.

In Chapter 4 a bivariate Poisson model for home and away goal counts based on a model proposed by Dixon and Coles (1997) was employed to perform further analyses into the origins and effects of home advantage in association football. Home advantage could be seen to reduce over time for a long term historical data set. This reduction might be explained by the evolution of new rules and strategy. For example, the introduction of three points for a win changed the way the game was played significantly at a league level. However, under modelling more recent data, no significant change in home advantage over time was found, suggesting that strategies and playing styles may have stabilised to some extent.

Although home advantage was found to vary significantly over teams, the predictive power of a model which included them was reduced. This is thought to be because of the reduction in information used to estimate these parameters. It was also found that home and away teams experience an advantage dependent on the number of red cards issued. However, prediction of cards was out of scope of the analysis, and as such couldn't be used in a prediction model.

Chapter 5 accounted an analysis of external covariates on home advantage. Various exploratory models, including piecewise constant regression and smooth splines, were used to understand the effect of these covariates on home advantage. Due to the possibility of defining a closed form expression for piecewise constant home advantage with distance, using many data from multiple leagues it was found that increasing distance seems to increase home advantage considerably. This increase begins to tail off at extreme distances. Distance has been hypothesised as a covariate of home advantage in literature (Pollard, 1986; Brown et al., 2002; Pollard, 2006), and there seems to be some evidence to indicate that there is a relationship between them (Brown et al., 2002; Clarke and Norman, 1995). However, previous studies are usually limited to single leagues or countries. As such, this is a new and innovative result from this piece of research.

Other covariates were tested including match attendance, referee experience and pitch dimensions. There was some evidence to suggest that increasing match attendance decreases home advantage, though a consistently performing model wasn't found. This could be better modelled using home and away attendance figures if they were made available. No significant relationship between home advantage and referee experience was found.

A significant negative trend was found between pitch length and home advantage, whilst no such relationship was found with pitch width. This might suggest that a reduction in the play area for the away team from that which they are more familiar makes their usual

playing style difficult. The most statistically significant combination of covariates tested for home advantage was found to be distance and pitch length. This could be used in a prediction model to formulate a betting strategy that would give an edge over the market.

Over or under dispersion was tested for, to identify whether another distribution could more effectively fit goal counts, and in any way better model the differing distributions between home and away goals. Under testing, a negative binomial model, which could account for increased variability of goal counts, did not more effectively model the data. Threshold mixture models, based around the idea of censoring, were used to explore methods of creating a better fitting model for the body and upper tails. Various combinations of count distributions were tested. However, a Poisson-Poisson mixture model proved to be the best fit for the data. The parameterisation of this model suggested that the right tail of the distribution was under dispersed. Even though the fit was better, the predictive power was found to be lower than the model laid out in Dixon and Coles (1997). This could be due to the limited information describing extreme value goal counts in the data. Due to their nature, transition mixture models suffer from an inorganic step change between the two distributions. A smooth change is desirable to describe the probability of events more naturally.

Changepoints, where an abrupt change in structure divides data into two (or more) homogeneous sections, occur in many aspects of statistics other than threshold mixture models, such as time series. The abruptness of such a change may not be instantaneous, and as such a method is required to model smooth changes. Such a method was defined as smoothly distribution transitions (SDTs), which consider the reduction in information distant from the change, using a weighting function. These developments are entirely original with respect to changepoints. This method can also be used to predict the final value of a smooth change outside of the sample bounds, where the change will end and where the change started, as well as the centre of the change. It may also regress to the standard changepoint model, and as such can be seen as an extension or evolution into new territory for changepoints.

The final research chapter discussed the concept of changes in performance in golf. These changes could be due to rule changes or technological advancement. However, it was found that, although a perceived decrease in consistency was seen, this reduction may be due to strength in depth. Even so, this is not conclusive as the reduction in standard deviation of round scores observed between 1975 and 2010 for the European Tour could support a counter argument that as players become more consistent and hit the ball further, the courses are made more difficult by increasing their length.

Bibliography

Aero Space Web (2006). Golf Ball Dimples & Drag. `http://www.aerospaceweb.org/question/aerodynamics/q0215.shtml`. [Online; accessed 21-Aug-2012].

Ahlberg, J., Nilson, E., Walsh, J., and Bellman, R. (1967). *The Theory of Splines and Their Applications*. Academic Press.

Arlot, S. and Celisse, A. (2009). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

ATASS Sports (2012). Statistical Modelling in Action. `http://www.atass-sports.co.uk`. [Online; accessed 27-Aug-2012].

Atherton, J., Charbonneau, B., Wolfson, D. B., Joseph, L., Zhou, X., and Vandal, A. C. (2009). Bayesian optimal design for changepoint problems. *Canadian Journal of Statistics*, 37(4):495–513.

Barnett, V. and Hilditch, S. (1993). The effect of an artificial pitch surface on home team performance in football (soccer). *Jounral of the Royal Statistical Sociaety A*, 159(1):39–50.

Bassett, D. R., Kyle, C. R., Passfield, L., Broker, J. P., and Burke, E. R. (1999). Comparing cycling world hour records, 1967-1996: Modeling with empirical data. *Medicine and Science in Sports and Exercise*, 31(11):1665–1676.

BBC (2012). BBC Sport Football. `http://www.bbc.co.uk/sport/0/football/`. [Online; accessed 21-Aug-2012].

Bell, W., Holan, S., and McElroy, T. (2012). *Economic Time Series: Modeling and Seasonality*. Taylor & Francis.

Blevins, R. (1992). *Applied Fluid Dynamics Handbook*. Krieger Publishing Company.

Boyko, R. H., Boyko, A. R., and Boyko, M. G. (2007). Referee bias contributes to home advantage in English Premiership football. *J Sports Sci*, 25(11):1185–1194.

Brouillette, M. (2010). Putter features that influence the rolling motion of a golf ball. *Procedia Engineering*, 2(2):3223 – 3229.

Brown, T., Van Raalte, J. L., Brewer, B. W., Winter, C. R., Cornelius, A. E., and Anderson, M. B. (2002). World Cup soccer home advantage. *Journal of Sports Behaviour*, 25(2):134–144.

Callaway Golf Company (2003). 2003 annual report. `http://www.annualreports.com/HostedData/AnnualReports/PDFArchive/ely2003.pdf`. [Online; accessed 23-July-2012].

Chen, J. and Gupta, A. (2011). *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*. Birkhäuser Boston.

Clarke, A. R. (1996). Home advantage in balanced competitions - English soccer 1990-1996. *Mathematics and Computers in Sport*, pages 111–116.

Clarke, S. R. (1993). Computer forecasting of Australian rules football for a daily newspaper. *The Journal of the Operational Research Society*, 44(8):753–759.

Clarke, S. R. and Norman, J. M. (1995). Home ground advantage of individual clubs in English soccer. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 44(4):509–521.

Courneya, K. S. and Carron, A. V. (1990). Batting first versus last: Implications for the home advantage. *Journal of Sport & Exercise Psychology*, 12:312–316.

Courneya, K. S. and Carron, A. V. (1992). The home advantage in sport competitions: a literature review. *Journal of Sport & Exercise Psychology*, 14(1):13–27.

Darrell Research (2012). Darrell Research Survey. `http://www.darrell-survey.com`. [Online; accessed 21-Aug-2012].

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternatives. *Biometrika*, 74(1):33–43.

Diaze, J. (2011). A tee too far. `http://www.golfdigest.com/story/golf-barney-adams-forward-tees`. [Online; accessed 25-09-2015].

Dierckx, P. (1995). *Curve and Surface Fitting with Splines*. Monographs on numerical analysis. Clarendon Press.

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.

Dowie, J. (1982). Why Spain should win the World Cup. *New Scientist*, 94:693–695.

Eckley, I., Fearnhead, P., and Killick, R. (2011). Analysis of changepoint models. chapter 10 in *Bayesian Time Series Models*, eds. D. Barber, A.T. Cemgil and S. Chiappa, Cambridge University Press. pages 205–222.

Edwards, J. (1989). *The home field advantage*. In J.H. Goldstein (Ed.), Sports, games and play: Social and Psychological Viewpoints (2nd ed., pp. 333-370). Hillsdale, NJ: Erlbaum.

ESPN (2012). PGA Leaderboard 2001. `http://espn.go.com/golf/leaderboard`. [Online; accessed 26-Aug-2012].

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

FIFA (2010). Almost half the world tuned in at home to watch 2010 FIFA World Cup South Africa^{TM}. `http://www.fifa.com/worldcup/archive/southafrica2010/organisation/media/newsid=1473143/index.html`. [Online; accessed 27-Aug-2012].

Freund, R., Wilson, W., and Sa, P. (2006). *Regression Analysis*. Regression Analysis Series. Elsevier Science.

Glamser, F. D. (1990). Contest location, player misconduct, and race: a case from English soccer. *Journal of Sport Behavior*, 13(1):41–49.

Gleitman, H., Gross, J., and Reisberg, D. (2010). *Psychology*. W. W. Norton & Company.

GOLF Today (2012). 2012 Open Championship prize money breakdown. `http://www.golftoday.co.uk/tours/2012/The_Open/prize_money.html`. [Online; accessed 23-July-2012].

Golfshake (2015). European tour shows strength in depth at Whistling Straits. `http://www.golfshake.com/news/view/9081/European_Tour_Shows_Strength_in_Depth_at_Whistling_Straits.htm`. [Online; accessed 25-09-2015].

Greer, D. L. (1983). Spectator booing and the home advantage: A study of social influence in the basketball arena. *Social Psychology Quarterly*, 46(3):252–261.

Gwyn, R. G., Ormond, F., and Patch, C. E. (1996). Comparing putters with a conventional blade and a cylindrically shaped clubhead. *Perceptual and Motor Skills*, 82(1):31–34.

Gwyn, R. G. and Patch, C. E. (1993). Comparing two putting styles for putting accuracy. *Perceptual and Motor Skills*, 76:387–390.

Haake, S. J. (2009). The impact of technology on sporting performance in Olympic sports. *J Sports Sci*, 27(13):1421–1431.

Hall, P. and Opsomer, J. D. (2005). Theory for penalised spline regression. *Biometrika*, 92(1):105–118.

Hawkins, D. M. (2001). Fitting multiple change-point models to data. *Computational Statistics & Data Analysis*, 37(3):323 – 341.

Hilbe, J. (2011). *Negative Binomial Regression*. Cambridge University Press.

IHHF (2012). International ice hockey federation rule book. `http://www.iihf.com/iihf-home/sport/iihf-rule-book.html`. [Online; accessed 08-Aug-2012].

Irving, P. G. and Goldstein, S. R. (1990). Effect of home-field advantage on peak performance of baseball pitchers. *Journal of Sports Behavior*, 13(1):23–28.

Johnson, E. M. (2001). *Mortally wound-ed? Hot, new solid-core balls have nearly KO'd their wound ball rivals - popularity of new solid-core multilayer golf balls*. Golf Digest June.

Jurkovac, T. (1995). Collegiate basketball players' perceptions of the home advantage. *Unpublished Master's thesis, Bowling Green State University*, Bowling Green, OH.

Karlsen, J. (2003). Golf putting: An analysis of elite-players technique and performance. *Master's thesis from the Norwegian University of Sport and Physical Education*, Oslo, Norway.

Koppet, L. (1972). Home court: Winning edge. *New York Times, January 9*. pp. 1, 3.

Lane, E. (1976). Home sweet home, for athletes. *Chicago Sun Times, August 20*. p. 54.

Larsen, R. and Marx, M. (2012). *An Introduction to Mathematical Statistics and Its Applications*. Prentice Hall.

Lee, Y., Nelder, J., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

Lefebvre, L. and Passer, M. (1974). The effects of game location and importance on aggression in team sport. *International Journal of Sport Psychology*, 5(2):102–110.

Lehman, D. R. and Reifman, A. (1987). Spectator influence on basketball officiating. *Journal of Social Psychology*, 127(66):673–675.

Lili, L. (2005). Golf statistics. *Unpublished Masters thesis. Lancaster University, Management School*. Lancaster, UK.

Lincoln, G., Guinness, F., and Short, R. (1972). The way in which testosterone controls the social and sexual behavior of the red deer stag (cervus elaphus). *Hormones and Behavior*, 3(4):375 – 396.

Lindsay, C. (2010). Colin Montgomerie hails European strength in depth. `http://news.bbc.co.uk/sport1/hi/golf/8797804.stm`. [Online; accessed 25-09-2015].

Lukes, R. A. (2006). Improving track cycling performance using computational fluid dynamics. *PhD thesis*. University of Sheffield, Sheffield.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36:109–118.

Masataka, K. (2008). Science and engineering technology behind Bridgestone Tour golf balls. *Sports Technology*, 1(1):57–64.

Mazur, A. and Booth, A. (1998). Testosterone and dominance in men. *Behav Brain Sci*, 21(3):353–363.

Monoghan, E. P. and Glickman, S. E. (1992). *Hormones and aggressive behaviour*. In Becker, J., Breedlove, A. and Crews, D. (Eds.) Behavioural endocrinology. MIT Press, **9**, pp. 261-85.

Morris, D. (1981). *The soccer tribe*. Cape.

Muller, H.-G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist.*, 20(2):737–761.

