Predicting Outcomes in Australian Rules Football

Richard Ryall
B.App.Sci(Stats)(Hons)

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

School of Mathematical and
Geospatial Sciences
RMIT University
January 2011

Statement of Authorship

The candidate hereby declares that:

- except where due acknowledgement has been made, the work is that of the candidate alone;
- the work has not been submitted previously, in whole or in part, to qualify for any other academic award;
- the content of the thesis is the result of the work which has been carried out since the official commencement date of the approved research program;
- any editorial work, paid or unpaid, carried out by a third party is acknowledged.

Richard Ryall

January 2011

“*Mathematics, rightly viewed, possesses not only truth, but supreme beauty - a beauty cold and austere, like that of sculpture.*”

- Bertrand Russell

Acknowledgements

I wish to acknowledge the following people and organisations, without whose assistance this dissertation would not have been possible.

Dr. Anthony Bedford - Thank you for being such an inspiration throughout this journey, I could not have asked for a better senior supervisor. Your passion for statistics in sport will never be forgotten.

RMIT Sports Statistics Research Group - To all the members past and present, it has been great to share this experience with like minded people and watch the reputation of this group grow each year. Special mention to **Dr. Adrian Schembri** and **Dr. Cliff Da Costa** for astute comments on earlier versions of this dissertation.

Dr. Mark Stewart - My career in sports statistics first started working under your guidance as a Research Assistant. This research turned into an ongoing project which I felt fortunate to have been a part of.

Mr. Jason Ferris - For teaching me amongst other things, how to write “good code” and for always making time available to answer any questions.

Prowess Sports - For providing data used at various stages throughout this dissertation.

Anonymous Referees - Your astute comments and advice on journal articles and conference proceedings helped improve the foundation of this dissertation.

Examiners - Thank you for your kind words and insightful suggestions which polished the overall quality of this dissertation.

Australian Postgraduate Award - For providing financial support without which this dissertation would not have been possible.

To my family - Mum, James and William; I'm always astonished by the obstacles we have overcome and I'm extremely proud of the people we are today. I draw strength from you all.

Finally, I wish to dedicate this dissertation to the memory of my father

Thomas Gordon Ryall

(17/09/1947 - 21/09/2004)

Summary

The primary aim of this dissertation was to utilise mathematical models and computer programming techniques to provide further insight in relation to predicting outcomes in Australian Rules football (AFL). This thesis comprises a collection of research problems relating to home advantage, match prediction and the efficiency of betting markets in AFL. Firstly, a new paradigm was proposed for predicting home advantage in AFL by separately evaluating a number of psychological (crowd intimidation), physiological (travel fatigue) and tactical (ground familiarity) factors. This novel method for quantifying home advantage was utilised for match prediction using a variant of the Elo ratings system. These predictions were applied to betting markets to see if consistent profits were attainable using betting strategies based around the Kelly criterion. Due to a severe lack of accessible in-play betting data, a computer program was developed using the programming language Perl to integrate with the Betfair Application Programming Interface (API) to automatically record in-play betting data for AFL matches. This information was updated in a MySQL database which could then be easily exported as a CSV file for manipulation in Excel. The in-play betting data was transformed to provide a visual representation of who is going to win the match and with what level of certainty. Tests of semi-strong efficiency were performed on the in-play betting data for the 2009 AFL season using logistic regression to see whether teams with certain characteristics are underbet or overbet relative to their chances of winning. A real time prediction model was developed using a Generalised Logistic Model which accounts for the interdependence, if any, between team quality and score difference as the match progresses. These predictions were applied to in-play betting markets to see if consistent profits were attainable using betting strategies based around the Kelly criterion. If home advantage in AFL is comprised of a combination of psychological, physiological and tactical factors then it's plausible that home advantage is dependent upon the current state of the game (score) since the crowd, for example, react to performance. Therefore, home advantage

was modelled at various stages during the game to see the difference, if any, between home teams with certain pre-game characteristics (favourite/underdog) and in-game characteristics (ahead/behind). Finally, a macro was written in Excel to automate the transformation of a mass of "live-streaming" performance data into a single web-based phases of play plot. Statistically, the plot provides an effective representation of the state of the game at any point in time, illustrating which team is playing a style of football highly correlated with winning. Graphically the plot is enhanced by adding images of a player's guernsey when a goal is scored.

Contents

List of Tables

List of Figures

List of Abbreviations

ACT	Australian Capital Territory
AFL	Australian Football League
AFNC	Australian Football National Council
API	Application Programming Interface
CS	Carrara Stadium
D	Docklands Stadium
EMH	Efficient Market Hypothesis
FIDE	Fédération Internationale des Échecs
FP	Football Park
G	Gabba
KP	Kardinia Park
MCG	Melbourne Cricket Ground
MKO	Manuka Oval
MLE	Maximum Likelihood Estimation
MO	Marrara Stadium
NSW	New South Wales
NT	Northern Territory
OLS	Ordinary Least Squares
PP	Princes Park
QLD	Queensland
S	Subiaco
SA	South Australia
SA	Stadium Australia
SAFL	South Australian Football League
SANFL	South Australian National Football League
SCG	Sydney Cricket Ground
TAS	Tasmania
USCF	United States Chess Federation
VBA	Visual Basic for Applications
VFA	Victorian Football Association
VFL	Victorian Football League
VIC	Victoria
WA	Western Australia
WAFA	West Australian Football Association
WAFL	West Australian Football League
YP	York Park

Chapter 1

Introduction

Why are AFL teams at a disadvantage when they travel interstate? Is Essendon a better football team than Carlton? Which team during a match represents good value from a betting perspective? With what level of certainty should Richmond beat Collingwood when the scores are level midway through the third quarter? Are some AFL teams lucky winners (or unlucky losers)? What role does home advantage play before and during an AFL game? These questions are constantly asked by spectators, commentators, coaches and the football public. Many will give subjective answers with no empirical evidence. For example, "The Dockers, willed on by a frenzied home crowd, charged home from a 14-point three-quarter time deficit to win" stated Braden Quartermaine, journalist for The Age newspaper (10th April 2010). "We were very lucky to win today, but we will take it" said Mark Thompson (Geelong coach) in a press conference after the 2009 Grand final. Statistics, or more specifically statistical modelling, can help provide us with objective answers to these questions.

In most sporting competitions, the performance of teams or individuals are measured on an objective criteria, both during and at the conclusion of the match. For example, during a boxing bout a player's performance can be evaluated based on the number of rounds they have won (and lost) during the bout. Similarly, at the conclusion of an AFL game a team's performance can be measured not only on whether they won, lost or drew but also by what magnitude. However, there are many factors that should be taken into account when measuring a team's performance objectively, such as any difference in team quality and the existence of home advantage. Notably, many of these factors are typically measured in isolation which contributes to misleading results. For example, the goal scoring accuracy of forwards in AFL can be measured by the number of goals they kick relative to the total number of shots that player has on goal. However, while a low conversion rate indicates the forward was an average kick at goal, it could also indicate that the position the shots were taken from (distance and angle) influenced the conversion rate.

An important aspect of analysis within this dissertation was to ask the appropriate questions when analysing AFL from both a pre-game and in-game perspective. Throughout my candidature I have constantly questioned previous research findings by sports research statisticians. Is there a better methodology? Are their results caused by another factor? Why did they only look at a single season? Did they explore all possible avenues? How can I build upon this research? If there is one thing that I have learnt during this period, it is that although numbers never lie they can often be misleading.

1.1 Why Australian Rules Football?

Like many Victorians, I have always had a strong passion for sport. Tennis was my sport of choice as a teenager, training numerous times a week at Melbourne Park (where the Australian Open is held) in preparation for competition at the crack of dawn every Saturday. Every year I would head to the Australian Open in January to watch the worlds best players play up to five sets on scorching hot days. Cricket was another sport I thoroughly enjoyed, and although I never played at a competitive level, I had all the gear and played some pretty serious front yard cricket with friends. Although I thoroughly enjoy watching and playing numerous other sports, in recent years AFL has piqued my interest over these sports. There is nothing quite like heading to the MCG on a cold winters night to watch your team play in front of up to 100,000 spectators. Unfortunately, my increased enthusiasm for AFL has been inversely proportional to the success of my beloved Essendon football club!

During my honours year as a statistics undergraduate, I undertook the subject "Regression Models in Econometrics" and the lecturer (Ms Kaye Marion) allowed us to model outcomes using our own data in an area of interest. The choice was obvious to me, so I tried to predict the best and fairest player for Essendon using basic historical statistics (kicks, handballs etc) which were available via a public domain. The final model performed very poorly ($R^2 = 0.03$), however I learnt that demonstrating there is no association between a number of predictor variables and an outcome variable is just as important as demonstrating an association. Several weeks after this project I received an email from a colleague of Kaye's, Dr. Mark Stewart, who required a research assistant to work on an AFL recruitment project. I literally jumped at the opportunity and my career in sports statistics snowballed. I attended a session for prospective PhD students where potential PhD supervisors could give a spiel about PhD topics. This led me to get in contact with Dr. Anthony Bedford whose research interests revolved around sport, and as they say, the rest is history. To this day I still collaborate with Dr. Mark Stewart and his colleagues on many applications of economics in sport.

1.2 Applications of Sports Statistics

The depth and breadth of sports statistics has grown rapidly over recent years. I have listed some below, however there are numerous others.

(i) Sports betting is a lucrative business with a plethora of sports bookmakers world wide. In 2008-09, industry revenue from horse and sports betting in Australia was

$22,674 million (IBISWorld, 2009). With the advent of internet betting, punters can bet on the outcome of AFL matches both pre-game and during the game (in-play); future medals (Brownlow, Coleman, rising star, etc.); match results and even head to head fantasy football scores! Statistical analysis can aid in the prediction of outcomes and assist punters to exploit inefficiencies in betting markets. Throughout my PhD candidature I have received many emails from professional punters to develop profitable mathematical models for sports betting.

(ii) Fantasy football (AFL) is a fantasy sports game in which participants take the role of a manager and select real players in different positions. Each manager has a salary cap which they must adhere too, players then earn points based on their actual match performance. Participants are placed in leagues and compete against one another for bragging rights, although now most competitions have prizes for overall winners which can be upwards of $50,000. There are many competitions with different scoring systems, with the majority of the larger competitions comprising several 100,000's of competitors. Sports statistics plays a pivotal role in the scoring system and the valuation methods based on player performance.

(iii) Sport broadcasting is big business. Currently, three American television networks CBS, NBC and Fox, and cable television's ESPN are paying a combined total of US$20.4 billion to broadcast NFL games. With CBS, Fox and NBC paying US$3.73, US$3.6 US$4.27 respectively until 2011 and ESPN paying US$8.8 billion until 2013 (nfl, 2007). Although the AFL is still well behind the NFL in terms of television revenue, the AFL contribute in excess of A$1 billion annually to the Australian economy (afl, 2009a). In 2006, the television rights for seasons 2007 to 2011 were sold for A$780 million (afl, 2006). Notably, television broadcasters are constantly trying to provide viewers with new metrics and visuals to provide further insight into the performance of teams and individual players, an area for statisticians to add value.

(iv) Sporting organizations also spend a great deal on sport science and performance analysis to both optimise athlete performance, and gain a competitive edge over their direct competition. Sporting success on a national scale is thought to boost national pride, increase “grass roots” (or suburban level) participation in sport, and increase employment opportunities in the business of sport. Current funding for high performance Olympic and Paralympic sport from all sources was $128.3 million per annum in 2010. Similarly in AFL, on field success leads to an increase in club memberships, gate receipts and club merchandise. This in turn allows clubs to spend more on football department spending. In 2006, the average non-player football department spending for each team in the AFL was approximately $4.1 million, with Collingwood spending the most ($5.8 million) and Kangaroos spending the least ($3.1 million) (afl, 2007). Incidently, Collingwood have finished in the final eight in seasons 2006 to 2010, eventually winning the premiership in 2010, meanwhile the Kangaroos finished in the final eight in only two of the previous five seasons (2007 and 2008).

(v) Many Australians are in tipping competitions, whether it be in a competition at work, the local pub, university or amongst friends. Every participant has their own strategy be it tipping the favourite, the home team, a random selection or a statistical model based on historical data. Some strategies are more sophisticated than others. Prizes for winning a tipping competition can vary from personal satisfaction and bragging rights to $100,000’s in some of the larger online competitions which are typically free to enter.

(vi) In Australia, there are radio stations and television channels that are dedicated to sport despite the substantially smaller population in comparison to the United Kingdom and United States of America. Throughout my PhD candidature, I have participated in many radio interviews and have appeared in newspapers explaining my latest research on statistics in sport to a more general audience.

(vii) Several journals are devoted to sport in areas of mathematics, statistics, economics and finance to name a few. Several journals which I have published in throughout my PhD candidature include: *Journal of Sport Sciences*, *Journal of Sports Finance* and *Journal of Quantitative Analysis in Sport*. Many other journals also feature special editions devoted to sport including the *International Journal of Forecasting* and *IMA Journal of Management Mathematics*.

(viii) Finally, sport is a great way for teachers to show students the application of statistics to the real world in an area that they could show a real interest. After all that is how my career in sports statistics first began.

1.3 Literature Review

This section explores previous research in sports statistics with a specific focus on home advantage, match prediction and market efficiency. Although there is a specific focus on AFL in this literature review, due to the sparsity of research in AFL, previous research in these areas has also been extended to other high scoring sports. Furthermore, the rationale for this dissertation is provided by detailing how this research builds upon and improves previous research, either methodologically or performance wise.

1.3.1 Home Advantage

Home advantage has long been recognised as a contributing factor to success in team and, more recently, individual sports. Home advantage typically refers to the net advantage of several factors which, generally speaking, have a positive effect on the home team and a negative effect on the away team (Harville and Smith, 1994). The much acclaimed paper by Schwartz and Barsky (1977) provides evidence of home advantage in four American sports namely baseball, football, ice hockey and college basketball. Since then, subsequent research in home advantage has been extended to many other professional sports including Pollard (1986) and Clarke and Norman (1995) in football (soccer), whilst Holder and Nevill (1997) and Bailey et al. (2010) investigated home advantage in individual sports encompassing tennis and golf. More relevant studies on home advantage in Australian Rules football include Bailey and Clarke (2004) and Clarke (2005). A comprehensive literature review on home advantage is given in Nevill and Holder (1999). Courneya and Carron (1992) provided a helpful taxonomy which integrated the previous findings of research on home advantage in soccer, hockey, baseball, basketball and grid iron.

Schwartz and Barsky (1977) proposed three explanations as to why home advantage may exist; learning/familiarity factors, travel factors and crowd factors. These can be classified as tactical, physiological and psychological respectively. Courneya and Carron (1992) build upon these and suggested referee bias as another factor for consideration. Although these factors are usually cited as the *cause* of home advantage in team sports, the precise *contribution* of each factor still remains relatively unknown (Pollard, 2008). Courneya and Carron (1992) suggested that future research endeavours should be directed towards explaining the possible *cause* of home advantage. The seminal paper by Clarke (2005) found evidence that non-Victorian teams have larger home advantage than Victorian teams (see Section 4.2 of Chapter 4 for further details). Furthermore, there was evidence that Victorian teams that shared the Melbourne Cricket Ground (MCG) received a smaller home advantage than Victorian teams that played their home matches at their respective training venue. Interestingly, the author cites crowd intimidation and ground familiarity as the contributing factors of home advantage in AFL *without* any empirical evidence. Such subjective statements about the precise cause of home advantage in sport is commonplace throughout the literature. For example, Stefani (2008) states that "the large size playing oval in Australian Rules Football probably reduces the crowd's psychological influence, compared to rugby union, soccer and the NBA which also have a large percentage of the ball being in play." (Stefani, 2008, p. 212).

Bailey and Clarke (2004) realised this deficiency in the literature and endeavoured to attribute the relative contribution of travel and familiarity factors towards home advantage in Australian Rules football. The authors found performance of the nominated home team increased as the difference in matches ever played at the home venue increased (ground familiarity). Similarly, when the nominated away team travels interstate, the further they travelled the greater the disadvantage was found. However, authors such as Courneya and Carron (1991) and Pace and Carron (1992) suggested that team ability, ground familiarity, travel fatigue and crowd intimidation affect performance simultaneously. For this reason, one aim of this dissertation is to disentangle and quantify the independent effects of home advantage in AFL, and further advance previous research by numerically quantifying the effect of crowd intimidation. Bailey and Clarke (2004) state that information on crowd numbers and more specifically crowd passion is not readily available making it difficult to quantify the effect of crowd intimidation. However, Biddle (1993) showed amongst other things, that team success was highly correlated with spectators attending matches. By extending the work of Biddle (1993) it is possible to predict not only total crowd numbers but also the mix of crowd support, that is, the breakdown of home, away and neutral supporters. The influence and relative weighting of psychological factors (crowd support and stadium density), physiological factors (distance travelled and origin of away team) and tactical factors (ground familiarity) can then be quantified controlling for other factors including team ability.

1.3.2 Rating Systems in Sport

One of the most fascinating aspects of sporting competitions is that the team of greater quality does not always win. It is this uncertainty that draws spectators to matches since there is the belief that any outcome (win, loss, draw) is possible on any given day. Fans, media, coaches and even players constantly argue about which team is superior, a question that typically remains unresolved even after a contest. Rating systems on the other hand, provide an objective measure separated from the passionate subjectivity of supporters. Here, it is important to distinguish between a rating and a ranking. A rating is measured on an arbitrary continuous scale which should describe a team's ability in relation to the competition, whereas a ranking is the ordering of teams (1st, 2nd, . . . , 2nd last, last) based on their respective rating. Stefani (1998) provided a comprehensive taxonomy and survey of sports rating systems in 83 different sports, classifying each sport as either combat sports, object sports and independent sports. Stefani (2010) extended this taxonomy to 156 different sports. In the earlier paper, Stefani (1998) classified AFL as an object team sport and compared several different rating methods to evaluate previous performances and predict future performances. These rating systems included a least squares method by Stefani (1987) originally developed for football and basketball predictions. This model was then implemented with and without a regression towards the mean (James and Stein, 1961). Other models included an exponential method (Clarke, 1993) known as "Tinhead the Tipster" (Clarke, 1988) and a probabilistic approach (Harville, 1980). Stefani and Clarke (1992) demonstrated that dissimilar ratings systems using the same inputs (information) tend to converge to a limited accuracy level in terms of the percentage of games correctly classified. Put simply, a ratings system is only as good as the information incorporated into the model. Akin to Bailey and Clarke (2004), one aim of this dissertation was to quantify home advantage more accurately by identifying the independent effects that contribute to home advantage in AFL. Consequently, this should improve the forecasting capabilities of a ratings system.

Rating systems have been described as either adaptive or accumulative (Stefani and Clarke, 1992). An accumulative rating system is where teams accumulate points that never diminish on which teams are subsequently ranked upon. For example, in the majority of soccer leagues, teams are awarded three points for a win, one point for a draw and no points for a loss. An adaptive rating system is where a team's rating rises or falls depending upon whether their performance is above or below a predicted level. For example, the World Chess Federation uses the Elo rating system (Elo, 1978). It was thought that performance in chess could not be inferred by a sequence of moves but rather from a series of wins, losses and draws. This system compares the number of games a player is expected to win with the observed number of games that player actually wins. If a player exceeds these expectations they receive a rating increase, similarly if a player falls short of these expectations they receive a rating decrease. However, in a sporting competition, team performance is not only measured by wins, draws and losses, but by the magnitude of those results. Applying Elo ratings to sporting competitions has typically focused on Association football (Hvattum and Arntzen, 2010; Leitner et al., 2009). Therefore, in this dissertation, a novel variation of the Elo ratings model is developed to predict the outcome of AFL matches.

The seminal paper by Clarke (1993), which utilised an exponential smoothing technique, remains the standard for match prediction in AFL. Weekly predictions are published at http://www.swinburne.edu.au/lss/statistics/footytips.html which have generated great media attention over the years (Clarke, 1993). Previous studies have shown that this program consistently predicts as many winners as the best expert tipster and outperforms them in predicting margins. More recently, the application of these predictions to betting markets (Clarke et al., 2008) have assisted punters to exploit betting inefficiencies via a subscription service (http://www.smartgambler.com.au/afl/intro.html). Interestingly, Clarke's first ratings algorithm outperformed the improved version in their first year of prediction for seasons 1981 and 1991 respectively. An explanation given by Clarke is that an even competition makes predicting winners far more difficult. Therefore, when comparing ratings systems across different eras, a new method for evaluating the evenness of the competition was introduced.

Various measures can be used to evaluate the performance of prediction models in game sports. Some commonly used measures in the literature include Average Absolute margin of Error (AAE), number of predicted winners and Return on Investment (Bailey and Clarke, 2004). However, a common limitation of existing literature is to evaluate the performance of a prediction model based on a single season. For example, Bailey (2000) used a multiple linear regression using seasons 1997 and 1998 as a training set in the forward prediction of the 1999 season. Similarly, Flitman (2006) used genetically defined neural networks and linear programming using seasons 1992 to 1995 in the forward prediction of season 2002. Stefani and Clarke (1992) showed that the number of predicted winners can vary by up to 10% from one season to the next. This is not so much a deficiency of prediction models, but rather showcases the ebbs and flows of the competitiveness of the AFL competition. Therefore, eight seasons of data were used in forward prediction in this dissertation. This removes the subjectivity of choosing fewer seasons in order to inflate (intentionally or unintentionally) the predictive power of the model.

1.3.3 Market Efficiency

In finance, market efficiency, or more specifically the Efficient Market Hypothesis (*EMH*), is the supposition that financial markets are "informationally correct". The much acclaimed paper on *EMH* by Fama (1970) defined market efficiency into three subsets: Weak Form Efficiency, whereby future prices can not be predicted by past prices; Semi-Strong Efficiency, whereby future prices cannot be predicted by publicly available information; and Strong Form Efficiency, whereby prices reflect all information, both public and private.

The efficiency of both financial and betting markets has received great attention in the literature. The fundamental question in both these markets is whether price incorporates all publicly available information. A direct test of market efficiency in financial markets is complicated, as the true worth of a share in a company and the expected payoff is always unknown. Betting markets on the other hand, provide the perfect opportunity to test for market efficiency. The expected payoff (betting odds) for each wager is fixed and the outcome of each wager is settled at the conclusion of an event.

A consistent finding in the literature on the efficiency of racetrack betting markets is the existence of “favourite-longshot bias” (Thaler and Ziemba, 1988). This is the tendency for favourites to be underbet and longshots to be overbet relative to their chances of winning. Although the expected return of betting on the favourite is significantly greater than that of the underdog, both methods typically yield negative expected returns once accounting for transaction costs. Previous research on the efficiency of the National Football League (NFL) include Zuber et al. (1985), Gandar et al. (1988), Golec and Tamarkin (1991), Dare and MacDonald (1996), Gray and Gray (1997) and Dare and Holland (2004). In those works, although profitable betting strategies were shown to exist in the NFL, predominantly by betting on the home-team underdog, any biases that did exist dissipated over time, a sign of increased efficiency. Woodland and Woodland (1994) investigated the efficiency of major baseball betting markets and found the favourite-longshot bias existed in reverse. However, a more recent study by Gandar et al. (2002) using a revised test for unbiasedness, found no evidence to reject the hypothesis that the baseball betting market is efficient. Woodland and Woodland (2001) investigated the efficiency of the National Hockey League (NHL) fixed odds betting markets and demonstrated the existence of reverse-longshot bias, a bias which increased when the underdog was also playing away from home.

More relevant studies on the efficiency of sports betting markets include Brailsford et al. (1995) and Schnytzer and Weinberg (2008) focusing on Australian Rules football. Brailsford et al. (1995) found evidence of a favourite-longshot bias. Schnytzer and Weinberg (2008) utilised the fact that many AFL games are played on a neutral ground and thus were able to disentangle the home team and favourite-longshot bias. They found evidence of a significant bias in favour of home teams.

Although considerable research has been conducted on the efficiency of sports betting markets, the amount of research dedicated to the efficiency of in-play sports betting markets is minuscule, primarily due to the infancy of in-play sports betting. Debnath et al. (2003) is perhaps the first to investigate information incorporation for in-play sports betting markets. The authors found that prices were highly correlated with score for both soccer and basketball. Interestingly, due to the infrequent nature of scoring in soccer, price changes are less frequent but more dramatic than basketball. Also, meaningful goals late in the match in soccer affect price significantly more than earlier goals. Gil and Levitt (2007) found visual evidence that immediately after a goal is scored the market does not fully incorporate new information as prices trend upwards for approximately 15 minutes after a goal is scored. However, this trend is likely to be attributed to the goal becoming more valuable as the match progresses. Easton and Uylangco (2007) measured the efficiency of one day cricket matches on a play-by-play basis (ball-by-ball). They found evidence that in-play betting markets in one day cricket matches incorporate "good news" (runs) or "bad news" (wickets) rapidly in the betting odds. Easton and Uylangco (2010) test the efficiency of in-play betting markets in tennis by comparing the implied probabilities deduced from the in-play betting odds with a previous model developed by Klaassen and Magnus (2003). Based on the 49 singles matches they analysed, an extremely high correlation was found between the model and betting market.

In summary, the research by Debnath et al. (2003); Gil and Levitt (2007); Easton and Uylangco (2007, 2010) investigated market behaviour of in-play betting markets in sport from a visual perspective, that is, how the betting market reacts to critical events such as a goal in soccer or a lost wicket in cricket. However, no statistical analysis was executed to objectively assess the *EMH* criterion controlling for all other variables including pre-game characteristics. Therefore, in this dissertation a new method is developed to test for specific biases utilising in-play betting markets in AFL. Furthermore, any specific biases found are only of practical importance if the bias is significant enough to be exploited via a profitable betting strategy in excess of commissions. Therefore, common betting strategies were implemented to determine if a profit can be derived by betting on teams with certain characteristics.

1.3.4 Intra-match Home Advantage

If home advantage in AFL is comprised of a combination of psychological, physiological and tactical factors, then it is plausible that home advantage is dependent upon the current state of the game (score) since the crowd, for example, react to performance. The most notable research in this area was the work by Jones (2007) who investigated home advantage in the NBA as a game-long process. The results from this study suggested that home advantage in NBA is strongly frontloaded, that is, two thirds of the home advantage at the end of the match is accumulated in the first period while the remaining is dispersed in small increments over the rest of the game. Furthermore, Jones found that after the first period, home advantage was greater when the home team was behind at the end of the previous quarter. However, several key problems plagued this investigation and led to conflicting findings and conclusions. Jones concluded that:

> “Before the game starts the home team can expect to win the game roughly 62.0% of the time. If the home team is behind at the end of the first quarter, that percentage drops to 44.4% in 2002-03 and 43.8% in 2003-04. The home advantage is not something that the home team retains regardless of how it performs during the game. If the home team lets itself be outscored in the first quarter, then the advantage it had when the game started is lost.” (Jones, 2007, p. 11).

This concluding remark contradicts the finding that home advantage is greatest when the home team is behind on the scoreboard. The decrease in home win percentage from 62% pre-game to 44% at the end of the first quarter if the home team is behind is most likely going to be *caused* by the difference in team quality. For this reason, the current research builds on the work by Jones by modelling the effects of home advantage in AFL during the match controlling for team quality.

A commonly described theory revolves around the idea of a home field disadvantage in Championship “Play Offs”. For example, Baumeister and Steinhilber (1984) investigated the home field disadvantage in the baseball World Series; similarly Wright et al. (1991) in golf championships; and Wright et al. (1995) in ice hockey championships. Baumeister and Steinhilber (1984) first introduced the notion that home teams close to victory appear to “choke” in the final game of a series. They found that home teams won 39% of matches in the decisive seventh game of the baseball World series between 1924 and 1982. Even though the results are somewhat counterintuitive, this has been well supported by subsequent laboratory experiments by Butler and Baumeister (1998). In their study, performers believed that supportive audiences were more helpful and less stressful. However, the results indicated that when respondents were required to perform a difficult task in front of supportive audiences, they elicited cautious behaviour, that is, speed decreased without improving accuracy. However, a number of studies have questioned the concept of the home field disadvantage in Championship “Play Offs” due primarily to the small sample the analysis is based on Courneya and Carron (1992). Another study by Wolfson et al. (2005) showed that 11% of supporters believed home advantage could be detrimental to the home team due to players feeling more pressure at home. The same theoretical questions can be applied to AFL during the match, that is, do home teams perform poorly when the match is there to be won? Or more specifically, do home teams perform poorly in the final quarter when the scores are close? Therefore, in this dissertation, the intra-match home advantage (or disadvantage) for teams with certain pre-game characteristics (favourite, underdog) and in-game characteristics (score difference) is investigated to determine if there is any statistically significant difference.

1.3.5 Real Time Predictions in Sport

Real time prediction in sport is a growing area of research as various researchers attempt to gain an edge for in-play betting markets (Glasson, 2006; Bailey and Clarke, 2006); provide a visual representation of the match over time (Westfall, 1990; Stern, 1994); or simple interest in explaining the variation during a game (Falter and Perignon, 2000; Klaassen and Magnus, 2003). The methodology used typically depends on the sport in question and the required frequency of probability estimates. For example, the Brownian motion model (Stern, 1994; Glasson, 2006) for modelling high scoring sports requires few inputs which are a function of time remaining. This results in a model producing a probability estimate at all stages of a match. Similarly, Klaassen and Magnus (2003) developed *TENNISPROB*, a computer algorithm which instantaneously calculates the in-game probability of either player winning based upon the current score in the match (game score, set score and match score) and the probability of player A or player B winning a point on serve. Bailey and Clarke (2008) incorporated a pre-game expected Margin of Victory (MOV), in conjunction with the Duckworth-Lewis method, to provide an updated MOV at the conclusion of each over. Similarly, Falter and Perignon (2000) provided a binary-probit model for in-game match prediction in soccer. Due to the general nature of regression models, probability estimates were only permissible at specific intervals during the match (15 minutes).

Although increasingly more research is being directed towards real-time predictions in team sports, no such literature to date takes into account the possibility of interdependence between opponent quality, current score and time remaining in the match. For example, the binary-probit model developed by Falter and Perignon (2000) does not allow for any interaction between objective pre-game variables (home advantage, team ratings) and current score. Similarly, if a pre-game favourite is expected to win by μ points, the Brownian motion model assumes the pre-game favourite will outscore the opposition by $\mu/4$ points each quarter, irrespective of current score. Although Glasson (2006) showed the errors between the bookmakers lines (μ) and score difference at each of the quarter time breaks are approximately equal to zero, this error term will later be shown to be biased since it measures the average of the errors. Therefore, another aim of this dissertation is to develop more statistically robust non-linear (S-shaped) functions for real-time prediction in AFL that provide a superior fit and account for the interdependence between score difference and difference in team quality at each of the quarter time breaks.

1.3.6 Phases of Play

Phases of play is the concept that two teams or players interact in a dynamic system, that is, in an active-reactive nature (McGarry et al., 2002). Examples of this concept may include the advantage (or disadvantage) a player has in a single point in squash in terms of their physical displacement (McGarry et al., 2002), the collective actions which lead to a goal in soccer (Grehaigne et al., 1997), or a measure to describe the performance of teams in NHL during the match (Bedford and Baglin, 2009). Borrie et al. (2002) suggested that simple frequency data is not able to capture the complex series of interrelationships between a wide variety of performance variables. Bedford and Baglin (2009) noted this and proposed that the sum of all teams adaptive winning behaviours along with their maladaptive losing behaviours could explain outcomes for NHL matches during the game. In their example, phases of play posits that teams fluctuate between periods of “high (in) phase” and “low (out of) phase”. High phase is a characteristic of winning teams and low phase is a characteristic of losing teams, with both teams being able to be in either state at any point in time. However, the authors noted that teams were typically “anti-phase stable”, that is, if one team was in high phase the other team would be in low phase and vice versa. Here “relative phase” describes the difference between the team phases.