Munson, B. R., Young, D. F., and Okiishi, T. H. (2006). *Fundamentals of Fluid Mechanics*. John Wiley & Sons, 5th edition.

Neave, N. and Wolfson, S. (2003). Testosterone, territoriality, and the home advantage. *Physiology & Behavior*, 78(2):269 – 275.

Nevill, A. M., Newell, S. M., and Gale, S. (1996). Factors associated with home advantage in English and Scottish soccer matches. *J Sports Sci*, 14(2):181–186.

Nilsson, J. and Karlsen, J. (2006). A new device for evaluating distance and directional performance of golf putters. *J Sports Sci*, 24(2):143–147.

Pelz, D. (1990). The long putter. *The Pelz Report*, 1:3.

Pollard, R. (1986). Home advantage in soccer: a retrospective analysis. *Journal of Sports Sciences*, 4(3):237–48.

Pollard, R. (2002). Evidence of a reduced home advantage when a team moves to a new stadium. *J Sports Sci*, 20(12):969–973.

Pollard, R. (2006). Home advantage in soccer: variations in its magnitude and a literature review of the inter-related factors associated with its existence. *Journal of Sports Behavior*, 29(2):169–189.

Pollard, R. (2008). Home advantage in football: A current review of an unsolved puzzle. *The Open Sports Sciences Journal*, 1(1):12–14.

Pollard, R. and Pollard, G. (2005). Long-term trends in home advantage in professional team sports in North America and England (1876-2003). *J Sports Sci*, 23(4):337–350.

Roth, A. (1957). You have to win on the road. *Sport, September*. p. 74.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Schwartz, B. and Barsky, S. F. (1977). The home advantage. *Social Forces*, 55(3):641–661.

Seber, G. and Lee, A. (2012). *Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley.

ShotLink (2015). Shotlink by cdw. `http://www.shotlink.com/`. [Online; accessed 28-09-2015].

Siegmund, D. (1988). Confidence sets in change-point problems. *International Statistical Review / Revue Internationale de Statistique*, 56(1):31–48.

Snyder, E. E. and Purdy, D. A. (1985). The home advantage in collegiate basketball. *Sociology of Sport Journal*, 2(4):352–356.

Stefani, R. and Clarke, S. (1987). Predictions and home advantage for Australian rules football. *Journal of Applied Statistics*, 19(2):251–261.

Stefani, R. T. (1983). Observed betting tendencies and suggested betting strategies for European football pools. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(3):319–329.

Sumner, J. and Mobley, M. (1981). Are cricket umpires biased? *New Scientist*, 91:29–31.

thefootballarchives.com (2014). The football archives. `http://www.thefootballarchives.com/`. Accessed: 30/10/2014.

Tierney, D. E. and Coop, R. H. (1998). A bivariate probability model for putting proficiency. In, M. R. Farrally & A. J. Cochran (Eds.). *Science and Golf III; Proceedings of the World Scientific Congress of Golf*. (pp. 385-394). Champaign, IL, Human Kinetics.

Titman, A. C., Costain, D. A., Ridall, P. G., and Gregory, K. (2015). Joint modelling of goals and bookings in association football. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):659–683.

Varca, P. E. (1980). An analysis of home and away game performance of male college basketball teams. *Journal of Sport Psychology*, 2(3):245–257.

Wager Minds (2012). Centaur Galileo Sports Betting Hedge Fund Collapses. `http://www.wagerminds.com/blog/uncategorized/centaur-galileo-sports-betting-hedge-fund-collapses-3944`. [Online; accessed 27-Aug-2012].

Watkins, J. R. (2008). Drive for show, putt for dough: Rates of return to golf skills, events played, and age on the PGA tour. *Michigan Journal of Business*, 1:35–60.

Wolfson, S., Wakelin, D., and Lewis, M. (2005). Football supporters' perceptions of their role in the home advantage. *J Sports Sci*, 23(4):365–374.

Yes! (2012). Innovation. `http://yesgolf.com/innovation`. [Online; accessed 23-Aug-2012].

Zhang, N. R., Siegmund, D. O., Ji, H., and Li, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika*, 97(3):631–645.

Appendix A

Derivations for Clarke and Norman (1995) Model

Clarke and Norman (1995) provided derivations for the overall home advantage, H, individual home advantage, h_i, and the team ability u_i as given in equation (3.1):

By considering the sum of the squared errors subject to the condition that $\sum_{i=1}^{N} u_i = 0$, and by use of the usual Lagrange multiplier expression, equation (A.1) must be minimised:

$$S = \sum_{i=1}^{N} \sum_{j=1(j\neq i)}^{N} (w_{ij} - u_i + u_j - h_i)^2 + \lambda \sum_{i=1}^{N} u_i. \tag{A.1}$$

Equations (A.2) and (A.3) can now be derived by partial differentiation with respect to u_I, $I = 1, \ldots, N$ and λ:

$$\sum_{j=1(j\neq i)}^{N} 2(w_{Ij} - u_I + u_j - h_I)(-1) + \sum_{i=1(j\neq i)}^{N} 2(w_{iI} - u_i + u_I - h_i) + \lambda = 0, \qquad I = 1, \ldots, N, \tag{A.2}$$

$$\sum_{j=1(j\neq I)}^{N} 2(w_{Ij} - u_I + u_j - h_I)(-1) = 0, \qquad I = 1, \ldots, N. \tag{A.3}$$

Expanding equation (A.3) gives

$$\sum_{j=1(j\neq I)}^{N} w_{ij} = (N-1)u_I + (N-1)h_I - \sum_{j=1(j\neq I)}^{N} u_j,$$

i.e.

$$\begin{aligned} HGD_I &= Nu_I + (N-1)h_I + \sum_{j=1(j\neq I)}^{N} u_j, \\ &= Nu_I + (N-1)h_I. \end{aligned} \tag{A.4}$$

Summing equation (A.4) over $I = 1, \ldots, N$ gives

$$\sum_{I=1}^{N} HGD_I = N \sum_{I=1}^{N} u_I + (N-1) \sum_{I=1}^{N} h_I,$$
$$HGD = (N-1)H. \tag{A.5}$$

where $H = \sum_{i=1}^{N} h_i$ is the total home advantage of all individual teams.

Substituting expression (A.3) into equation (A.2) removes the first summation term, then

$$\begin{aligned} -\lambda/2 &= \sum_{i=1(i\neq I)}^{N} (w_{iI} - u_i + u_I - h_i) \\ &= \sum_{i=1(i\neq I)}^{N} w_{iI} - \sum_{i=1(i\neq I)}^{N} u_i - \sum_{i=1(i\neq I)}^{N} h_i + (N-1)u_I \\ &= -AGD_I + u_I - H + h_I + (N-1)u_I \\ &= -AGD_I - H + h_I + Nu_I, \end{aligned} \tag{A.6}$$

so

$$\begin{aligned} \sum_{I=1}^{N} -\lambda/2 &= \sum_{I=1}^{N} -AGD_I - NH + \sum_{I=1}^{N} h_I + N \sum_{I=1}^{N} u_I, \\ -N\lambda/2 &= \sum_{I=1}^{N} HGD_I - (N-1)H + 0 \\ &= 0, \end{aligned} \tag{A.7}$$

from equation (A.3). This gives the result that $\lambda = 0$, which can be substituted into equation (A.6) as shown by

$$AGD_I = -H + h_I + Nu_I. \tag{A.8}$$

Subtracting equation (A.8) from equation (A.4) gives

$$\begin{aligned} HGD_I - AGD_I &= Nu_I + (N-1)h_I + H - h_I - Nu_I \\ &= H + (N-2)h_I. \end{aligned} \tag{A.9}$$

Finally, H can be calculated from equation (A.5), h_i from equation (A.9) and u_i from equation (A.4).

Appendix B

Home Advantage and Other Parameter Estimates

Team	Year															
	2000	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
Arsenal	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Aston Villa	0.73	0.59	0.50	0.67	0.53	0.63	0.69	0.98	0.81	0.62	0.67	0.51	0.67	0.58	0.45	0.43
Birmingham			0.49	0.60	0.46	0.42		0.64		0.46	0.53					
Blackburn		0.70	0.61	0.71	0.37	0.76	0.84	0.69	0.61	0.51	0.65	0.65				
Blackpool											0.80					
Bolton		0.57	0.48	0.68	0.56	0.73	0.76	0.50	0.60	0.52	0.72	0.64				
Bournemouth																0.71
Bradford	0.49															
Burnley										0.52					0.40	
Cardiff														0.48		
Charlton	0.81	0.49	0.54	0.71	0.50	0.62	0.55									
Chelsea	1.09	0.85	0.79	0.92	0.81	1.05	1.00	0.87	0.98	1.23	0.94	0.87	1.04	1.03	1.02	0.92
Coventry	0.59															
Crystal Palace					0.48									0.49	0.67	0.60
Derby	0.60	0.42						0.28								
Everton	0.73	0.58	0.56	0.63	0.52	0.51	0.83	0.74	0.81	0.73	0.70	0.65	0.77	0.89	0.69	0.93
Fulham		0.46	0.49	0.73	0.61	0.73	0.62	0.53	0.56	0.47	0.69	0.65	0.70	0.62		
Hull									0.58	0.43				0.57	0.47	
Ipswich	0.90	0.53														
Leeds	1.02	0.68	0.70	0.57												
Leicester	0.63	0.39		0.69											0.66	1.04
Liverpool	1.13	0.84	0.72	0.76	0.59	0.83	0.90	0.90	1.12	0.73	0.81	0.62	1.00	1.51	0.74	0.98
Man City	0.67		0.56	0.77	0.54	0.65	0.46	0.62	0.85	0.88	0.83	1.21	0.91	1.50	1.18	1.10
Man United	1.24	1.10	0.87	0.86	0.66	1.06	1.31	1.07	0.98	1.02	1.09	1.16	1.19	0.95	0.88	0.75
Middlesboro			0.57													
Middlesbrough	0.70	0.44		0.62	0.62	0.73	0.71	0.59	0.42							
Newcastle	0.70	0.96	0.75	0.71	0.55	0.70	0.61	0.63	0.60		0.79	0.76	0.64	0.65	0.58	0.69
Norwich					0.51							0.70	0.58	0.42		0.61
Portsmouth				0.66	0.50	0.56	0.72	0.65	0.57	0.42						
QPR												0.58	0.43		0.62	
Reading							0.84	0.57					0.62			
Sheffield United							0.52									
Southampton	0.64	0.59	0.50	0.61	0.54								0.69	0.80	0.76	0.91
Stoke									0.57	0.41	0.63	0.48	0.47	0.67	0.69	0.64
Sunderland	0.73	0.37	0.25			0.40		0.50	0.50	0.58	0.64	0.60	0.57	0.61	0.44	0.75
Swansea												0.59	0.67	0.82	0.66	0.65
Tottenham	0.77	0.62	0.61	0.65	0.54	0.78	0.92	0.92	0.66	0.80	0.77	0.88	0.91	0.83	0.83	1.06
Watford							0.47									0.62
West Brom			0.35		0.43	0.47			0.54		0.79	0.60	0.74	0.65	0.54	0.53
West Ham	0.72	0.62	0.50			0.78	0.57	0.57	0.62	0.58	0.61		0.64	0.59	0.63	1.02
Wigan						0.68	0.60	0.47	0.50	0.46	0.57	0.57	0.67			
Wolves				0.55						0.39	0.65	0.56				

Table B.1: Estimates of α parameters for the seasons 2000/2001 - 2015/2016, under the model defined in equation (4.4) (optimising over indendence function alongside other parameters).