Lames (2006), McGarry et al. (1999) and Palut and Zanone (2005) investigated the effect lateral displacement of squash and tennis players had on the outcome of points. The centre of the baseline, commonly referred to as the ‘T’, was used as a point of reference as it was seen as an advantageous position. Therefore, when players deviated significantly from this position they were deemed out of position and thus out of phase. The authors found that immediately preceding a point a disturbance (or perturbation) was typically observed. For example, a well placed shot which left the opposing player out of position for their next shot. These results suggest that players in racquet sports can fluctuate between in and out of phase, with any significant disturbance to this phase usually resulting in the conclusion of a point.

In this dissertation, the work of Bedford and Baglin (2009) is built upon by adapting the concept behind phases of play to AFL. Part of the output from this work is a plot of the phases of play which is visually enhanced in this dissertation by adding images of players guernsey when a goal is scored. This will enable viewers to not only identify when goals were scored but which players scored them. Furthermore, this procedure is automated by utilising a macro in Excel which generates the relative phase plot from raw "live-streaming" performance data. Finally, by integrating interchange data, the plot "comes to life" as it allows viewers to watch the interactive relative phase plot as if the game were live with the plot, interchange bench and scoreboard all updating in real time. This allows coaches to to objectively assess the performance of their team during the course of the game, whilst identifying which players are on the field when critical events occur (i.e. a goal is scored).

1.4 Research Questions and Publications

Consequently, in collating the array of topics encompassed in this dissertation, the following research questions will be tackled. Each research question forms the foundation of a section/chapter in this dissertation, with sections of each chapter previously being published (or accepted/in review for publication) in a peer-reviewed journal or fully refereed conference proceedings. Note that the number labels and titles of the research questions correspond to the respective chapters in the dissertation.

1.4.1 Research Questions

An outline of research questions and corresponding relevant chapters are now detailed.

Chapter 4 Home advantage

(i) How are travel and familiarity factors quantified?

(ii) Can crowd passion, that is, the breakdown of home and away supporters be accurately quantified prior to the start of the match?

(iii) When quantifying home advantage in AFL, do travel factors, familiarity factors and crowd factors exist independently of one another?

Chapter 5 Ratings

(i) Can ELO ratings, originally developed to rate chess players, be adapted to rate AFL teams?

(ii) Do profitable betting strategies exist by identifying value bets?

(iii) Is it possible to evaluate the evenness of the AFL competition from year to year?

Chapter 6 Collecting in-play betting data

(i) What is the most efficient method for collecting in-play betting data for AFL matches and how can this be implemented?

Chapter 7 In-play betting markets as a measure of expectation

(i) How can in-play betting data for AFL matches be transformed into a real-time measure of which team is going to win the match and with what level of certainty?

(ii) Does this measure have any visual appeal and how can this procedure be automated?

(iii) How accurate is this measure and are there any parallels with score difference?

Chapter 8 The efficiency of in-play betting markets

(i) Do in-play betting markets in AFL satisfy the Efficient Market Hypothesis (*EMH*) criterion?

(ii) Are profits attainable by betting on teams with certain characteristics?

Chapter 9 Intra-match home advantage

(i) Does home advantage in AFL occur at different stages of the match?

(ii) What role do pre-game characteristics (favourite/underdog) and in-game characteristics (ahead/behind) play on home advantage during the game?

Chapter 10 In-play predictions

(i) Is there any interaction between score difference and difference in team quality during an AFL match?

(ii) Does accounting for this interdependence, if any, increase the reliability of the predictions?

(iii) Do profitable betting strategies exist by identifying value bets?

(iv) Are the betting results consistent for specific in-game intervals?

Chapter 11 Phases of Play

(i) Is it possible to present statistical predictions in AFL that are simultaneously representative of a team's likelihood of winning and graphically simple enough to be widely interpretable?

(ii) How can this procedure be automated and implemented?

(iii) How can interchange data be integrated to provide an objective measure of individual player performance?

1.4.2 Publications

Chapter 4 Home Advantage

Ryall, R. and Bedford, A. (2011). Independent effects that augment home ground advantage. *Journal of Sports Sciences.* Manuscript in review.

Chapter 5 Ratings

Ryall, R. and Bedford, A. (2010). An optimized ratings-based model to forecast Australian Rules football. *International Journal of Forecasting*, 26(3):511-517.

Chapter 6 Collecting In-Play Betting Data

Ryall, R. and Bedford, A. (2009). An automated approach to compare in-the-run markets with score in evaluation of team performance. In Lyons, K., Baca, A., and Lebedew, A., editors, *Seventh International Symposium on Computer Science in Sport*, pages 155-162.

Chapter 7 In-Play Betting Data as a Measure of Expectation

Ryall, R. and Bedford, A. (2009). An automated approach to compare in-the-run markets with score in evaluation of team performance. In Lyons, K., Baca, A., and Lebedew, A., editors, *Seventh International Symposium on Computer Science in Sport*, pages 155-162.

Chapter 8 The Efficiency of In-Play Betting Markets

Ryall, R. and Bedford, A. (2010). The efficiency of in-play betting markets in Australian Rules football. *International Journal of Sports Finance*, 5(3):193-207.

Chapter 9 Intra-Match Home Advantage

Ryall, R. and Bedford, B. (2011). The intra-match home advantage in Australian Rules football. *Journal of Quantitative Analysis in Sports.* Manuscript accepted for publication 27th January 2011.

Chapter 10 In-Play Predictions

Ryall, R. and Bedford, A. (2010). Fitting probability distributions to real-time AFL data for match prediction. In Bedford, A. and Ovens, M., editors, *Tenth Australasian Conference on Mathematics and Computers in Sport*, pages 121-128.

Chapter 11 Phases of Play

Ryall, R. and Bedford, B. (2008). An algorithm to plot an AFL teams performance in real-time using interactive phases of play. In Hammond, J., editor, *Ninth Australasian Conference on Mathematics and Computers in Sport*, pages 108-114.

Chapter 2

Australian Rules Football

This chapter details the many facets of Australian Rules football. In Section 2.1, the history of the game is described in chronological order. Section 2.2 details the current teams, field and playing positions, scoring system, objectives and rules, the fixture, and the ladder which is used to determine which teams play in the finals series at the conclusion of the regular season. Section 2.3 explains how players are currently recruited and how this has changed over time. Section 2.4 discusses the major providers of Australian Rules football statistics. The importance of each section will be realised in the latter chapters of this dissertation. To avoid confusion with the world game of football, which is typically referred to as soccer in Australia, Australian Rules Football will be referred to as AFL for the remainder of this dissertation, except where otherwise stated. Please note that full definitions of the colloquial terms used in this chapter are provided in the glossary. Those that are familiar with AFL should proceed to the following chapter.

2.1 History

This section details the history of Australian Rules football in chronological order, including its origin, evolution and development. This will detail aspects of interstate rivalry, tribalism and changes. The information provided in this chapter including any direct quotes was summarised from Hess et al. (2008).

2.1.1 Pre 1890's

The year 1859 is often referred to as being the landmark of Australian Rules football. However, football in one form or another, existed well before this period and was often recognized as an "amusement of the military". The Melbourne Football Club was the first established club occurring in May 1859. A committee meeting comprising of Tom Willis, William Hammersly and Thomas Smith took place in the afternoon of 17 May 1859. This meeting is recognized as arguably the most significant meeting in the history of Australian Sport. The meeting resulted in the document "Rules of the Melbourne Football Club, May 1859" which comprised ten simple rules which resulted in a game that was remarkably adaptable and relatively easy to understand for newcomers. The lack of an offside rule as well as goal-behind scoring both come from Sheffield Rules football that competed with FA rules until the last Sheffield teams joined the English FA in the 1870's to make Association Football the dominant non-running code. Under Sheffield, there was no offside rule (other than not playing behind the goal keeper). Also, there were four vertical posts as with AFL which created behind posts in 1866. Under Sheffield rules there was a cross bar forming three boxes with the four poles. A score in the middle box was called a goal while a shot into either outer box was called a "rouge". Only goals counted but if the goals were tied, the most "rouges" won. Of course AFL preserves that scoring with goals and behinds today. In the following years, several new teams were formed and competed against one another. However, there was no league structure, especially in terms of who or when teams were to play.

In 1877 the Victorian Football Association (VFA) was established. Foundation senior clubs included Albert Park, Carlton, East Melbourne, Essendon, Hotham, Melbourne, St. Kilda and West Melbourne. While foundation junior clubs included Ballarat, Hawthorn, Northcote, South Melbourne, Standard, Victoria United, Victorian Railways and Williamstown. The VFA wasn't the first established football association, in fact the South Australian Football Association (SAFA) later renamed the South Australian National Football League (SANFL) was established earlier that year in 1877. The SANFL is not only the oldest surviving football league of any kind in Australia, but also one of the oldest football leagues world wide. The foundation clubs included Adelaide, Bankers, Kensington, South Park, Victorians, Willunga, Woodville, Kapunda and Gawler. The West Australian Football Association (WAFA) was established several decades later (1885) and was later renamed the West Australian Football League (WAFL). Foundation clubs included Fremantle, Rovers and the Victorians.

Australian Rules football was (and remains) the dominant sport in the southern states, which includes Victoria, South Australia, West Australia and Tasmania. However, in New South Wales and Queensland it was only a minor code as Rugby League and Rugby Union dominated these northern states. An Australian historian named Ian Turner labeled this divide as the "Barassi-line" after Ron Barassi, a legend of Australian Rules football. Figure 2.1 shows the Barassi-line where the States to the right of the line are dominated by rugby league and union, and the States to the left were dominated by Australian Rules football. Although the States to the right of the "Barassi-line" occupied approximately 25% of Australia's land mass, it supported more than 50% of Australia's population.

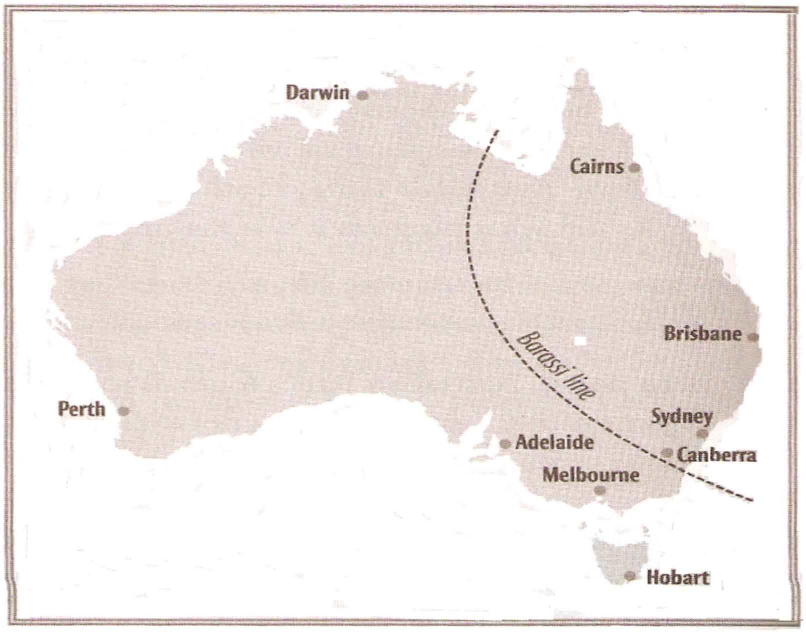


Figure 2.1: The Barassi-line

2.1.2 1890's to 1910's

The Australian Football Council (AFC) was established in 1906 and comprised delegates from the New South Wales Football League, Queensland Football League, South Australian Football League, Tasmanian Football League, Victorian Football League and West Australian Football League. The principal aim of the AFC was to promote the "Australasian game of football".

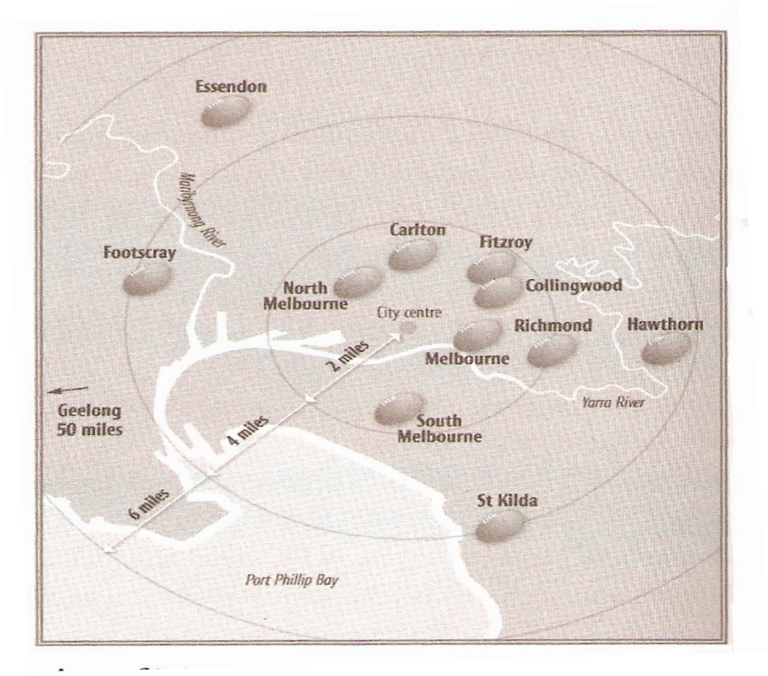


Figure 2.2: Location of Victorian Football League clubs, 1925

The Victorian Football League (VFL) was established in 1896 when several clubs broke away from the Victorian Football Association (VFA). The VFL was initially an eight team competition comprising Carlton, Collingwood, Essendon, Fitzroy, Geelong, Melbourne, St Kilda and South Melbourne. In 1908, the competition increased to 12 teams with the introduction of Richmond and University. However, University only lasted six seasons eventually disbanded at the end of the 1914 season. The location of the Victorian Football League teams in 1925 is revealed in Figure 2.2.

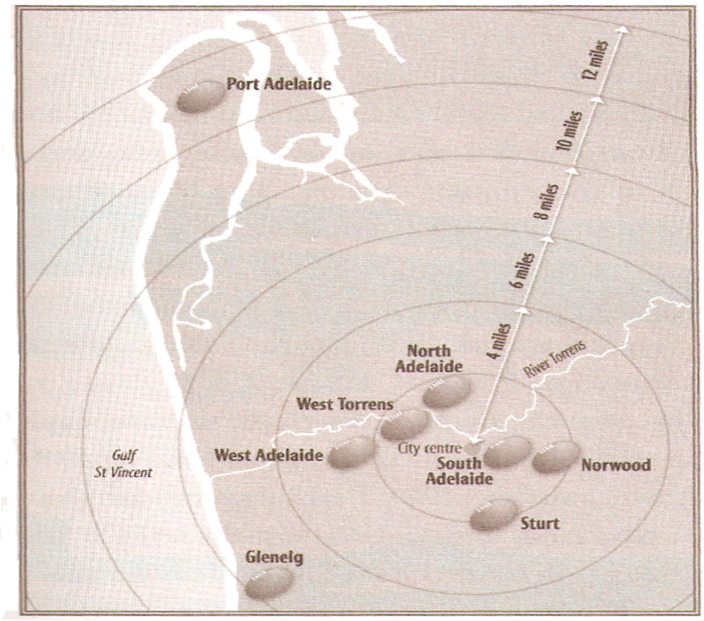


Figure 2.3: Location of South Australian Football League clubs, 1925

The South Australian National Football League (SANFL) during this period was an eight team competition which included Norwood, Port Adelaide, West Torrens, North Adelaide, West Adelaide, Sturt, South Adelaide and Glenelg. The location of the South Australian Football League teams in 1925 is revealed in Figure 2.3.

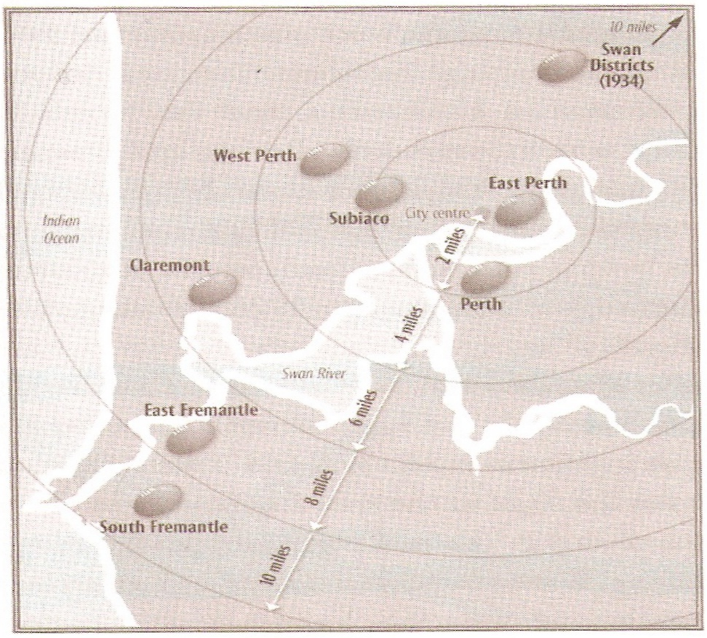


Figure 2.4: Location of West Australian Football League clubs, 1925

The West Australian Football League (WAFL) during this period was a six team competition comprising Perth, East Perth, West Perth, East Fremantle, South Fremantle and Subiaco. In 1926, the competition increased to a seven team competition with the induction of Claremont. The location of the West Australian Football League teams in 1925 is revealed in Figure 2.4.

2.1.3 1920's to 1940's

A national agenda was pursued throughout the 1920's by the Australian Football Council (AFC). This was somewhat difficult given the football divide. Interstate carnivals were thought to be the best way of showcasing the game to the country and were held every three years. All States and Territories participated in the carnival with Victoria, South

Australia, Western Australia and Tasmania in section 1 and Queensland, New South Wales, Northern Territory and Australian Capital Territory in section 2. The carnival was rotated throughout the nation and was dominated by Victoria throughout its existence.

In 1925, the VFL expanded to a twelve team competition, admitting Footscray, Hawthorn and North Melbourne, all of which were distinguished teams in the VFA. Between 1927 and 1930, Collingwood won four consecutive premierships and became the first team to finish a season without losing a game, with both outcomes remaining as records that still stand today. During this period of sustained success they possessed the largest supporter base in the country. Collingwood became known as "The Machine" due to the systematic way they played the game. Although they competed in five grand finals in the 1930's they were later overtaken by South Melbourne as they recruited the best talent from across the nation and played in four consecutive grand finals between 1933 and 1936. In the SANFL, Port Adelaide was known as the team to beat. However, throughout the 1930's six of the eight teams won a premiership demonstrating the evenness of the SANFL competition. Successful teams in the WAFL fluctuated considerably during the 1930's with East Fremantle (four premierships), West Perth (three), and Subiaco (three) eventually being overtaken by Claremont who proceeded to play in five consecutive grand finals between 1936 and 1940.

2.1.4 1950's and 1960's

In 1956, the first television was introduced into Australia during a period of sustained economic and national growth. This coincided with the 1956 Olympic games which were held in Melbourne, and it was noted that this was likely to stimulate the purchase of television sets. The number of television license holders increased exponentially in its first few years of operation before stabilizing during the mid 1960's. Australian Rules football, cricket and tennis all attracted significant television interest, so much so, that radio broadcasters were poached by TV stations with the promise of more pay and greater media exposure.

Although the VFL were keen to promote their game via television, they were cautious, as they had concerns that direct telecast would be to the detriment of match day attendances. Therefore, from 1957 to 1960 only the final 30 minutes of VFL matches were telecast live to Melbourne households. However, in late 1960 it was concluded that live broadcasts were not in the best interests of the game and a ban was placed on live broadcasts. This ban lasted until 1977, when the grand final between Melbourne and Collingwood was telecast live to a nation-wide audience.

During the 1950's the VFL became the wealthiest sports league in the nation through ever growing public interest, media coverage and large attendances at matches. However, it was still only a semi-professional game nationwide at the top level. Throughout the 1960's match payments were set at approximately £6 under the Coulter Law which translated to approximately one-third of average weekly earnings. However, it was common knowledge that the richer clubs like Carlton, Colllingwood and Essendon were paying their players significantly more than the less financially stable clubs like Fitzroy, North Melbourne and South Melbourne. Notably, the Coulter Law was eventually scrapped as most players were put on contracts.

In 1967, a zoning system was established in the VFL to ensure fairness and equity for the recruitment of players from country Victoria. The state of Victoria was split into twelve sections and each club was allocated a specific section. Although the zones were randomly allocated and the VFL had the intention of rotating the zones every couple of seasons, this did not happen and allowed some teams to prosper (Essendon and Geelong) while other teams perished (South Melbourne and Fitzroy).

West of the Victorian border, the SANFL was dominated by Port Adelaide in the 1950's winning six consecutive premierships between 1954 to 1959. This overwhelming dominance by Port Adelaide continued until the mid 1960's when Sturt played in six consecutive grand finals between 1965 to 1970, winning five. In the WAFL, South Fremantle not only dominated the state competition winning six premierships between 1947 to 1954, but also defeating many visiting teams from Victoria and South Australia.

2.1.5 1970's

The 1970's marked the beginning of the commercialisation of sport worldwide. For example, in the United States, Pete Rozelle the National Football League commissioner signed an unprecedented four year TV rights deal worth US$656 million. Corporate sponsorship in the VFL began in 1968 when the cigarette-manufacturing company W.D. & H.O. Willis (Amatil) provided prize money to the four finalists. In the late 1970's Marlboro (Philip Morris) and Escort (Amatil/Wills) provided $165,000 and $375,000 respectively towards sponsorship of the VFL. North Melbourne was at the forefront of the commercialisation of football by diversifying its business interests which were rumored to include pubs and discos. The revenue the VFL received from TV rights was approximately $200,000 annually, however this grew substantially over time. By the end of the decade the league received $600,000 for all home and away matches, $120,000 for the live telecast of the grand final and $200,000 for the live telecast of all night matches.

During the 1970's it was clear that the VFL, WAFL and SANFL were in a league of their own in regards to player talent. Throughout the 1970's, unlike the WAFL and SANFL, the VFL competition was dominated by a handful of teams. Hawthorn and Carlton each won three premierships, while North Melbourne and Richmond won two each. The decade began the same way it finished with Carlton defeating Collingwood in memorable Grand Finals. In 1970, the VFL's own stadium Waverly Park was opened, with the inaugural match being played between Geelong and Fitzroy. The need for Waverly Park was prompted by all other grounds being owned and therefore controlled by local municipalities, or in the case of the MCG a board of trustees nominated by the Victorian state government. Waverly Park was originally going to hold 157,000 spectators all of whom would have an unrestricted view of the game. However, the eventual capacity was 100,000, and crowds rarely reached this due primarily to its location (20 kilometres from the centre of Melbourne), and that the MCG was the home of Victorian football. Although, Waverly Park (or VFL Park as it became known) was becoming increasingly popular, the majority of finals matches were still played at the MCG. In 1977, the league installed lighting at a cost of more than $1 million which not only allowed VFL Park to host night football, but also concerts and the World Series Cricket (WSC).

The SANFL followed suit and announced that it was completing arrangements for a stadium "Football Park", with a capacity to hold 70,000 spectators, to be opened at West Lakes. The inaugural match was held between Central District and North Adelaide on the 4th of May 1974. Football Park was a huge success and drew 450,000 spectators to 25 matches, attendances similar to Victorian crowds. It also gave the SANFL control over its own destiny.

In Western Australia, unlike its Victorian and South Australian counterparts, the WAFL did not have its own stadium. Subiaco Oval was the home of WAFL football, however the league leased the ground and had to share with the Subiaco Football Club. Incidentally, Western Australia achieved financial security of Subiaco Oval in 1991 when it signed a 99 year lease.

Several rule changes were also implemented during this period in order to increase the attractiveness of the game. These changes included: (1) A final five system; (2) a centre square (initially a diamond) of which a limit of four players per team are permitted within at the centre bounce; (3) two-umpire system, since one umpire was unable to keep up with the speed of the game and (4) an interchange system was introduced in 1978 to allow players to be interchanged as frequently as required rather than permanently replaced.

Although the VFL, WAFL and SANFL revenue increased significantly throughout the 1970's this had come at a cost, as greater power had been given to corporate sponsors, television stations and players through their player associations. The commercialization of football was seen as a double edged sword, as it created as many threats as opportunities.

2.1.6 1980’s

The 1980’s began with South Melbourne, one of the VFL’s foundation clubs, relocating in 1982 to Sydney to avoid financial ruin. The South Melbourne football club had struggled both on and off the field for many years. Between 1946 to 1981 the club had an atrocious on-field run, reaching the finals only twice (in the 1970’s). Prior to its demise in 1982 it had been reported that the club had an operating loss of $150,000 for the previous five seasons. The club made a proposal to the VFL seeking permission to play all 11 away games in Sydney and all 11 home games at VFL Park, effectively giving South Melbourne 22 home games. This proposal was latter amended and a revised proposal to play all its home games in Sydney was accepted. However, this revised proposal marked the beginning of the South Melbourne relocation saga. The Keep South At South (KSAS) group was formed which strongly opposed the relocation. The group took legal action and gained control of the club at an extraordinary meeting on 22 September 1981. The club plunged further into crises as the VFL directors refused to revoke their decision and allow South Melbourne to play 11 home games in Sydney. The players went on strike in early November and would not budge on their demands, particularly wanting the board to commit to a long term future in Sydney. Eventually the KSAS board of managment resigned and Bill Collins was declared President. The clubs fortunes did not improve after the relocation, failing to reach finals in the first three years of the competition.

In early 1985, medical entrepreneur and millionaire Dr. Geoffrey Edelsten lodged a proposal with the VFL to buy the Sydney Swans. Although there was competition from other bidders, Edelsten was granted the license for $6.5 million. However, Edelsten later revealed that he was unable to make all the required payments and was interested in acquiring financial interest from investors. The club went on a spending spree to buy their way to success, it was reported they spent $2 million on players during the 1986 season, almost twice the League’s newly introduced salary cap. This strategy reaped immediate rewards with the club finishing fourth in 1986 and 1987. During this same period the Swans were branded a “showbiz” team since they had the flamboyant tight shorted full-forward Warwick Capper, cheerleaders and the antics of Edelsten and his wife Leanne. However, the party did not last long due largely to the stock market crash in October 1987, which resulted in financial ruin for the club owner Edelsten. The club was forced to sell many of their players to pay off their debts and the club was eventually sold back to the VFL for a measly $10 in the middle of 1988.

During the South Melbourne relocation debacle during the 1980’s, another problem was created in relation to player transfer rules. Silvio Foschini, a rover for South Melbourne requested a transfer to St Kilda as he wanted to remain in Melbourne and not relocate permanently to Sydney. This request was refused by South Melbourne and Foschini instigated legal action which was heard by the Victorian Supreme Court. On 15 April 1983, Justice Crockett stated that “the VFL rules and regulations were in restraint of trade” and thus Foschini was entitled to play with St Kilda. This ruling resulted in the VFL instigating new rules and regulations to aid in the equitable interests of players, teams and the league. A salary cap was also implemented in 1985 in an attempt to provide a ceiling of total player payments for each team in order to ensure the competition remained viable. The WAFL quickly followed the VFL’s initiative and also introduced a salary cap that same year.

In late 1980, the East Perth Football Club submitted an application to join the VFL. Although the application was eventually rejected, it raised the question as to the longevity of the WAFL, as many clubs were under severe financial pressure, not to mention the talent drain to Victoria. Similarly, in 1981, the SANFL directors would seek to “make a formal submission and application to the VFL for the entry of a SANFL corporately managed team(s) in the VFL competition at the earliest possible time”. Yet again an application from another state league to enter the VFL was rejected. However, in 1986 the SANFL received an invitation to take up a license and enter an extended VFL for the 1987 season. In late 1986, a meeting was held with the VFL club presidents who voted unanimously to award a licence to a private Brisbane corporation. However, the presidents were split on a West Australian team entering the competition as many of the financially solvent clubs saw Western Australia as a valuable recruiting zone. The West Australian team eventually received the go ahead from the club presidents. Both the Queensland and West Australian team agreed to pay the $4 million licence fee which was divided equally by the current VFL clubs, many of whom used this money to relieve their debt.

The newly formed teams experienced financial difficulties almost immediately. The infamous Christopher Skase who was a property developer at the time, was the major shareholder of the Brisbane team. However, within three years the club had accumulated $27 million in debt, not to mention Skase's Quintex corporation collapsing after the sharemarket crash in late 1987. The Brisbane team was eventually sold to Reuben Pelerman who proceeded to lose $4 million in two years and was forced to relinquish ownership. Brisbane's on-field performance was dismal, failing to finish higher than tenth until 1995. Indian Pacific Limited (IPL) was a public company which was created in order to raise funds for the newly formed West Coast team in Western Australia. When IPL was floated it was a disaster as many supporters were foreign to the idea with only a handful of wealthy investors making up the shortfall. Eventually the West Australian Football Commission reclaimed ownership.

At the end of the 1989 season the Footscray Football Club was in dire straits, having an accumulated debt of approximately $1.5 million. The club struggled to attract supporters and sponsors, and its home ground the Western Oval was in incredibly poor condition. There were rumours that the Footscray and Fitzroy presidents had agreed to a merger. The Save The Bulldogs (STB) campaign gathered significant momentum including a Supreme Court injunction which temporally stopped the merger. Footscray were told they needed an income of $5 million for the 1990 season which was double what it earned in the previous season. A fund was created to prevent the merger which was bolstered by football supporters, local industries and the state government. The VFL later announced it had abandoned the merger and the Footscray Football Club was free to play in the 1990 season and beyond.

2.1.7 1990's

It wasn't until 1990 that the VFL was renamed the Australian Football League (AFL) to reflect that it was now indeed a national competition. The 1990 premiership was won by Collingwood, its first in over 30 years after reaching the grand final eight times during that same period.

Since the early 1980's the SANFL (South Australian National Football League) had toyed with the idea of entering a team in the VFL. This later came to fruition in 1990 when the SANFL advised the AFL that it was considering entering a team for the 1993 season. However, this was conditional on not having to pay a license fee and there being no more than fourteen teams in the competition. The SANFL was dominated by Port Adelaide winning 30 premierships between 1877 and 1990. It's closest rival was Norwood who had won the majority of its premierships prior to World War II. In 1990, Port Adelaide attempted to go it alone and enter the 1990 AFL season without the knowledge or consent of the SANFL. According to the SANFL president Max Basheer, Port Adelaide's action divided the SANFL community and caused "emotions to run high". It was widely thought that Port Adelaide leaving the SANFL would cause irreversible damage to the state competition, since AFL revenue would be distributed to Port Adelaide and not the SANFL. The AFL eventually opted for the team managed by the SANFL in 1990. Notably, as the Adelaide Crows entered the national competition in 1991 the total number of home and away attendances in the SANFL dropped by almost a third. The finals system was also restructured in 1991 with the AFL opting for a finals six rather than a finals five in order to increase revenue and accommodate the extra clubs.

The entry of the Adelaide Crows created a 15 team competition, which was less than satisfactory, since byes needed to be scheduled, which was seen by the AFL administration as revenue lost. The dominance of the West Coast Eagles in the early 1990's created discussion of an additional team in Western Australia. In 1994, the Fremantle football club was formed and it entered the national competition in 1995. Akin to the SANFL, the introduction of two teams to the national competition resulted in a significant decrease in WAFL attendances. However, the combined attendance of WAFL and AFL matches in Western Australia was at an all time record. The introduction of the Fremantle Dockers also created a great rivalry between the two West Australian teams, Fremantle and West Coast.

In 1996, the VFL/AFL celebrated its centenary. Incidentally, the centenary year marked the end of the Fitzroy football club, one of the founding members of the VFA. Fitzroy had an extremely poor on field record in the VFL/AFL which was mainly attributed to its financial instability and declining membership base. In the 30 years prior to 1996, the Fitzroy football club changed the location of its home ground on many occasions in order to develop a strong social club and improve its declining and ageing membership. In 1996, it was widely thought that North Melbourne and Fitzroy were going to merge. However, many of the AFL clubs thought this would create a "super club" since North Melbourne had lost only three games that season and went on to win the premiership. On the 4th July 1996, the AFL presidents voted against the merger by fourteen votes to one. Another merger deal was then proposed by Brisbane which included a playing list of forty four players (eight Fitzroy players) and a minimum of six games in Melbourne. The AFL presidents voted in favour of the merger and Brisbane Lions were born.

The cash strapped Hawthorn Football Club also initiated merger talks with Melbourne, encouraged by the AFL Commission's $6 million incentive package. Several former players and supporters of both clubs reacted angrily to the merger discussions. Melbourne members were in favour of the proposed merge, however it did not proceed as Hawthorn members voted strongly against the proposal. This led the AFL commission to eventually withdraw the $6 million merger incentive.

In 1996, Port Adelaide won its third consecutive premiership in the SANFL. The following year Port Adelaide were admitted to the AFL seven years after its initial bid. Ironically in 1997, Port Adelaide's first year in the competition, the premiership was won by Adelaide. The next year Adelaide achieved back-to-back premierships, with Andrew

McLeod receiving the “Norm Smith Medal” awarded to best player in the Grand Final, in both Grand Finals. Akin to the two West Australian teams, a rivalry developed between Adelaide and Port Adelaide for many years to come. In 1994, the finals system was also changed to a final eight instead of a final six.

In 1997, Footscray and North Melbourne changed their name to the Western Bulldogs and Kangaroos respectively in order to broaden the clubs appeal. Table 2.1 shows that although club memberships increased significantly from 1998 to 2008, teams that languished at the bottom of total club memberships in 1998 were still at the bottom of the ladder in 2008 excluding a few anomalies.

	1998		2003		2008	
	Rank	Members	Rank	Members	Rank	Members
Adelaide	1	41,985	1	47,097	1	48,720
Brisbane	16	16,108	10	24,365	16	22,737
Carlton	9	25,402	5	33,525	7	39,360
Collingwood	7	27,099	2	40,455	4	42,498
Essendon	6	27,099	6	31,970	5	41,947
Fremantle	11	22,186	8	25,368	3	43,366
Geelong	14	19,971	11	24,017	8	36,850
Hawthorn	5	27,649	7	31,500	6	41,436
Kangaroos	12	20,196	13	21,403	13	29,619
Melbourne	15	17,870	16	20,555	10	32,600
Port Adelaide	2	38,305	4	35,425	9	34,185
Richmond	8	27,092	9	25,101	11	30,820
St Kilda	10	23,204	12	23,626	12	30,063
Sydney	4	31,089	14	21,270	15	26,721
West Coast	3	37,496	3	36,234	2	44,863
Western Bulldogs	13	20,064	15	21,260	14	28,306

Table 2.1: Club membership figures and rank, 1998, 2003 and 2008

2.1.8 2000's

The first decade of the new millennium started with Essendon embarking on a winning streak of twenty consecutive matches, eventually going on to defeat Melbourne in the Grand Final to win the premiership. It was a very successful year for Essendon also winning the Pre-season Cup and James Hird the Essendon Captain winning the "Norm Smith" and

"Brownlow" medal that year. The first match was also held at Docklands stadium in round 1 with Essendon defeating Port Adelaide by 94 points. This stadium was known for its retractable roof, the first of its kind in the AFL. Essendon was coached by the infamous Kevin Sheedy who was known for his publicity stunts and his promoting of AFL to a far and wide audience. Notably, 2007 was the end of an era of the Essendon football club as James Hird (253 AFL games) retired and Kevin Sheedy departed after an illustrious playing coaching career spanning over 850 senior matches.

Prior to 2007, the competition was dominated by Non-Victorian clubs. The Brisbane Lions established themselves as one of the greatest teams of all time winning three consecutive premierships in 2001, 2002 and 2003. They also made the Grand Final in 2004 succumbing to Port Adelaide. Between 2005 and 2007 a great rivalry was established between West Coast and Sydney with all of their six matches during this period being decided by four points or less. Notably, Sydney defeated West Coast in the Grand Final in 2005 by four points to win their first premiership since 1933 prior to relocating from South Melbourne in 1982. Leo Barry (Sydney) received great media attention after the victory due to a contested mark he took in the dying seconds of the match which prevented West Coast an opportunity to win the match. This mark is showcased in Figure 2.5. Then in 2006 West Coast returned the favour, defeating Sydney by one point in the Grand Final. More recently the competition has been dominated by the dynasty which is the Geelong football club, making three consecutive Grand Finals (2007 to 2009) winning premierships in 2007 and 2009. Ironically, 2008 was their most successful home and away season losing only one match to Collingwood. Yet in 2009 they defeated St Kilda in the Grand Final who matched Essendon's record of 20 consecutive matches during the home and away season.

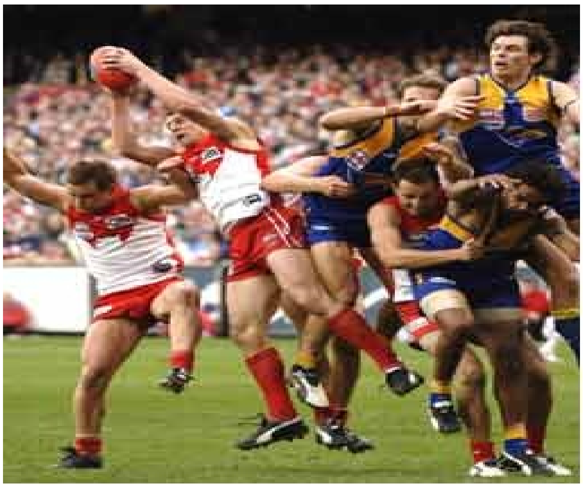


Figure 2.5: The infamous Leo Barry contested mark in the dying seconds of the 2005 AFL Grand Final

In December 2007, there were discussions with the Kangaroos who were under several financial pressure to relocate to the Gold Coast. This offer was eventually rejected by the Kangaroos and expansion plans for two new teams entering the competition were then discussed at a meeting in early 2008. The AFL awarded a license to Gold Coast (*GC17*) and Greater Western Sydney (*GWS*) to enter the competition in 2011 and 2012 respectively. There has been much controversy over the expansion plans, as both new teams receive a greater salary cap in their first few years in the competition in order to entice players from other clubs. Both *GC17* and *GWS* also receive a horde of priority draft selections in the upcoming drafts, as the AFL wanted both teams to have immediate success. However, this was likely to have dire consequences on poorly performing teams such as Richmond who would no longer receive the best young talent. Furthermore, they could also lose their current young stars as the new teams could afford to pay them considerably more.

Both *GC17* and *GWS* have gone to great lengths to promote their clubs in their respective locations. In 2009, *GC17* announced the recruitment of Karmichael Hunt who was lured away from his Rugby League team the Brisbane Broncos. Since much of Queensland

and New South Wales is dominated by Rugby League, this was seen as an experiment which could attract Rugby followers to AFL. In 2010, *GWS* followed suit and acquired Israel Falou who also defected from Rugby League (Brisbane Broncos). The recruitment of both players has created much controversy as both are estimated to earn approximately one million dollars per season, with both players having minimal AFL experience. *GWS* also employed Kevin Sheedy as their inaugural coach, who is seen as a great publicity magnet as he is well known for his marketing of AFL.

Several new rules were also integrated in the 2006 season in order to increase the speed of the game and reduce congestion. The number of interchanges increased significantly over this period, with teams averaging 23 in 2003 and exceeding 100 in 2010. By increasing the number of interchanges clubs are able to maximise the physical output of their players by giving them short periods of rest. There is now the suggestion that the game is becoming too quick and as a result a number of soft tissue injuries (i.e. hamstrings) are occurring more frequently. Therefore, there is an indication the AFL will limit the number of interchanges in subsequent seasons.

2.2 The Game

This section discusses the many facets of AFL to provide a better background on the game itself. Topics of interest include the current teams as of season 2010, field and player positions, scoring system, objectives and rules, fixture, and the ladder which is used to determine which teams play in the finals series at the conclusion of the regular season.

2.2.1 Teams

The Australian Football League currently consists of 16 teams (as of season 2010). There are currently 10 teams based in Victoria, two in each of Western Australia and South Australia and one in New South Wales and Queensland. Although teams are not represented in the other States and Territories, games are occasionally played there to increase the popularity and exposure of AFL. Table 2.2 displays some background information on these clubs. Details include an image of the teams match day jumper guernseys, team name, team nickname, location, training ground, VFL/AFL debut and total number of VFL/AFL premierships. This information summarizes the clubs historical background including their relative success in the VFL/AFL.

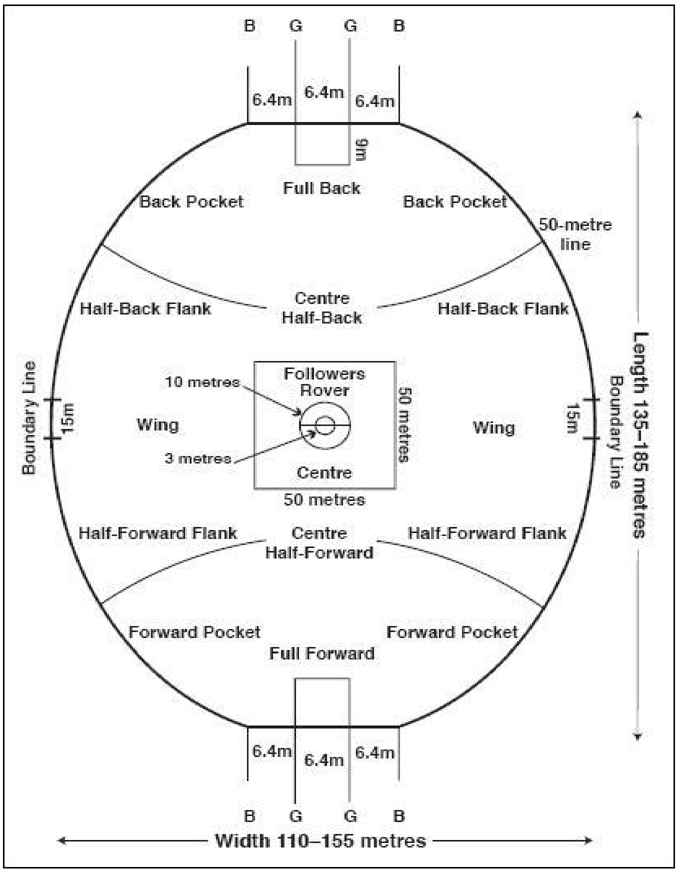
Jumper	Club	Nickname	Location	Training Ground	Debut*	Premierships*
	Adelaide	Crows	Adelaide, South Australia	Football Park	1991	2
	Brisbane Lions	Lions	Brisbane, Queensland	Gabba	1997	3
	Carlton	Blues	Melbourne, Victory Park	Princes Park	1897	16
	Collingwood	Magpies	Melbourne, Victoria	Westpac Centre	1897	14
	Essendon	Bombers	Melbourne, Victoria	Windy Hill	1897	3
	Fremantle	Dockers	Fremantle, Western Australia	Subiaco	1995	0
	Geelong	Cats	Geelong, Victoria	Kardinia Park	1897	8
	Hawthorn	Hawks	Melbourne, Victoria	Waverly Park	1925	10
	Kangaroos	Kangaroos	Melbourne, Victoria	Arden Street Oval	1925	4
	Melbourne	Demons	Melbourne, Victoria	Melbourne Rectangular Stadium	1897	12
	Port Adelaide	Power	Adelaide, South Australia	Alberton Oval	1997	1
	Richmond	Tigers	Melbourne, Victoria	Punt Road Oval	1908	10
	St Kilda	Saints	Melbourne, Victoria	Moorabbin Oval	1897	1
	Brisbane Lions	Lions	Brisbane, Queensland	Gabba	1997	3
	West Coast	Eagles	Perth, Western Australia	Subiaco	1987	3
	Western Bulldogs	Bulldogs	Melbourne, Victoria	Whitten Oval	1954	1

*VFL/AFL

Table 2.2: Background and history of the 16 currently listed AFL clubs, 2010

2.2.2 Field and Player Positions

In a regular game of AFL there are two teams of 22 players, of which only 18 are permitted on the field at any one time, with the remaining four players on the interchange bench. Players are then rotated on and off the bench at the coaches request for many reasons, including rotating "fresh legs", and players coming off the field due to coach instigated disciplinary actions (e.g. conceding a 50-metre penalty). Each team historically comprises three backs (2 × Back Pocket, 1 × Full Back), three half-backs (2 × Half-Back Flank, 1 × Centre-Half Back), three midfielders (2 × Wing. 1 × Centre), three followers/rovers (2 × followers, 1 × rover), three half-forwards (2 × Half-Forward Flank, 1 × Centre-Half Forward), three forwards (2 × Forward Flank, 1 × Full Forward) and four players on the bench. Note that there are no restrictions on where players can move which can often result in matches that are free flowing or more recently very congested. Furthermore, there are currently no limits on the numbers of interchanges permissible during a match. Figure 2.6 displays the AFL field and typical player positions. The ball used in AFL matches, which is also known as a sherrin, is made from leather and is showcased in Figure 2.7.



Source. "Laws of Australian Football 2009."

Figure 2.6: AFL playing field and playing positions

2.2.3 Scoring System

Points are scored in several ways including a goal worth six points and a behind worth one point. A goal is awarded to the attacking team when the football is kicked completely over the goal line, regardless whether or not it bounces, provided it has not touched an opposition player in any way. A behind is awarded to the attacking team when the football passes over the behind line; or the football strikes any part of the goal post; or prior to the football passing over the behind or goal line it is touched by another player; the defending

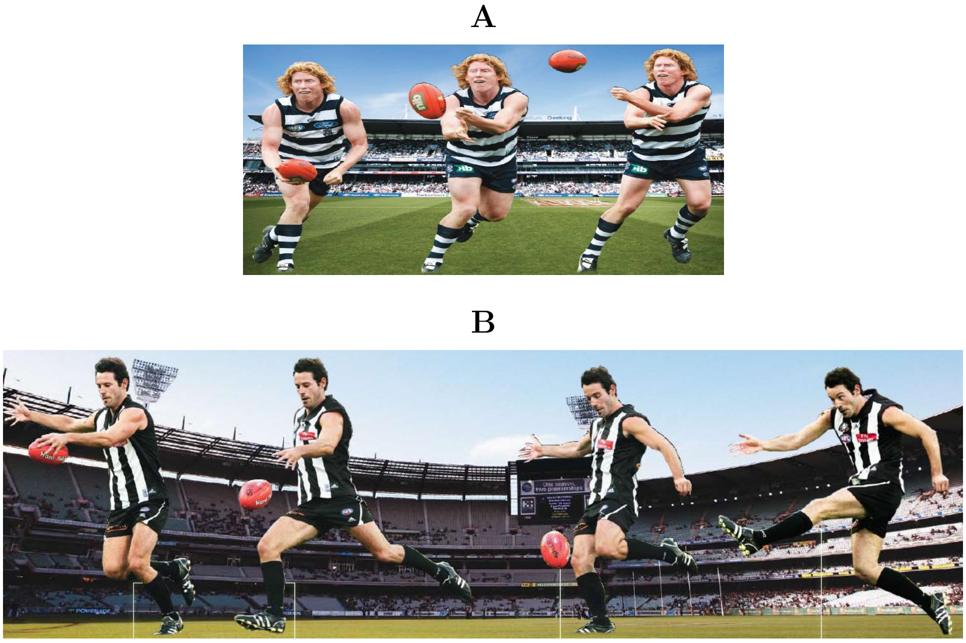
Figure 2.7: The ball ("Sherrin") used in AFL

team that deliberately plays the ball over the behind or goal line concedes a rushed behind. A typical score in a game might be 16 goals, 16 behinds, 112 points to 14 goals 8 behinds, 92 points, for a final winning margin of 20 points. Due to the high scoring nature of the game, draws are rare, occurring approximately once every 125 matches.

2.2.4 Objectives and Rules

Each game consists of four 20 minute quarters plus approximately 10 minutes of extra time (time on) per quarter. Time keepers stop the clock and call time on when the goal-umpire signals a goal or a behind has been scored; the boundary-umpire signals the ball is out of bounds or out of bounds on the full; the field umpire crosses their arms and indicates they are going to perform a ball up; the field umpire signals to do so for another reason such as the blood rule.

The primary objective of each team is to outscore their direct opposition. Typically this is achieved by scoring as many goals as possible while minimising the number of scoring opportunities for the opposition. The team which is ahead at the final siren wins the match. The ball can be moved in several ways namely via a kick, handball or run and bounce, failure to dispose of the football correctly results in a free kick to the opposition. Examples of an incorrect disposal include throwing and dropping the ball. Figure 2.8 shows an example of the techniques involved in a kick and a handball.

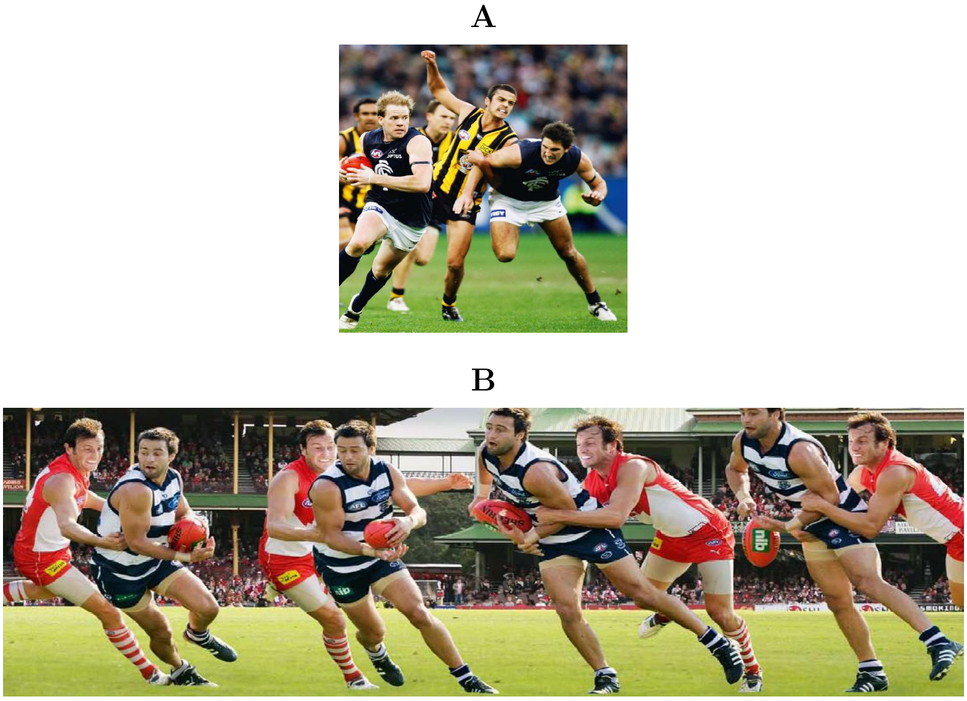


Source: 'AFL Record' Part 2 of a series on how to play the game - Handball;
'AFL Record' Part 1 of a series on how to play the game - Kicking

Figure 2.8: Disposals. (A) Handball. (B) Kick.

AFL is considered a contact sport, with players allowed to tackle and shepherd opposition players. A tackle is where a player uses their arms to prevent an opposition player from disposing of the football correctly. This is typically achieved by pinning both arms of the opponent to their body and not allowing the football to easily spill out. A tackle must make contact below the shoulders and above the knees and can only be performed when the opposition player has possession of the football. If a tackle is performed correctly, and the player being tackled had a reasonable amount of time to dispose of the ball, a free kick is awarded to the player who performed the tackle. A shepherd on the other hand, is a push, bump or block on an opposition player who is within a five metre radius of the football. The idea behind this manoeuvre is it allows players from the attacking team space to run into and prevents players from the defending team an opportunity to tackle or at the very least apply pressure. Dangerous physical contact (such as a tackle above the shoulders) are discouraged with a free kick, 50 metre penalty or suspension depending upon the severity of the incident. Infringements are also awarded (free kick and 50 metre penalty) for interference when marking, deliberate slowing of play, pushing an opponent in the back and many others.

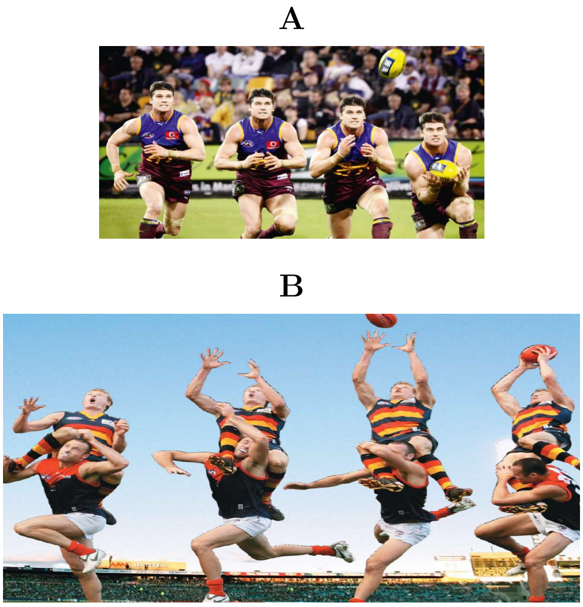
Figure 2.9 shows an example of the techniques involved in a tackle and shepherd. In the top pane of Figure 2.9, Jarrad Waite (Carlton) on the far right is performing a shepherd on Jordan McMahon (Richmond). This allows his team mate Nick Stevens (Carlton), who is in possession of the football, space to run into. In the bottom pane of Figure 2.9, Jude Bolton (Sydney) is performing a tackle on Jimmy Bartel (Geelong). As the tackle plays out, Jude Bolton has pinned Jimmy Bartels left arm to his body, which makes it increasingly difficult for Jimmy Bartel to correctly dispose of the football. In the last frame Jimmy Bartel has not preformed the handball correctly which would typically result in a free kick to Jude Bolton from where the tackle took place.



Source: 'AFL Record' Part 15 of a series on how to play the game - Shepherding;
'AFL Record' Part 6 of a series on how to play the game - Tackling

Figure 2.9: Contact. (A) Shepherd. (B) Tackle.

A unique feature of AFL is the mark, whereby a player catches the kicked ball which was deemed to have travelled at least 15 metres by a field umpire. The player who marked the ball is then entitled to an unimpeded free kick. Figure 2.10 shows an example of the techniques involved in a chest and overhead mark. An overhead mark on top of the shoulders (or back) of another player is a spectacular mark which is commonly referred to as a speccy, screamer or hanger. In the top pane of Figure 2.10 Jonathan Brown (Brisbane Lions) has his eyes solely on the football and drops down to one knee to take a chest mark. Similarly, in the bottom pane of Figure 2.10, Brett Burton (Adelaide) who is known as "Birdman" for his high flying marks, takes an overhead mark on top of the shoulders of Matthew Warnock (Melbourne). When taking an overhead mark, care must be taken not to put your hands in the back of an opponent for leverage or a free kick will result.

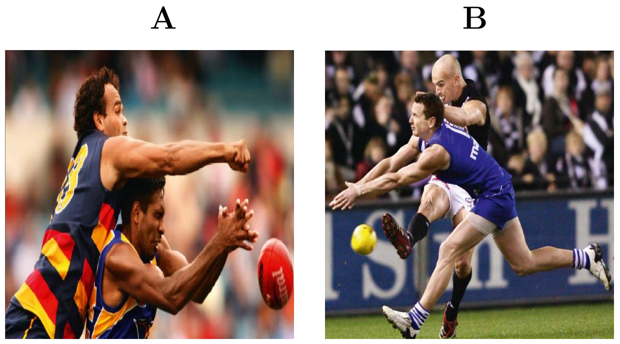


Source: 'AFL Record' Part 8 of a series on how to play the game - Chest Marking; 'AFL Record' Part 3 of a series on how to play the game - Overhead Marking

Figure 2.10: Marking. (A) Chest mark. (B) Overhead mark.

From a defensive point of view, players are also permitted to spoil and smother the football. These are typically referred to as one percenter's, primarily due to their infrequency and defensive nature. A team which continually measures high on one percenter's, essentially means they are applying defensive pressure on opponents by doing the little extra efforts which make turnovers more likely. The objective of a spoil is to stop an opposition player from taking possession (usually via a mark) by punching the incoming football with a clenched fist. A smother is performed by blocking the football with outstretched hands immediately after it has been kicked by an opposition player and prevents the ball from traveling to its initial destination

Figure 2.11 shows an example of the techniques involved in a spoil and a smother. In the left pane of Figure 2.11, Graham Johncock (Adelaide) performs a spoil which prevents David Wirrpanda (West Coast) from taking a mark by punching the football with a clenched fist. In the right pane of Figure 2.11, Adam Simpson (Kangaroos) performs a smother by outstretching both arms immediately after John Anthony (Collingwood) has kicked the ball, which prevents the ball traveling to where John Anthony had intended.

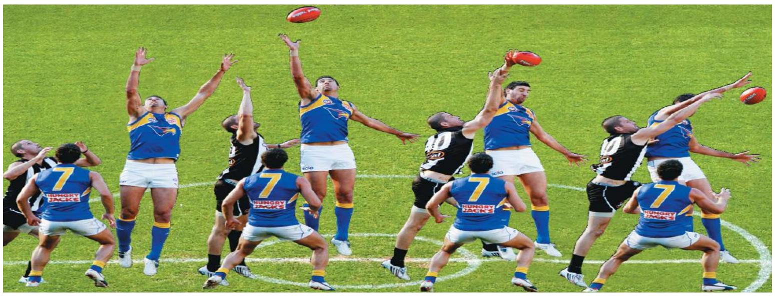


Source: 'AFL Record' Part 9 of a series on how to play the game - Spoiling; 'AFL Record' Part 13 of a series on how to play the game - Smothering

Figure 2.11: One percenter's. (A) Spoil. (B) Smother.

Each quarter commences with a centre bounce where a field umpire restarts play by bouncing the ball into the centre of the ground or propelling the ball into the air. Once the ball is in the air the two opposing ruckmen from both teams compete in a ruck dual. The primary objective of each ruckman in the ruck contest is to tap the ball to the advantage of a team mate, or gain a significant amount of ground by knocking the ball into the general direction of their respective goals. This event is somewhat akin to a tip-off in basketball.

Figure 2.12 shows an example of the techniques involved in a ruck contest. In this example, Dean Cox (West Coast) on the right is up against Chris Bryan (Collingwood) on the left in a ruck contest. Dean Cox wins the hitout and taps the ball to the advantage of his teammate Daniel Kerr (West Coast) in the number seven guernsey.



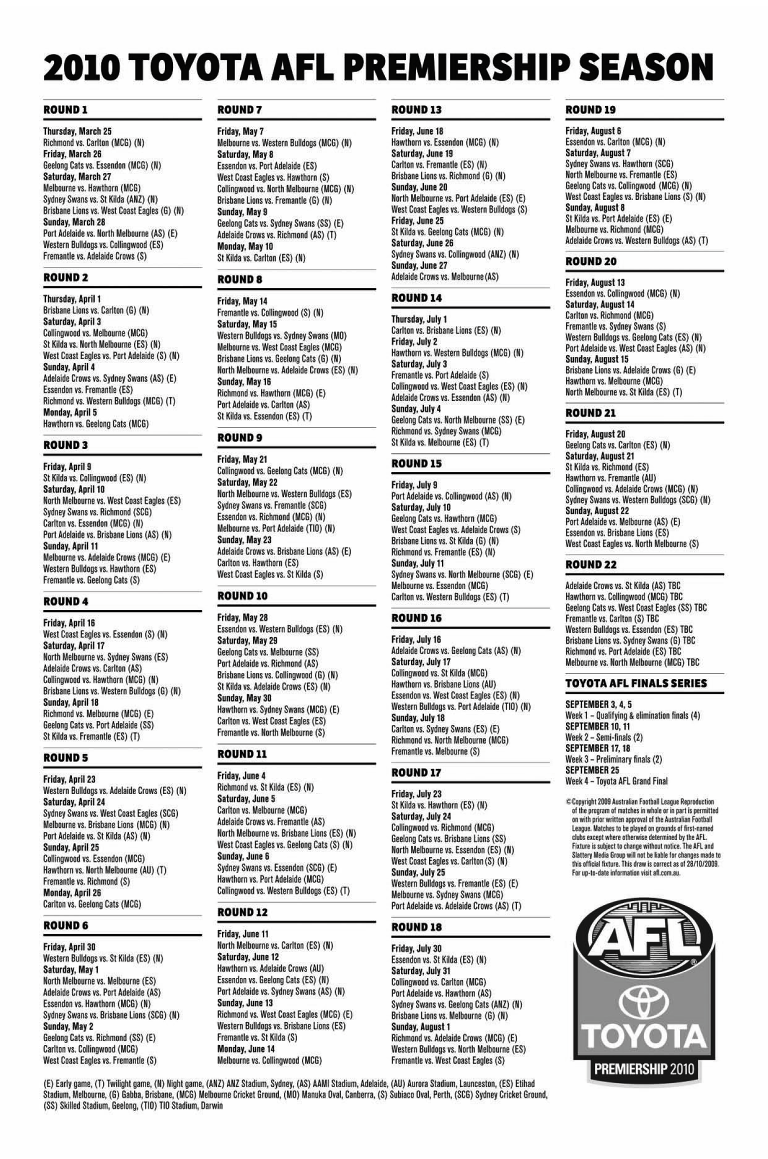
Source: 'AFL Record' Part 5 of a series on how to play the game - Ruck

Figure 2.12: The ruck contest.

For a more detailed description of the game including video highlights, player profiles and the latest news on the game visit www.afl.com.au.

2.2.5 The Fixture

A unique feature of AFL is the unbalanced nature of the competition in respect to team quality and home advantage. From 1996 to 2010 the fixture has been unbalanced, with 22 rounds and 16 teams, which results in any given team playing eight opponents once and seven opponents twice. Furthermore, historical traditions (and marketing matches for maximum crowds) have an effect on the choice of teams which play each other twice. The schedule itself is also not naturally sequential. For example, just because Team A plays Team B in Round 1, their next meeting will not necessarily be scheduled in Round 16 (after the possibility of playing all other teams once); they could meet again prior to Round 16, or they may only play each other once for the entire season. Non-Victorian teams play at least half of their matches at home, while the Victorian teams play more than half of their games at home. This is due to other Victorian teams sharing the same home ground. In the AFL draw, the team which is named first is the nominal home team. However, the nominal home team will not always have a distinct home advantage; and in some cases the game will be played on the opposition's home ground. An example of this was Melbourne in 2008 who were under significant financial pressure and sold a home game to Sydney for financial gain. In effect, they bank on a larger gate taking and surrender their home advantage. Some matches are occasionally moved to bigger venues to maximise crowd capacity, which may also impact on (usually the loss of) home advantage. Figure 2.13 shows the 2010 AFL premiership fixture.