Team	Year															
	2000	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
Arsenal	1.04	1.33	1.68	0.87	1.43	0.97	1.02	1.02	1.26	1.35	1.30	1.59	1.17	1.21	1.19	1.08
Aston Villa	1.16	1.73	1.77	1.44	2.01	1.68	1.17	1.69	1.60	1.24	1.72	1.69	2.14	1.73	1.81	2.19
Birmingham			1.86	1.56	1.76	1.51		1.99		1.48	1.69					
Blackburn		1.88	1.66	1.86	1.62	1.30	1.56	1.55	1.99	1.75	1.74	2.47				
Blackpool											2.32					
Bolton		2.26	1.93	1.81	1.68	1.27	1.49	1.72	1.75	2.12	1.66	2.45				
Bournemouth																1.97
Bradford	1.86															
Burnley										2.58					1.67	
Cardiff														2.10		
Charlton	1.54	1.76	2.14	1.63	2.23	1.68	1.70									
Chelsea	1.24	1.44	1.46	0.97	0.59	0.70	0.70	0.84	0.81	1.07	0.99	1.50	1.24	0.79	1.06	1.58
Coventry	1.68															
Crystal Palace					2.38									1.39	1.63	1.49
Derby	1.57	2.23						2.78								
Everton	1.59	2.08	1.87	1.85	1.74	1.50	1.04	1.07	1.23	1.58	1.33	1.25	1.26	1.13	1.61	1.65
Fulham		1.60	1.90	1.53	2.32	1.79	1.71	1.91	1.12	1.46	1.28	1.61	1.86	2.47		
Hull									2.08	2.37				1.52	1.61	
Ipswich	1.14	2.32														
Leeds	1.18	1.36	2.20	2.54												
Leicester	1.37	2.30		2.12											1.77	1.08
Liverpool	1.08	1.09	1.59	1.21	1.58	0.78	0.78	0.93	0.93	1.14	1.29	1.28	1.38	1.54	1.55	1.49
Man City	1.74		2.06	1.78	1.50	1.47	1.24	1.70	1.66	1.48	0.98	0.95	1.08	1.11	1.28	1.25
Man United	0.86	1.69	1.35	1.14	1.02	1.07	0.81	0.72	0.80	0.92	1.13	1.07	1.37	1.25	1.20	1.03
Middlesboro			1.68													
Middlesbrough	1.19	1.67		1.68	1.78	1.79	1.40	1.69	1.85							
Newcastle	1.35	1.93	1.89	1.31	2.21	1.29	1.34	2.09	1.95		1.69	1.63	2.10	1.71	2.02	1.91
Norwich					2.96							2.11	1.80	1.75		1.95
Portsmouth				1.74	2.25	1.90	1.20	1.29	1.88	2.08						
QPR												2.05	1.84		2.33	
Reading							1.36	2.11					2.27			
Sheffield United							1.55									
Southampton	1.28	1.95	1.75	1.47	2.56								1.87	1.35	1.07	1.22
Stoke									1.79	1.50	1.41	1.65	1.38	1.48	1.46	1.60
Sunderland	1.10	1.80	2.42			2.08		1.86	1.76	1.77	1.64	1.46	1.66	1.73	1.66	1.83
Swansea												1.62	1.61	1.57	1.58	1.52
Tottenham	1.47	1.91	2.38	1.84	1.57	1.18	1.57	2.00	1.48	1.33	1.37	1.33	1.44	1.50	1.72	1.06
Watford							1.66									1.46
West Brom			2.44		2.34	1.75			2.19		2.10	1.63	1.78	1.70	1.63	1.39
West Ham	1.34	2.05	2.25			1.70	1.67	1.59	1.48	2.11	2.06		1.65	1.43	1.50	1.53
Wigan						1.61	1.68	1.61	1.46	2.48	1.77	1.95	2.25			
Wolves				2.48						1.75	1.91	2.60				

Table B.2: Estimates of β parameters for the seasons 2000/2001 - 2015/2016, under the model defined in equation (4.4) (optimising over indendence function alongside other parameters).

Year	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
$\hat{\gamma}$	1.45	1.25	1.32	1.30	1.41	1.42	1.46	1.38	1.30	1.58	1.39	1.31	1.26	1.31	1.35	1.23

Table B.3: Estimates of γ parameters for the seasons 2000/2001 - 2015/2016 under the model defined in equation (4.4).

Year	$\hat{\rho}$	**Deviance**	**p-value**
2000	-0.05	0.51	0.48
2001	-0.11	2.41	0.12
2002	0.11	2.12	0.15
2003	-0.14	4.22	0.04
2004	-0.07	0.97	0.32
2005	0.04	0.26	0.61
2006	0.00	0.00	0.98
2007	-0.07	1.00	0.32
2008	-0.12	2.18	0.14
2009	-0.08	1.14	0.29
2010	-0.13	3.27	0.07
2011	-0.13	3.26	0.07
2012	-0.11	2.71	0.10
2013	0.14	3.82	0.05
2014	0.06	0.70	0.40
2015	-0.06	0.84	0.36

Table B.4: Estimates of ρ parameters for the seasons 2000/2001 - 2015/2016, along with the associated deviance and p-values from a hypothesis test between a null hypothesis of $\rho = 0$ and an alternative hypothesis of $\rho \neq 0$, under the model defined in equation (4.4).

Team	$\hat{\gamma}_i$
Arsenal	1.37
Aston Villa	1.16
Barnsley	2.09
Birmingham City	1.46
Blackburn Rovers	1.42
Blackpool	1.20
Bolton Wanderers	1.25
Bradford City	2.10
Burnley	1.48
Cardiff City	1.66
Charlton Athletic	1.37
Chelsea	1.41
Coventry City	1.39
Crystal Palace	0.96
Derby County	1.38
Everton	1.43
Fulham	1.58
Hull City	1.18
Ipswich Town	1.09
Leeds United	1.14
Leicester City	1.09
Liverpool	1.38
Manchester City	1.38
Manchester United	1.27
Middlesbrough	1.52
Newcastle United	1.49
Norwich City	1.52
Nottingham Forest	1.15
Portsmouth	1.68
Queens Park Rangers	1.28
Reading	1.10
Sheffield United	3.02
Sheffield Wednesday	1.23
Southampton	1.51
Stoke City	1.81
Sunderland	1.26
Swansea City	1.55
Tottenham Hotspur	1.44
Watford	2.06
West Bromwich Albion	1.33
West Ham United	1.66
Wigan Athletic	1.15
Wimbledon	1.26
Wolverhampton Wanderers	1.20

Table B.5: Team home advantage parameter estimates for teams present in the Premiership between 1995/1996 - 2013/2014

Team	Year																			σ^2
	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	
Arsenal	1.60	1.38	1.72	1.37	1.35	2.50	1.14	1.24	1.21	1.59	2.42	2.16	1.00	0.84	1.38	0.85	1.11	1.89	1.13	0.48
Aston Villa	1.61	1.36	1.13	1.81	0.99	1.43	0.91	1.46	1.00	1.35	0.91	0.86	0.92	1.01	1.25	1.18	1.16	0.96	1.29	0.26
Barnsley	-	-	2.10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	NA
Birmingham City	-	-	-	-	-	-	-	1.57	1.51	1.53	2.10	-	1.87	-	1.01	1.06	-	-	-	0.39
Blackburn Rovers	2.60	1.98	2.34	1.25	-	-	1.50	0.86	0.96	1.53	1.56	1.47	1.07	1.22	2.14	0.92	1.17	-	-	0.54
Blackpool	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.20	-	-	-	NA
Bolton Wanderers	0.70	-	1.57	-	-	-	0.83	1.92	1.00	1.08	1.45	1.24	1.78	1.06	1.62	1.91	1.01	-	-	0.41
Bradford City	-	-	-	-	2.20	1.98	-	-	-	-	-	-	-	-	-	-	-	-	-	0.15
Burnley	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.48	-	-	-	-	NA
Cardiff City	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.66	NA
Charlton Athletic	-	-	-	0.95	-	1.61	1.52	1.37	1.31	2.01	1.15	1.26	-	-	-	-	-	-	-	0.32
Chelsea	1.89	1.32	1.09	1.03	1.96	1.83	1.85	1.52	1.03	1.00	1.90	1.37	1.25	0.95	1.94	1.30	1.70	1.20	1.54	0.36
Coventry City	1.00	1.01	1.30	2.00	4.17	0.65	-	-	-	-	-	-	-	-	-	-	-	-	-	1.30
Crystal Palace	-	-	0.68	-	-	-	-	-	-	1.09	-	-	-	-	-	-	-	-	1.20	0.27
Derby County	-	1.24	1.74	1.23	0.99	1.63	1.54	-	-	-	-	-	1.51	-	-	-	-	-	-	0.27
Everton	1.21	1.19	1.55	1.11	1.57	1.83	1.37	1.41	1.49	1.04	1.82	1.74	1.63	1.29	1.40	1.56	1.26	1.49	1.65	0.23
Fulham	-	-	-	-	-	-	1.40	1.71	1.25	1.31	1.83	0.89	1.36	2.53	2.23	1.58	3.00	1.27	1.50	0.59
Hull City	-	-	-	-	-	-	-	-	-	-	-	-	-	0.86	1.83	-	-	-	1.12	0.50
Ipswich Town	-	-	-	-	-	1.19	0.95	-	-	-	-	-	-	-	-	-	-	-	-	0.17
Leeds United	1.09	1.16	1.19	1.07	0.99	1.29	1.40	0.75	1.68	-	-	-	-	-	-	-	-	-	-	0.26
Leicester City	-	0.91	0.71	1.67	1.28	2.57	1.01	-	0.66	-	-	-	-	-	-	-	-	-	-	0.67
Liverpool	1.94	1.60	1.61	1.83	1.22	1.29	0.97	0.98	1.11	1.54	1.28	2.18	1.78	1.14	2.39	1.68	1.05	0.87	1.10	0.43
Manchester City	1.76	-	-	-	-	0.94	-	1.47	1.30	0.88	1.53	0.53	1.66	2.23	1.27	1.31	1.46	1.63	1.61	0.42
Manchester United	0.97	1.00	1.35	1.28	1.55	1.63	0.84	1.32	1.39	1.21	1.06	1.24	1.43	1.72	1.53	1.68	1.41	1.10	0.82	0.27
Middlesbrough	3.37	1.99	-	1.09	1.01	0.69	1.93	3.04	1.31	1.11	1.41	2.35	1.68	1.56	-	-	-	-	-	0.80
Newcastle United	1.36	2.82	1.70	1.18	1.98	1.44	1.17	1.34	1.74	1.17	1.48	1.55	1.26	1.51	-	2.75	1.08	1.13	1.16	0.51
Norwich City	-	-	-	-	-	-	-	-	-	2.00	-	-	-	-	-	-	1.17	1.58	1.55	0.34
Nottingham Forest	1.40	0.94	-	1.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.24
Portsmouth	-	-	-	-	-	-	-	-	2.92	2.13	0.86	1.64	1.00	2.13	2.39	-	-	-	-	0.75
Queens Park Rangers	1.95	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.28	0.77	-	0.59
Reading	-	-	-	-	-	-	-	-	-	-	-	1.26	0.86	-	-	-	-	1.18	-	0.21
Sheffield United	-	-	-	-	-	-	-	-	-	-	-	3.03	-	-	-	-	-	-	-	NA
Sheffield Wednesday	1.68	1.01	1.39	0.96	1.24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.29
Southampton	1.64	1.78	1.28	3.63	1.36	2.08	1.00	1.40	1.20	1.96	-	-	-	-	-	-	-	1.13	1.45	0.70
Stoke City	-	-	-	-	-	-	-	-	-	-	-	-	-	1.37	2.38	2.06	2.29	1.63	1.52	0.42
Sunderland	-	1.34	-	-	0.97	1.10	1.64	1.09	-	-	0.85	-	1.77	1.61	2.03	1.25	1.37	0.95	1.04	0.36
Swansea City	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.59	1.48	1.58	0.06
Tottenham Hotspur	1.07	0.76	1.10	1.47	2.36	1.95	1.90	1.44	2.36	2.96	1.42	1.48	2.30	0.88	1.48	1.21	1.44	0.78	1.20	0.60
Watford	-	-	-	-	2.21	-	-	-	-	-	-	1.90	-	-	-	-	-	-	-	0.22
West Bromwich Albion	-	-	-	-	-	-	-	1.43	-	1.05	2.09	-	-	2.58	-	1.16	0.88	1.52	1.26	0.57
West Ham United	1.39	2.24	2.51	2.26	1.61	1.14	2.00	1.00	-	-	1.37	2.18	1.34	1.21	1.76	1.27	-	3.09	1.66	0.58
Wigan Athletic	-	-	-	-	-	-	-	-	-	-	1.14	0.95	1.63	1.00	1.06	1.22	1.10	1.24	-	0.21
Wimbledon	0.96	1.33	1.11	1.22	1.86	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.34
Wolverhampton Wanderers	-	-	-	-	-	-	-	-	1.52	-	-	-	-	-	0.69	1.88	0.90	-	-	0.55

Table B.6: Seasonally varying, team dependent home advantage MLEs, $\hat{\gamma}_i$, for each season between 1995/1996 and 2013/2014.

Appendix C

ATASS Dataset Description

League	Start Year	End Year
Australian A-League	2001	2012
Australian Bundesliga	2001	2012
Belgian Pro League	2001	2012
Brazilian Compeonato Serie A	2002	2012
Brazilian Serie B	2007	2012
English Premier League	2006	2012
English Championship	2001	2012
English League 1	2001	2012
English League 2	2001	2012
English Conference	2001	2012
French Ligue 1	2001	2012
French Ligue 2	2001	2012
German Bundesliga	2001	2012
German 2. Bundesliga	2001	2012
German 3. Liga	2008	2012
German	2001	2012
German Regionalliga Nord	2008	2012
German Regionalliga South	2008	2012
German Regionalliga West	2008	2012
Irish Premiership	2004	2012
Irish Cup	2001	2012
Italian Serie A	2001	2012
Mexican Liga MX	2001	2012
Scottish Premiership	2001	2012
Scottish Championship	2001	2012
Scottish League 1	2001	2012
Scottish League 2	2001	2012
Spanish La Liga	2001	2012
Spanish Copa del Rey	2001	2012
Swedish Allsvenskan	2002	2012
Swedish Superetan	2009	2012

Table C.1: Usable leagues from ATASS data set, with first season start year and last season end year. Note some seasons missing or incomplete.