2010 TOYOTA AFL PREMIERSHIP SEASON

ROUND 1

Thursday, March 25
Richmond vs. Carlton (MCG) (N)
Friday, March 26
Geelong Cats vs. Essendon (MCG) (N)
Saturday, March 27
Melbourne vs. Hawthorn (MCG)
Sydney Swans vs. St Kilda (ANZ) (N)
Brisbane Lions vs. West Coast Eagles (G) (N)
Sunday, March 28
Port Adelaide vs. North Melbourne (AS) (E)
Western Bulldogs vs. Collingwood (ES)
Fremantle vs. Adelaide Crows (S)

ROUND 2

Thursday, April 1
Brisbane Lions vs. Carlton (G) (N)
Saturday, April 3
Collingwood vs. Melbourne (MCG)
St Kilda vs. North Melbourne (ES) (N)
West Coast Eagles vs. Port Adelaide (S) (N)
Sunday, April 4
Adelaide Crows vs. Sydney Swans (AS) (E)
Essendon vs. Fremantle (ES)
Richmond vs. Western Bulldogs (MCG) (T)
Monday, April 5
Hawthorn vs. Geelong Cats (MCG)

ROUND 3

Friday, April 9
St Kilda vs. Collingwood (ES) (N)
Saturday, April 10
North Melbourne vs. West Coast Eagles (ES)
Sydney Swans vs. Richmond (SCG)
Carlton vs. Essendon (MCG) (N)
Port Adelaide vs. Brisbane Lions (AS) (N)
Sunday, April 11
Melbourne vs. Adelaide Crows (MCG) (E)
Western Bulldogs vs. Hawthorn (ES)
Fremantle vs. Geelong Cats (S)

ROUND 4

Friday, April 16
West Coast Eagles vs. Essendon (S) (N)
Saturday, April 17
North Melbourne vs. Sydney Swans (ES)
Adelaide Crows vs. Carlton (AS)
Collingwood vs. Hawthorn (MCG) (N)
Brisbane Lions vs. Western Bulldogs (G) (N)
Sunday, April 18
Richmond vs. Melbourne (MCG) (E)
Geelong Cats vs. Port Adelaide (SS)
St Kilda vs. Fremantle (ES) (T)

ROUND 5

Friday, April 23
Western Bulldogs vs. Adelaide Crows (ES) (N)
Saturday, April 24
Sydney Swans vs. West Coast Eagles (SCG)
Melbourne vs. Brisbane Lions (MCG) (N)
Port Adelaide vs. St Kilda (AS) (N)
Sunday, April 25
Collingwood vs. Essendon (MCG)
Hawthorn vs. North Melbourne (AU) (T)
Fremantle vs. Richmond (S)
Monday, April 26
Carlton vs. Geelong Cats (MCG)

ROUND 6

Friday, April 30
Western Bulldogs vs. St Kilda (ES) (N)
Saturday, May 1
North Melbourne vs. Melbourne (ES)
Adelaide Crows vs. Port Adelaide (AS)
Essendon vs. Hawthorn (MCG) (N)
Sydney Swans vs. Brisbane Lions (SCG) (N)
Sunday, May 2
Geelong Cats vs. Richmond (SS) (E)
Carlton vs. Collingwood (MCG)
West Coast Eagles vs. Fremantle (S)

ROUND 7

Friday, May 7
Melbourne vs. Western Bulldogs (MCG) (N)
Saturday, May 8
Essendon vs. Port Adelaide (ES)
West Coast Eagles vs. Hawthorn (S)
Collingwood vs. North Melbourne (MCG) (N)
Brisbane Lions vs. Fremantle (G) (N)
Sunday, May 9
Geelong Cats vs. Sydney Swans (SS) (E)
Adelaide Crows vs. Richmond (AS) (T)
Monday, May 10
St Kilda vs. Carlton (ES) (N)

ROUND 8

Friday, May 14
Fremantle vs. Collingwood (S) (N)
Saturday, May 15
Western Bulldogs vs. Sydney Swans (MO)
Melbourne vs. West Coast Eagles (MCG)
Brisbane Lions vs. Geelong Cats (G) (N)
North Melbourne vs. Adelaide Crows (ES) (N)
Sunday, May 16
Richmond vs. Hawthorn (MCG) (E)
Port Adelaide vs. Carlton (AS)
St Kilda vs. Essendon (ES) (T)

ROUND 9

Friday, May 21
Collingwood vs. Geelong Cats (MCG) (N)
Saturday, May 22
North Melbourne vs. Western Bulldogs (ES)
Sydney Swans vs. Fremantle (SCG)
Essendon vs. Richmond (MCG) (N)
Melbourne vs. Port Adelaide (TIO) (N)
Sunday, May 23
Adelaide Crows vs. Brisbane Lions (AS) (E)
Carlton vs. Hawthorn (ES)
West Coast Eagles vs. St Kilda (S)

ROUND 10

Friday, May 28
Essendon vs. Western Bulldogs (ES) (N)
Saturday, May 29
Geelong Cats vs. Melbourne (SS)
Port Adelaide vs. Richmond (AS)
Brisbane Lions vs. Collingwood (G) (N)
St Kilda vs. Adelaide Crows (ES) (N)
Sunday, May 30
Hawthorn vs. Sydney Swans (MCG) (E)
Carlton vs. West Coast Eagles (ES)
Fremantle vs. North Melbourne (S)

ROUND 11

Friday, June 4
Richmond vs. St Kilda (ES) (N)
Saturday, June 5
Carlton vs. Melbourne (MCG)
Adelaide Crows vs. Fremantle (AS)
North Melbourne vs. Brisbane Lions (ES) (N)
West Coast Eagles vs. Geelong Cats (S) (N)
Sunday, June 6
Sydney Swans vs. Essendon (SCG) (E)
Hawthorn vs. Port Adelaide (MCG)
Collingwood vs. Western Bulldogs (ES) (T)

ROUND 12

Friday, June 11
North Melbourne vs. Carlton (ES) (N)
Saturday, June 12
Hawthorn vs. Adelaide Crows (AU)
Essendon vs. Geelong Cats (ES) (N)
Port Adelaide vs. Sydney Swans (AS) (N)
Sunday, June 13
Richmond vs. West Coast Eagles (MCG) (E)
Western Bulldogs vs. Brisbane Lions (ES)
Fremantle vs. St Kilda (S)
Monday, June 14
Melbourne vs. Collingwood (MCG)

ROUND 13

Friday, June 18
Hawthorn vs. Essendon (MCG) (N)
Saturday, June 19
Carlton vs. Fremantle (ES) (N)
Brisbane Lions vs. Richmond (G) (N)
Sunday, June 20
North Melbourne vs. Port Adelaide (ES) (E)
West Coast Eagles vs. Western Bulldogs (S)
Friday, June 25
St Kilda vs. Geelong Cats (MCG) (N)
Saturday, June 26
Sydney Swans vs. Collingwood (ANZ) (N)
Sunday, June 27
Adelaide Crows vs. Melbourne (AS)

ROUND 14

Thursday, July 1
Carlton vs. Brisbane Lions (ES) (N)
Friday, July 2
Hawthorn vs. Western Bulldogs (MCG) (N)
Saturday, July 3
Fremantle vs. Port Adelaide (S)
Collingwood vs. West Coast Eagles (ES) (N)
Adelaide Crows vs. Essendon (AS) (N)
Sunday, July 4
Geelong Cats vs. North Melbourne (SS) (E)
Richmond vs. Sydney Swans (MCG)
St Kilda vs. Melbourne (ES) (T)

ROUND 15

Friday, July 9
Port Adelaide vs. Collingwood (AS) (N)
Saturday, July 10
Geelong Cats vs. Hawthorn (MCG)
West Coast Eagles vs. Adelaide Crows (S)
Brisbane Lions vs. St Kilda (G) (N)
Richmond vs. Fremantle (ES) (N)
Sunday, July 11
Sydney Swans vs. North Melbourne (SCG) (E)
Melbourne vs. Essendon (MCG)
Carlton vs. Western Bulldogs (ES) (T)

ROUND 16

Friday, July 16
Adelaide Crows vs. Geelong Cats (AS) (N)
Saturday, July 17
Collingwood vs. St Kilda (MCG)
Hawthorn vs. Brisbane Lions (AU)
Essendon vs. West Coast Eagles (ES) (N)
Western Bulldogs vs. Port Adelaide (TIO) (N)
Sunday, July 18
Carlton vs. Sydney Swans (ES) (E)
Richmond vs. North Melbourne (MCG)
Fremantle vs. Melbourne (S)

ROUND 17

Friday, July 23
St Kilda vs. Hawthorn (ES) (N)
Saturday, July 24
Collingwood vs. Richmond (MCG)
Geelong Cats vs. Brisbane Lions (SS)
North Melbourne vs. Essendon (ES) (N)
West Coast Eagles vs. Carlton (S) (N)
Sunday, July 25
Western Bulldogs vs. Fremantle (ES) (E)
Melbourne vs. Sydney Swans (MCG)
Port Adelaide vs. Adelaide Crows (AS) (T)

ROUND 18

Friday, July 30
Essendon vs. St Kilda (ES) (N)
Saturday, July 31
Collingwood vs. Carlton (MCG)
Port Adelaide vs. Hawthorn (AS)
Sydney Swans vs. Geelong Cats (ANZ) (N)
Brisbane Lions vs. Melbourne (G) (N)
Sunday, August 1
Richmond vs. Adelaide Crows (MCG) (E)
Western Bulldogs vs. North Melbourne (ES)
Fremantle vs. West Coast Eagles (S)

ROUND 19

Friday, August 6
Essendon vs. Carlton (MCG) (N)
Saturday, August 7
Sydney Swans vs. Hawthorn (SCG)
North Melbourne vs. Fremantle (ES)
Geelong Cats vs. Collingwood (MCG) (N)
West Coast Eagles vs. Brisbane Lions (S) (N)
Sunday, August 8
St Kilda vs. Port Adelaide (ES) (E)
Melbourne vs. Richmond (MCG)
Adelaide Crows vs. Western Bulldogs (AS) (T)

ROUND 20

Friday, August 13
Essendon vs. Collingwood (MCG) (N)
Saturday, August 14
Carlton vs. Richmond (MCG)
Fremantle vs. Sydney Swans (S)
Western Bulldogs vs. Geelong Cats (ES) (N)
Port Adelaide vs. West Coast Eagles (AS) (N)
Sunday, August 15
Brisbane Lions vs. Adelaide Crows (G) (E)
Hawthorn vs. Melbourne (MCG)
North Melbourne vs. St Kilda (ES) (T)

ROUND 21

Friday, August 20
Geelong Cats vs. Carlton (ES) (N)
Saturday, August 21
St Kilda vs. Richmond (ES)
Hawthorn vs. Fremantle (AU)
Collingwood vs. Adelaide Crows (MCG) (N)
Sydney Swans vs. Western Bulldogs (SCG) (N)
Sunday, August 22
Port Adelaide vs. Melbourne (AS) (E)
Essendon vs. Brisbane Lions (ES)
West Coast Eagles vs. North Melbourne (S)

ROUND 22

Adelaide Crows vs. St Kilda (AS) TBC
Hawthorn vs. Collingwood (MCG) TBC
Geelong Cats vs. West Coast Eagles (SS) TBC
Fremantle vs. Carlton (S) TBC
Western Bulldogs vs. Essendon (ES) TBC
Brisbane Lions vs. Sydney Swans (G) TBC
Richmond vs. Port Adelaide (ES) TBC
Melbourne vs. North Melbourne (MCG) TBC

TOYOTA AFL FINALS SERIES

SEPTEMBER 3, 4, 5
Week 1 – Qualifying & elimination finals (4)
SEPTEMBER 10, 11
Week 2 – Semi-finals (2)
SEPTEMBER 17, 18
Week 3 – Preliminary finals (2)
SEPTEMBER 25
Week 4 – Toyota AFL Grand Final

©Copyright 2009 Australian Football League Reproduction of the program of matches in whole or in part is permitted on with prior written approval of the Australian Football League. Matches to be played on grounds of first-named clubs except where otherwise determined by the AFL. Fixture is subject to change without notice. The AFL and Slattery Media Group will not be liable for changes made to this official fixture. This draw is correct as of 28/10/2009. For up-to-date information visit afl.com.au.

(E) Early game, (T) Twilight game, (N) Night game, (ANZ) ANZ Stadium, Sydney, (AS) AAMI Stadium, Adelaide, (AU) Aurora Stadium, Launceston, (ES) Etihad Stadium, Melbourne, (G) Gabba, Brisbane, (MCG) Melbourne Cricket Ground, (MO) Manuka Oval, Canberra, (S) Subiaco Oval, Perth, (SCG) Sydney Cricket Ground, (SS) Skilled Stadium, Geelong, (TIO) TIO Stadium, Darwin

Source: http://www.afl.com.au/portals/0/afl_docs/fixture_document.pdf

Figure 2.13: AFL premiership fixture, 2010

2.2.6 The Ladder

Teams are awarded four premiership points for a win, two premiership points for a draw and zero premiership points for a loss during the Home and Away Season. At the conclusion of each round, teams are ranked based on their cumulative premiership points and in the case of two or more teams having an equal number of premiership points, their ladder position is further determined by their percentage which is defined by:

$$\% = \frac{\sum_{i=1}^{n} PF_i}{\sum_{i=1}^{n} PA_i} \times 100 \tag{2.1}$$

where PF_i = points scored for in game i, PA_i = points scored against in game i and n = number of home and away games.

Table 2.3 shows an example of the premiership ladder at the conclusion of the 2000 home and away season. Interestingly, the ladder in 2009 is almost the complete reversal of the ladder in 2000, with St Kilda and Collingwood both occupying top four positions, and Melbourne and Kangaroos both finishing in the bottom four. The almost complete reversal of the ladder from 2000 to 2009 indicates that the reverse drafting system promotes competitive balance.

Rank	Team	P	W	L	D	PF	PA	%	PTS
1	Essendon	22	21	1	0	2816	1770	159.1	84
2	Carlton	22	16	6	0	2667	1979	134.77	64
3	Melbourne	22	14	8	0	2557	2159	118.43	56
4	Kangaroos	22	14	8	0	2447	2304	106.21	56
5	Geelong	22	12	9	1	2234	2306	96.88	50
6	Brisbane Lions	22	12	10	0	2602	2222	117.1	48
7	Western Bulldogs	22	12	10	0	2321	2241	103.57	48
8	Hawthorn	22	12	10	0	2198	2251	97.65	48
9	Richmond	22	11	11	0	2068	2221	93.11	44
10	Sydney	22	10	12	0	2254	2219	101.58	40
11	Adelaide	22	9	13	0	2255	2347	96.08	36
12	Fremantle	22	8	14	0	1886	2618	72.04	32
13	West Coast	22	7	14	1	2216	2399	92.37	30
14	Port Adelaide	22	7	14	1	1928	2295	84.01	30
15	Collingwood	22	7	15	0	2089	2431	85.93	28
16	St Kilda	22	2	19	1	1855	2631	70.51	10

Note. P = Played, W = Won, L = Lost, D = Drawn, PF = Points For, PA = Points Against, PTS = Premiership Points

Table 2.3: AFL ladder, 2000

2.3 Recruitment of Players

2.3.1 History

The national draft was first held in 1986 prior to Brisbane and West Coast entering the competition the following year. Currently, the drafting process consists of four distinct phases at the conclusion of each AFL season. These phases are (1) the Trading period; (2) the National Draft; (3) the Pre-Season Draft and (4) the Rookie Draft. In 1990 to 1993 the drafting period also included a mid-season draft.

2.3.2 Trading

The trading period is the first phase of the drafting process and occurs shortly after the conclusion of each season. This allows teams to trade senior players on that years list to other clubs in exchange for other players, national draft selections or a combination of both. The simplest form of a deal includes trading player A for national draft selection x, or trading player A for player B. However, this is not always the norm. Many trades involve a combination of clubs, national draft selections and players. All players involved in the trade must consent to the trade before it can be finalised. Players typically initiate the trade due to a lack of opportunity at their current club.

2.3.3 National Draft

The order of selection is based on the reverse ladder positions of the previous season. For example, Melbourne finished last in the 2009 AFL season and thus received the first selection in the National Draft. The priority draft selection was first introduced in 1993 to award poorly performing teams special assistance with an additional early draft selection.

The eligibility criteria and the draft pick number for the priority selection has changed considerably over the years. This is due to speculation that teams deliberately lose matches when they are out of finals contention in order to receive a priority selection. To be eligible for the National Draft players must be 17 years of age on or before the 30th April the year they will potentially be drafted. This age based criteria has slowly been lifted as there was concern about players as young as 15 or 16 having to move interstate to play AFL. Potential AFL players must nominate for the National Draft prior to the cutoff date. The National Draft forms the foundation of many AFL clubs as approximately 100 players are selected across all clubs.

2.3.4 Pre-Season Draft

The AFL Pre-Season draft is for the recruitment of uncontracted players and it occurs after the National Draft and at approximately the same time as the Rookie Draft. Akin to the National Draft, the order of selection is based on the reverse ladder positions of the previous season. The importance of the Pre-Season Draft has diminished greatly from its conception in 1989 with over 50 selections that year to just 8 selections in 2009. Unlike the National Draft, not all clubs participate in the Pre-Season draft, since many clubs fill their senior list in the National Draft.

2.3.5 Rookie Draft

The Rookie Draft is the final phase of the the drafting process and is limited to the recruitment of players under the age of 23. Akin to the National Draft, the order of selection is based on the reverse ladder positions of the previous season. Rookies are typically young players which require significant development (i.e. very athletic with minimal football skills). These players are not permitted to play in the team unless they are promoted to the senior list due to a long term injury or the retirement of a senior player. At the end of the year, rookie listed players can be officially upgraded to the senior list, stay as a second year rookie or be delisted. The majority of teams have six rookies, with very few making it on the senior list. Brisbane and Sydney are permitted to have additional rookies to encourage investment in local players due to their local leagues (QAFL and AFL Sydney) being of lower standard in comparison to Victoria (VFL), Adelaide (SANFL) and Perth (WAFL).

2.4 AFL Statistics Providers

The Australian Football League consists of two major statistical providers to AFL clubs and third parties, namely Champion Data and ProWess Sports. This section provides a brief background on these companies.

2.4.1 Champion Data

Champion Data was established in 1995 and received a licence from the AFL in 1999. Their client base includes the Australian Football League (AFL), National Rugby League (NRL), Rugby Union, Netball, Cricket and provide summaries of other major Australian sports. They provide information to all AFL clubs, TV stations, radio, various websites including AFL (www.afl.com.au) and Superfooty (www.heraldsun.com.au/sport/afl), directly or indirectly most major telecommunication companies, all major News print groups and related AFL publications and Stadiums. AFL clubs get live services, vision services, recruitment and advanced analytical reports and post match analysis tools.

2.4.2 ProWess Sports

ProWess Sports (Gundy Computer Services Pty Ltd) has been in operation since 1982 and has provided technological applications and services to sport teams, leagues, coaches and analysts both in Australia and internationally. Their client base, both past and present, include but not limited to the National Basketball League (NBL), Womens National Basketball League (WNBL), Victorian Netball Association (WNA), Australian Football League (AFL) and the West Australian Football League (WAFL).

2.4.3 Summary

Although a plethora of data are recorded by Champion Data and ProWess sports, the amount of in-depth statistical analysis conducted by both companies and AFL clubs for that matter is minimal. Prior to starting my PhD and throughout my candidature I have worked directly and indirectly with Champion Data and ProWess sports. In 2007 an ongoing research collaboration agreement was set up between the RMIT Sports Statistics Research Group and ProWess Sports. Under this agreement, the group obtained detailed in-game and post-match statistics for AFL matches to analyse as the data became available.

Chapter 3

Methods

3.1 Introduction

In this chapter, the methodology used in subsequent chapters is described. In Section 3.2, linear regression methodology is discussed, which is utilised in Chapter 4 and Chapter 8. Similarly, Section 3.3 covers logistic regression and is incorporated in Chapter 11. Section 3.4 details optimisation algorithms which are implemented in Chapter 5 and Chapter 10. In Section 3.5, Elo ratings are described which are adapted for AFL in Chapter 5. The final section, Section 3.6, covers the computer programming component of this dissertation which is utilised in Chapter 6, 7 and 11.

3.2 Linear Regression

This section covers the linear regression component of this dissertation. To begin, the mathematical formulation of the linear regression model is described. This is followed by the various assumptions that must be satisfied in order to make inferences about the coefficients of the regression model. Then the estimation of parameters using two methods namely *Ordinary Least Squares* (OLS) and *Maximum Likelihood Estimation* (MLE) is showcased. After that, the calculation of the residuals (errors) are specified which are required to satisfy the various assumptions of linear regression. Finally, the coefficient of determination is explained which is used to assess the adequacy of the fitted model. It should be noted that the methodology described in this Section has been extracted from Greene (2002).

3.2.1 Introduction

Linear regression remains one of the most widely used statistical techniques across many disciplines including econometrics, environmental sciences and biostatistics to name a few. Linear regression models are extremely powerful since they have the power to empirically tease out very complicated relationships between variables. However, they are only appropriate under certain assumptions (discussed later) and are often misused, even in published journal articles.

3.2.2 The Linear Regression Model

In statistics, the multiple linear regression model is used to model the relationship between one or more independent variables and a dependent variable. The generic form of the model is given by

$$\begin{aligned} y &= f\left(1, x_1, x_2, \ldots, x_K\right) + \epsilon \\ &= \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_K\beta_K + \epsilon \qquad (3.1) \end{aligned}$$

where y is the explained (or dependent) variable, x=$(1, x_1, \ldots, x_K)'$ is a column vector of explanatory (or independent) variables, $\beta = (\beta_0, \beta_1, \ldots, \beta_K)'$ is a column vector of coefficients and ϵ allows for random variation in y for a fixed value of x. The following n by p matrix denotes the observed values for x.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \ldots x_{1p} \\ x_{21} & x_{22} \ldots x_{2p} \\ \vdots & \ddots \quad \vdots \\ x_{n1} & x_{n2} \ldots x_{np} \end{bmatrix} \tag{3.2}$$

The objective of the linear regression model is to estimate the unknown parameters $\beta_0, \beta_1, \ldots, \beta_K$ which provide a "best fit" to a series of data points.

For example, suppose we were interested in predicting the Brownlow medal in AFL, which is the best and fairest award for all players for a given year. The scoring system for the Brownlow medal is a "3-2-1" voting system. In each match of the season the best player is awarded three votes, the next best player two votes and the next best player one vote. These votes are determined by the field umpires at the conclusion of the match. The player that has the most votes at the conclusion of the season, provided they have not been suspended (this is the "fairest" component) is awarded the Brownlow medal. In this example, the dependent variable would be the number of votes for each player for each round. The independent variables could be almost anything from the number of disposals a player receives to whether a player's team won or lost (since it is widely believed that three votes is awarded to a player on the winning team). For more information regarding Brownlow medal prediction see Bailey and Clarke (2002) and Bailey and Clarke (2008).

3.2.3 Assumptions

There are several assumptions of the linear regression model which must be satisfied in order to make inferences about the coefficients derived from the model. These assumptions are listed below.

A1. Linearity: $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \epsilon_i$. The model specifies a linear relationship between y and $x_1, \ldots, x_K$.

A2. Full Rank: None of the independent variables is a perfect linear combination of the other independent variables.

A3. Exogeneity of the independent variables: $E\left[\epsilon_i | x_{j1}, x_{j2}, \ldots, x_{jK}\right] = 0$. This states that the expected value of the disturbance at observation i in the sample is not a function of the independent variables observed at any observation, including this one. This means that independent variables will not carry useful information for prediction of ϵ_i.

A4. Homoscedasticity and non-autocorrelation: Each disturbance ϵ_i, has the same finite variance σ^2 and is uncorrelated with every other disturbance ϵ_j.

A5. Exogenously generated data: The data in $(x_{j1}, x_{j2}, \ldots, x_{jK})$ may be any mixture of constants and random variables. The process generating the data operates outside the assumptions of the model, that is, independently of the process that generates ϵ_i. Note that this extends **A3**. Analysis is done conditionally on the observed X.

A6. Normal Distribution: The disturbances are normally distributed.

3.2.4 Least Squares Regression

Linear regression approximates the unknown parameters of the stochastic relation $y_i = \mathbf{x}_i^{'}\boldsymbol{\beta} + \epsilon_i$. Firstly, it is important to distinguish between population quantities $\boldsymbol{\beta}$ and ϵ_i and sample estimates, denoted $\mathbf{b}$ and e_i. Here the population regression is $E[y_i|\mathbf{x}_i] = \mathbf{x}_i^{'}\boldsymbol{\beta}$ while the estimate of $E[y_i|\mathbf{x}_i]$ is given by

$$\hat{y}_i = \mathbf{x}_i^{'}\boldsymbol{\beta} \tag{3.3}$$

Here the disturbance associated with the i^{th} data point is denoted

$$\epsilon_i = y_i - \mathbf{x}_i^{'}\boldsymbol{\beta} \tag{3.4}$$

for a given value of $\mathbf{b}$, the estimate of ϵ_i is given by the residual

$$e_i = y_i - \mathbf{x}_i^{'}\mathbf{b}. \tag{3.5}$$

Therefore, based on these definitions,

$$y_i = \mathbf{x}_i^{'}\boldsymbol{\beta} + \epsilon_i = \mathbf{x}_i^{'}\mathbf{b} + e_i \tag{3.6}$$

There are numerous methods for estimating the unknown vector $\boldsymbol{\beta}$ of population quantities; most commonly used methods include *Ordinary Least Squares* (OLS) and *Maximum Likelihood Estimation* (MLE).

Ordinary Least Squares

Ordinary Least Squares (OLS) is perhaps the most frequently used method for estimating the unknown vector $\boldsymbol{\beta}$. The least squares coefficient vector minimizes the sum of the squared errors, which is given by

$$\begin{aligned}\sum_{i=1}^{n} e_{i0}^2 &= \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{'}\boldsymbol{\beta})^2 \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{'}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^{'}\mathbf{y} - 2\mathbf{y}^{'}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^{'}\mathbf{X}^{'}\mathbf{X}\boldsymbol{\beta} \qquad (3.7)\end{aligned}$$

where $\boldsymbol{\beta}$ denotes the choice of the coefficient vector.

To calculate the minimum, the partial derivative of (3.7) with respect to $\boldsymbol{\beta}$ is set to zero and solved.

$$\frac{\partial}{\partial\boldsymbol{\beta}} = -2\mathbf{X}^{'}\mathbf{y} + 2\mathbf{X}^{'}\mathbf{X}\boldsymbol{\beta} = 0 \qquad (3.8)$$

Let $\mathbf{b}$ be the solution, therefore $\mathbf{b}$ satisfies the least squares normal equations given by

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y} \qquad (3.9)$$

Since the inverse of $\mathbf{X}^{'}\mathbf{X}$ exists based on the full rank assumption (**A2.**), then the solution is given by

$$\mathbf{b} = (\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{y} \qquad (3.10)$$

Maximum Likelihood Estimation

Another method of estimating the unknown vector $\boldsymbol{\beta}$ is that of *Maximum Likelihood Estimation* (MLE). The probability density function (pdf) of a random variable y conditional on a given set of parameters $\boldsymbol{\theta}$ is denoted $f(y|\boldsymbol{\theta})$. The joint density (or likelihood function) of n *independent identically distributed* (iid) observations from given pdf is denoted.

$$f(y_1,\dots,y_n|\theta) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\mathbf{y}) \qquad (3.11)$$

It is more convenient to write the likelihood function after a log transformation. Also, the likelihood is written more conveniently as L. Therefore,

$$L = \text{ln}L(\boldsymbol{\theta}|y) = \sum_{i=1}^{n}\text{ln}f(y_i|\boldsymbol{\theta}) \qquad (3.12)$$

In linear regression, the likelihood function for a sample of n independent, identically and normally distributed disturbances is given by

$$L = (2\pi\sigma^2)^{-n/2} e^{\mathbf{e}'\mathbf{e}/(2\sigma^2)} \tag{3.13}$$

The transformation from ϵ_i to y_i is $\epsilon_i = y_i - \mathbf{x}_i\boldsymbol{\beta}$, such that the Jacobian for each observation $[\partial\epsilon_i/\partial y_i]$ equals one. Therefore, the likelihood function can now be written

$$L = (2\pi\sigma^2)^{-n/2} e^{-1/(2\sigma^2)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})} \tag{3.14}$$

and the log-likelihood is given by

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \tag{3.15}$$

To maximise the log-likelihood the partial derivative of (3.15) is taken with respect to $\boldsymbol{\beta}$ and σ^2 which is given by

$$\begin{bmatrix} \frac{\partial \ln L}{\partial \beta} \\ \frac{\partial \ln L}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} \\ \frac{-n}{2\sigma^2} + \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix} \tag{3.16}$$

The values which satisfy these equations are

$$\hat{\beta}_{ML} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{b} \quad \text{and} \quad \hat{\sigma}^2_{ML} = \frac{\mathbf{e}'\mathbf{e}}{n} \tag{3.17}$$

3.2.5 Analysis of Residuals

One such method to test the assumptions defined in Section 3.2.3, and thus the adequacy of the linear regression model, is plotting the residuals. The residual for the ith case is given by

$$\hat{z}_i = y_i - \hat{y}_i \tag{3.18}$$

where y_i is the observed outcome and $\hat{y}_i$ is the predicted outcome.

If a relationship exists between the residuals $\hat{z}_i$ and any variable then there is an effect from that variable which has not yet been accounted for. A common plot includes the residuals $\hat{z}_i$ against the fitted values $\hat{y}_i$ which reveals outliers and whether the assumption of constant variance and linearity are appropriate. Additional measures used to detect outliers include Mahalanobis Distance and Cook's Distance. Another common plot is the residuals $\hat{z}_i$ against a time dependent predictor variable or the order number of the experiment, a smooth plot will show that the assumption of independence is not valid. Residual independence can also be checked using the Durbin-Watson statistic.

3.2.6 Goodness of Fit

The coefficient of determination, commonly denoted by R^2 is used to assess the goodness-of-fit of the linear regression model. R^2 is described as the amount of variation that can be explained by the regressors where $R^2 \in [0, 1]$. If $R^2 = 1$ the values of $\mathbf{x}$ and y all lie on the same hyperplane such that all the residuals are zero. On the contrary, if $R^2 = 0$ the fitted values correspond to a horizontal line such that all the elements of $\mathbf{b}$ except the constant term are zero. The "variability" of the data is measured through different sum of squares where

$$\begin{aligned} SS_T &= \sum_{i=1}^{n}(y_i - \bar{y})^2, \text{total sum of squares} \\ SS_R &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2, \text{regression sum of squares} \\ SS_E &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \text{residual sum of squares} \qquad (3.19) \end{aligned}$$

where,

$$R^2 = 1 - \frac{SS_E}{SS_T} = \frac{SS_R}{SS_T} \qquad (3.20)$$

Additional measures of goodness of fit which account for the complexity of the model include the adjusted R^2 and Akaike Information Criterion (AIC).

3.3 Binary Logistic Regression

This section covers the logistic regression component of this dissertation. Firstly, a brief introduction is given describing the importance of logistic regression over linear regression when the outcome variable is dichotomous. The next section covers fitting the logistic regression model including the estimation of parameters using *Maximum Likelihood Estimation* (MLE). This is followed by the analysis of residuals and examining the adequacy of the fitted model. Finally, the interpretation of the coefficients of the fitted model using odds ratios is explained. It should be noted that this methodology has been summarised from Hosmer and Lemeshow (2000).

3.3.1 Introduction

What distinguishes a logistic regression model from a linear regression model is that the dependent variable in logistic regression is *always* dichotomous. For example, the presence/absence of a disease in epidemiology, success/failure of an operation in medical sciences and win/loss in sport to name just a few. The difference between the two models is reflected in the assumptions and the choice of the parametric model. Once these differences are accounted for, the techniques used in linear regression are implemented and built upon in logistic regression.

The first difference concerns the relationship between the dependent and independent variables. Recall the population regression (or conditional mean) for linear regression is $E[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$, which implies $E[y_i|\mathbf{x}_i]$ can take on any value of x between $-\infty$ and $+\infty$.

However, in logistic regression the dependent variable (Y) is dichotomous taking values 0 or 1, which implies the conditional mean must be greater than or equal to zero and less than or equal to one. There are many distribution functions that have been proposed to deal with the analysis of dichotomous dependent variables (Cox and Snell, 1989). Logistic regression utilises the logistic distribution primarily due to flexibility of the function and the meaningful interpretations of the results.

Let $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$ denote a collection of p independent variables. The conditional probability that the outcome is present is denoted by $P(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$. The logit of the multiple logistic regression model is given by

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{3.21}$$

The logit transformation which is central to the study of logistic regression is given by (3.22) and is showcased in Figure 3.1.

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} \tag{3.22}$$

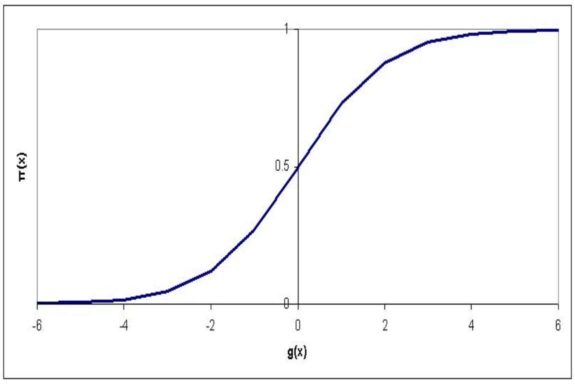


Figure 3.1: The logit transformation

Interestingly, $g(x)$ has many of the desirable properties of the linear regression model. That is, it is linear in its parameters, it can be continuous and can range between $-\infty$ and

$+\infty$.

The second crucial difference between linear and logistic regression concerns the conditional distribution of the dependent variable. Recall in the linear regression $y_i = E[y_i|\mathbf{x}_i]+\epsilon_i$, where ϵ_i follows a normal distribution with mean zero and unknown variance σ^2. However, when the dependent variable is dichotomous this is not the case. Since when $y = 1$, $\epsilon = 1-\pi(x)$ with probability $\pi(x)$, similarly when $y = 0$, $\epsilon = -\pi(x)$ with probability $1-\pi(x)$. Therefore, ϵ follows a binomial distribution with mean zero and variance $\pi(x)[1-\pi(x)]$.

3.3.2 Fitting the Logistic Regression Model

Let n denote a sample of independent observations (x_i, y_i), $i = 1, 2, \ldots, n$, where y_i denotes the value of dichotomous dependent variable. In linear regression, the unknown parameters can be estimated via Ordinary Least Squares and Maximum Likelihood Estimation. However, in logistic regression the unknown parameters can only be estimated numerically (as opposed to analytically), therefore Maximum Likelihood Estimation is utilised.

In logistic regression, the likelihood function for a sample of n independent, identically and Bernoulli distributed disturbances is given by

$$L = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i} \left[1-\pi(\mathbf{x}_i)\right]^{1-y_i} \tag{3.23}$$

Therefore, the log-likelihood is given by

$$\ln L = \sum_{i=1}^{n} y_i \ln[\pi(\mathbf{x}_i)] + (1-y_i)\ln[1-\pi(\mathbf{x}_i)] \tag{3.24}$$

To maximise the log-likelihood function, we differentiate (3.24) with respect to the $p+1$ coefficients (β_i) and set the resulting equations equal to zero. This results in $p+1$ likelihood equations which can be expressed as follows

$$\sum [y_i - \pi(\mathbf{x}_i)] = 0 \tag{3.25}$$

and

$$\sum x_{ij}\left[y_i - \pi(\mathbf{x}_i)\right] = 0 \tag{3.26}$$

for $j = 1, 2, \ldots, p$.

In linear regression, the likelihood equations are linear, thus making solving the unknown parameters easy. However, in logistic regression the likelihood equations are nonlinear and require special iterative methods for their solutions. Statistical programs such as SPSS and Stata routinely derive these solutions.

3.3.3 Analysis of Residuals

In linear regression, as in most statistics, the residual is defined as the difference between the observed and the fitted value $(y - \hat{y})$. However, in logistic regression there are several different methods for measuring the difference between the observed and fitted values. The primary purpose for the analysis of residuals in logistic regression is to identify cases for which the model poorly fits, or cases that have a significant influence on the estimated parameters of the model. In linear regression, we can assume that the error is independent of the conditional mean of Y. However, in logistic regression the error variance is a function of the conditional mean. Consequently, the residuals are standardized by adjusting them for their standard errors.

Here we will consider two measures for the difference between the observed and the fitted value, namely the Pearson residual and the Deviance residual. The Pearson residual for given covariate pattern is denoted

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{m_j\hat{\pi}_j\,(1 - \hat{\pi}_j)} \tag{3.27}$$

where m_j denotes the number of subjects with $\mathbf{x} = \mathbf{x}_j$.

Furthermore, the deviance residual is given by

$$d\left(y_j, \hat{\pi}_j\right) = \pm \left\{ y_j \text{ln} \left[\frac{y_j}{m_j \hat{\pi}_j} + (m_j - y_j) \text{ln} \left(\frac{m_j - y_j}{m_j(1 - \hat{\pi}_j)} \right) \right] \right\}^{\frac{1}{2}} \tag{3.28}$$

where ± is the same sign as $(y_j - m_j \hat{\pi}_j)$.

When $y_j = 0$ the deviance residual is given by

$$d\left(y_j, \hat{\pi}_j\right) = -\sqrt{2m_j |\text{ln}(1 - \hat{\pi}_j)|} \tag{3.29}$$

and the deviance residual when $y_j = m_j$ is denoted

$$d\left(y_j, \hat{\pi}_j\right) = \sqrt{2m_j |\text{ln}(\hat{\pi}_j)|} \tag{3.30}$$

3.3.4 Goodness of Fit

In linear regression, the sum of squared errors is the criterion for selecting parameters. However, in logistic regression the log likelihood defined in (3.24) is the criterion for selecting parameters. A model with more parameters will always fit at least as well (have a greater log-likelihood) as a similar model with fewer coefficients. Therefore, a likelihood ratio test is used to compare the fit of two models where one model is always nested inside the other which is given by

$$G = -2\text{ln} \frac{(\text{likelihood without the variable})}{(\text{likelihood with the variable})} \tag{3.31}$$

In linear regression, the R^2 statistic defined in (3.20) is used to explain how much variation in the dependent variables can be explained by the independent variables. However, in logistic regression an equivalent R^2 statistic does not exist. Therefore, to evaluate the goodness of fit of logistic models, several pseudo R^2 measures have been developed. These are defined "pseudo" R^2 because they have the same characteristics as R^2 in the sense that they are on a similar scale, ranging from 0 to 1 with higher values indicating better model fit and vice versa, but they can't be interpreted as one would interpret an OLS R^2. Therefore, great caution should be taken when interpreting this statistic. Other goodness of fit measures include the Hosmer-Lemeshow statistic (Hosmer and Lemeshow, 2000).

3.3.5 Odds Ratios

In logistic regression, assessing the model adequacy (or goodness-of-fit) should precede any attempt at interpreting the coefficients of the model. The interpretation of any fitted model is of great practical importance since we can draw meaningful inferences from the estimated coefficients in the model. That is, what do the estimated coefficients tell us about the initial questions that motivated the study in the first place? This firstly involves determining the functional relationship between the dependent variable and the independent variable in question, then appropriately defining the unit of change in the independent variable.

Firstly, it is important to determine what function of the dependent variable yields a linear relationship of the independent variables. This is denoted by, or as a link function (Dobson, 1990). In linear regression, this is simply the identity function, since the dependent variable is linear in its parameters by its definition. However, in logistic regression the link function is the logit transformation.

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \tag{3.32}$$

In the linear regression model, the slope coefficient β_1 is equal to the difference between the value of the dependent variable at $x+1$ and x for any value of x. For example, suppose $y(x) = \beta_0 + \beta_1 x$, therefore $\beta_1 = y(x+1) - y(x)$. The interpretation of the coefficients in linear regression is relatively straightforward since it expresses the resulting change in dependent variable for a unit change in the independent variable. However, in logistic regression, the slope coefficient represents the change in the logit for a one unit change in the independent variable [i.e. $\beta_1 = g(x+1) - g(x)$].

Odds ratios come from transforming the logistic regression coefficients such that the independent variables affect the odds instead of the logged odds of the dependent variable. This is calculated by simply taking the exponent of the logistic regression coefficients. For the exponentiated coefficients, a coefficient of 1 leaves the odds unchanged, a coefficient greater than 1 increases the odds and a coefficient which is less than 1 reduces the odds.

3.4 Optimisation and Simulation

This section covers the optimisation component of this dissertation. To begin, a brief introduction describing the many applications of optimisation and some preliminary terminology is stated. This is followed by the mathematical formulation of optimisation problems. Then some important considerations are canvassed. Finally, two optimisation algorithms (simulations) are discussed. It should be noted that this methodology has been summarised from Nocedal and Wright (1999).

3.4.1 Introduction

Optimisation is widely used in many disciplines. Investors wish to maximise their returns whilst avoiding excessive risks; manufacturers aim to minimise the cost of their products without jeopardising quality; supermarkets schedule staff so costs are minimised; and couriers take the optimal route which minimises their petrol consumption.

To utilise optimisation techniques, the objective must first be formulated, that is, a quantitative measure of performance of the system. For example, the objective in the previously mentioned examples would be to maximise returns (or profits), minimise cost, minimise person time and minimise petrol consumption respectively. This measure is commonly referred to as the objective function, which is typically either minimised or maximised. The objective function is dependent on certain characteristics of the system known as variables. The goal of any optimisation problem is to find values of the unknown variables which optimise the objective function. Typically these unknown variables have constraints placed on them. For example, supermarkets that are trying to schedule their staff so costs are minimised are constrained by the maximum number of hours a staff member can work on a particular shift.

3.4.2 Mathematical Formulation

The mathematical formulation of an optimisation problem may use the following notation. Let

- $\mathbf{x}$ be a vector of unknown parameters
- $f(\mathbf{x})$ be the objective function, which is to be maximised or minimised
- $\mathbf{c}$ is be a vector of constraints which the unknown parameters must satisfy

Now the optimisation problem can be written as:

$$\begin{array}{l} \min_{x\in\mathbb{R}} f(x) \quad \text{or} \quad \max_{x\in\mathbb{R}} f(x) \\ \text{subject to} \\ c_i(x) \leq 0, \quad i = 1, 2, \ldots, m. \end{array}$$

For example, a sporting stadium might wish to maximise the daily profit for a given football match. Let x_1 and x_2 denote the number of spectators and cost per ticket respectively. Let z denote the daily profit for a given football match measured in thousands of dollars. Therefore, let

$$z = x_1 x_2 \tag{3.33}$$

The stadium can accommodate a maximum of 80,000 spectators and because of government legislation they are not permitted to sell tickets for more than \$30 each. Furthermore, the cost per ticket and number of spectators must both be nonnegative. Integrating the objective and the constraints we have

$$\begin{aligned} \max\ z = x_1 x_2 & \\ \text{subject to} & \\ x_1 &\leq 80,000 \\ x_2 &\leq 30 \\ x_1, x_2 &\geq 0 \end{aligned} \tag{3.34}$$

3.4.3 Important Considerations

There are many important considerations that should be taken into account when defining an optimisation problem. Let us consider four general issues which may arise.

(i) Is the optimisation problem *discrete* or *continuous* or a combination of the two? Discrete optimisation usually refers to problems in which the optimal solution is derived from a finite set of feasible solutions, that is, a vector of integers. However, continuous optimisation problems refer to problems in which the optimal solution is derived from an infinite set of feasible solutions, that is, a vector of real numbers. Typically speaking, continuous optimisation models are easier to solve since the behaviour of the function at all points close to $\mathbf{x}$ are similar due to the smoothness of the function. However, the same can not be said about discrete optimisation models due to their discrete nature. optimisation models that have both discrete and continuous variables are referred to as *mixed integer programming* problems.

(ii) Is the optimisation problem *stochastic* or *deterministic*? Stochastic optimisation problems arise when the model is not fully specified, that is, there is some unknown quantity at time of formulation. For example, in economics and finance an important characteristic of companies is future cash flow which is always unknown but can be estimated. Deterministic models on the other hand, are models that are fully specified, that is, there is no unknown quantity at time of formulation.

(iii) Is the optimisation problem *constrained* or *unconstrained*? A constrained optimisation model has explicit constraints on the unknown parameters which must be met in order for the objective function to be feasible. A constraint could simply be a bound place on a variable $a \leq x_1 \leq b$; declaring a variable must take integer values $x_2 \in \mathbb{Z}$; a more general linear constraint $\sum_i^n x_i \leq c$; or a nonlinear inequality which is a complex function comprising several variables. For unconstrained optimisation models every possible solution is feasible.

(iv) Is the *local* solution also the *global* solution? Many computer algorithms seek only a local solution, that is, the objective function is smaller than all other values within its vicinity. Furthermore, many computer algorithms have no in-built functions to check for local/global solutions. However, many non-linear functions have several local minimums in which case one would be interested in which one of these local minimums is also the global minimum, that is, the best solution of all such minima.

3.4.4 Optimisation Algorithms

An optimisation algorithm is an iterative numerical procedure for finding the values of the vector $\mathbf{x}$ that maximises (or minimises) the objective function $f(\mathbf{x})$ subject to the constraints $\mathbf{c}$. The algorithm begins with an initial estimate of the unknown parameters $\mathbf{x}_0$ then a sequence of improved estimates $(\mathbf{x}_i)_{i=1}^{\infty}$ are generated until no more improvements can be made or a solution is approximated with sufficient accuracy. The strategy of going from one iteration to the next is what separates the algorithms from one another. Some of the most common optimisation algorithms include Monte Carlo Sampling and Latin Hypercube Sampling.

Monte Carlo Sampling

Monte Carlo methods are a class of computational algorithms that utilise repeated random or pseudo-random numbers. These methods are typically used when computing an exact solution is unfeasible or impossible. Although there is not one definitive Monte Carlo method, the approach of many Monte Carlo methods are similar. Typically, a domain of possible inputs is defined of which inputs are generated randomly, then a deterministic computation is performed using these inputs and finally the results of the individual computations are aggregated into a final result.

Latin Hypercube Sampling

To understand the statistical method of Latin Hypercube Sampling it is crucial to comprehend the Latin Hypercube. Firstly, a Latin square is $n \times n$ square filled with n different colours such that each colour is represented only once in each row and each column. Similarly, a Latin Hypercube is the generalization of this concept to an arbitrary number of dimensions. Latin Hypercube sampling uses a statistical technique known as “stratified sampling without replacement”, whereby sampling is undertaken from a function of N variables with each variable being split into M equally probable intervals. The M sample points are then placed such that the Latin Hypercube is satisfied.

3.5 Elo Ratings

This section covers the Elo ratings component of this dissertation. To begin, a brief introduction on the history and background of Elo ratings is discussed. This is followed by the mathematical formulation of Elo ratings. Then the application of Elo ratings to world football is examined. It should be noted that this methodology has been extracted from Elo (1978).

3.5.1 Introduction

Elo ratings were originally developed by Arpad Elo to calculate the relative skill of chess players. The system entered official use in 1960 by the US Chess federation and was published later in 1978. The Elo method as originally conceived for chess has been used officially by international sports federations in mind sports: specifically in FIDE chess, FMJD Draughts and IGF Go. Of additional interest are applications in physical sports: ISF Sumo Wrestling and WCF croquet. The sumo wrestling application is unique in that the adjustment factor and player standard deviation are adjusted dependently. Elo ratings are a numerical system in which differences in ratings can be converted into scoring or winning probabilities. He states that the many performances of an individual when evaluated over an appropriate scale will be normally distributed. Furthermore, he says that performance in chess can't be quantified absolutely, it can only be inferred by the numbers of wins, losses and draws. In simple terms, the Elo ratings system calculates the expected number of games a player is expected to win in a given tournament. If a player exceeds these expectations they receive a ratings increase, while a player that falls short of these expectations receive a rating decrease. A powerful attribute of the Elo ratings system is a player can win a tournament and still receive a rating decrease if that player loses more games than expected.

The relative difference in ratings between two players is used to determine the winning probabilities, however the average rating and spread of ratings are typically arbitrarily chosen. Elo suggests scaling ratings such that a difference of 200 rating points would mean that the stronger player has an expected win of 0.76, and the United States Chess Federation (USCF) initially aimed for an average club player to have a rating of 1500.

3.5.2 Mathematical Formulation

The exact formula for calculating player A's probability of winning using the logistic curve is given by

$$W_e = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{3.35}$$

where W_e is the expected game result, R_A is the rating of player A and R_B is the rating of player B.

The formula for updating a new rating is given by

$$R_n = R_o + K(W - W_e) \tag{3.36}$$

where R_n is the new rating, R_o is the old rating, W is the observed game result (loss=0, draw=0.5 and win=1), W_e is the expected game result and K is the change in ratings multiplier.

The coefficient K reflects the relative weights attributed to the pre-game rating and the event performance rating. For example, a high K gives greater weight to more recent performances. Similarly, a low K gives more weight to earlier performances. This new rating can be updated after a single match or at the conclusion of a tournament. Typically, K ranges between 10 and 32. The United States Chess Federation (USCF) and the Fédération Internationale des Échecs (FIDE) using varying levels of K dependent upon the magnitude of a players rating, the greater the rating the smaller the value of K and vice versa.

An example might help to clarify how ratings are updated. Suppose Player A has a rating of 1781, and plays in a three-round tournament. Here W and W_e are replaced by cumulative actual wins and expected wins respectively. He loses his first match to a player rated 1943, wins his second match to a player rated 1721 and draws his final match to a player rated 2019. His actual wins, counting draws as half wins, are $(0 + 1 + 0.5) = 1.5$. His expected wins, calculated according to (3.35), was $(0.282+0.585+0.203) = 1.07$. Therefore, his new rating is $1781 + 32 \times (1.5 - 1.07) = 1795$, assuming that a K factor of 32 is used.

3.5.3 The World Football Elo Rating System

The World Football Elo rating system is used to rate international football teams. In 1997, Bob Runyan adapted the Elo rating system to international football and posted the results on the internet (www.eloratings.net). The system was adapted to football by weighting the importance of the match, making an adjustment for home advantage and an adjustment for goal difference in the match result. Here, the smoothing coefficient K is weighted dependent upon the type of tournament played. Where, K equals

- 60 for World Cup finals;
- 50 for continental championship finals and major intercontinental tournaments;
- 40 for World Cup and continental qualifiers and major tournaments;
- 30 for all other tournaments;
- 20 for friendly matches.

K is then adjusted for the goal difference (GD) in the game, such that K increases as the goal difference (GD) increases. Let A denote the adjustment which is multiplied to K. Here,

$$A = \begin{cases} 1, & \text{if GD=1} \\ \frac{3}{2}, & \text{if } GD\text{=}2 \\ \frac{11+GD}{8}, & \text{if } GD > 2 \end{cases} \tag{3.37}$$

A variant of Elo ratings has also been adapted to FIFA women's Association football. A noteworthy feature of the FIFA women's system is that all wins are not scored as 1 and all losses are not score as 0; score difference is converted to a 1-0 scale of fractional win and loss. For more information on Elo ratings in sport see Stefani (2010).

3.6 Computer Programming

This section covers the computer programming component of this dissertation. The first section covers VBA programming which is used at various stages of this dissertation. The next section covers the programming languages Perl and MySQL which are used in Chapter 6.

3.6.1 VBA Programming

Microsoft Excel is a commonly used spreadsheet application which has many features including, but not limited to, calculation, graphing tools and pivot tables. However, of greater importance to this dissertation is the macro programming language known as Visual Basic for Applications (*VBA*). This allows users to manipulate spreadsheets in ways that is not possible via manual spreadsheet techniques. One of the easiest ways to generate *VBA* code is utilising the macro recorder, this records all interactions and converts them into *VBA* code contained within a macro. However, there are many other features that can't be recorded and must be manually entered into the *VBA* module directly by the programmer. For example, loop functions, screen prompts and many graphical display items. These macros can then be implemented via a button or keyboard shortcut.

(Note: Although I was relatively proficient user of EXCEL, *VBA* programming was completely unfamiliar territory when I first started my PhD. I taught myself *VBA* programming through trial and error in small “baby steps”. Throughout my PhD candidature I realised how little I previously new about Excel and the power of the programming and graphing capabilities in Excel. The fruits of this labour are predominately showcased in Chapter 11)

3.6.2 Perl and MySQL

Perl is a high level, general purpose, open source (software), dynamic programming language with over 20 years of development. Perl includes smart tools for text processing that make it ideal for working with HTML, XML, and all other mark-up and natural languages. Perl’s Database Integration Interface (DBI) supports third-party databases including Oracle, Sybase, Postgres, MySQL and many others.

MySQL is arguably the worlds most popular Relational Database Management System (RDBMS). MySQL works on many different system platforms including Linux, Mac OS X and Microsoft Windows to name just a few. All major programming languages with language-specific APIs include Libraries for accessing MySQL databases. The MySQL server and official libraries are mostly implemented in ANSI C/ANSI C++.

(Note: Throughout early 2008, my supervisor and I started collecting in-play betting odds for AFL matches. This required manually recording the odds at quarter time, half time and three quarter time, this was not only time consuming but extremely frustrating when you are watching a delayed telecast (the odds typically give a good indication of who is winning). Therefore, we both thought there must be a more efficient method for collecting the data we required. My supervisor stumbled upon a book (Magee, 2008) which explained how to automate the collection of horse racing betting odds. So I proceeded to read this book from cover to cover (several times), bought an EEE PC with a linux operating system and enhanced my extremely limited knowledge of Perl and MySQL. Four months later I had achieved what I set out to achieve which was to develop a program to automatically record in-play betting data for AFL matches with minimal human intervention. See Chapter 6 for more details.)

3.7 Summary

In summary, this chapter has defined methodology that is utilised in subsequent chapters. In Chapter 4, a linear regression model was fitted to individual match margins to quantify home advantage in AFL. In Chapter 5, an optimised Elo ratings model was used to forecast match results. In Chapter 6, a computer program was developed using Perl to integrate seamlessly with Betfair's Application Programming Interface (API) to automatically record in-play betting data to a MySQL database. In Chapter 7, the in-play betting data obtained in Chapter 6 is transformed to normalized implied probabilities and plotted against time to give a graphical real-time measure of expectation. Furthermore, the procedure for generating this graph is automated using macros (VBA programming) in Excel. In Chapter 7, the efficiency of in-play betting markets were tested using the *Efficient Market Hypothesis* (EMH) which incorporates a logistic regression component. In Chapter 9, Analysis of Variance (ANOVA) was used to quantify the intra-match home advantage in AFL. In Chapter 10 a generalised logistic function was optimised for in-game prediction of AFL matches. In Chapter 11, logistic regression was used to transform a mass of real-time performance variables to a single probability assessment which was plotted against time. Furthermore, the procedure for generating this graph was automated using macros (VBA programming) in Excel.

Part I

Pre-Game

Chapter 4

Home Advantage

In this chapter, a new paradigm is proposed to quantify the precise cause of home advantage in AFL. In Section 4.1, a brief introduction on home advantage in sport is given. Section 4.2 details the independent factors that are thought to contribute towards home advantage in AFL. These factors include psychological (crowd support and stadium density), physiological (distance travelled and origin of away team) and tactical (ground familiarity). Territoriality effects (i.e. hormonal increase playing at home) and referee bias are difficult to quantify due to their subjective nature, therefore their effect are subsumed under tactical and psychological factors. In Section 4.3, a multiple linear regression model is proposed to quantify the contribution of each factor towards home advantage. Then in Section 4.4 the results are discussed including quantifying the average home advantage (and disadvantage) each team received over the previous decade. Material from this chapter has been published in Ryall and Bedford (2011a).

4.1 Introduction

In predicting the outcome of AFL matches it has been shown that home advantage plays an important role as well as the quality of the two competing teams (Stefani and Clarke, 1992). Home advantage typically refers to the net advantage of several factors which, generally speaking, have a positive effect on the home team and a negative effect on the away team (Harville and Smith, 1994). The much acclaimed paper by Schwartz and Barsky (1977) on home advantage in team sports (major league baseball, college and professional football, professional ice hockey, and college basketball) showed its existence and how it varied from one sport to another. Since their work, home advantage has been extended to other sports (for example, Pollard (1986) on soccer; Holder and Nevill (1997) on tennis and golf; Jones et al. (2005) on Rugby; Clarke (2005) in the AFL). A comprehensive literature review on home advantage in sport is provided in Nevill and Holder (1999).

The seminal paper by Clarke (2005) quantified home advantage in AFL by fitting various linear regression models to individual match margins. The results suggest that although a unique home advantage for each team may not be necessary, there was overwhelming evidence to suggest there is a difference between home advantage for Victorian and non-Victorian teams. The author suggests that this lends support to the notion that ground familiarity and crowd support are major determinants of home advantage in AFL, however there is no empirical evidence to support this subjective statement.

Since the analysis of Clarke (2005), which was based on seasons 1980 to 1998, much has changed in AFL. For example, a new Victorian venue "Docklands" has been introduced which has a retractable roof which is closed at the AFL's discretion. Furthermore, the training venues of Victorian teams (excluding Kardinia Park) have all been phased out with the intent to maximise crowd capacity at the MCG and Docklands. Additionally, in order to increase the popularity of AFL, matches are occasionally played outside of a team's home state or territory. These matches are predetermined by the league and the AFL clubs involved. For example, in 2007 and 2008 the Kangaroos sold a total of six home matches to be played at Carrara Stadium in Darwin in a deal believed to be worth $400,000 a match. It should be noted that teams which sell home matches to the AFL (or other clubs) are still the nominated home in these matches. In essence these teams surrender any home advantage and are rewarded financially by the league (or other clubs). This suggests that quantifying home advantage in AFL is more complex than ever before.

4.2 Independent Effects

Schwartz and Barsky (1977) proposed three explanations as to why home advantage may exist: learning/familiarity (tactical) factors, travel (physiological) factors and crowd (psychological) factors. Courneya and Carron (1992) build on this suggesting referee bias as another factor to consider. Although these factors are usually cited as the cause of home advantage in team sports, the precise contribution of each factor still remains relatively unknown (Pollard, 2008).

Bailey and Clarke (2004) realised this deficiency in the literature by endeavouring to attribute the relative contribution of travel and familiarity factors towards home advantage in AFL. The authors found performance of the nominated home team increased as the difference in all matches ever played at the home venue increased (ground familiarity). Furthermore, when the nominal away team traveled interstate the authors found the further they traveled the greater the disadvantage. However, Courneya and Carron (1991) and Pace and Carron (1992) discussed how team ability, ground familiarity, travel fatigue and crowd intimidation affected performance simultaneously. To overcome this problem of confounding variables, they used a multiple regression model with each of the predictor variables (such as number of time zones crossed and distance travelled) entered as both main effects and two way interactions. Akin to Courneya and Carron (1991) and Pace and Carron (1992), this chapter attempts to disentangle the contributing factors of home advantage in

AFL. This is undertaken by firstly defining factors that can be quantified and could contribute to home advantage in AFL. These factors include psychological (crowd support and stadium density), physiological (distance travelled and origin of away team) and tactical (ground familiarity). Then the contribution of each factor towards home advantage is deduced by utilising a multiple linear regression model on margin of victory which is adjusted for any difference in team quality.

4.2.1 Ground Familiarity

The concept of ground familiarity and its existence in team sports is typically contextualized as the percentage of games won at venues with dissimilar attributes. Dowie (1982) commented on the significant variation in Association football pitches, however the teams with the largest (Manchester City and Carlisle) and smallest (Bristol Rovers and Halifax Town) playing areas yielded a similar advantage to the rest of the competition. However, in AFL, ground familiarity is typically referred to as how many games each team play at a specific venue (Bailey and Clarke, 2004) since each team can play multiple games at the same away venue.

Pollard (2008) also suggests territoriality as a factor which can influence home advantage. He states that humans (like animals) respond to a real or perceived threat of an invasion of their home territory and this in turn responds to an increase in hormonal activity (Neave and Wolfson, 2003). Therefore, if territoriality did indeed exist in AFL it would fall under ground familiarity. For example, the more games a team plays at a specific venue the greater the sense of ownership and possible increase in hormonal activity.

A novel feature in the paper by Clarke (2005) was the Melbourne Cricket Ground (MCG) teams effect for Victorian clubs that used the MCG as their home ground. It was thought that clubs using the MCG as their home ground should be at a *disadvantage* to other Victorian clubs for a number of reasons. One explanation is that the mix of supporters in the crowd is likely to be much more equal (more neutral supporters) due to a significant portion of the ground being allocated to Melbourne Cricket Club (MCC) and AFL members. Another explanation is the number of matches played at the MCG is more than any other venue (thus other clubs become familiar with the ground). They found Victorian clubs which used the MCG as their home ground received an average 2.8 point advantage (over other Victorian teams) whilst the other Victorian clubs received an average 9.5 point advantage (over other Victorian teams) which was statistically significant ($p = 0.01$). This suggests there is a ground familiarity factor to consider in AFL.

Bailey and Clarke (2004) defined ground familiarity in AFL as three subsets of difference in experience between the two teams at a given venue. This difference in experience is based on the *historical* difference in the number of times the two competing teams have played at the given venue. For example, less than 10 matches experience was worth +3.8 points, between 10 and 50 matches difference was worth +7.1 points, and greater than 50 matches difference was worth +10.2 points. However, using this methodology over a number of years would result in some Victorian teams who play frequently at the MCG (Collingwood) or Docklands (St Kilda) accumulating greater experience at the venue even though the makeup of the team (i.e. the players) is likely to be vastly different. Therefore, in this section a new method is proposed to quantify the contribution of ground familiarity towards home advantage in AFL. To build upon the work of Bailey and Clarke (2004), the number of games each team plays at a given venue is compared to the opposition *within* each season. The ground familiarity factor (GF) is given by

$$GF = f_{i,j,k,l}(home) = \frac{\sum_{k=1}^{22} g_{i,k,l}(home)}{\sum_{k=1}^{22} g_{i,k,l}(home) + \sum_{k=1}^{22} g_{i,k,l}(away)} \tag{4.1}$$

where k is the weekly index of games team i plays at venue g relative to team j for each season l.

For example, in 2009 Essendon played nine matches at Docklands Stadium and Melbourne played only three. Therefore, in round 12 when they played against each other at Docklands Stadium, Essendon received a 0.5α advantage $\left(\frac{9-3}{9+3}\right)$, where α is the unknown GF

coefficient.

Although the schedule varies considerably from year to year, the number of games each team plays in their home state remains largely unchanged. For example, each Victorian team typically has a primary home ground which they play the majority of their home games (6+) and a secondary home ground where the play their remaining home games (2+). Therefore, a teams experience at a specific venue will not increase from one season to the next unless they play more games at that venue in a single season. Table 4.1 displays which venue the 16 teams played all of their games (nominated home games in brackets), which state/territory the venue is in and the home state/territory of each team for the 2007 AFL season.

Team (State)	VIC			NSW		QLD		SA	WA	ACT	TAS	NT
	MCG	D	KP	SCG	SA	G	CS	FP	S	MKO	YP	MRO
Adelaide (SA)	2	4	1	0	0	1	1	12(11)	1	0	0	0
Brisbane Lions (QLD)	2	3	1	1	0	12(11)	1	1	1	0	0	0
Carlton (VIC)	8(5)	8(6)	0	1	0	1	1	1	2	0	0	0
Collingwood (VIC)	15(9)	3(2)	0	0	1	1	0	1	1	0	0	0
Essendon (VIC)	8(4)	9(7)	0	1	0	0	0	2	2	0	0	0
Fremantle (WA)	2	2	1	1	0	1	0	2	12(11)	0	0	1
Geelong (VIC)	2(1)	7(2)	8(8)	0	0	1	0	2	1	0	1	0
Hawthorn (VIC)	8(7)	5	0	1	0	1	0	2	1	0	4(4)	0
Kangaroos (VIC)	3(2)	10(6)	1	0	0	1	3(3)	1	2	0	1	0
Melbourne (VIC)	12(8)	4(1)	0	1	0	1(1)	0	1	2	1(1)	0	0
Port Adelaide (SA)	3	2	1	1	0	1	0	12(11)	1	0	1	0
Richmond (VIC)	12(8)	4(3)	1	1	0	0	0	2	2	0	0	0
St. Kilda (VIC)	4	13(11)	0	0	1	1	0	1	2	0	0	0
Sydney (NSW)	3	2	1	8(8)	3(3)	1	0	1	1	2	0	0
West Coast (WA)	1	4	1	0	1	0	0	2	12(11)	0	1	0
Western Bulldogs (VIC)	5(1)	12(8)	0	0	0	1	0	1	1	1(1)	0	1(1)

Note. MCG = Melbourne Cricket Ground, D = Docklands, KP = Kardinia Park , SCG = Sydney Cricket Ground, SA = Stadium Australia, G = Gabba, CS = Carrara Stadium, FP = Football Park, S = Subiaco, MKO = Manuka Oval, YP = York Park, MRO = Marrara Stadium VIC = Victoria, NSW = New South Wales, QLD = Queensland, SA = South Australia, WA = Western Australia, ACT = Australian Capital Territory, TAS = Tasmania, NT = Northern Territory.