Appendix D

Distance as a Covariate of Home Advantage

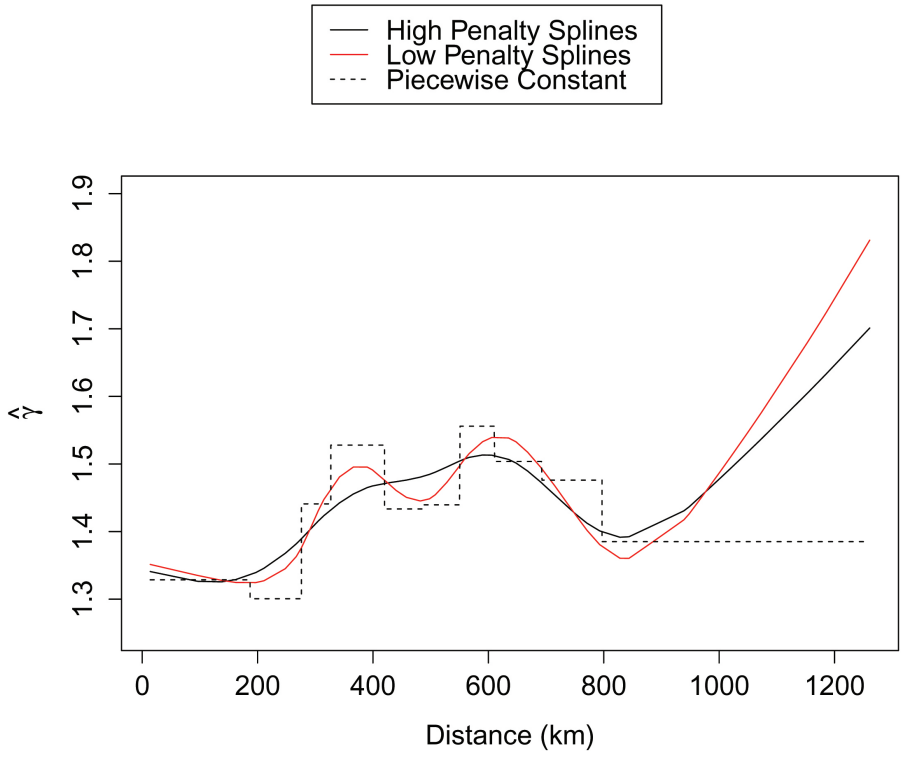


Figure D.1: Ligue 1 (France), 2003/2004 - 2011/2012: Penalised spline smooth curves describing the relationship of home advantage with distance, with high ($\Psi = 10$) and low ($\Psi = 1$) penalties, compared to a piecewise constant regression.

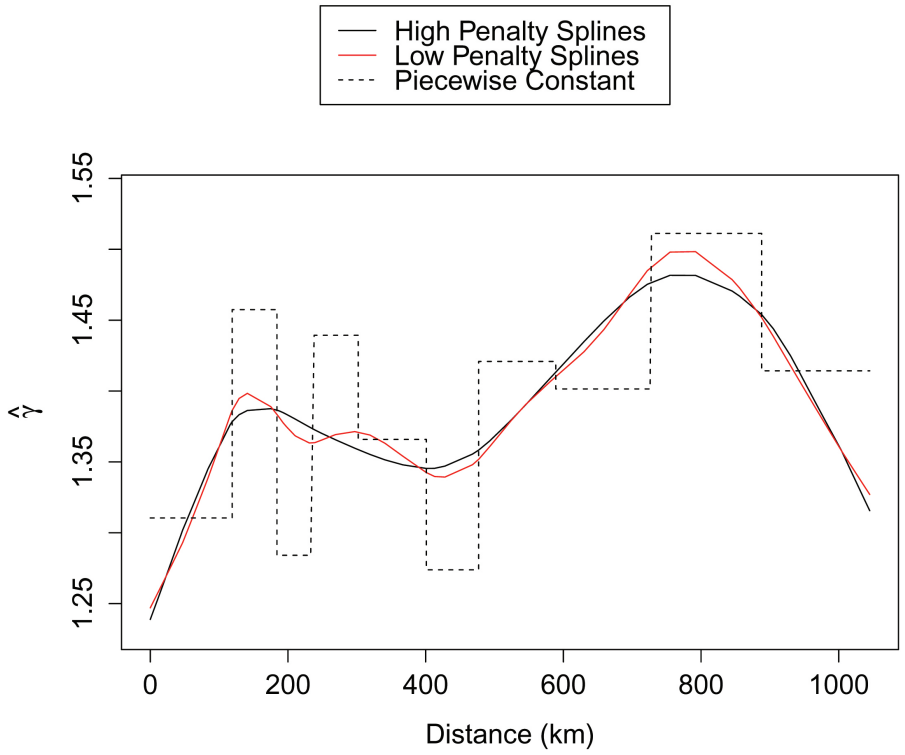


Figure D.2: Serie A (Italy), 2004/2005 - 2011/2012: Penalised spline smooth curves describing the relationship of home advantage with distance, with high ($\Psi = 10$) and low ($\Psi = 1$) penalties, compared to a piecewise constant regression.

	a_0	a_1	a_2	a_3
Log-Linear	3.11×10^{-1}	1.08×10^{-4}	NA	NA
Log-Quadratic	2.40×10^{-1}	4.68×10^{-4}	-3.61×10^{-7}	NA
Log-Cubic	2.39×10^{-1}	4.68×10^{-4}	-3.82×10^{-7}	3.03×10^{-11}

Table D.1: Ligue 1 (France) 2003/2004- 2011/2012: Parameter values for first, second and third order polynomial regressions relating a regressor of distance between teams to home advantage.

	a_0	a_1	a_2	a_3
Log-Linear	2.91×10^{-1}	8.02×10^{-5}	NA	NA
Log-Quadratic	2.74×10^{-1}	1.80×10^{-4}	-9.75×10^{-8}	NA
Log-Cubic	2.93×10^{-1}	-3.26×10^{-5}	4.33×10^{-7}	-3.52×10^{-10}

Table D.2: Serie A (Italy) 2004/2005 - 2011/2012: Parameter values for first, second and third order polynomial regressions relating a regressor of distance between teams to home advantage.

	a_0	a_1	a_2	a_3
Log-Linear	2.94×10^{-1}	1.05×10^{-4}	NA	NA
Log-Quadratic	2.67×10^{-1}	3.00×10^{-4}	-2.18×10^{-7}	NA
Log-Cubic	2.68×10^{-1}	2.36×10^{-4}	-2.50×10^{-8}	-1.41×10^{-10}

Table D.3: Combined data from Premier League (England), Ligue 1 (France) and Serie A (Italy): Parameter values for first, second and third order polynomial regressions relating a regressor of distance between teams to home advantage.

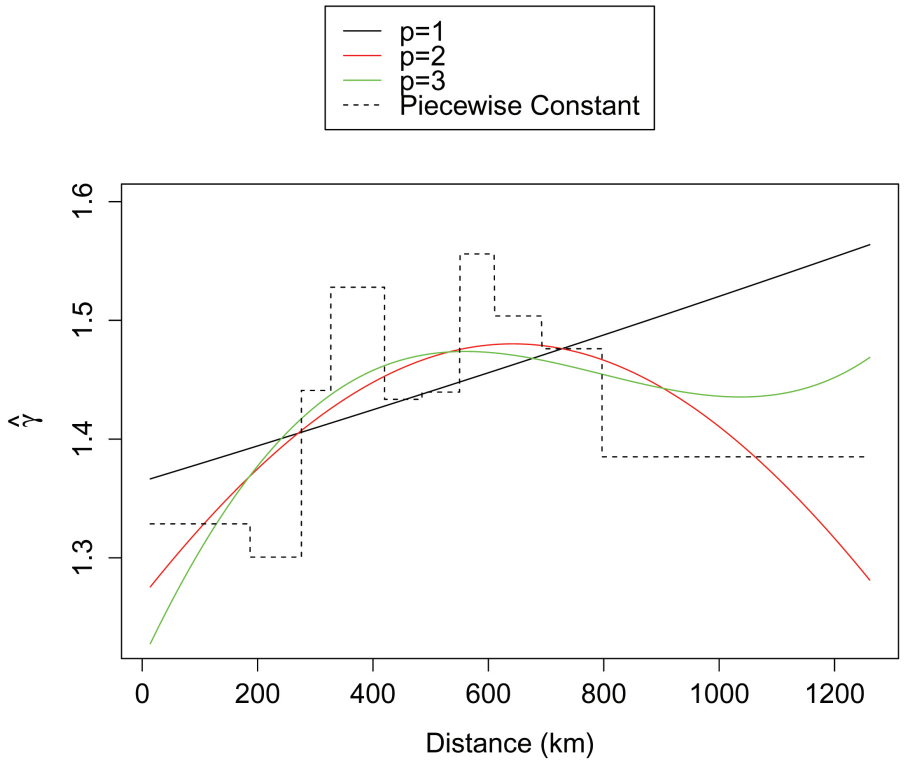


Figure D.3: Ligue 1 (France), 2003/2004 - 2011/2012: Comparison of first to third order polynomial regression models for distance as a regressor for home advantage.

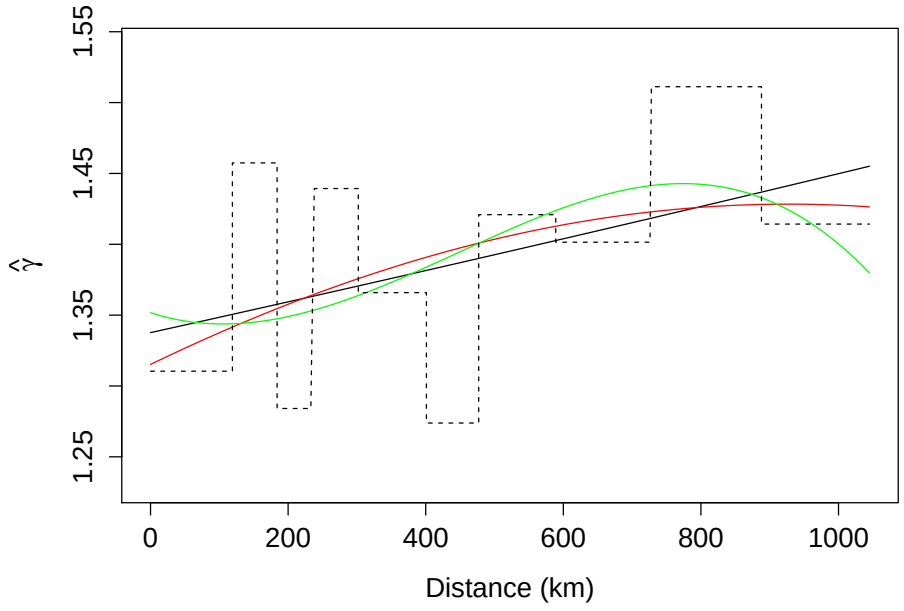


Figure D.4: Serie A (Italy), 2004/2005 - 2011/2012: Comparison of first to third order polynomial regression models for distance as a regressor for home advantage.

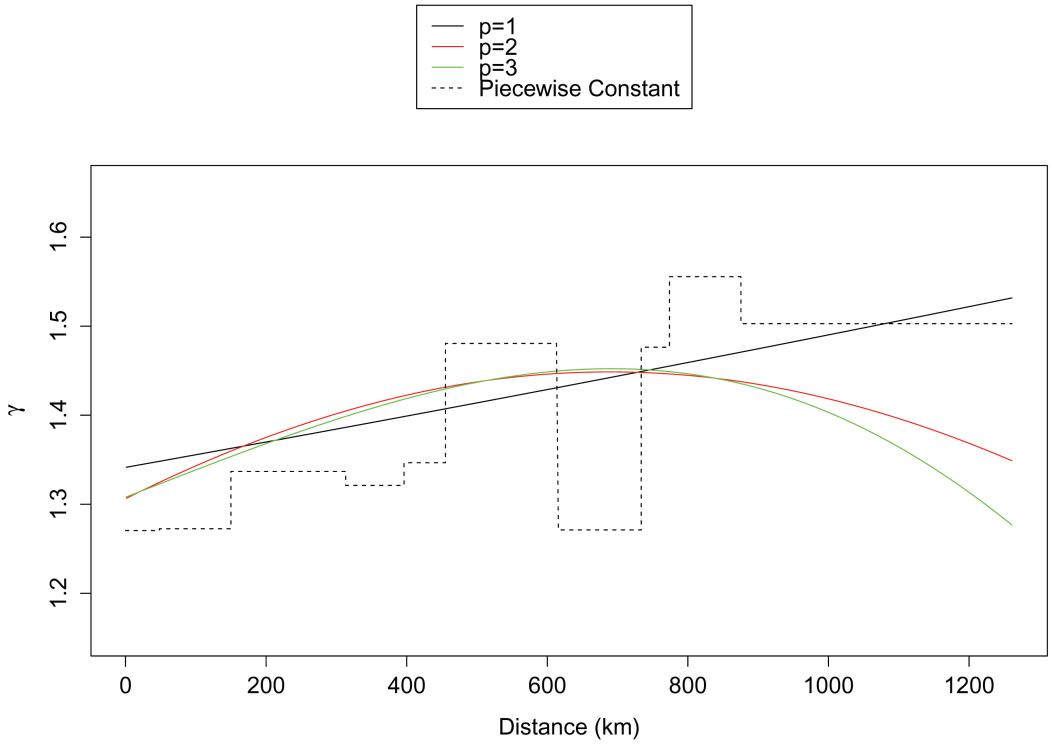


Figure D.5: Combined data from Premier League (England), Ligue 1 (France) and Serie A (Italy): Comparison of first to third order polynomial regression models for distance as a regressor for home advantage.