Table 4.1: Home grounds of AFL teams, 2007

4.2.2 Travel Fatigue

The concept of travel in AFL is somewhat different to other international team sports due to some "visiting" teams not having to travel interstate to the required destination (for example, two Victorian teams, West Coast vs Fremantle or Adelaide vs. Port Adelaide). Therefore, the common "on the road" advantage (or disadvantage) which accounts for the current sequence of home or away games in American sports (Bedford and Baglin, 2009) is virtually redundant, since only a few non-Victorian team travel back to back in consecutive rounds. There has been conflicting evidence to support the notion that as distance travelled increases, so too does the detrimental effect of home advantage. Pollard (1986) found distance traveled to be insignificant, with home teams in basketball having a similar home winning percentage for visiting teams travelling, regardless of whether they travelled more or less than 200 miles. However, Bailey and Clarke (2004) found visiting teams in AFL that travelled more than 1500kms were disadvantage by an additional seven points compared to teams travelling less than 1500kms.

It is widely assumed that West Coast and Fremantle have the greatest home advantage in AFL, since any visiting team has to travel at least 2160 kilometres. To test whether there is a difference in the magnitude of home advantage when the away team is from a different State to the home team, a binary variable TRAVEL was introduced ($TRAVEL$=0: away team is from the same State, *TRAVEL=1*: away team is from a different State). Furthermore, since non-Victorian teams travel interstate nearly three times as frequently as Victorian teams they might become accustomed to travelling. Therefore, the binary variable VIC is used to differentiate between Victorian teams travelling interstate (VIC=1) and non-Victorian teams travelling interstate (VIC=0). Additionally, to see what effect total distance travelled has on home advantage the variable $DIST$ is introduced. This is simply the distance travelled between the major cities of the home and away team's state or territory. In the case of both teams travelling interstate $DIST$ is simply the difference between the distance travelled by the home and away side. Since the difference between the shortest distance (250kms) and

State	ACT	NSW	NT	QLD	SA	TAS	VIC	WA
ACT	-							
NSW	250	-						
NT	3133	3147	-					
QLD	1643	733	2846	-				
SA	958	1161	2616	1600	-			
TAS	857	914	3734	1643	1039	-		
VIC	462	712	3147	1373	654	443	-	
WA	3087	3288	2651	3604	2130	2961	2721	-

Table 4.2: Distance in kilometres between each state/territory in Australia

the largest distance (3604kms) travelled is a multiple of more than ten, a log transformation was applied to *DIST*. Table 4.2 displays the direct distance between each state/territory measured from the major city of the respective state/territory.

4.2.3 Crowd Intimidation

The importance crowd intimidation plays in contributing to home advantage in team sports has received positive and negative support. Schwartz and Barsky (1977) found crowd density in Major League Baseball (MLB) increased home winning percentage from 48% when the crowd density was small (less than 20% full capacity) to 57% when crowd density was large (greater than 40% full capacity). Two other studies on crowd density (Pollard, 1986) and absolute crowd size (Dowie, 1982) found no significant difference in home advantage across four divisions in the English soccer league. A study by Thirer and Rampey (1979) investigated the effect crowd support had on the number of fouls and turnovers in college basketball. They found normal crowd behaviour resulted in the away team committing more infractions (i.e. committed more fouls and lost more possessions or turnovers), however antisocial crowd behaviour (i.e. swearing) was detrimental as the home team committed more infractions. However, there is likely to be a causal relationship between in-game team performance and the behaviour of the crowd. For example, a crowd will likely chant obscenities *because* their team is performing poorly.

Another important aspect to consider in calculating the magnitude and significance of crowd intimidation in any team sport is the mix of crowd support; that is, the breakdown of home, away and neutral supporters. This is particularly important in AFL due to the likelihood of crowd support being much more even when the home and away team are from the same state. Any discrepancies in crowd support between two teams from the same state could be attributed to the difference between club members from each team and how well each team is performing in the current year. Biddle (1993) showed that team success was highly correlated with attending matches. However, crowd support is likely to be highly biased towards the home team in all other cases. For example, one would expect non-Victorian teams to have a bigger following in Victoria than Victorian teams have interstate for two reasons. Firstly, non-Victorian teams play more games in Victoria than a Victorian team plays in any other one state or territory. Furthermore, a Victorian team supporter living interstate could follow any one of the 10 Victorian teams whereas a non-Victorian team supporter living in Victoria has a maximum of two teams to choose from any other state.

Although the breakdown of crowd support is always unknown even after a match (unless the crowd is audited), one can estimate the expected number of home and away supporters using a regression model. Borland and Lye (1992) predicted the attendance of AFL matches for seasons 1981 to 1986 (note that all teams were based in Victoria during this period), the only factor which was team dependent and significant was the rating of each team. Akin to Borland and Lye (1992), three team dependent factors were considered: the number of members from the previous year, the rating of each team (defined later) and the number of games each team plays at each state (Geelong is not grouped with Victoria due to Kardina Park being approximately 80 kilometers outside of Melbourne) as a function of total games. The state variable is replaced with 1 if either team is playing in their home state (Geelong is defined as playing in their home state in both Geelong and Victoria due to a large following in both cities). Each of these factors are split by home/away and travel/no travel, since the majority of club members for example, will not travel interstate to watch their team play. Table 4.3 shows the results of the model defined in (4.2) where the home team is defined as the nominated home team according to the AFL schedule.

$$\begin{aligned} CROWD &= \beta_1 SH + \beta_2 RNTH + \beta_3 RTH + \beta_4 MNTH + \beta_5 MTH \\ &+ \beta_6 SA + \beta_7 RNTA + \beta_8 RTA + \beta_9 MNTA + \beta_1 0 MTA \end{aligned} \tag{4.2}$$

State	Coefficient	p-value
SH	7223.346	<0.001
RNTH	137.789	<0.001
RTH	45.053	0.493
MNTH	0.607	<0.001
MTH	0.303	<0.001
SA	9615.742	<0.001
RNTA	165.875	<0.001
RTA	69.296	<0.001
MNTA	0.361	<0.001
MTA	0.100	0.001

Note. *SH* = state home; *RNTH* = rating no travel home; *RTH* = rating travel home; *MNTH* = members no travel home; *MTH* = members travel home; *SA* = state away; *RNTA* = rating no travel away; *RTA* = rating travel away; *MNTA* = members no travel away; *MTA* = members travel away.

Table 4.3: Linear regression results: Estimated home/away supporters, 1997 to 2008

The regression model explained an astonishing 93.41% of the variation in crowd attendance, with only the rating of the home team if they are travelling statistically insignificant. Therefore, the estimated number of home supporters is simply the sum product of the home coefficients and variables. Similarly, the estimated number of away supporters is simply the sum product of the away coefficients and variables. Now the crowd intimidation factors used in the full home advantage model are defined as the difference between the estimated number of home and away supporters ($CROWD$); and the difference between the estimated number of home and away supporters divided by crowd capacity of the designated venue ($DENS$). Restrictions are placed on $CROWD$ such that it does not exceed the crowd capacity of the venue ($CROWD <= CAPACITY$), similarly $DENS$ can not exceed one ($DENS <= 1$). An assumption with this method is the number of neutral supporters does not increase when the ratings of either team increase, however it is acknowledged that this is not always likely to be the case.

4.2.4 Referee Bias

It is widely perceived that referees have a tendency to favour the home side in team sports. Some examples might include the referee being coming from the same city or country as the home team, and possible intimidation from the home crowd for favourable decisions. A study by Dohmen (2008) showed compelling evidence that referees may be crowd pleasers who, for example, award more extra time at the end of each half if the home team is not winning. However, in AFL, extra time (time on) for each quarter is not at the umpire's discretion. Every time the ball is not in-play (i.e. a goal is scored) the clock is paused and restarted by the time keeper once the ball is back in-play and this time is not known by the umpire. Nevill et al. (1996) showed that the officials in English and Scottish soccer leagues favoured the home team when awarding free kicks. However, the mere fact that the home team receives more free kicks than that of the away team does not prove referee bias exists. A number of studies including the work of Sumner and Mobley (1981) recognised that this association could be attributed to the differing playing styles of the two competing teams. For example, the away team might spend more time defending and thus naturally incur more penalties. Contrary to other sports, officials in AFL are not based in their home city; they are in fact rotated throughout the country hence attempting to remove any favouritism. Therefore, if referee bias does indeed exist, it will be attributed to *CROWD* or *DENS* (i.e. pressure from the home crowd for favourable decisions).

4.3 Methods

This chapters analysis is based on seasons 1997 to 2008. AFL data was collected from AFL tables (http://stats.rleague.com/afl/) which consisted of year, round, (nominal) home team, away team, ground and home team winning margin. The distance between each state/territory measured from the major city was extracted from Geoscience Australia (http://www.ga.gov.au), and the membership numbers of all AFL clubs during this period was taken from AFL tables (http://stats.rleague.com/afl/).

There are several models that can be used to analyse the results of games between two teams. Clarke (2005) lists several of these models which are common in the literature. Perhaps the most common model for predicting match outcomes allows for a common home advantage for all teams (Stefani, 1983, 1987; Stefani and Clarke, 1992; Clarke, 1993) which is given in (4.3):

$$a_{ij} = r_i - r_j + h_{ij} + e_{ij} \tag{4.3}$$

where a_{ij} is the actual margin of victory of team i against team j, r_i is the rating of team i, r_j is the rating of team j, h_{ij} is the home advantage team i receives against team j and e_{ij} is a zero mean random error.

Assuming home advantage is constant for all teams, it can be derived by minimizing the sum of the squared errors in (4.3). Therefore, the least squares value for a single home advantage (h) which minimizes (4.3) for the M games at home is given in (4.4) (Stefani and Clarke, 1992).

$$h = \frac{1}{M}\left[\sum_{i=1}^{n}\sum_{m=1}^{k} a_{ij}(m) + \sum_{i=1}^{n}\sum_{m=1}^{k}(r_j^m - r_i^m)\right] \tag{4.4}$$

In a balanced schedule where each team plays each other team an equal number of times home and away, the right hand double summation in (4.4) will tend towards zero. An example of a balanced schedule is the English Premier League (EPL) where there are 20 teams and 38 rounds so every team plays each other team once at home and once away. Therefore, in the case of a balanced schedule, home advantage can be calculated independently of team ratings. However, the AFL schedule is unbalanced (16 teams and 22 rounds) so it is important to control for the quality of the two competing teams (Clarke, 2005).

Since this research focuses on quantifying home advantage rather than the development of a ratings system, the ratings for all teams are simply based on the average margin of victory for each team split by season. These ratings are then retrospectively fitted in (4.3). For example, in round 7, season 2000, Essendon defeated Collingwood by +40 points (a_{ij}), Essendon's average winning margin in season 2000 was +47.5 points (r_i), similarly Collingwood's average winning margin in season 2000 was -15.5 points (r_j). Table 4.6 shows the average margin of victory for each team across seasons 1997-2008.

Stewart et al. (2007) used ordinary least squares regression in an attempt to identify elite AFL players using margin of victory as the response variable and 51 predictor variables. The initial model used all 51 variables, groups of variables that were found to be insignificant were then removed and the regression model was re-run. This was completed a number of times until all the remaining variables were significant.

A similar approach to that of Stewart et al. (2007) was used in this analysis. The initial home advantage model accounts for ground familiarity (GF), travel fatigue [$TRAVEL$, VIC and ln($DIST$)], crowd intimidation ($CROWD$ and $DENS$) and the ratings of the two competing teams all of which were defined previously, is given given below:

$$h_{i,j} = \alpha_1 GF + \alpha_2 TRAVEL + \alpha_3 VIC + \alpha_4 ln(DIST) + \alpha_5 CROWD + \alpha_6 DENS \tag{4.5}$$

The regression model is then re-run, removing the most insignificant variable from the previous stage until all remaining variables are significant ($p < 0.05$). Firstly, it is important to have an understanding of the relationship between each of the independent variables due to the high level of multicollinearity. Table 4.4 shows a correlation matrix of the independent variables.

	GF	TRAVEL	VIC	ln(DIST)	CROWD	DENS
GF	1.0000					
TRAVEL	0.6708	1.0000				
VIC	0.4396	0.5014	1.0000			
ln(DIST)	0.6906	0.9334	0.4531	1.0000		
CROWD	0.6558	0.7992	0.5601	0.7770	1.0000	
DENS	0.6547	0.6950	0.5111	0.7018	0.8688	1.0000

Table 4.4: Correlation matrix of independent predictor variables of home advantage

The results are as expected with a strong correlation between the crowd factors ($CROWD$ and $DENS$) and also the travel factors ($TRAVEL$ and $ln(DIST)$). Table 4.5 shows the results of the regression model defined in (4.5).

Stage	State	Coefficient	p-value
1	GF	5.950	0.005
	TRAVEL	-1.130	0.610
	VIC	6.243	0.185
	ln(DIST)	0.663	0.254
	CROWD	-0.000	0.276
	DENS	1.500	0.808
2	GF	6.013	0.005
	TRAVEL	6.087	0.192
	VIC	-1.087	0.623
	ln(DIST)	0.684	0.234
	CROWD	-0.000	0.210
3	GF	6.103	0.004
	TRAVEL	5.772	0.211
	ln(DIST)	0.709	0.216
	CROWD	-0.000	0.143
4	GF	4.687	0.013
	TRAVEL	3.677	0.402
	ln(DIST)	0.617	0.279
5	GF	5.133	0.005
	ln(DIST)	1.047	0.001

Table 4.5: Stepwise regression results: Predictors of home advantage, 1998 to 2008

Firstly, note the coefficient of $CROWD$ is extremely small in all stages where it is present. This is due to the variable $CROWD$ representing the difference between two teams predicted number of supporters (which could be upwards of 40,000 in some instances). Therefore, a small coefficient for $CROWD$ is logical since the outcome variable is the adjusted margin of victory given in (4.5). Secondly, due to many variables having similar p-values in Stage 3, different combinations of the predictor variables were trialed in Stage 4 which resulted in the two most significant predictors in Stage 5.

The final model explains 6.82% of the variation in margin of victory adjusted for any differences in team quality. At first thought, the amount of variation explain seems quite insignificant, however it is a significant improvement over other home advantage models. Courneya and Carron (1991) found home advantage to explain less than 1.2% of the variation in win/loss for basketball. Applying a non-Victorian and Victorian home advantage as stated in Clarke (2005) explains 5.3% of the variation, yielding a 6.5 point advantage for Victorian teams; and a 12 point advantage for non-Victorian teams (similar to the results in Table 4.7). A direct comparison of the two models is appropriate since models use the same number of predictor variables. The model defined in Clarke (2005) can be criticised in that it rewards a Victorian team playing at home equally, regardless of which state the visiting team is coming from.

Team	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Adelaide	17.4	18.6	-15	-4.2	2.7	13.7	16.4	-4	24.9	31.5	7.7	8.1
Brisbane Lions	4.7	-26.7	34.1	17.3	25	30.8	18.8	30.2	-1.1	-18	4.6	-2
Carlton	-3	-4	2.7	31.3	23.4	-28.1	-40.5	-18.6	-29.7	-28.4	-33.8	-6.2
Collingwood	10	-9	-16	-15.5	6.5	8.4	18.2	-8.3	-24.6	17.3	0.9	10.4
Essendon	-7.5	8.1	22.5	47.5	29.7	4.2	10.4	2.5	-8.4	-20.4	-9.5	-21.7
Fremantle	-7	-24.5	-19.2	-33.3	-31.7	-11.4	2.9	0.5	0.4	8.5	2.5	-6
Geelong	14.5	-8.5	-5.7	-3.3	-5.8	-4.4	-9.4	15.8	10.4	-0.9	39.9	46.4
Hawthorn	-12.3	-4.1	-3.9	-2.4	4.9	-7.7	0.5	-32.1	-18.8	-13.9	11	26.7
Kangaroos	9.8	16.7	15.2	6.5	-9.5	-1.3	-1.7	0.3	-0.7	-18.8	8.4	-3
Melbourne	-43.5	2.4	-20.1	16.5	-10.4	-0.1	-20.2	10.3	-4.3	8.6	-24	-44.2
Port Adelaide	-7.5	-4	-9.2	-16.7	25.2	26.2	21.7	26.8	-1.7	-10.9	12.5	-4.1
Richmond	-16.8	4.2	-8.8	-7	7	-16.9	-10.5	-34.2	-7.6	-14.2	-26.3	-2.7
St. Kilda	17.1	2	-2	-35.3	-33	-22.1	-4.2	24.3	27.3	14.6	-3	9.2
Sydney	13.5	6.1	2.5	1.6	13.1	6.7	12.7	6.1	12.6	21.3	15.1	10.5
Western Bulldogs	1.7	15.2	16.8	3.6	-7	4	-30.8	-22.8	1.5	6.3	-16.3	17.9
West Coast	9	7.6	6	-6.7	-40.1	-2.1	15.6	3.4	19.9	17.4	10.3	-39.3

Table 4.6: Average winning margins, 1997 to 2008

4.4 Results

The results suggest there are only two statistically significant predictors of home advantage in AFL, ground familiarity and distance travelled by the visiting team. This confirms the popular hypothesis that West Coast and Fremantle could have the greatest home advantage. This chapter provides objective agreement with the subjective statement made by Stefani (2008) that “the large size playing oval in Australian Rules Football probably reduces the crowd’s psychological influence, compared to rugby union, soccer and the NBA which also have a large percentage of the ball being in play” (Stefani 2008, p. 212); and objective disagreement with the subjective statement made by Clarke (2005) that due to non-Victorian teams having a greater home advantage than Victorian teams. This lends support to the notion that crowd effects and ground familiarity are the major determinants of home advantage. It is worth noting that crowd intimidation in realistic terms is not an inter-game measure but rather an intra-game measure which may depend upon current state (score) in the match. This idea is further explored in Chapter 9. Table 4.7 displays the average home advantage for the nominated home team.

Team	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	Average
Adelaide	10.73	10.64	10.86	10.81	10.78	10.78	10.72	10.86	10.75	10.89	10.75	10.78
Brisbane Lions	12.33	12.48	12.48	12.36	12.44	12.36	12.36	12.35	12.44	12.36	12.41	12.40
Carlton	6.23	6.22	6.71	7.00	5.27	6.34	5.67	5.61	5.16	4.67	6.65	5.96
Collingwood	6.03	6.03	6.01	5.99	5.40	6.87	6.75	5.90	7.07	6.58	6.22	6.26
Essendon	7.13	5.68	5.66	6.52	5.74	5.78	6.58	4.82	5.86	5.79	4.54	5.83
Fremantle	11.57	11.67	11.66	11.87	11.84	11.88	11.85	11.89	11.84	11.84	11.78	11.79
Geelong	7.57	6.82	7.06	5.58	6.74	6.85	6.72	6.77	6.56	8.54	7.30	6.96
Hawthorn	6.19	6.14	6.22	6.87	5.99	6.76	6.76	5.75	4.42	6.08	6.78	6.18
Kangaroos	6.23	5.39	5.16	5.01	4.62	4.68	5.29	4.39	5.22	4.41	4.50	4.99
Melbourne	6.04	7.00	5.93	4.30	4.88	5.29	4.33	5.05	5.06	3.88	5.09	5.17
Port Adelaide	9.29	10.75	10.69	10.72	10.72	10.86	10.69	10.69	10.89	10.75	10.78	10.62
Richmond	6.30	5.63	5.93	5.91	6.86	4.77	5.33	6.76	5.62	5.83	5.20	5.83
St. Kilda	6.73	7.60	6.90	6.36	6.44	5.86	6.92	6.67	6.53	6.34	6.40	6.61
Sydney	11.82	11.87	11.70	12.00	11.53	11.63	11.48	11.67	11.53	11.59	11.62	11.68
Western Bulldogs	6.82	6.33	4.62	4.60	5.04	4.72	2.97	4.24	3.14	4.18	4.05	4.61
West Coast	11.56	11.66	11.61	11.91	11.88	11.81	11.85	11.88	11.91	11.93	11.88	11.81

Table 4.7: Home advantage by club, 1998 to 2008

Firstly, note the consistency of the non-Victorian teams home advantage. This is primarily due to non-Victorian teams nominal home ground having stayed constant over time (ground familiarity), and in the case of having a secondary home ground, this has always been in the same state (no travel). Similarly, the inconsistencies of the Victorian teams' home advantage can be attributed to the constant changing of venue(s) of their nominal home ground(s). Interestingly, Brisbane has the highest home advantage due to not having to share their nominated home ground with any other team. Although the non-Victorian teams have a greater home advantage they also travel interstate approximately every second week where they are at a significant disadvantage. Table 4.8 shows the away advantage for the nominated away team, that this, the average number of points each team is disadvantaged when they are the nominated away team.

Team	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	Average
Adelaide	-10.80	-10.66	-10.32	-10.45	-10.19	-10.74	-10.59	-10.08	-10.16	-10.06	-10.52	-10.42
Brisbane Lions	-11.88	-12.04	-11.84	-10.41	-9.96	-10.17	-10.30	-10.06	-8.46	-8.69	-10.28	-10.37
Carlton	-6.38	-5.43	-6.64	-6.12	-6.22	-6.49	-7.14	-6.16	-6.88	-7.19	-6.08	-6.43
Collingwood	-4.89	-6.23	-6.12	-5.17	-6.01	-5.80	-5.97	-5.81	-5.79	-6.07	-5.99	-5.80
Essendon	-4.41	-6.86	-6.09	-6.14	-6.74	-6.09	-5.87	-6.76	-5.98	-6.95	-5.41	-6.12
Fremantle	-11.43	-11.84	-11.06	-10.93	-11.37	-11.27	-11.47	-11.40	-11.49	-10.31	-10.13	-11.15
Geelong	-5.84	-6.96	-7.47	-8.21	-7.37	-6.66	-7.21	-6.37	-6.55	-6.26	-7.37	-6.93
Hawthorn	-8.15	-8.33	-7.05	-5.65	-6.11	-6.27	-5.37	-7.00	-6.43	-7.05	-6.60	-6.73
Kangaroos	-7.48	-4.99	-5.51	-6.50	-6.44	-6.79	-6.66	-6.74	-6.56	-6.49	-6.52	-6.43
Melbourne	-8.13	-5.79	-7.18	-7.12	-6.67	-6.20	-6.90	-6.40	-6.34	-6.27	-6.31	-6.66
Port Adelaide	-9.28	-10.82	-10.61	-10.69	-10.39	-10.46	-9.13	-10.61	-8.84	-10.62	-9.19	-10.06
Richmond	-6.52	-5.84	-6.02	-6.47	-5.82	-7.05	-6.91	-5.41	-7.24	-6.97	-6.76	-6.46
St. Kilda	-7.35	-6.99	-6.48	-7.73	-7.16	-6.46	-6.20	-5.48	-7.05	-7.23	-7.19	-6.85
Sydney	-11.49	-9.94	-8.42	-7.97	-8.11	-8.69	-8.79	-8.35	-8.56	-8.71	-8.47	-8.86
Western Bulldogs	-6.80	-7.47	-7.13	-6.81	-6.36	-6.46	-6.38	-7.23	-6.15	-5.16	-6.35	-6.57
West Coast	-11.72	-11.71	-11.29	-11.44	-11.27	-11.6	-11.43	-11.45	-11.50	-11.46	-11.68	-11.50

Table 4.8: Away advantage by club, 1998 to 2008

At first glance the away advantage seems be the exact opposite (i.e. negative) of the home advantage. For example, Adelaide had an average $+10.78$ advantage across seasons 1998 to 2008 when there were the nominal home team, and had an -10.42 disadvantage across seasons 1998 to 2008 when there were the nominal away team. Therefore, the total advantage the ground location had on Adelaide across seasons 1998 to 2008 was $+0.36$, which is a negligible advantage. However, this is not always the case. To isolate this differential, Table 4.9 shows the total advantage (= home advantage - home disadvantage) across seasons 1998 to 2008.

Team	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	Average
Adelaide	-0.07	-0.03	0.54	0.36	0.59	0.04	0.13	0.78	0.59	0.83	0.23	0.36
Brisbane Lions	0.45	0.44	0.64	1.95	2.48	2.19	2.07	2.29	3.98	3.68	2.12	2.03
Carlton	-0.15	0.79	0.07	0.88	-0.94	-0.15	-1.47	-0.55	-1.72	-2.51	0.57	-0.47
Collingwood	1.14	-0.20	-0.11	0.82	-0.60	1.07	0.78	0.10	1.28	0.52	0.22	0.46
Essendon	2.72	-1.18	-0.42	0.38	-1.00	-0.31	0.71	-1.94	-0.12	-1.16	-0.87	-0.29
Fremantle	0.13	-0.16	0.60	0.93	0.47	0.60	0.38	0.49	0.35	1.53	1.65	0.63
Geelong	1.73	-0.14	-0.41	-2.63	-0.62	0.19	-0.49	0.39	0.01	2.28	-0.07	0.02
Hawthorn	-1.96	-2.19	-0.83	1.22	-0.12	0.49	1.40	-1.25	-2.01	-0.97	0.18	-0.55
Kangaroos	-1.25	0.41	-0.35	-1.49	-1.82	-2.11	-1.37	-2.35	-1.34	-2.08	-2.03	-1.44
Melbourne	-2.10	1.21	-1.25	-2.82	-1.78	-0.92	-2.57	-1.34	-1.29	-2.38	-1.22	-1.50
Port Adelaide	0.01	-0.07	0.09	0.04	0.33	0.41	1.56	0.09	2.05	0.13	1.59	0.56
Richmond	-0.23	-0.21	-0.09	-0.56	1.03	-2.29	-1.58	1.35	-1.62	-1.14	-1.56	-0.63
St. Kilda	-0.62	0.61	0.42	-1.37	-0.72	-0.60	0.73	1.19	-0.52	-0.89	-0.78	-0.23
Sydney	0.33	1.93	3.29	4.03	3.41	2.94	2.69	3.32	2.96	2.87	3.14	2.81
Western Bulldogs	0.02	-1.14	-2.51	-2.20	-1.32	-1.74	-3.40	-2.99	-3.01	-0.98	-2.30	-1.96
West Coast	-0.16	-0.06	0.32	0.47	0.61	0.20	0.43	0.43	0.41	0.47	0.20	0.30

Table 4.9: Total advantage by club, 1998 to 2008

It becomes immediately evident that certain teams have an unfair schedule in terms of where the matches are played (i.e. what ground). This could be a deficiency of this model or indicate that the schedule is biased towards certain teams. For example, the Western Bulldogs surrender a -1.96 point deficit on average in each game which is attributed to them playing numerous nominal home games interstate for financial gain. Conversely, the Sydney Swans have gain a +2.81 point advantage each game that is attributed to them playing numerous nominal away games interstate, where the nominal home team is also playing interstate. For example, in 2007 Sydney played two matches at Manuka Oval in the ACT against the Western Bulldogs and Melbourne. These examples are regular occurrences in the AFL schedule. Therefore, it suggests that there are some clear deficiencies in the AFL schedule. Interestingly, the AFL has equalization policies in place (i.e. salary cap and drafting system) to make for a more even competition, however clubs that struggle financially are forced to sell home games, which has clearly been shown to be to the detriment of team performance in this Chapter. For more information on the analysis of the AFL schedule see Clarke (1998).

Chapter 5

Ratings

In this chapter, an Elo ratings model is adapted and then optimised to forecast AFL matches. To begin, Section 5.1 provides a brief introduction on rating systems with applications to sport. In Section 5.2, the importance of modifying previous seasons' ratings at the beginning of a new season is discussed. Section 5.3 details how the Elo ratings are adapted for AFL. Furthermore, Section 5.4 evaluates the results of the model based on various measures of performance including the reliability of the probability forecasts, number of predicted winners and average absolute margin of error. Then the applications to betting markets are discussed in Section 5.5. Material from this chapter has been published in Ryall and Bedford (2010c).

5.1 Introduction

In sporting competitions, it is of great interest to develop ratings models which accurately describe previous performances. In most team sports, crude systems are used to order teams based on absolute objective measures. For example, in Association Football teams are awarded zero points for a loss, one point for a draw and three points for a win. Teams are then ranked accordingly with teams on an equal number of wins (counting draws as a $\frac{1}{3}$ win) split by their respective goal differential. However, these win-loss systems are often subjected to criticism in their effectiveness of forecasting future results. This deficiency is primarily attributed to the assumption that each team plays a similar schedule which is not reflected in win-loss standings. This suggests a mathematically based system, which incorporates the quality of opposition, home advantage and the magnitude of victories (or defeats), is an appropriate method for measuring a team's true ability.

5.2 Initial Ratings

The computation of rating systems in sport requires some initial rating for each team at the beginning of each season. If the initial values are the same for all teams, then the ratings can't be expected to be reliable until a sufficient number of past results have been incorporated into the model (Hvattum and Arntzen, 2010). The next decision is to determine the initial ratings in subsequent seasons which are crucial for a variety of reasons. Firstly, the chosen values are virtually (excluding any home ground advantage) the sole predictor in the opening round of a competition (Clarke, 1993). In the English Premier League (EPL) there are no team or individual salary caps, which indicates every team should perform at a similar level from one season to the next. More specifically, although there is likely to be player transfers between seasons, the quality and depth of specific teams should remain somewhat similar. Therefore, the rating of each team at the conclusion of season n can be used as an initial rating for season $n + 1$.

However, in AFL there is a salary cap for total player payments for each club, and the drafting system helps reward poorly performing teams by awarding them early draft choices. These equalization policies assist in making a more even competition. Therefore, on average, we expect the stronger teams to get weaker and the weaker teams to get stronger. Authors such as James and Stein (1961) suggest using team ratings from the end of the previous year but shrinking them towards the mean. For example, suppose Essendon had a rating of 1900 at the end of season 2000 where 1500 is the league average. Essendon's initial rating at the beginning of season 2001 using a factor of $\frac{1}{2}$ would be $(1900 - 1500)/2 + 1500 = 1700$.

The Elo rating system is designed to be self correcting, which means its easier for a higher rated team to lose points than gain points and vice versa for low rated teams. Therefore, if the initial ratings are not an accurate measure of a teams ability, as the season progresses the Elo ratings system will self correct towards a teams true ability. However, this can be a very slow process particularly when a team follows a very good season with a very poor season (or vice versa). For example, in 2007 West Coast won 15 matches (out of 22) and finished third (out of 16) at the conclusion of the regular season. However, in the following season West Coast won only four matches and finished second last. Although the Elo ratings will self correct, the choice of initial ratings in this instance is likely to overpower predicted match results, particularly early in the season. Bailey (2000) stated that the predictive power of his model increased with the number of rounds played and therefore betting was restricted until the commencement of the fourth round of the season.

Therefore, in this section a separate ratings algorithm for every round of the season is proposed, whereby the initial ratings for each round are subject to a decay model whilst all other variables are kept constant. So, as the season progresses, the initial ratings are smoothed out entirely, leaving ratings based solely on current season results. Applying this to the AFL, premiership points (win = 4 points, draw = 2 points and loss = 0 points) from the previous year are utilised as the input subject to the following decay model:

$$S_k = N_0 e^{\nu k} \tag{5.1}$$

where S_k is the season multiplier for the initial ratings for round k ($k = 1, 2, \ldots, 22$ rounds), N_0 and ν ($\nu \in [-\infty, 0)$) are unknown coefficients which need to be optimised.