Appendix E

Relative Attendance as a Covariate of Home Advantage

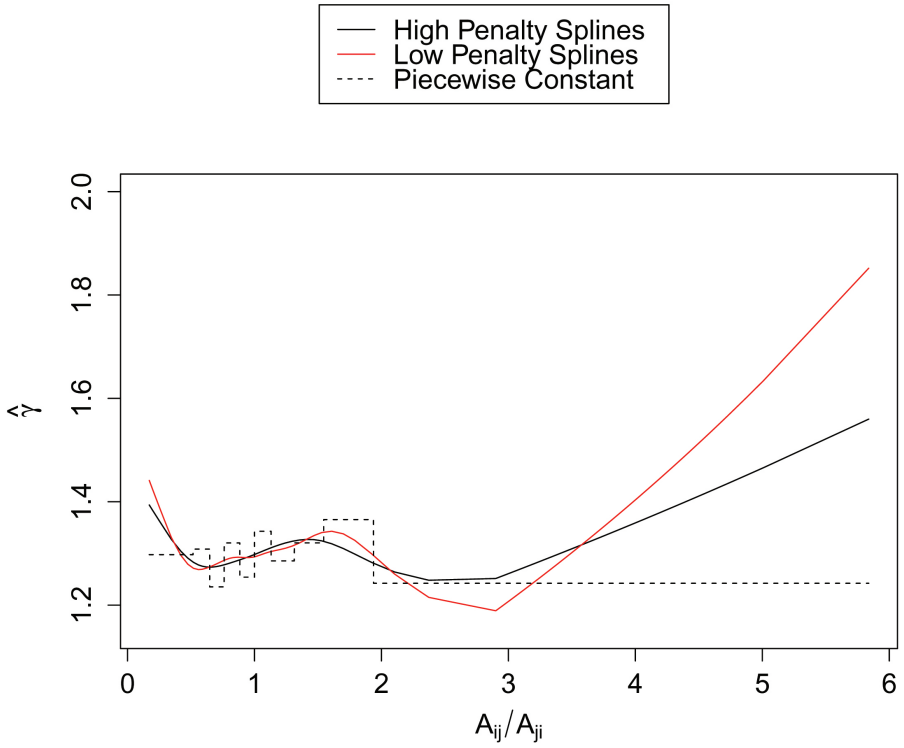


Figure E.1: Championship, 2004/2005 - 2013/2014: Penalised spline smooth curves describing the relationship of home advantage with relative attendance, with high ($\Psi = 10$) and low ($\Psi = 1$) penalties, compared to a piecewise constant regression.

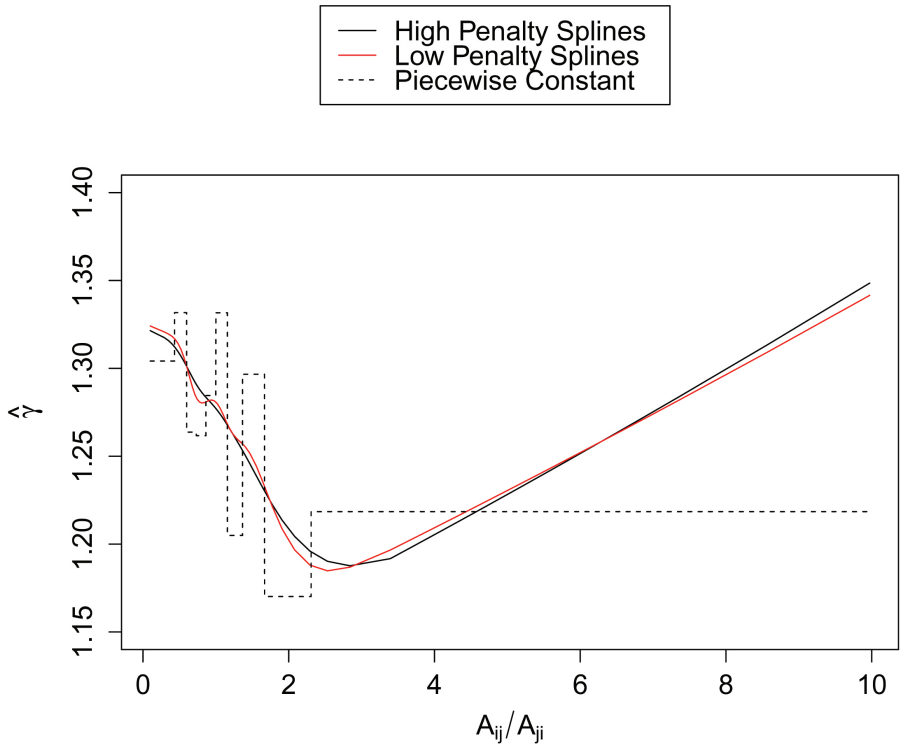


Figure E.2: League 1, 2004/2005 - 2013/2014:Penalised spline smooth curves describing the relationship of home advantage with relative attendance, with high ($\Psi = 10$) and low ($\Psi = 1$) penalties, compared to a piecewise constant regression

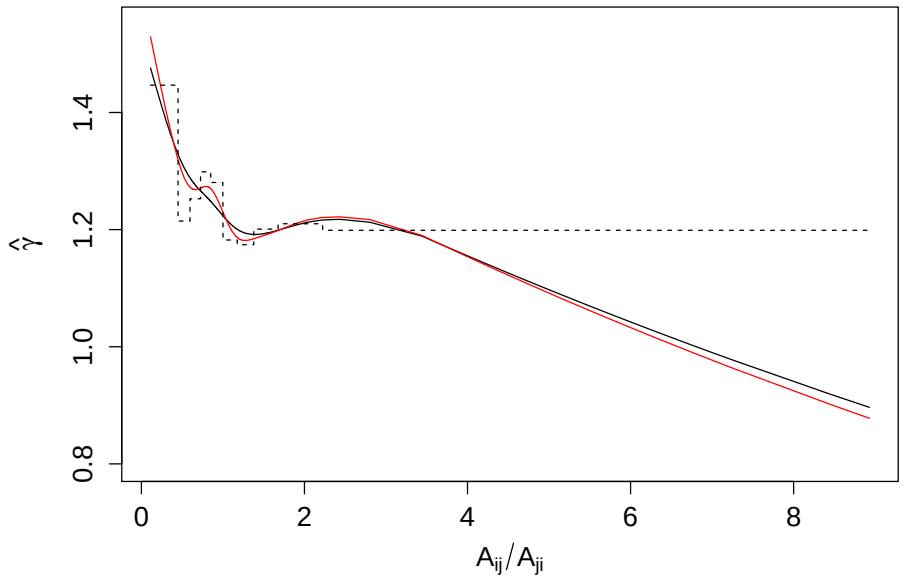


Figure E.3: League 2, 2004/2005 - 2013/2014: Penalised spline smooth curves describing the relationship of home advantage with relative attendance, with high ($\Psi = 10$) and low ($\Psi = 1$) penalties, compared to a piecewise constant regression

	a_0	a_1	a_2	a_3
Log-Linear	2.57×10^{-1}	1.89×10^{-3}	NA	NA
Log-Quadratic	2.99×10^{-1}	-5.82×10^{-2}	1.59×10^{-2}	NA
Log-Cubic	1.29×10^{-1}	3.01×10^{-1}	-1.82×10^{-1}	2.93×10^{-2}

Table E.1: Championship 2004/2005 - 2013/2014: Parameter values for first, second and third order polynomial regressions relating a regressor of A_{ij}/A_{ji} to home advantage.

	a_0	a_1	a_2	a_3
Log-Linear	2.61×10^{-1}	-3.60×10^{-2}	NA	NA
Log-Quadratic	2.83×10^{-1}	-6.43×10^{-2}	6.02×10^{-3}	NA
Log-Cubic	3.35×10^{-1}	-1.62×10^{-1}	4.95×10^{-2}	-4.78×10^{-3}

Table E.2: League 2 2004/2005 - 2013/2014: Parameter values for first, second and third order polynomial regressions relating a regressor of A_{ij}/A_{ji} to home advantage.

	a_0	a_1	a_2	a_3
Log-Linear	2.66×10^{-1}	-2.53×10^{-2}	NA	NA
Log-Quadratic	3.06×10^{-1}	-7.53×10^{-2}	1.02×10^{-2}	NA
Log-Cubic	3.06×10^{-1}	-7.53×10^{-2}	1.02×10^{-2}	3.01×10^{-8}

Table E.3: League 1 2004/2005 - 2031/2014: Parameter values for first, second and third order polynomial regressions relating a regressor of A_{ij}/A_{ji} to home advantage.

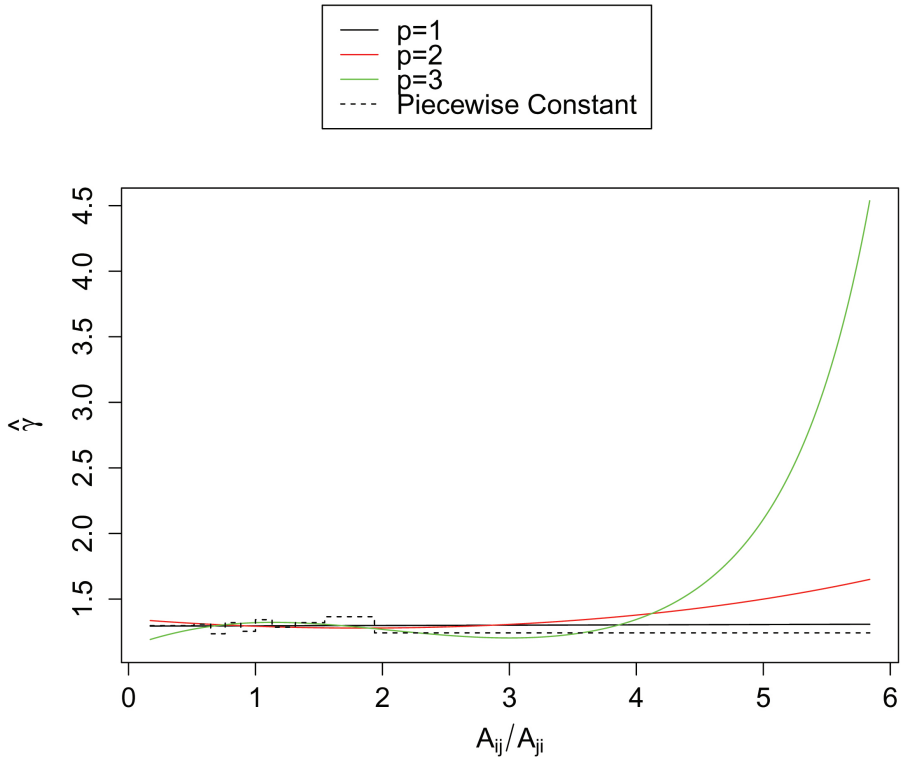


Figure E.4: Championship, 2004/2005 - 2013/2014: Comparison of first to third order polynomial regression models for $A_{i,j}/A_{j,i}$ as a regressor for home advantage.

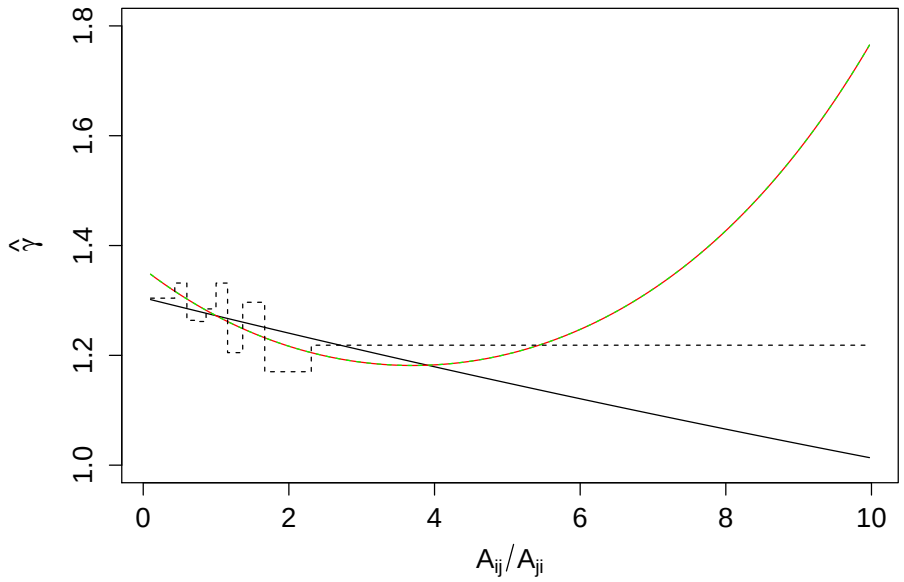


Figure E.5: League 1, 2004/2005 - 2013/2014: Comparison of first to third order polynomial regression models for $A_{i,j}/A_{j,i}$ as a regressor for home advantage.