Returning to the previous example where Essendon had a rating of 1900 at the end of season 2000, now let $S_1 = 0.8$ and $S_{18} = 0.1$. Therefore, the initial rating for round 1 season

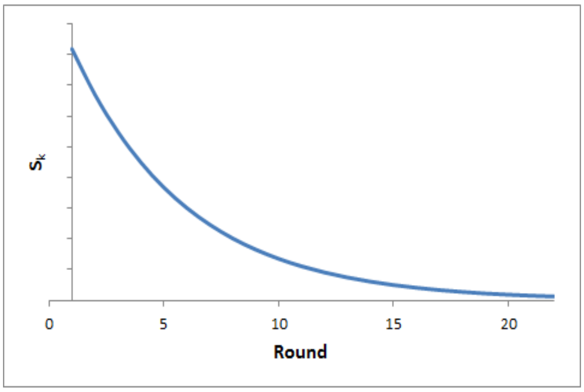


Figure 5.1: Initial ratings decay model

2001 is $0.8 \times (1900 - 1500) + 1500 = 1820$ and the initial rating for round 18 season 2001 is $0.1 \times (1900 - 1500) + 1500 = 1540$. It is clear that as the current season progresses, less weight is given to the previous seasons results. Figure 5.1 illustrates how the initial ratings multiplier, as defined in (5.1) decreases as the season progresses.

The importance of adjusting the initial ratings are showcased in Table 5.1, where the Elo ratings for West Coast's 2008 season using two different methods is displayed. Note that home advantage in Table 5.1 is measured in Elo points. The first model, which adjusts each team's initial rating, correctly classified 16 out of 22 matches, and had an AAE of 36.6 points. The second model, which keeps the initial rating of each team constant correctly predicted 14 matches and had an AAE of 37.7. It's interesting to note that the absolute error for the first six matches is smaller for model 1 than model 2; this indicates that adjusting a team's initial rating makes the current rating arguably more indicative of their true rating since less emphasis is placed on the previous season results.

		Score		Model 1: Adjusted initial rating						Model 2: Unadjusted initial rating					
Round	Opposition	For	Against	For (Initial)	Against	Home Advantage	Predicted Margin	Correct	Absolute Error	For	Against	Home Advantage	Predicted Margin	Correct	Absolute Error
1	Br	92	76	592	477	113	34	1	18	592	477	113	34	1	18
2	Ad	57	133	589 (580)	507	-107	-4	1	72	601	507	-107	-2	1	74
3	Fr	73	87	533 (569)	447	0	13	0	27	552	447	0	16	0	30
4	Sy	45	107	500 (560)	573	-102	-26	1	36	525	573	-102	-23	1	39
5	Po	104	128	471 (552)	449	102	19	0	43	499	449	102	23	0	47
6	Wb	74	134	432 (545)	589	-105	-39	1	21	463	589	-105	-34	1	26
7	Ca	74	111	413 (539)	402	110	19	0	56	446	402	110	23	0	60
8	Ka	83	89	367 (534)	516	-94	-36	1	30	400	516	-94	-32	1	26
9	Ad	97	47	359 (529)	594	107	-20	0	70	393	594	107	-14	0	64
10	Co	73	173	410 (525)	603	-102	-43	1	57	442	603	-102	-39	1	61
11	Sy	78	83	391 (522)	604	112	-15	1	10	421	604	112	-11	1	6
12	Es	91	113	378 (519)	303	-104	-4	1	18	409	303	-104	0	0	22
13	Ge	47	182	355 (516)	665	105	-31	1	104	384	665	105	-27	1	108
14	Ha	69	126	319 (514)	641	-98	-59	1	2	343	641	-98	-56	1	1
15	Ri	75	152	312 (512)	425	105	-1	1	76	336	425	105	3	0	80
16	Br	67	113	261 (511)	502	-112	-51	1	5	282	502	-112	-48	1	2
17	St	103	86	253 (509)	564	105	-31	0	48	274	564	105	-28	0	45
18	Fr	83	116	287 (508)	406	0	-18	1	15	307	406	0	-15	1	18
19	Es	113	103	267 (507)	482	105	-17	0	27	286	482	105	-14	0	24
20	Me	45	79	292 (506)	235	-101	-7	1	27	309	235	-101	-4	1	30
21	Ha	62	133	265 (505)	618	105	-37	1	34	281	618	105	-35	1	36
22	Ge	65	164	248 (504)	834	-107	-88	1	11	263	834	-107	-87	1	12
Total/Average								16	36.6					14	37.7

Note. Ad = Adelaide, Br = Brisbane, Ca = Carlton, Co = Collingwood, Es = Essendon, Fr = Fremantle, Ge = Geelong, Ha = Hawthorn, Ka = Kangaroos, Me = Melbourne, Po = Port Adelaide, Ri = Richmond, St = St Kilda, Sy = Sydney, Wb = Western Bulldogs

Table 5.1: Adjusted and unadjusted Elo ratings, home advantage, predicted margins and absolute errors for West Coast's matches, 2008

5.3 Methods

Elo ratings were originally developed by Arpad Elo to calculate the relative skill of chess players. It is a numerical system in which differences in ratings can be converted into scoring or winning probabilities (Elo, 1978). In this seminal book, Elo states that the many performances of an individual, when evaluated over an appropriate scale, will be normally distributed. Furthermore, he says that performance in chess can't be quantified absolutely, it can only be inferred by the numbers of wins, losses and draws. In simple terms, the Elo ratings system calculates the expected number of games a player is expected to win in a given tournament. If a player exceeds these expectations they receive a ratings increase, while a player that falls short of these expectation receive a rating decrease. A powerful attribute of the Elo ratings system is that a player can win a tournament and receive a rating decrease if that player loses more games than expected.

However, in AFL, the performance of each team cannot only be measured by win/loss, but also by the magnitude of that win (or loss). Akin to the world Football Elo rating system, a goal difference index is introduced here and adapted for AFL to magnify rating increases (and decreases) for strong wins (or losses). Recall in Section 3.5 of Chapter 3, the exact formula for calculating a teams expected game score adjusted for home advantage using the logistic curve is given by

$$W_e = \frac{1}{1 + 10^{(R_j - R_i + H_{ij})/400}} \tag{5.2}$$

where W_e is the expected game result for team i, R_i is the rating of team i, R_j is the rating of team j and H_{ij} is the home advantage team i receives against team j at home.

The results of Chapter 4 suggested that home advantage in AFL could be more accurately estimated by quantifying the independent effects that comprise home advantage. Recall the statistically significant factors of home advantage in AFL were ground familiarity and distance traveled. Therefore, the home advantage H_{ij} expressed in (5.2) can be defined as

$$H_{ij} = \alpha GF_{ij} + \beta ln(DIST_{ij}) \tag{5.3}$$

where $DIST$ is distanced traveled by the away team, α and β are unknown coefficients and

$$GF_{ij} = \frac{\sum_{k=1}^{22} g_{i,k,l}(home)}{\sum_{k=1}^{22} g_{i,k,l}(home) + \sum_{k=1}^{22} g_{i,k,l}(away)} \tag{5.4}$$

where k is the weekly index of games team i plays at venue g relative to team j for each season l.

Recall in Section 3.5 of Chapter 3, the formula for updating a new rating was given by

$$R_n = R_o + KA(W - W_e) \tag{5.5}$$

where R_n is the new rating, R_o is the old rating, W is the observed game result (win=1, draw=0.5, loss=0), W_e is the expected game result, K is the change in ratings multiplier and A the adjustment for the goal difference index. Here A is defined by

$$A = \phi + \rho\frac{|SF_i - SF_j|}{6} \tag{5.6}$$

where SF_i is the score for team i, SF_j is the score for team j and ϕ and ρ are unknown parameters which need to be optimised.

The denominator in (5.6) represents the value of a goal in AFL, therefore A becomes a linear function of margin in terms of goals. Furthermore, the change in ratings multiplier (K) becomes redundant since all matches in the home and away season are assumed to be of equal importance.

Bringing together (5.1) - (5.6) yields a substantial optimisation problem, totaling six unknown coefficients. These are S_K and N_0 for the initial ratings, α and β for home advantage, and ϕ and ρ for the goal difference index. There are several different loss functions that can be utilised for evaluating prediction models (Witten and Frank, 2005). In this chapter, we will concentrate on quadratic loss or more specifically the Brier Score (Brier, 1950) which is given by

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \tag{5.7}$$

	Brier Score	
Probability Forecast	Loss	Win
0.00	0.00	1.00
0.10	0.01	0.81
0.20	0.04	0.64
0.30	0.09	0.49
0.40	0.16	0.36
0.50	0.25	0.25
0.60	0.36	0.16
0.70	0.49	0.09
0.80	0.64	0.04
0.90	0.81	0.01
1.00	1.00	0.00

Table 5.2: How does the Brier score work for given probability forecasts?

where p_i is the forecast probability given in equation (3.35), o_i is the outcome variable (loss=0, draw=0.5, win=1) and N represents the total number of matches used in the training set (352).

Table 5.2 shows how the Brier Score is calculated for given probability forecasts. It is clearly evident that the further the probability forecast is from the actual outcome, the harsher the penalty, due to the quadratic nature of the Brier Score.

Seasons 2000 and 2001 were used as a training set in the forward prediction of seasons 2002 to 2009. Practical bounds were also placed on the parameters to reduce the total possible number of combinations and speed up convergence. Simulations were carried out utilising the Monte Carlo algorithm within Riskoptimiser, an add-in for Excel.

5.4 Results

Various measures can be used to evaluate the performance of prediction models in teams sports. Some commonly used measures in the literature include AAE, number of predicted winners and Return on Investment (ROI) (Bailey and Clarke, 2004). In this section the first two measures to evaluate the performance of the Elo ratings model are utlised. Furthermore, akin to Stefani and Clarke (1992), comparisons of the probability forecasts against the relative frequency of a win is analysed to check the reliability of the probability forecasts. Note that the application to betting markets is discussed in the following section.

Firstly, the reliability of the probability forecasts are investigated. This is to verify whether the probability forecasts of teams with certain characteristics are under (or over) inflated relative to their true probability of winning. For example, if a team is predicted to have a 75% chance of victory according to the model will they win, on average, 75% of the time? Akin to Stefani and Clarke (1992), this assumption can be validated by comparing the probability forecasts against the actual probability of winning. However, in order to do this the probability forecasts must be banded into subgroups to increase the sample sizes when calculating the actual probability of winning. Therefore, the probability forecasts of the favourite winning is banded into five subgroups. The number of games and the actual probability of winning for each subgroup of probability forecasts are shown in Table 5.3. For example, for matches in which the Elo ratings model predicted the favourite to have a 50-59% chance of victory ($n = 429$), on average, they actually won 54.6% of the time. Therefore, provided the probability forecasts are evenly distributed in each subgroup, the actual probability of winning should approximately be the midpoint of the probability forecast range. The results suggest that the probability forecasts mirror closely the actual probability of winning.

Predicted Probability of Winning	Games	Actual Probability of Winning
0.50 - 0.59	429	0.546
0.60 - 0.69	383	0.600
0.70 - 0.79	336	0.750
0.80 - 0.89	205	0.869
0.90 - 1.00	55	0.918
Total	1408	0.671

Table 5.3: Predicted and actual probability of winning, 2002 to 2009

The average percentage of games correctly classified is slightly worse than Stefani and Clarke (1992) (Table 5.4) across seasons 1980-1989 (including finals matches). However, a direct comparison of results of two ratings system across different eras is inappropriate for a number of reasons. Firstly, the competition consisted of fewer teams in 1980-89, and secondly, the style of football was changed substantially, now leading to arguably harder to pick results. It is interesting to note that Clarke's first ratings algorithm outperformed his improved version in the first year of prediction in season 1981 and 1991 respectively. A possible explanation given by Clarke is that an even competition makes predicting winners far more difficult. Bailey and Clarke (2004) showed that the AAE using simple exponential smoothing increased from approximately 20 in 1897 to well above 30 over the last three decades. This suggests AFL is becoming increasingly harder to predict.

To gauge the evenness of the competition, the standard deviation of premiership points at seasons end is derived. A high standard deviation indicates a greater difference between the low and high quality teams, thereby leading to an uneven competition, which makes predicting winners easier. Similarly, a low standard deviation indicates an even competition, making predicting winners challenging. Table 5.4 displays the standard deviation of premiership points against the number of predicted winners for the two rating systems.

Year	Std Dev	% Correct using 0.75 power method Clarke (1993)	Year	Std Dev	% Correct using adjusted Elo ratings
1980	20.29	69.2	2002	15.59	68.2
1981	23.07	75.4	2003	17.22	67.9
1982	21.68	67.0	2004	17.1	69.0
1983	15.26	66.7	2005	15.61	64.8
1984	15.91	64.4	2006	17.03	65.3
1985	19.72	63.4	2007	16.26	63.9
1986	18.13	65.9	2008	18.9	68.8
1987	15.73	73.1	2009	18.59	68.8
1988	16.21	65.9			
1989	17.28	64.7			
Ave	18.33	67.60	Ave	16.82	67.10

Table 5.4: Standard deviation of premiership points and percentage of games correctly classified by two different models, 1980 to 1989 and 2002 to 2009

The correlation between the standard deviation of premiership points and the percentage of games correctly classified is evident for both ratings systems ($r_{power} = 0.37$ and $p_{power} = 0.29$, $r_{Elo} = 0.57$ and $p_{Elo} = 0.14$), excluding a few anomalies. To more clearly see this correlation, a visual representation of the percentage of games correctly classified against the standard deviation for both methods is showcased in Figure 5.2. The difference in standard deviations across the two different eras supports the notion that AFL matches are more difficult to predict in the 2000's than the 1980's.

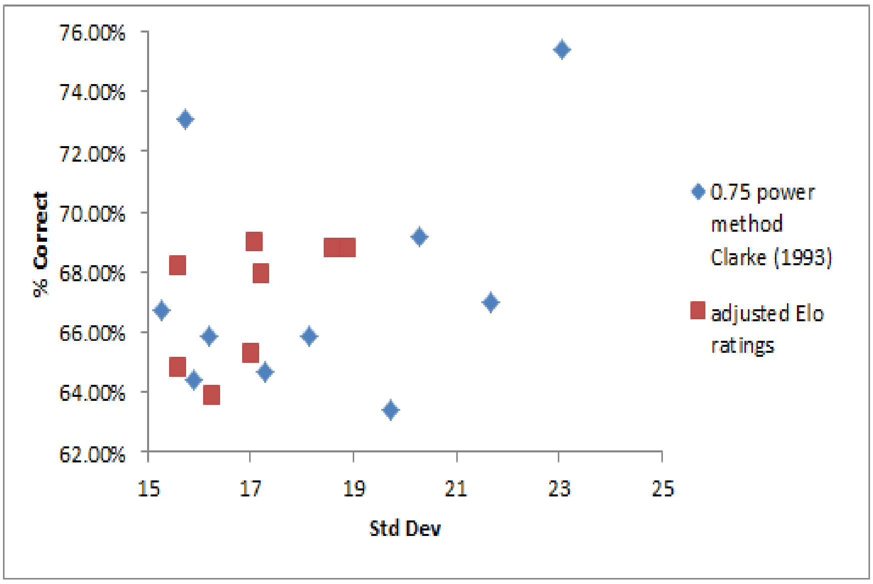


Figure 5.2: Scatter plot of standard deviation of premiership points against percentage of games correctly classified by two ratings systems, 1980 to 1989 and 2002 to 2009

To investigate the AAE, the probability forecasts need to be converted into estimated point margins. Assuming the ratings are normally distributed with mean 0 and standard deviation σ, the estimated point margin x is given by:

$$x = \sigma \left[\Phi^{-1} W_e \right] \tag{5.8}$$

A Kolmogorov-Smirnov test for normality was conducted between the estimated margins and actual margin of victory ($p = 0.079, n = 1408$), indicating normality. Table 5.5 displays the AAE, where $\sigma = 40$, noting that varying σ by a few points has little effect on the estimated point margin x. An AAE of 29.7 compares favourably to the three model comparison (Bailey and Clarke, 2004) across seasons 1997 to 2003. Here the benchmark model (Clarke, 1993) had an AAE of 30.5, the team model (multiple linear regression model) had an AAE of 30.2 and the individual player model (multiple linear regression model) had an AAE of 29.8. However, again it is important to note that this comparison is across two

Year	AAE
2002	29.32
2003	27.30
2004	30.12
2005	31.20
2006	32.42
2007	29.08
2008	30.71
2009	27.62
Average	29.72

Table 5.5: Average absolute error of adjusted Elo ratings, 2002 to 2009

different eras.

One cause for concern is the potential for the model to over inflate large wins relative to their actual importance. For example, the adjustment for the goal difference index in (5.6) is a linear function of score. Therefore, the adjustment of a win by 100 points is valued by the model as ten times more important than if the same team beat the same opponent by 10 points. It could be argued that teams that are behind by large margins during the game will effectively give up, leading to an even greater margin.

An interesting case study of ratings over-inflation due to large wins occurred in round 21 season 2010 when Hawthorn hosted Fremantle at York Park in Tasmania. Fremantle went into the game needing to win one of their last two matches to play a home final. Fremantle's opponents in these last two rounds were Hawthorn and Carlton, both were fringe top eight teams needing to win there last two games to receive a home final. The distance Fremantle had to travel in round 21 to York Park was approximately 3000kms, one of the longest trips in the AFL. The Fremantle coaching staff made the decision to rest half of their best team in round 21, in order to concentrate on winning their round 22 match against Carlton. Much to the delight of punters who had heavily backed Hawthorn to beat the line, the Fremantle side was smashed in round 21 by 116 points. The Fremantle side was ridiculed for their tactics in the media. However, it paid dividends, as they defeated Carlton by six points the next week giving Fremantle a home final (which was ironically against Hawthorn). The model predicted Hawthorn as a three point favourite (even though they had to travel nearly 3000kms) which was largely due to the thrashing they gave Fremantle just two weeks earlier. Fremantle defeated Hawthorn by 30 points, an incredible 146 point turnaround in just two weeks! Although examples such as this are few and far between, large wins do occur, and their importance, particularly against lesser opponents, needs to be investigated.

The opposite is also true, the model can potentially under-inflate small wins relative to their importance. Table 5.6 shows the Elo ratings (and rankings) and AFL rankings at the conclusion of the 2009 AFL home and away season. Interestingly, the biggest discrepancy between the Elo and AFL rankings is Geelong who were the fifth best team according to the Elo ratings yet finished 2nd on the AFL ladder. The primary reason why Geelong's Elo rating was not truly representative of their ability was their amount of weak wins (five of their 18 wins were by eight points or less). Ironically, Geelong were dominant in the finals series, defeating St Kilda in the Grand Final by 12 points to win the premiership.

Team	Elo Rating	Elo Rank	Premiership Points	%	AFL rank
Adelaide	665.97	3	56	117.61	5
Brisbane Lions	571.09	6	54	106.72	6
Carlton	544.95	7	52	110.46	7
Collingwood	659.98	4	60	122.27	4
Essendon	479.25	8	42	97.79	8
Fremantle	354.15	13	24	77.34	14
Geelong	652.22	5	72	127.38	2
Hawthorn	468.48	9	36	92.55	9
Kangaroos	396.53	12	30	83.37	13
Melbourne	293.02	15	16	74.66	16
Port Adelaide	317.28	14	36	88.68	10
Richmond	278.81	16	22	74.29	15
St Kilda	772.86	1	80	155.71	1
Sydney	419.83	11	32	93.14	12
Western Bulldogs	702.18	2	60	122.58	3
West Coast	428.51	10	32	93.3	11

Table 5.6: Elo and AFL rankings, 2009

5.5 Applications to Betting Markets

Fixed odds betting is a form of wagering against set odds offered by an individual, bookmaker or betting exchange. In Australia, the fixed odds betting markets are expressed as decimal odds. In fixed odds betting, the punter must part with there initial stake and if successful they would receive the initial stake multiplied by the quoted odds. For example, in round 18, season 2010, St Kilda started favourites ($1.19) against Essendon ($5.42), with

Essendon going on to win by 33 points. If an initial stake of \$100 was bet on St Kilda to win, the bettor would lose \$100, however if the same amount was bet on Essendon to win the bettor would receive \$542 (100 × 5.42) which includes the initial stake. Ideally, bookmakers price an event (i.e. a football match) such that the net outcome is always in their favour. For example, the sum of the probabilities (1/price) for all possible outcomes of a single event is in excess of 100%. This excess over 100%, commonly referred to as the bookmakers markup or overround, represents the profit to the bookmakers for a balanced book. That is, the amount bet on both teams is distributed evenly relative to their probability of winning (1/price). However, in the case of a unbalanced book, the bookmaker will either stand to collect more winnings than what is mathematically expected or have to pay out more that what was initially staked.

Leitch and Tanner (1991) suggest that forecasting models can be evaluated based on their return on investment. For examples in team sports see Bailey (2000) and Bailey and Clarke (2004) in AFL and Dixon and Pope (2004) in Association football. The magnitude of any perceived market imbalance can be quantified by multiplying the predicted probability by the market price and subtracting the initial stake. That is,

$$A = (P \times M) - 1 \tag{5.9}$$

where A = advantage, P = predicted probability of winning and M = market price.

The concept of value betting stipulates that a bettor must have a positive expectation on a single bet. This arises when the odds estimated by a punter are more accurate than those estimated by a bookmaker. For example, in round 3 of the 2009 season, Carlton hosted Essendon at the MCG. According to the Elo ratings Carlton had a 67% chance of victory, but at \$3.90 with one bookmaker instead of the fair price of \$3.03, the model flagged a 28.70% ($0.33 \times 3.90 - 1 = 0.287$) overlay on Essendon. Therefore, although Essendon were the underdog they represented good value relative to their chances of winning. Essendon actually won the game by four points, but that is besides the point. The moral of the story is that over time value bets are potentially profitable.

Kelly (1956) developed a betting strategy to maximise the long term growth of an initial bankroll subject to the size of this advantage. This system, left untamed, is extremely volatile since the bet size grows unbounded given any profit on the initial bankroll is reinvested. The formula can be simplified to

$$B = \frac{A}{M - 1} \tag{5.10}$$

where B=% of bankroll, A = advantage and M = market price.

In the previous example, where Essendon were estimated to have a 33% chance of victory, they were paying \$3.90 for a win, and there was an advantage of 28.7%, the betting fraction $B = 0.287/(3.90 - 1) = 0.10$, represents 10% of the total bankroll.

Due to the volatility of the Kelly system, authors such as MacLean et al. (1992) state that a fractional Kelly systems should be implemented to reduce the overall risk. A key assumption of the Kelly system is the result of bet n is known before bet $n + 1$ is placed. However in AFL, it is possible for up to three matches to be played simultaneously. A solution proposed by Bailey and Clarke (2004) is to use a constant Kelly, that is, to bet a proportion of a constant pool subject to the Kelly system.

Akin to Bailey and Clarke (2004), the betting strategy implemented in this section is that of a constant Kelly system using a constant pool of \$1000. The criterion for this system was a minimum positive advantage of 10% which was arbitrarily chosen. The bookmakers odds used in this analysis were provided by Pinnacle Sports; their overround is approximately 5%. Table 5.7 displays the results which include the total number of bets, total bets won, percentage of bets won, total bet, profit/loss and the return on investment (ROI). Figure 5.3 displays the cumulative profit/loss across seasons 2002 to 2009 which provides a clearer picture how the model performs within each season. The results compare favourably to professional betting tips providers who charge annual fees for mathematically based predictions in AFL (See, for example, www.sportspunter.com).

Year	# bets	# won	% Won	Total bet	Profit/Loss	ROI
2002	97	56	57.7%	$42,056.44	$7,592.62	18.1%
2003	92	42	45.7%	$30,872.63	$2,494.47	8.1%
2004	102	50	49.0%	$35,769.06	$4,824.16	13.5%
2005	101	52	51.5%	$35,461.21	$7,141.13	20.1%
2006	98	41	41.8%	$39,298.57	-$5,078.10	-12.9%
2007	117	66	56.4%	$44,636.19	$6,779.43	15.2%
2008	88	33	37.5%	$27,029.65	-$965.02	-3.6%
2009	79	33	41.8%	$22,193.15	$5,412.28	24.4%
Ave	99	49	48.50%	$36,446.25	$3,255.53	10.40%

Table 5.7: Head to head betting with Pinnacle Sports, 2002 to 2009

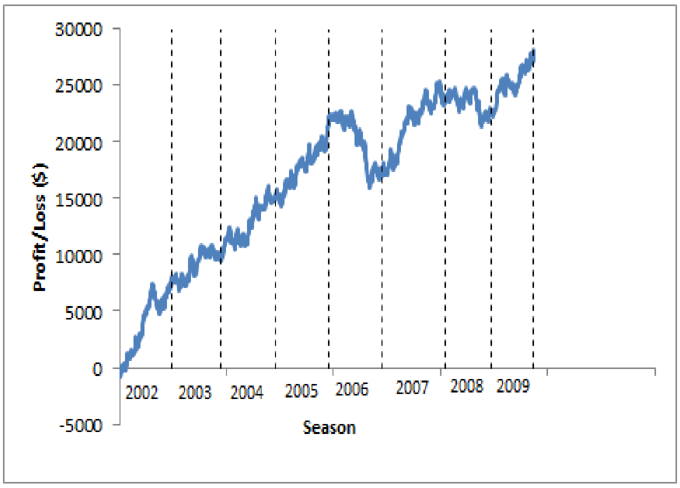


Figure 5.3: Head to head betting with Pinnacle Sports, 2002-2009

It is important to note that there are a plethora of other factors which are likely to influence match outcomes, and hence betting results. These can include injuries and/or suspensions to key players, the importance of the match or the departure of senior coaches mid-season. The nature of these factors are more important as the season progresses. For example, the departure of a coach mid-season is typically attributed to poor results during that season, which usually occurs towards season's end. Clearly the subjective input of such knowledge could increase the return on investment. These ratings should be interpreted as a base for profitable betting strategies.

In addition there are many different betting strategies which could be implemented that could potentially increase overall returns and/or reduce volatility. For example, betting only on home teams, favourites and betting at certain stages during the season could influence betting results. Interestingly, annual returns have decreased over the eight years which could indicate an increase in efficiency by bookmakers. Furthermore, work in this Chapter has focused purely on head to head betting, therefore there is potential in other betting markets with more complicated outcomes such as Premiership markets (Clarke, 1996).

Part II

In-Play

Chapter 6

Collecting In-Play Betting Data

In this chapter, a computer algorithm is developed to automate the collection of in-play betting data for AFL matches. To begin, Section 6.1 provides a brief introduction of in-play betting markets. Section 6.2 discusses betting exchanges with a specific focus towards the Betfair exchange. Section 6.3 describes the complex process of developing a computer program to integrate seamlessly with Betfair's Application Programming Interface (API) using the programming language Perl. This information is stored in a MySQL database which can then be easily exported as a CSV file for manipulation in Excel. The final section, Section 6.4 concludes the chapter by providing a sample of the in-play betting data. Material from this chapter has been published in Ryall and Bedford (2009).

6.1 Introduction

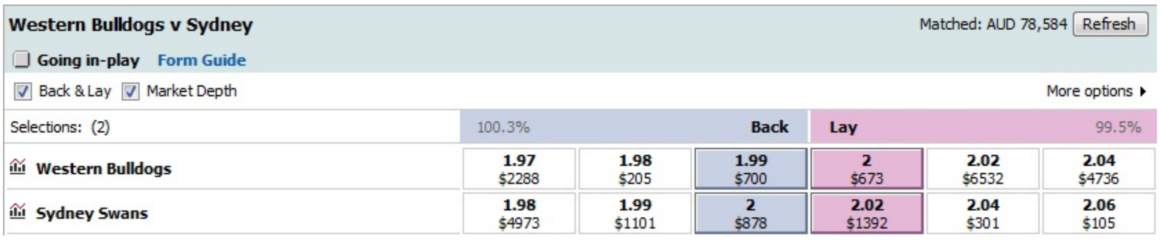
In-play betting is a relatively new phenomenon where punters bet on the outcome of an event that has already started. For example, the outcome of an election whereby the counting process has started, or a sporting event which is in progress. This is usually facilitated by betting exchanges such as Betfair which is discussed in the following section. In-play betting adds a new dimension to the betting experience as astute punters are now faced with the need to update their estimates (or predictions) as an event progresses. For example, prior to the start of an AFL match, the importance of team quality and home advantage were showcased in Chapter 4 and Chapter 5. In addition, during the game punters now have to weight the relative importance of team quality and home advantage as the match progresses and incorporate current score into their estimates. If mathematical models can aid in the prediction of pre-match outcomes, then it is reasonable to assume that these models will be of greater importance during the game, as punters are faced with even more factors to take into consideration.

6.2 The Betfair Exchange

Traditionally, betting markets have been restricted to licensed bookmakers. The most common form of betting markets that bookmakers offer are known as fixed odds, whereby the bookmaker offers wagering against set odds. For example, suppose party A wishes to back (bet on) some outcome and party B wishes to lay (bet against) the same outcome. In fixed odds betting, party A would agree to pay party B an initial stake if the outcome is not realised, and party B would pay party A the initial stake multiplied by odds that were agreed upon by the two different parties (hence fixed odds). For example, in round 18 season 2010, St Kilda started favourites ($1.19) against Essendon ($5.42). Suppose party A agreed to stake $100 on St Kilda to win the match. In this example, party A collects $119 from party B ($19 profit plus $100 stake) if St Kilda win, otherwise if St Kilda lose, party B collects the $100 initial stake from party A. Typically, punters that bet with bookmakers play the role of party A (bet on) and the bookmaker plays the role of party B (bet against).

However, with the advent of the internet, the arrival of betting exchanges in 2000 marked the beginning of a revolution in the industry. This revolution allowed individual punters to bet with each other directly on the outcome of events. Contrary to the standard bookmaker, punters that utilise betting exchanges can lay individual outcomes, that is, betting against a particular player or team. Betfair was established in June 2000 and has become the the largest betting exchange in the world. Betfair charges a commission of 5% on all winning bets which drops to as little as 2% for the heaviest users. Since the betting exchange does not otherwise impose any overround (bookmakers advantage for balanced books) the prices are very competitive for popular events. Betfair claims to have over two million clients and process over six million transactions a day which equates to a turnover in excess of £50m/week (bet, 2008). There are a plethora of betting markets available on Betfair which include most sporting events and many non-sporting events (e.g. political elections). Betting markets for the AFL include betting on the premiership, or the Coleman medal (equivalent to the golden boot in Association Football), or the Brownlow (best and fairest player) to name a few. Meanwhile "match odds" markets allow betting on the outcome of an individual game, by backing (betting on) or laying (betting against) a home or away win with fixed odds.

Figure 6.1 shows the order book of the Betfair Exchange between the Western Bulldogs and Sydney in a Semi-final from season 2010. An example might help to clarify how to interpret this order book. Suppose a punter wishes to back the Western Bulldogs to win, they can stake up to $700 at odds of $1.99, they can bet a further $200 at reduced odds of $1.98 and a further $2288 at even more reduced odds of $1.97. Therefore, a $100 wager on the Western Bulldogs to win would return $199 ($99 profit plus $100 stake), less Betfair's commission (which varies between 2% to 5%). All odds displayed are from the perspective of the backer. For example, the stake of $673 at odds of $2 to lay the Western Bulldogs indicates that someone (or some combination of clients) are hoping to back the Western Bulldogs at the asking price of $2 (i.e. slightly above the current market price). Thus if someone were to lay (bet against) $100 on the Western Bulldogs, they would be risking $200 in order to win $100.



Western Bulldogs v Sydney

Matched: AUD 78,584 Refresh

□ **Going in-play** Form Guide

☑ Back & Lay ☑ Market Depth

More options ▸

Selections: (2)	100.3%		Back	Lay		99.5%
Western Bulldogs	1.97 $2288	1.98 $205	1.99 $700	2 $673	2.02 $6532	2.04 $4736
Sydney Swans	1.98 $4973	1.99 $1101	2 $878	2.02 $1392	2.04 $301	2.06 $105

Figure 6.1: The Betfair exchange: Match odds for Western Bulldogs vs. Sydney, Semi-Final 2010

The real hurdle for betting exchanges is to achieve sufficient liquidity for all betting markets. For example, it is common that the back price does not match the lay price. This differential is usually smaller for the favourite than the underdog. The more volume that is bet on a particular event, the closer the back and lay price become.