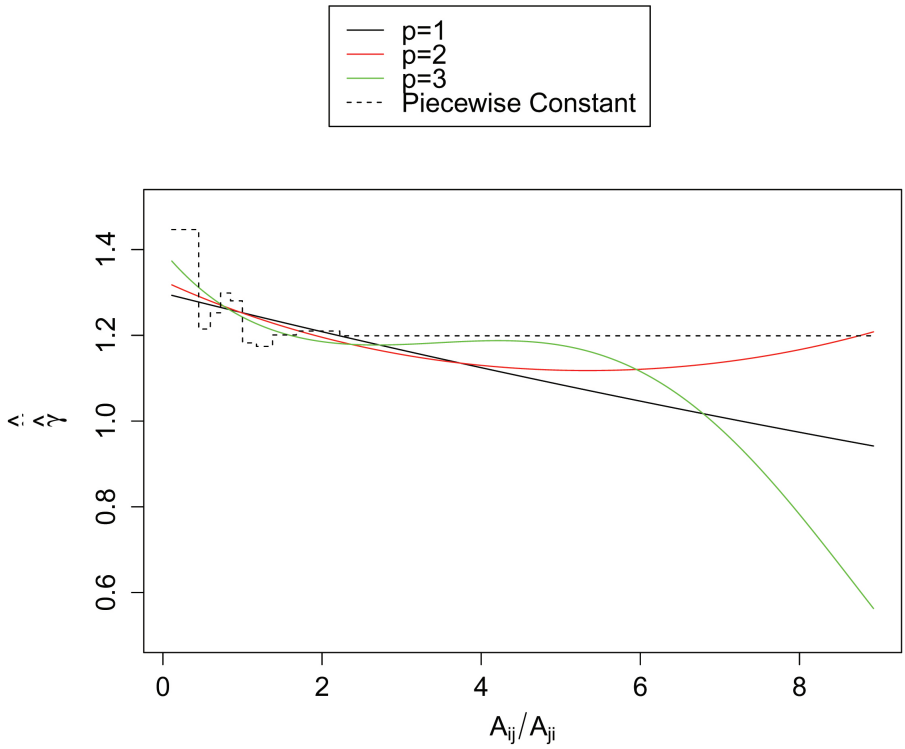


Figure E.6: League 2, 2004/2005 - 2013/2014: Comparison of first to third order polynomial regression models for $A_{i,j}/A_{j,i}$ as a regressor for home advantage.

Appendix F

Referee Experience as a Covariate of Home Advantage

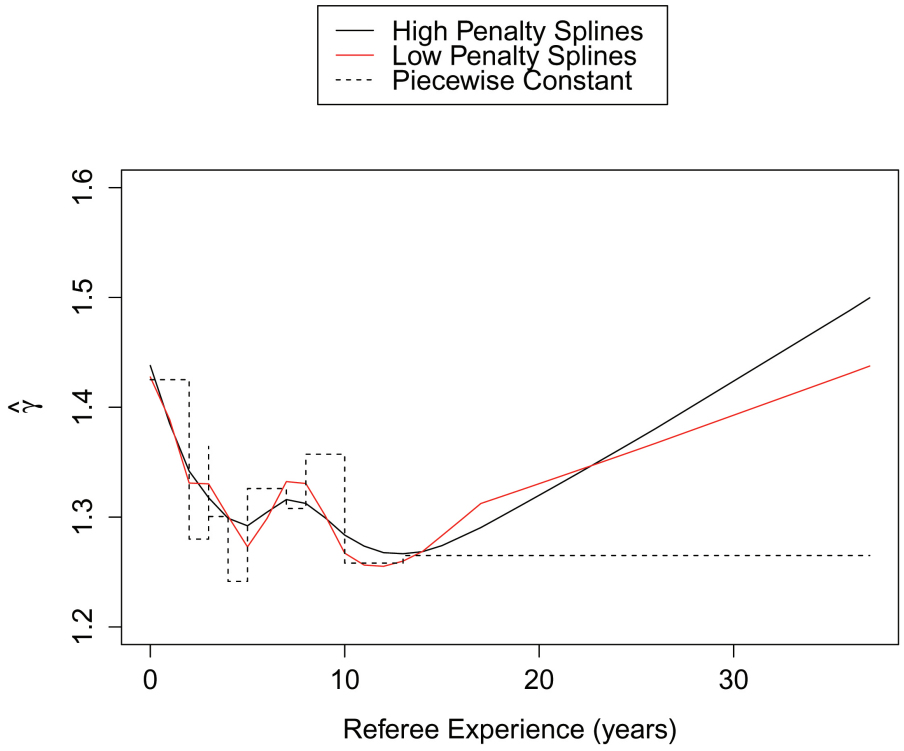


Figure F.1: Championship, 2004/2005 - 2013/2014: Penalised spline smooth curves describing the relationship of home advantage with referee experience, with high ($\Psi = 10$) and low ($\Psi = 1$) penalties, compared to a piecewise constant regression.

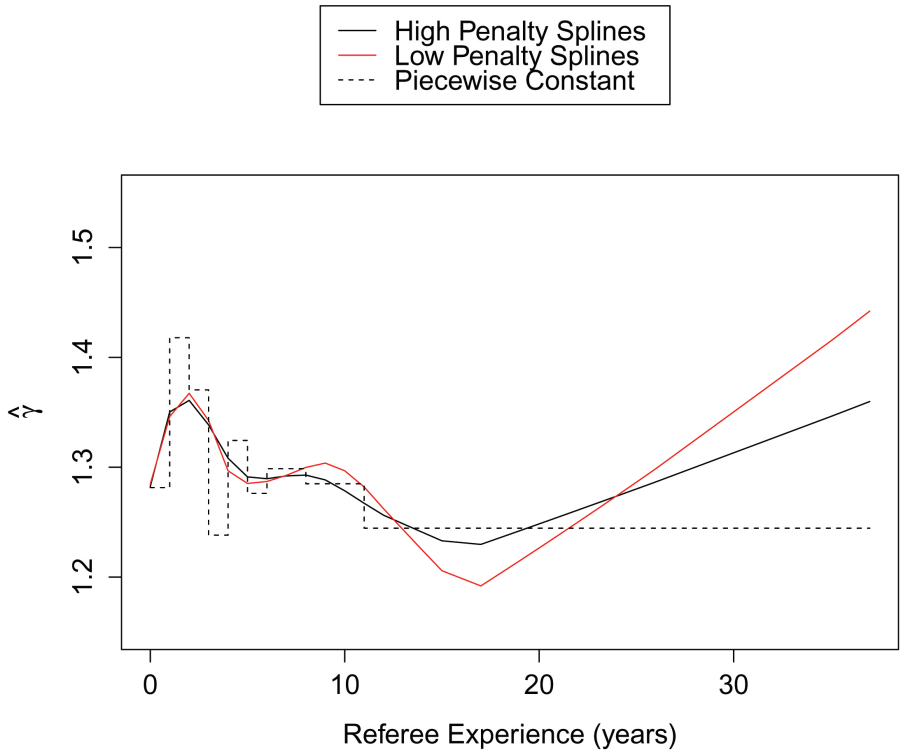


Figure F.2: League 1, 2004/2005 - 2013/2014:Penalised spline smooth curves describing the relationship of home advantage with referee experience, with high ($\Psi = 10$) and low ($\Psi = 1$) penalties, compared to a piecewise constant regression

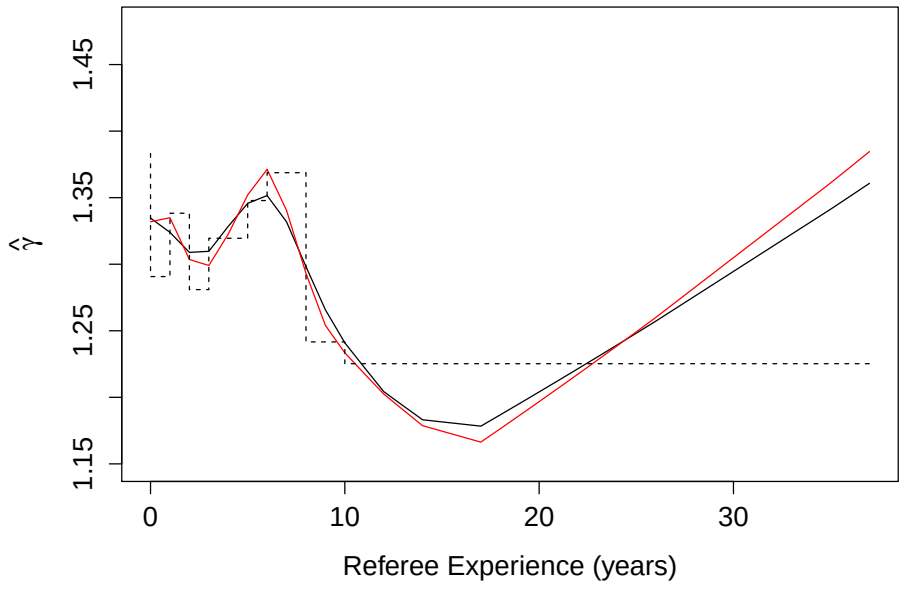


Figure F.3: League 2, 2004/2005 - 2013/2014: Penalised spline smooth curves describing the relationship of home advantage with referee experience, with high ($\Psi = 10$) and low ($\Psi = 1$) penalties, compared to a piecewise constant regression

	a_0	a_1	a_2	a_3
Log-Linear	2.93×10^{-1}	-3.25×10^{-3}	NA	NA
Log-Quadratic	3.14×10^{-1}	-9.98×10^{-3}	3.64×10^{-4}	NA
Log-Cubic	3.45×10^{-1}	2.52×10^{-2}	1.99×10^{-3}	-4.32×10^{-5}

Table F.1: Championship 2000/2001 - 2011/2012: Parameter values for first, second and third order polynomial regressions relating a regressor of referee experience to home advantage.

	a_0	a_1	a_2	a_3
Log-Linear	2.86×10^{-1}	-3.48×10^{-3}	NA	NA
Log-Quadratic	2.97×10^{-1}	-7.52×10^{-3}	2.01×10^{-4}	NA
Log-Cubic	2.87×10^{-1}	-7.80×10^{-4}	-5.30×10^{-4}	1.76×10^{-5}

Table F.2: League 2 2000/2001 - 2011/2012: Parameter values for first, second and third order polynomial regressions relating a regressor of referee experience to home advantage.

	a_0	a_1	a_2	a_3
Log-Linear	2.88×10^{-1}	-3.80×10^{-3}	NA	NA
Log-Quadratic	2.96×10^{-1}	-7.00×10^{-3}	1.50×10^{-4}	NA
Log-Cubic	2.84×10^{-1}	2.41×10^{-3}	-9.10×10^{-4}	2.43×10^{-5}

Table F.3: League 1 2000/2001 - 2011/2012: Parameter values for first, second and third order polynomial regressions relating a regressor of referee experience to home advantage.

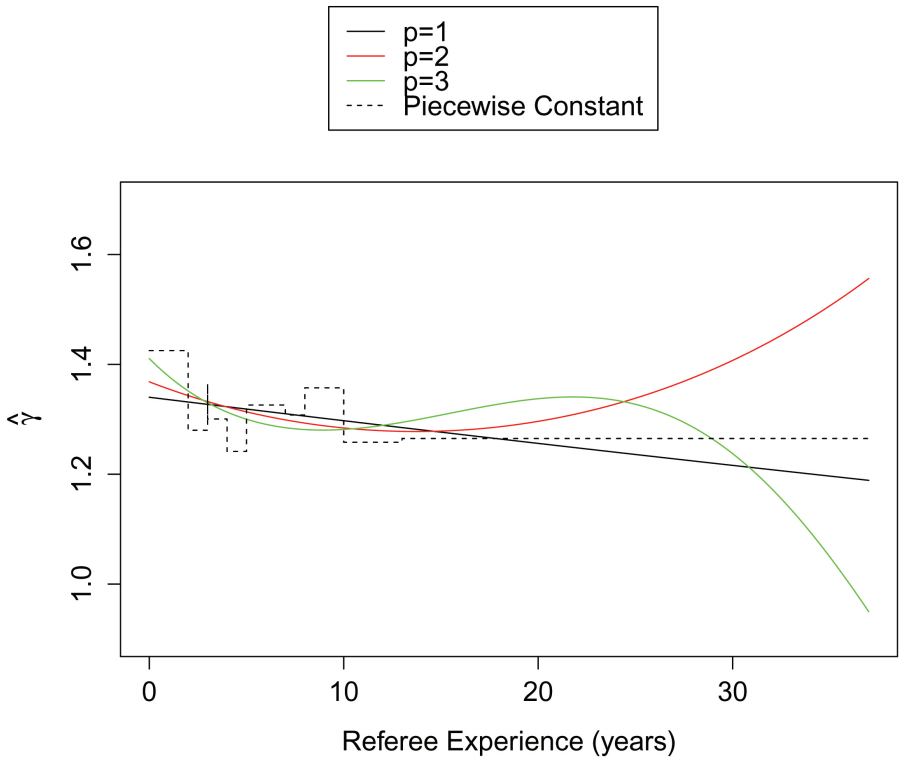


Figure F.4: Championship, 2004/2005 - 2013/2014: Comparison of first to third order polynomial regression models for referee experience as a regressor for home advantage.

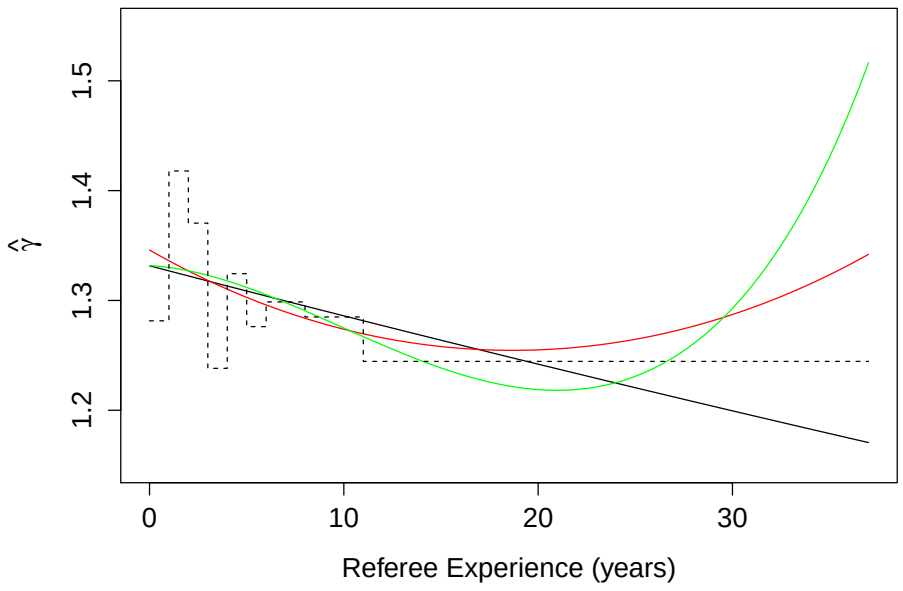


Figure F.5: League 1, 2004/2005 - 2013/2014: Comparison of first to third order polynomial regression models for referee experience as a regressor for home advantage.

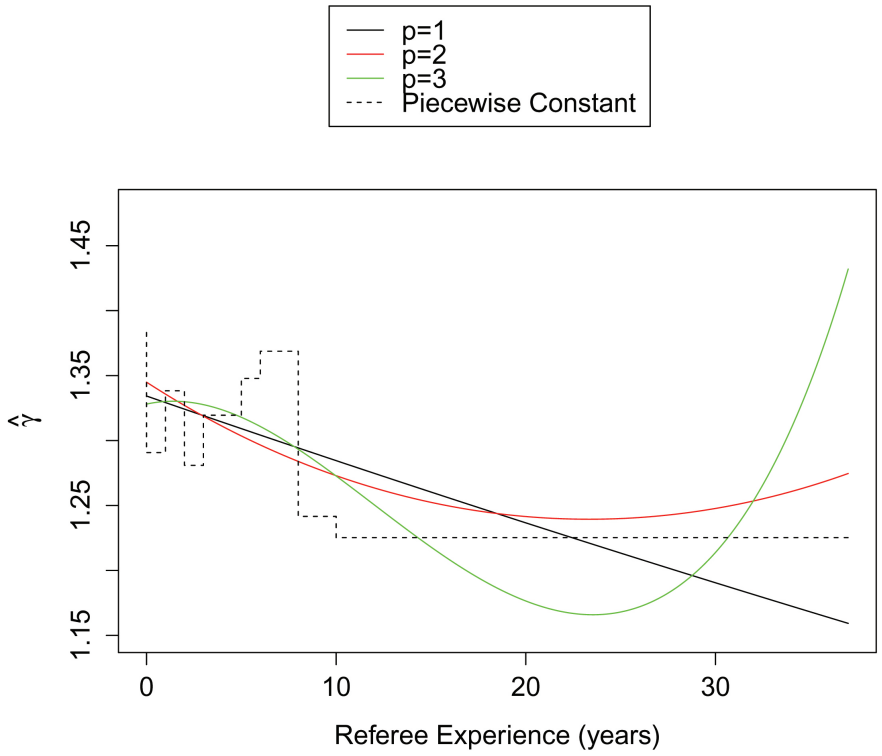


Figure F.6: League 2, 2004/2005 - 2013/2014: Comparison of first to third order polynomial regression models for referee experience as a regressor for home advantage.

Appendix G

Autotuning

G.1 Auto-tuning the Smoothing Parameter

Testing using this auto-tuning routine has shown it to be ineffectively slow. Therefore it is included only to show the methodology under testing and for completeness.

The Gaussian weighting function may be represented as in equation (G.1), where $h > 0$ represents the smoothing parameter.

$$\phi_h(t, c) = 1 - \Phi\left(\frac{t-c}{h}\right). \tag{G.1}$$

Selection of h may be achieved using cross validation techniques. The degenerate case of k-fold cross validation is referred to as 'leave one out' cross validation. This process removes one data point from the data set and maximises the likelihood over the remaining data. Some measure of fit of the resulting estimated model may then be carried out on the data point which has been removed to assertain the predictive capabilities of the model. The measure of fit which has been chosen is the mean square error (MSE).

The derivation of the expected value for X_t under the model is given by

$$\begin{aligned} \mathrm{E}(X_t) &= \int_0^\infty x_t f(x_t, \boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, c, h)\, dx_t \\ &= \int_0^\infty x_t \left[\phi_h(t, c) f(x_t, \boldsymbol{\lambda_1}) + (1 - \phi_h(t, c)) f(x_t, \boldsymbol{\lambda_2})\right] dx_t \\ &= g_t(\boldsymbol{\lambda}, c, h), \end{aligned} \tag{G.2}$$

where $\boldsymbol{\lambda} = (\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2})$ are the parameters of the respective distributions.

For a given h, the likelihood $L^{(-t)}(c, \boldsymbol{\lambda})$ (obtained by leaving out the data point at index position t from the likelihood given by equation (7.5)) is maxmised over c and $\boldsymbol{\lambda}$ with MLEs equal to $\hat{c}^{(-t)}$ and $\hat{\boldsymbol{\lambda}}^{(-t)}$ respectively. The estimated expected value at index

position t may then be calculated using the MLEs as $g_t\left(\hat{\boldsymbol{\lambda}}^{(-t)}, \hat{c}^{(-t)}, h\right)$. This process is repeated for $t = 1, \ldots, n$, allowing the calculation of MSE of the estimated expected value for each value of t. The optimal value of h may then be found by minimising the sum of the MSEs, as given by

$$\arg\min_{h} \sum_{t=1}^{n} \left[x_t - g_t\left(\hat{\boldsymbol{\lambda}}^{(-t)}, \hat{c}^{(-t)}, h\right)\right]^2 = h_{opt}. \tag{G.3}$$

G.2 Implications of Zero Bound on Smoothing Parameter on Model Comparison

When considering a two-sided hypothesis

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

where θ_0 is the true value of parameter θ, the p-value is defined as the probability, under the assumption of the null hypothesis, of obtaining a result equal to or more extreme than that which was actually observed.

Considering the distribution of parameter estimator $\hat{\theta}$ to be asymptotically normally distributed about the true value, θ_0, with standard deviation σ_0, a Pearson's chi-squared test (asymptotically equivalent to a likelihood ratio test) may be employed to evaluate how likely it is that any observed difference between $\hat{\theta}$ and the sample estimate $\hat{\theta}_{obs}$ occured due to chance. The p-value can be calculated as

$$p = P\left[\left(\frac{\hat{\theta} - \theta_0}{\sigma_0}\right)^2 > \left(\frac{\hat{\theta}_{obs} - \theta_0}{\sigma_0}\right)^2\right] = 2\left\{1 - \Phi\left(\left|\frac{\hat{\theta}_{obs} - \theta_0}{\sigma_0}\right|\right)\right\} \tag{G.4}$$

where $\Phi()$ is a standard normal cumulative distribution function.

If the estimator $\hat{\theta}$ has a lower bound of zero the restricted MLE is given by $\hat{\theta}^+ = \max\left\{\hat{\theta}, 0\right\}$. As a result, $\hat{\theta}^+$ coincides with $\hat{\theta}$ if $\hat{\theta} \geq 0$. However, a point mass at zero replaces the left tail of the distribution of $\hat{\theta}$ below zero. The size of the point mass on zero can be calculated as

$$q_0 = P(\hat{\theta}^+ = 0) = P(\hat{\theta} \leq 0) = \Phi(-\theta_0/\sigma_0).$$

The p-value relating to the restricted maximum likelihood estimate of an observed value, $\hat{\theta}^+_{obs}$ may then be given by

$$\begin{aligned} p &= P\left[\left(\frac{\hat{\theta}^{+}-\theta_0}{\sigma_0}\right)^2 > \left(\frac{\hat{\theta}^{+}_{obs}-\theta_0}{\sigma_0}\right)^2\right] \\ &= P\left(\frac{\hat{\theta}^{+}-\theta_0}{\sigma_0} < -\left|\frac{\hat{\theta}^{+}_{obs}-\theta_0}{\sigma_0}\right|\right) + P\left(\frac{\hat{\theta}^{+}-\theta_0}{\sigma_0} > \left|\frac{\hat{\theta}^{+}_{obs}-\theta_0}{\sigma_0}\right|\right) \\ &= \max\left\{0, \Phi\left(-\left|\frac{\hat{\theta}^{+}_{obs}-\theta_0}{\sigma_0}\right|\right) - \Phi\left(-\frac{\theta_0}{\sigma_0}\right)\right\} + 1 - \Phi\left(\left|\frac{\hat{\theta}^{+}_{obs}-\theta_0}{\sigma_0}\right|\right) \end{aligned}$$

If θ_0 is the bound of the parameter space, i.e. $\theta_0 = 0$, the distribution of possible estimates is altered to reflect the implications of the bound as given above and the cumulative distribution function (cdf) of the sampling distribution for $\hat{\theta}^{+}$ takes the form shown in Figure G.1 below.

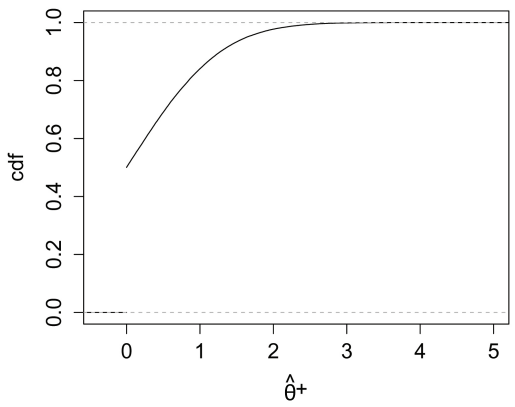


Figure G.1: Cumulative distribution function (cdf) of the sampling distribution relating to $\hat{\theta}^{+}$.

In the case of the null hypothesis, where $\theta_0 = 0$, the zero bounded p-value, p_b becomes

$$p_b = 1 - \Phi\left(-\frac{\hat{\theta}^{+}_{obs}}{\sigma_0}\right) \tag{G.5}$$

Now the significance level test has rejection region $\hat{\theta}^{+}_{obs} > 0$ and its actual error rate is smaller than that experienced when using the standard test as given by equation (G.4). Comparing equations describing the p-values of the bounded and unbounded case (equations (G.4) and (G.5) respectively) it is clear that a 0.05 significance level ($p = 0.05$) in the unbounded case becomes a 0.025 significance level ($p_b = 0.025$) in the bounded case.

Appendix H

Evolving Technologies in Golf

H.1 Introduction

The effect of technology on sporting performance is a wide and varied field of research. Many sports use some level of technology that impacts upon output. Some sports display the effect more than others, cycling for example uses a relatively high level of technology and engineering that can be manipulated to decrease track times (Bassett et al., 1999; Lukes, 2006; Haake, 2009). However, technology has only limited impact on some sports. Running categories feel little effect from technological input, the greatest of which was the introduction of starting blocks in 1948 (Haake, 2009).

Sports which have a greater public interest and (in most cases) greater attraction for investment, supply attractive research opportunities to the academic and industry sectors. Golf is one such sport, which utilises a host of patented technological devices (Yes! , 2012). The introduction of new materials and technologies has allowed all aspects of the game to be affected in some way, especially in ball and putter design.

The structure of professional golf tournaments incites a 'winner takes all' attitude to the sport. This commands a substantial motivation for players to perform to their best ability (Watkins, 2008). Technology can be thought of as an aid to the physical skill of a golfer. The margin of difference between the top and the bottom players that make the cut is generally relatively small. For example, the Open Championship 2012 at Royal Lytham and St. Annes resulted in a difference in total score of just 25, between Ernie Els at 273 and Andrés Romero at 298, a difference of just 8%. Els received £900,000 at the end of the four day stint, compared to £10,200 for Romero (GOLF Today, 2012). Money is the driving force behind the game (at a professional level), and the scale of investment into new technology that may increase player performance, even by a small amount, is increasingly large.

H.2 Literature Review

H.2.1 Putters

Putting accounts for approximately 40% of the strokes played in a game of professional golf and therefore, 40% of points (Brouillette, 2010). As putting performance relies on a single club, it is obvious that any technogical aid that can be brought to a player's putting stroke would be greatly beneficial. Putter design has attracted a considerable amount of attention from manufacturers over the last decade after the introduction of the Odyssey 2-Ball putter in 2001, which achieved a remarkable market share in the US of nearly 50% by 2003 (Callaway Golf Company, 2003). This led the way for a trend to more creative putter designs, giving professional golfers the possibility of achieving an advantage over their opponents.

Putter head design focusses on the alignment and weighting, varying the 'sweet spot' on the putter face (Gwyn et al., 1996). Material choice and cavity inserts can vary the 'feel' of the putter, distinctly changing the vibrations which travel up the shaft. The design of a putter head mainly focusses on the positioning and size of the sweet spot. Nilsson and Karlsen (2006) designed an experiment to test the effect of miss hits using a trio of putters. It was found that a wing shaped putter performed best in terms of distance and direction on miss hits, over blade and mallet types. This is shown in Figure H.1, which shows both the roll distance and impact point.

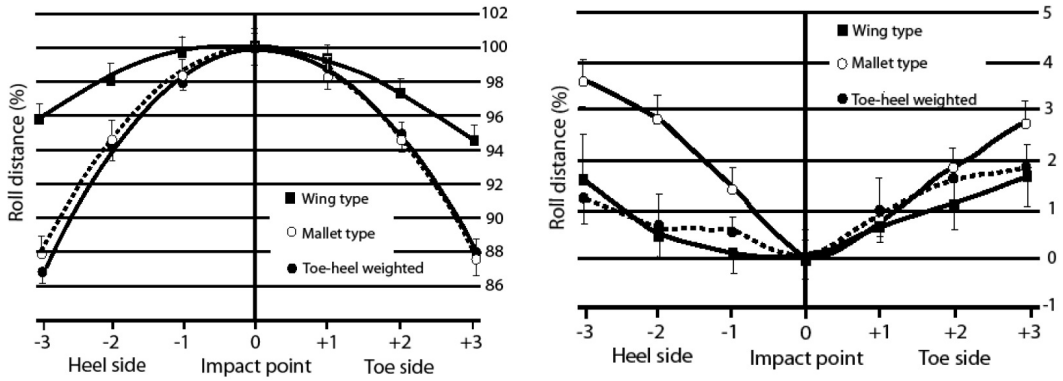


Figure H.1: (a) The relationship between roll distance (0% represents roll distance at impact point 0) and horizontal impact point deviation for three different putters. (b) The relationship between relative medio-lateral deviation as a percentage of roll distance after ball impact and horizontal impact point deviation for three different putters. Note, on both plots the sweet spot is identified as impact point = 0 and each data point represents the mean of 10 ball impacts (Nilsson and Karlsen, 2006).

Putter length has also been under investigation, although it is unclear to what extent, if any, the effect of increasing the shaft size has: Gwyn and Patch (1993) found little difference between performance whilst using long and standard putters, however, Pelz

(1990) showed long putters to be advantageous in putts under 0.9 m and worse on putts longer than 6.1 m.

Tierney and Coop (1998) used data from senior PGA Tour players to design and manufacture a "world class" putter. A comparison was drawn between their dataset and data from the ten best performing putters taking part in the 2009 PGA Tour and also from controlled tests on various putts at a practice green by elite Norwegian players (Karlsen, 2003). This is shown in Table H.1.

Distance	PGA Tour top 10 putters	"World class model"	Norwegian elite players
1 meter	93.10%	92.00%	89.70%
2 meter	64.20%	65.00%	56.90%
3 meter	43.90%	45.30%	37.30%
4 meter	30.70%	31.50%	25.10%
5 meter	22.60%	22.40%	11.50%

Table H.1: Putting test carried out by Karlsen (2003), regarding the percentage of succesful short putts by PGA Tour professionals, "World class model" (Tierney and Coop, 1998) and by Norwegien elite players.

The "world class model" compared well with the other groups tested for short putt distances. Tierney and Coop (1998) also estimated the percentage of successful putts for longer distances, the percentage of 3-putts and the expected number of putts per hole for a world class player. This is shown in Figure H.2.

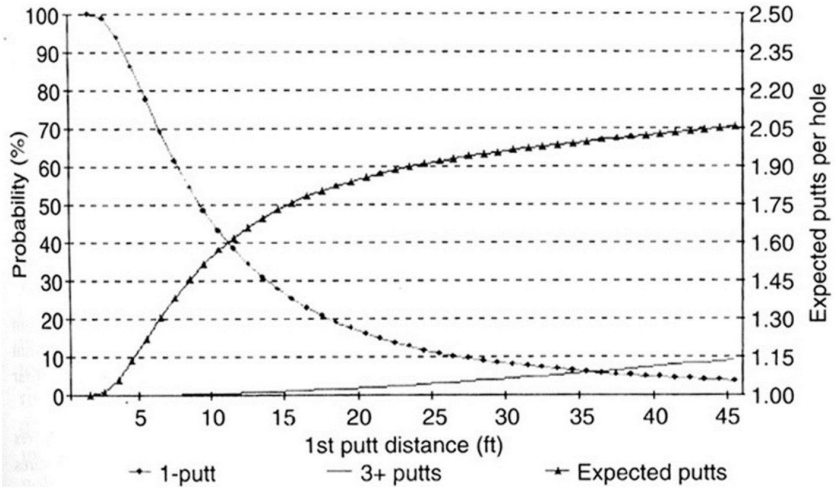


Figure H.2: 1-putt and 3-putt probabilities and expected numbers of putts taken from different distances by a world class player using the "world class putter".

Some research has been carried out using exact position and thus direction of shot in addition to distance from the target hole. Tierney and Coop (1998) analysed the deviation and estimated that a world-class player in tournament conditions would deviate

on average by 6.5° and 1.3% of the total putting distance. Karlsen (2003) found similar measurements using players with a handicap of +1.5 on a flat indoor green.

Club face grooves may have some effect on the roll of the ball during putting. Yes^{TM}, a major manufacturer of putters, uses a type of groove referred to as C-Grooves^{TM}. It is claimed that these grooves solve problems with unwanted side or back spin and ball skidding due to loft (Yes! , 2012). Their website states that: "Upon contact, these edges grip the ball surface and apply physical forces that simultaneously lift the ball out of its resting position and impart an over-the-top rolling motion".

H.2.2 Golf Ball Design

Wound gold balls, in which rubber thread is wound around two kinds of core (a liquid filled centre or a solid rubber core) and then wrapped in a balata or surlyn cover, were originally used by most tour players, for their spin and feel (see Figure H.3 (a)). Solid balls were more commonly used by amateurs because they travelled farther.

Solid core golf balls have now replaced wound balls, with many multilayer designs which offer different properties. One-piece balls (Figure H.3 (b)) are made of synthetic rubber. They are durable, but deform by a large amount upon impact, making them suitable for driving range use. Two piece solid balls (Figure H.3 (c)) are dual structured, with a high restitution core wrapped in a cover. These balls generate excellent distance as they allow a more efficient transfer of energy. Three and four piece multilayer balls (Figure H.3 (d)) wrap the core material in multiple covers, resulting in the ability to customise hardness, specific gravity, feel and other properties. This allows for compensation of a miss hit stroke (Masataka, 2008).

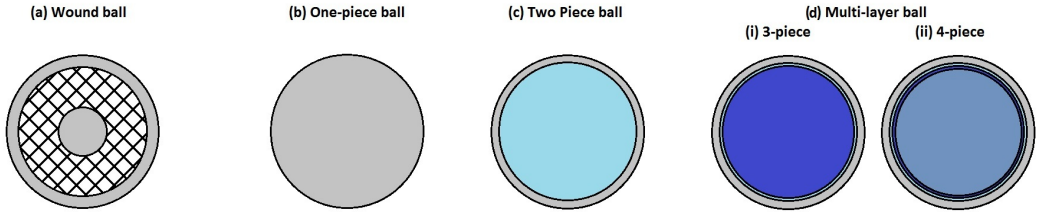


Figure H.3: Cross-sections of the most common golf ball designs (Masataka, 2008).

Nick Price won the British Open and the PGA Tour using solid golf balls in 1993 and 1994. This led to his success as the worlds number one player. Figure H.4 shows the transition from wound to solid balls during the 2001 season (Masataka, 2008). It can be seen that by the year of 2006, all PGA Tour players had made the switch to a version of the multi-piece golf ball. One possible driving factor for tour-wide change could have been Tiger Woods's choice of golf ball - changing from wound balls to solid core balls

made by Nike (Johnson, 2001).

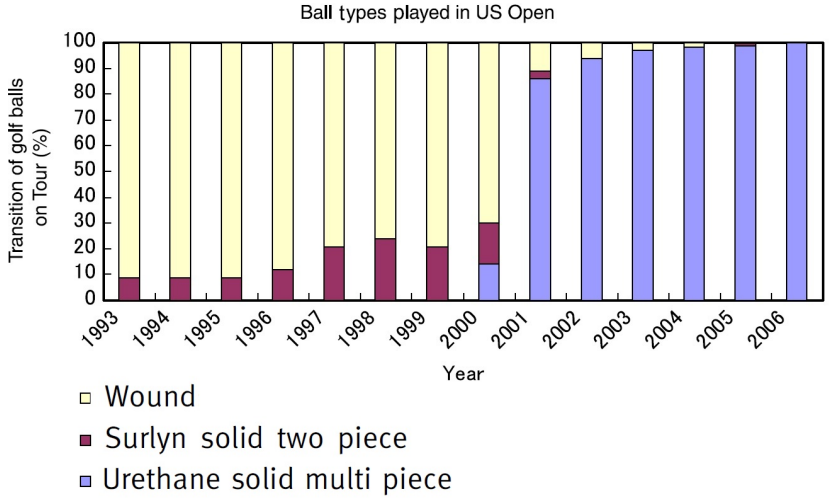


Figure H.4: Transition of golf balls on tour (Darrell Research, 2012).

The consistency of solid golf balls improved due to new production technologies and therefore their quality and properties have been widely noted as giving good distance, feel and spin (Masataka, 2008). Wound balls have the property of a reduction in initial velocity with decrease in temperature. Figure H.5 shows (a) the temperature dependence and (b) the amount of ball compression experienced by wound and solid balls. A typical wound ball will compress in a much more uneven and unpredictable nature than a solid core urethane ball.

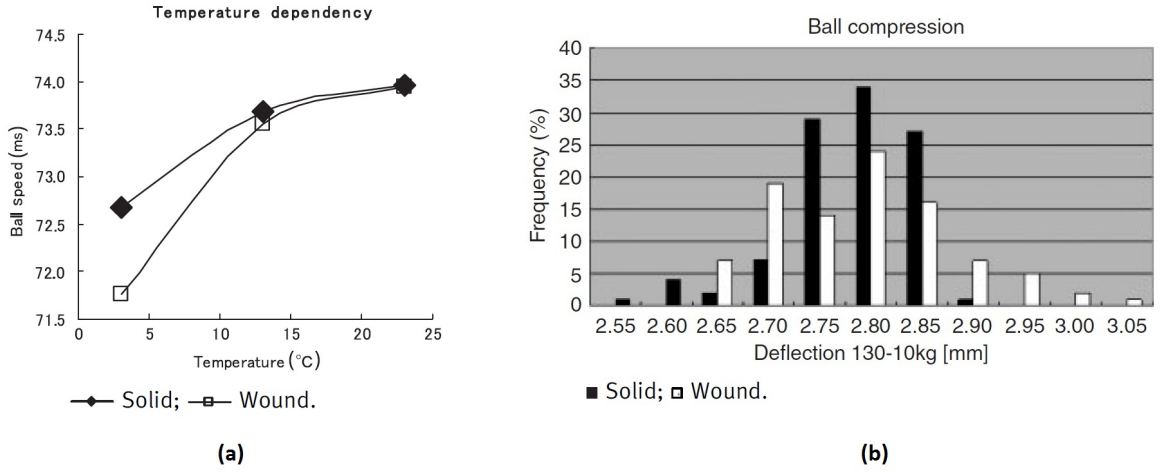


Figure H.5: (a) Temperature dependence and (b) compresion dispersion of wound and solid urethane core balls (Masataka, 2008).

Ball design is not limited solely to internal construction, the outside of the ball affects the way a ball travels through the air. During travel, particles of fluid move in the boundary

layer close to a balls surface. Dimples increase the amount of linear momentum and energy that those particles experience. Any solid body that is in relative motion with a surrounding fluid will experience this boundary layer activity due to the peculiar nature of fluid friction near the body's surface (Munson et al., 2006). Dimples on a golf ball trip close-moving air particles, causing them to be disturbed and vibrate sideways whilst progressing with forward velocity, instead of maintaining parallel *lanes of traffic*, as generally experienced in laminar flow. This vibration causes particles in adjacent *lanes* to make contact with each other and causes a transference of linear momentum from one particle to the next. Fast moving particles will be slowed and slow moving particles will increase in velocity increasing the kinetic energy and linear momentum within the boundary layer.

When the particles bump together vigorously, the airflow becomes turbulent instead of laminar, the former having the greater energy and linear momentum. This results in the ability to resist adverse pressures over larger distances along the surface of a golf ball. Eventually, particles will be forced out of the boundary layer due to adverse pressure ceasing forward motion of boundary particles, this is called the *point of separation*. Downstream and upstream of the point of separation, pressures that act on the ball are different, this difference being referred to as *pressure drag* or *form drag* (Blevins, 1992). As shown in Figure H.6 the point of separation is located closer to the rear of the ball with respect to direction of travel than laminar flow. This creates a smaller pressure drag when the boundary is turbulent, thereby increasing travel distance.

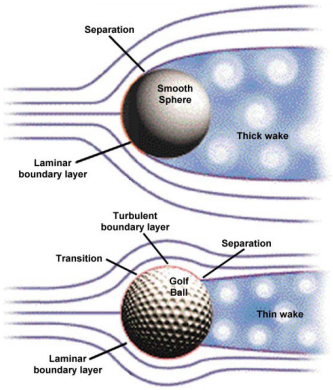


Figure H.6: Viscous wake and delayed separation (Aero Space Web, 2006).