Betting exchanges allow punters to place bets in-play, that is, once an event is underway. Unlike pre-game betting markets, in-play betting markets can swing rapidly, particularly in low scoring sports. For example, in 2008 there was a memorable match between Aston Villa and Everton in the English Premier League (EPL). With one minute of stoppage time remaining in the match, Aston Villa led 2-1 and were paying \$1.04 for the win and \$26 for the draw. Everton then equalized, and with less than one minute of stoppage time to go, \$280,000 was traded on the draw at \$1.01 and \$30 was traded on Aston Villa to win at astonishing odds of \$440. Aston Villa's Ashley Young found a gap in Everton's defence to run through and score a last second sealer, delivering an early Christmas present to all those who believed a victory was still possible for Aston Villa. Counter to this example, AFL price swings for in-play betting markets are likely to be considerably less volatile due to the high scoring nature of the game. According to an analyst at Betfair, approximately A\$80,000 in volume was bet in-play at Betfair alone during an average AFL game in 2009. Volume bet increases to over A\$140,000 for some of the "blockbuster" games (i.e. grand final) and gets as low as A\$1,000 for a game of less interest (i.e. two non-Victorian teams). Interestingly, the median volume bet in-play at Betfair for an AFL match was just over A\$35,000 suggesting the distribution of volume bet in-play is heavily skewed towards blockbuster games. Although over the previous three years the volume bet in-play relative to pre-game for AFL matches at Betfair has increased (14% to 19%), 80% of volume is still bet pre-game.

A unique feature of the Betfair exchange is the Application Programming Interface (API). The sports API enables users to develop programs which seamlessly integrate with the Betfair sports exchange.

6.3 Application Programming Interface

According to Wikipedia, an API is a "set of routines, data structures, object classes and/or protocols provided by libraries and/or operating system services in order to support the building of applications". The Betfair API is language independent, which means they can be called by several programming languages. These APIs are accessed via a Simple Object Access Protocol (SOAP) interface over a secure web connection.

There are three connection end-point URLs that access the Betfair sports betting API services:

- Global: http://api.betfair.com/global/v3/BFGlobalService
- UK Exchange: http://api.betfair.com/gexchange/v3/BFExchangeService
- AUS Exchange: http://api-au.betfair.com/gexchange/v3/BFExchangeService

The global services are used to log in and out, administer a client's Betfair account and funds, and navigate though the events hierarchy until the client reaches a particular market. The exchange services are used to view and bet on sports events. There is a separate exchange for the UK and Australia. A full list of all current global and exchange services is given in Betfair Sports Exchange API 6 Reference Guide (https://bdp.betfair.com/images/stories/downloads/BetfairSportsExchangeAPIReferenceGuidev6.pdf).

In the remainder of this section, comprehensive details of a computer program (developed *exclusively* for this dissertation), which integrates seamlessly with the Betfair API in order to automate the collection of in-play betting data for AFL matches are provided. There are many reasons why this path was taken instead of trying to obtain the data from external sources. Firstly, the data of interest (quarter by quarter odds) was not easy to obtain. (Note: I had emailed several bookmakers to no avail, the typical response I received was that this information was either not cached or was not available to the general public, especially academics!). Furthermore, companies such as Fracsoft (www.fracsoft.com) that provide betting data (including in-play) to interested parties for a fee, were not permitted to record betting data on the Australian exchange (i.e. AFL matches). Two options remained, either record the data manually or automate the procedure. (Note: As previously stated in Section 3.6 of Chapter 3, throughout early 2008 my supervisor and I manually recorded in-play betting odds for AFL matches. This was not only extremely time consuming (on weekends), but extremely frustrating when you are watching a delayed telecast since the odds typically give a good indication of who is winning). Due to the troublesome nature of manual collection, it was decided to automate the collection of in-play betting data for AFL matches.

The idea behind this novel program was adapted from Magee (2008), who amongst other things, detailed how to automate the collection of in-play betting data for UK horse racing. However, the program developed here varies considerably to Magee (2008). Firstly, the program is adapted to in-play fixed odds AFL matches using the Australian Betfair exchange. Additionally, one of the strengths of the program is the ability to record in-play betting data for AFL matches that are played simultaneously. For example, in a typical round in the AFL schedule, up to three matches can be played simultaneously. Therefore, it was important that the program was equipped to handle these subtleties as it was a regular occurrence each round. Furthermore, the amount of manual interference has been reduced as much as possible. For example, once the inputs have been entered into the program prior to the start of a round, the program will run (without interruption) and record the in-play betting for all matches of the current round.

Magee (2008) uses the the operating system Linux, the programming language Perl to interact with the Betfair API's and the relational database MySQL to temporally store the data. For convenience, the same operating system, programming language and relational database was utilised in this section. A separate EEE PC was purchased such that the program could run uninterrupted over the weekend. Figure 6.2 displays a flowchart which helps to explain how the program works. The remainder of this section describes the intricacies of the program. To begin, the database and tables are defined in the relational database MySQL.

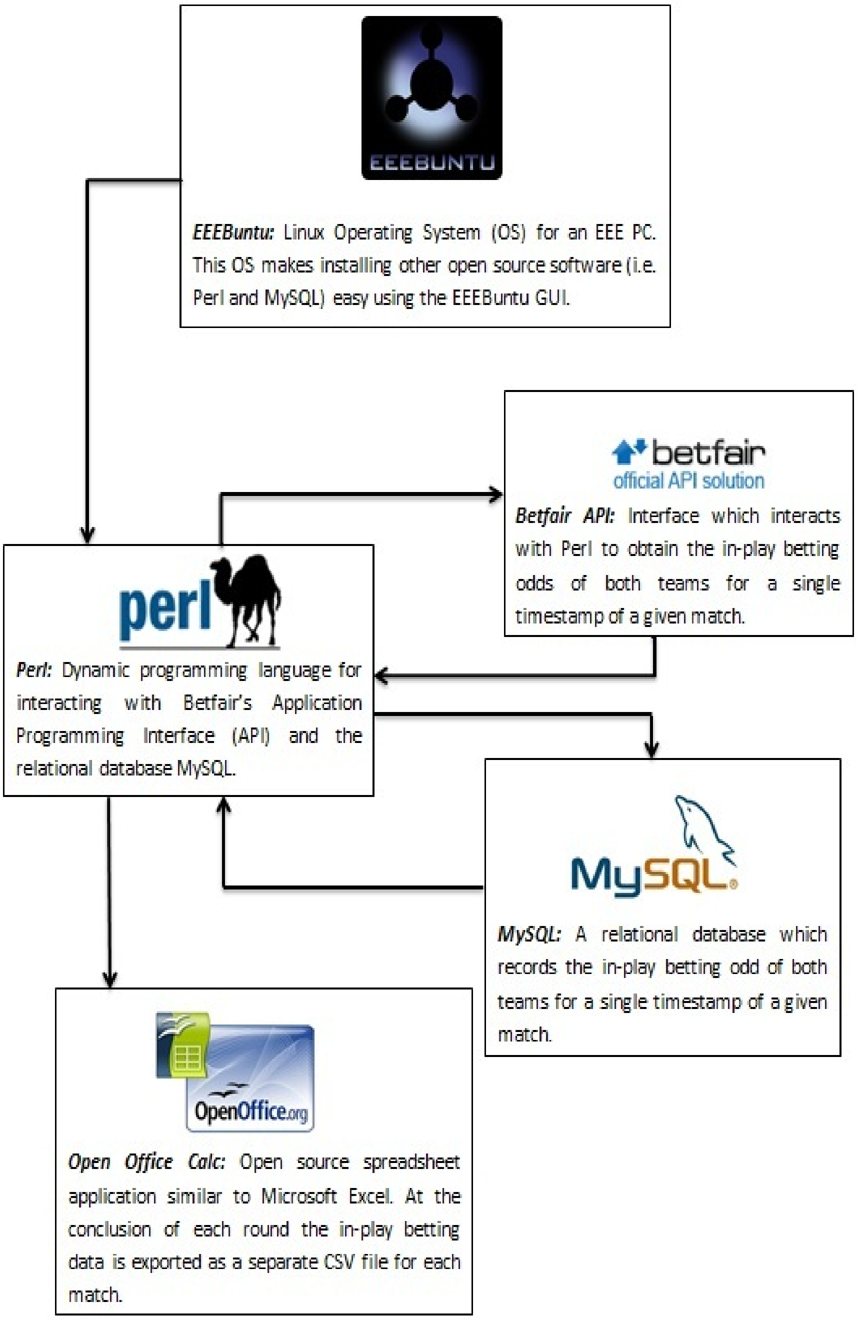

EEEBUNTU
EEEBuntu: Linux Operating System (OS) for an EEE PC. This OS makes installing other open source software (i.e. Perl and MySQL) easy using the EEEBuntu GUI.
betfair official API solution
Betfair API: Interface which interacts with Perl to obtain the in-play betting odds of both teams for a single timestamp of a given match.
perl
Perl: Dynamic programming language for interacting with Betfair's Application Programming Interface (API) and the relational database MySQL.
MySQL
MySQL: A relational database which records the in-play betting odd of both teams for a single timestamp of a given match.
OpenOffice.org
Open Office Calc: Open source spreadsheet application similar to Microsoft Excel. At the conclusion of each round the in-play betting data is exported as a separate CSV file for each match.

Figure 6.2: Flowchart of the procedures involved in automating the collection of in-play betting data for AFL matches

6.3.1 Define Database and Tables in MySQL

In MySQL, data is stored in tables which are specified within a database. Therefore, an empty database **autodb** was generated in MySQL. The next stage is to define tables within this database where the data will be stored. When defining a table in MySQL, each variable must be specified including the variable name and variable type (e.g. time, decimal, varchar). For the purpose of automating the collection of in-play betting data for AFL matches, six variables were required. These included a timestamp, team name, back price, back volume, lay price and lay volume. Note that it is possible to extract up to three levels of odds to back or lay a specific team (as discussed in Section 6.2), however due to small liquidity in this market, it was decided the first level of prices would be more than adequate. For convenience, a separate table is defined for each match of a specified round, namely **AFL_1, AFL_2, . . . , AFL_8**. The script to create these tables is given in Section 13.1.1 of the Appendix.

6.3.2 Define Pre-Requisite Modules and Variables

In this section, the pre-requisite modules and variables required for the Perl script are defined. Examples of the pre-requisite modules include BetfairAPI6Examples (Perl library for accessing Betfair API services), XML::Simple (easy to read and write XML files) and Data::Dumper (makes debugging easier). Full descriptions of the parameters available for return by each service defined in BetfairAPI6Examples can be found in the Betfair Sports Exchange API 6 Reference Guide (https://bdp.betfair.com/images/stories/downloads/BetfairSportsExchangeAPIReferenceGuidev6.pdf). The variables required for the Perl script include scalar variables (single value) which are preceded by the $ sign (for example, the login variables $username, $password and $productId), array variables (ordered collection of scalars) which are preceded by the @ sign (for example, @match_days which is an array which includes the dates for all matches for a specified round), and hash variables (is a map from strings to scalars; the strings are called keys, and the scalars are called values) which are preceded by the % sign (for example, for any given match %static_runner_data contains the name of the two teams and their corresponding id (keys), and the market data for those keys including back price, back volume, lay price and lay volume). A full list of the pre-requisite modules and variables are given at the beginning of Section 13.1.2 in the Appendix.

6.3.3 Login to Betfair

As previously stated, the Login service is a global service and requires three input parameters a `username`, `password` and `productId` (82 is Betfair's free access API). This service logs into the users Betfair account and, if logged in successfully, returns a parameter called `sessionToken` which is a unique code required for all other services.

The free access API is valid for all current Betfair clients, and in order to be considered a current Betfair client, a transaction must have occurred in the account holders name in the previous three months. It is possible to transfer money between a Betfair clients Australian wallet and Main wallet to maintain the current client status (i.e. you do not have to make a bet). Unfortunately, this is something that was found during the latter stages of this work. The free access API allows users to make up to 60 calls (discussed later) a minute and there is no online help provided, besides the Betfair forum.

6.3.4 Betfair Event Hierarchy

If a user wishes to identify the current fixed betting odds of an event using the Betfair web interface, they must cycle through the event hierarchy until they reach the said event and betting market. For example, suppose a punter wanted to bet (fixed odds) on the Essendon vs. Hawthorn match in round 6, season 2010, they would follow these simple steps.

Firstly, open the Betfair home page (www.betfair.com.au), click on the tab “All Sports”, on the left hand side click on the tab “Australian Rules”, then click on the tab “AFL 2010”, then click on the tab “Round 6 - 01 May” and finally click on the tab “Match Odds”. Figure 6.3 shows an extract from the Betfair web interface cycling through the event hierarchy in the previous example. Similarly, Figure 6.4 shows the match odds in the previous example using the betfair web interface.

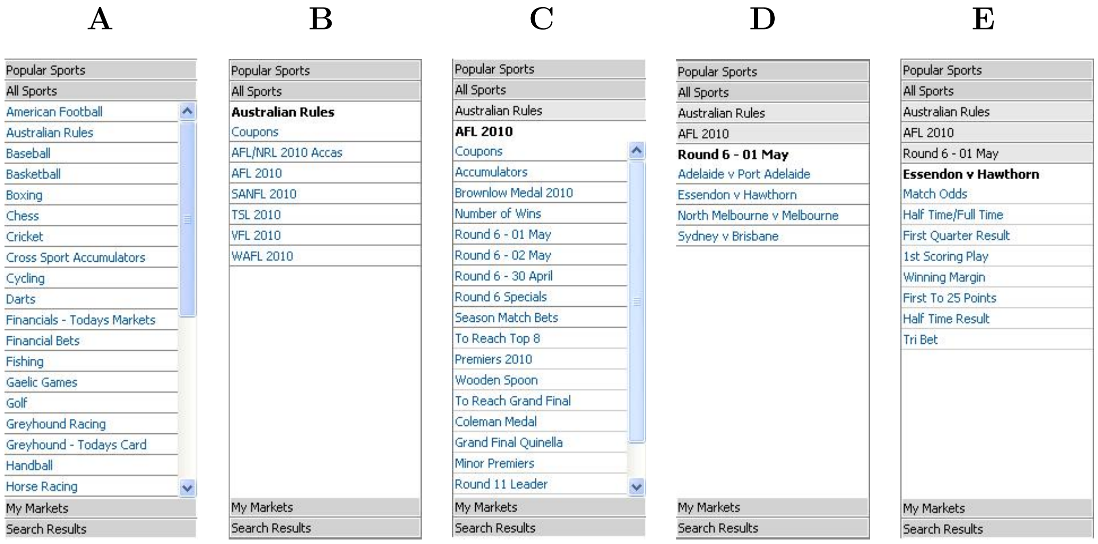
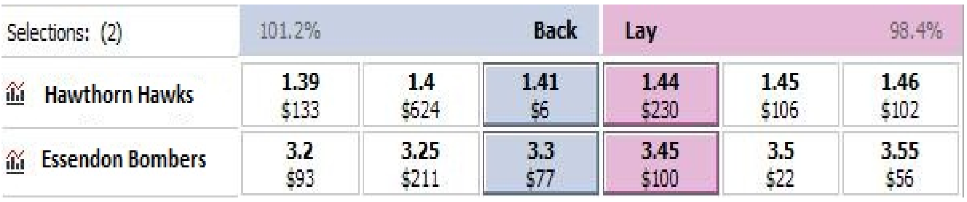


Figure 6.3: The Betfair web interface event hierarchy. (A) All Sports. (B) Australian Rules. (C) AFL 2010. (D) Round 6 - 01 May. (E) Match Odds.



Selections: (2)	101.2%		Back	Lay		98.4%
Hawthorn Hawks	1.39 $133	1.4 $624	1.41 $6	1.44 $230	1.45 $106	1.46 $102
Essendon Bombers	3.2 $93	3.25 $211	3.3 $77	3.45 $100	3.5 $22	3.55 $56

Figure 6.4: Match odds for Essendon vs. Hawthorn in round 6 season 2010 from the Betfair web interface.

To replicate the above procedure using the Betfair API a similar approach is taken. For instance, each event or market in the hierarchy corresponds to a unique id which can be retrieved by calling subroutines for accessing Betfair's API global and exchange services. To begin, the global service `getAllEventTypes` returns a list of *all* event names (`eventName`) and their corresponding id (`eventId`), similarly `getActiveEventTypes` returns a list of all *active* event names (`eventName`) and their corresponding id (`eventId`). These services require the input parameter `sessionToken`. Table 6.1 shows an extract from output of the `getActiveEventTypes` global service made on 27/04/2010. In this example only `eventName` and the corresponding `eventId` are returned, that is, there are no `marketName` and `marketId`. Therefore, in order to retrieve the current fixed odds for the previously mentioned Essendon vs Hawthorn example, we must continue cycling through the event hierarchy.

`EventName`	`EventId`	`MarketName`	`MarketId`
Ice Hockey	7524		
Horse Racing	7		
Tennis	2		
...	...		
Australian Rules	61420		
...	...		
Financial Bets	6231		
Horse Racing - Virtual	26397698		
Rugby Union	5		

Table 6.1: Sample output of `getActiveEventTypes` global service

The next stage is to cycle through the event "Australian Rules" to retrieve the `eventId` for the `eventName` "AFL 2010". This is achieved using the global service `GetEvents`. This service requires the input parameters `sessionToken` and `eventParentId` where `eventParentId` is the id returned by `GetActiveEventTypes` (or `GetAllEventTypes`) or an earlier `GetEvents` request. This service returns event items (such as `eventId` and `eventName`) and/or market items (such as `marketId` and `marketName`). Table 6.2 shows an extract from output of the `getEvents` global service made on 27/04/2010 using the `eventId` for Australian Rules (61420). Again no `marketName` and `marketId` are returned so we must continue cycling through the event hierarchy.

`EventName`	`EventId`	`MarketName`	`MarketId`
SANFL 2010	26556914		
Coupons	26684595		
AFL 2010	26502908		
VFL 2010	26556915		
WAFL 2010	26556916		
TSL 2010	26556913		

Table 6.2: Output of `getEvents` global service using the `eventId` "Australian Rules"

The next stage is to cycle through the event "AFL 2010" to retrieve the `eventId` for the `eventName` "Round 6 - 01 May". Akin to the previous stage, the `GetEvents` global service is implemented using the `eventId` for "AFL 2010" (26502908). Table 6.3 shows an extract of the output. Interestingly, now `MarketName` and `MarketId` are returned. However, these markets do not include the market of interest (match odds between Essendon and Hawthorn in round 6 season 2010). Therefore, we must continue cycling through the event hierarchy.

`EventName`	`EventId`	`MarketName`	`MarketId`
Brownlow Medal 2010	26512790	Premiers 2010	100106152
Round 6 - Specials	26581634	Coleman Medal	100143762
Number of Wins	26556579	Winning Region	100162488
Round 6 - 01 May	26581597	Minor Premiers	100162470
Season Match Bets	26556580	To Reach Top 4	100150539
Round 6 - 02 May	26581605	To Reach Grand Final	100179824
Coupons	26684596	To Reach Top 8	100114004
Round 6 - 30 April	26581577	Wooden Spoon	100114269
		Grand Final Quinella	100162468
		Round 11 Leader	100173804

Table 6.3: Output of `getEvents` global service using the `eventId` "AFL 2010"

The next stage is to cycle through the event "Round 6 - 01 May" to retrieve the `eventId` for the `eventName` "Essendon v Hawthorn". Akin to the previous stages, the `GetEvents` global service is implemented using the `eventId` for "Round 6 - 01 May" (26581597). Table 6.4 shows an extract of the output. Again no `marketNames` and `marketId` are returned so we must continue cycling through the event hierarchy.

`EventName`	`EventId`	`MarketName`	`MarketId`
Western Bulldogs v St Kilda	26581578		
North Melbourne v Melbourne	26581598		
Sydney v Brisbane	26581604		
Essendon v Hawthorn	26581601		
Adelaide v Port Adelaide	26581600		
Carlton v Collingwood	26581624		
West Coast v Fremantle	26581632		
Geelong v Richmond	26581612		

Table 6.4: Output of `getEvents` global service using the `eventId` "Round 6 - 30 April", "Round 6 - 01 May" and "Round 06 - 02 May"

The next stage is to cycle through the event "Essendon v Hawthorn" to retrieve the `marketId` for the `marketName` "Match Odds". Note that now the return parameters are markets not events since we have reached the end of the event hierarchy. Akin to the previous stages, the `GetEvents` global service is implemented using the `eventId` for "Essendon v Hawthorn" (26581601). Table 6.5 shows an extract of the output.

`eventName`	`eventId`	`marketName`	`marketId`
		Half Time Result	100184598
		Match Odds	100184592
		Half Time/Full Time	100184591
		1st Scoring Play	100184589
		Tri Bet	100184602
		First Quarter Result	100184590
		Winning Margin	100184593
		First to 25 points	100184605

Table 6.5: Output of `getEvents` global service using the `eventId` "Essendon vs. Hawthorn"

Now the match odds can be retrieved using the `GetMarkets` exchange service which requires the input parameters `sessionToken` and `marketId`. The service returns all static market data for the market requested. The format of the static market data are hash variables, whereby the team names are matched against their respective market data (i.e. back price, back volume, lay price and lay volume). The market data for each team is distinguished by a `runnerId`. Table 6.6 shows an extract of an `GetMarkets` exchange service for `marketId` "match odds"

Team Name	Back Price	Back Volume	Lay Price	Lay Volume
Essendon Bombers	1.41	6	1.44	230
Hawthorn Hawks	3.30	77	3.45	100

Table 6.6: Output of `GetMarkets` exchange service using the `marketId` "match odds"

This somewhat arduous process has finally retrieved the match odds for the round 6 match in season 2010 between Essendon and Hawthorn. The similarities between using the Betfair web interface and the Betfair's API to obtain the same result are now immediately evident. Figure 6.5 displays a flowchart of how the match odds are obtained using the Betfair web interface and the Betfair API. It has taken considerably more time using the API than the web interface so one might ask why would you go down this path? The answer lies in the ability of computer programs to automate certain procedures using the API. Now that a brief program has been written to display the match odds it can be adapted to suit a user's need. For example, the `marketId` for the match odds of a round of matches can easily be automated with the user only having to entire the current round number and the dates when the matches are played. Recall the purpose of this Chapter was to develop a computer program to automate the collection of in-play betting data. The remainder of this section details how the program is adapted to meet this criteria.

All Sports

⇓

Australian Rules

⇓

AFL 2010

⇓

Round 6 - 01 May

⇓

Essendon v Hawthorn

⇓

Match Odds

Figure 6.5: Flowchart of obtaining match odds for Essendon vs. Hawthorn in round 6 season 2010

6.3.5 Boolean Stopping Condition

The program was set to run anytime at the user's request prior to the start of the current round. Furthermore, the program automatically breaks when the current time reaches a pre-determined finish time. Although it is possible to use an endless loop (i.e. no stopping condition) this requires manual human intervention which can use up valuable internet resources (downloads). Therefore, the user enters the date and time (24 hour clock) when the program should automatically break. These parameters are then transformed to a scalar such that a conditional statement can check whether the current time has surpassed the finish time.

There are four parameters that the user enters in order to break the program at a specified time, namely a scalar for month (1 January, 2 February, . . . , 12 December), the day (1 to 31), the hour (0 to 23) and minute (0 to 59). Then these parameters are transformed to hours passed since the beginning of the year. This is achieved by multiplying month by the number of days in that month (28 to 31), then again by 24 (number of hours in a day). Similarly, the day is multiplied by 24 (number of hours in a day). Therefore, assuming minute equals zero, hours passed since the start of the year is given by:

$$\text{time} = \begin{cases} (\text{month} \times 28 \times 24) + (\text{day} \times 24) + \text{hour}, & \text{if month=Feb} \\ (\text{month} \times 30 \times 24) + (\text{day} \times 24) + \text{hour}, & \text{if month=Apr, Jul, Sep, Nov} \\ (\text{month} \times 31 \times 24) + (\text{day} \times 24) + \text{hour}, & \text{otherwise} \end{cases} \tag{6.1}$$

For example, suppose the user wishes to record the in-play betting data for round 6, season 2010. The final match is played on 2nd May at 4:40pm (AEST), therefore the match should conclude well before 10pm (AEST). In this example, month=5, day=2, hour=22 and minute=0. Therefore, hours passed since the beginning of the year is $(5 \times 31 \times 24) + (2 \times 24) + 22 = 3,790$.

The current time can be evaluated using the function localtime(). This function returns the current year, month, day, hour, minute and second as scalars. Akin to the finish time, the current time is transformed to a scalar such that a conditional statement can check whether the current time has surpassed the finish time. Since the program runs over several days the current time needs be be recalculated after every iteration within the loop. Once the scalar current time is greater than or equal to the scalar finish time, the program breaks.

6.3.6 Collect In-Play Betting Data

This section discusses the majority of the perl script used to automate the collection of in-play betting data. Firstly, suppose the `marketId` for match odds for are all eight matches of a given round is stored in the array @market. The first loop is a while loop that executes code repeatedly until the scalar current time is greater than or equal to the scalar finish time at which point the program breaks. Immediately after this boolean stopping condition within the loop, the current time is recalculated and transformed to the previously mentioned scalar current time.

The next stage is to determine which of the matches, if any, are currently in-play. The exchange service `getMarketPricesCompressed` requires the input parameters `sessionToken` and `marketId`. An important return parameter for this service is "delay" which returns a value greater than zero once the match has started. This is because there is a delay in matching bets for in-play markets to allow punters the chance to cancel their current bets (which are unmatched) after a significant event has occurred (i.e. a goal is scored). Therefore, a second loop (foreach) is nested within the first loop and cycles through the array @market and creates another array (@market_intherun) which contains the `marketId` of all matches in the array @market which are currently in-play. Note that it is necessary for @market_intherun to be an array as appose to a scalar due to some matches being played simultaneously.

The subsequent step is to cycle through all the matches which are currently in-play and retrieve the static market data. Therefore, a third loop which is nested inside the first loop but not the second, cycles through the array @market_intherun and obtains the associated static runner data using the exchange service `getMarkets`. If no markets are currently in-play the program prints "no current in-the-run market" on the screen. The `getMarkets` service returns a hash variable which contains an array of market data (i.e. back price, back volume, lay price and lay volume) for each `runnerName` (team) of a given market. Note that the `runnerId` for each `runnerName` which is currently in-play, is contained in the array @names.

The following detailed stage is to record the in-play betting data to a MySQL database. As such, a fourth loop, which is nested inside loop one and three, cycles through the array @names and obtains the in-play betting data for both teams of a single match using the exchange service `getMarketPricesCompressed`. This service returns the back price, back volume, lay price and lay volume. This information alongside the current time (24 clock hh:mm:ss) and the `runnerName`, is printed on screen and recorded in a separate MySQL table. In order to record the betting data in separate table for each match, a counter was required to differentiate between each match.

As previously mentioned, there are some limitations of the Betfair free access API. First and foremost, there is a limitation of 60 calls of the global and exchange services per minute. Therefore, although it is possible to record the in-play betting data for a match every second throughout the match, this time interval has been increased to approximately every 12 seconds to satisfy an upper limit of 60 calls per minute. It is also important to note that the program would break whenever the internet connection cut out. Therefore, the program would be checked at various stages throughout the weekend (i.e. once a day). It is plausible that this problem can be solved with better technology than a home setup. The full program is given in Section 13.1.2 of the Appendix.

6.3.7 Export Betting Data to Excel

At the conclusion of each round each table (**AFL_1**, **AFL_2** ... **AFL_8**) is exported as a CSV file for easy manipulation in Excel. The following is an example of how the table **AFL_1** would be exported to the folder '/tmp/'.

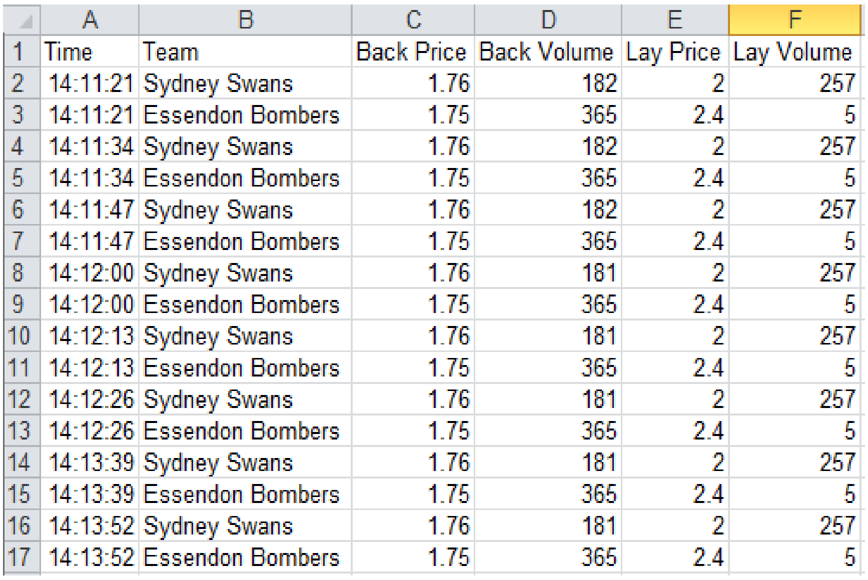
```
select * from afl_1 into outfile '/tmp/afl_1.csv';
```

Once all tables have being successfully exported, each tables data is cleared ready for the next round. The following code is an example of how the table **AFL_1** would be cleared.

```
delete * from afl_1.csv;
```

6.4 Results

Figure 6.6 shows an example of the output of the customized program. What is immediately evident (at least in this match) is that the volume bet is extremely limited to say the least. For example, at the first time stamp (14:11:21), the maximum volume permissible to bet on (back) Essendon and Sydney is $365 and $182 respectively. This minute volume has significant consequences on the actual odds offered. For example, prior to the start of the match Essendon were paying $2.19 while Sydney were paying $1.80 (Pinnacle Sports), however once the match was underway Essendon were paying $1.75 and Sydney were paying $1.76 (Betfair) which is a significant drop considering nothing has changed in regards to the outcome of the match. However, these data are still an extremely valuable commodity which has several practical applications which are discussed in subsequent chapters.



	A	B	C	D	E	F
1	Time	Team	Back Price	Back Volume	Lay Price	Lay Volume
2	14:11:21	Sydney Swans	1.76	182	2	257
3	14:11:21	Essendon Bombers	1.75	365	2.4	5
4	14:11:34	Sydney Swans	1.76	182	2	257
5	14:11:34	Essendon Bombers	1.75	365	2.4	5
6	14:11:47	Sydney Swans	1.76	182	2	257
7	14:11:47	Essendon Bombers	1.75	365	2.4	5
8	14:12:00	Sydney Swans	1.76	181	2	257
9	14:12:00	Essendon Bombers	1.75	365	2.4	5
10	14:12:13	Sydney Swans	1.76	181	2	257
11	14:12:13	Essendon Bombers	1.75	365	2.4	5
12	14:12:26	Sydney Swans	1.76	181	2	257
13	14:12:26	Essendon Bombers	1.75	365	2.4	5
14	14:13:39	Sydney Swans	1.76	181	2	257
15	14:13:39	Essendon Bombers	1.75	365	2.4	5
16	14:13:52	Sydney Swans	1.76	181	2	257
17	14:13:52	Essendon Bombers	1.75	365	2.4	5

Figure 6.6: Screen dump of match odds for Sydney vs. Essendon in round 15 season 2009 using the customized program

Chapter 7

In-Play Betting Data as a Measure of Expectation

In this chapter, the in-play betting data collected in Chapter 6, was transformed to normalized implied probabilities and plotted against time to give a graphical real-time measure of expectation. To begin, Section 7.1 provides a brief introduction on information incorporation of in-play betting odds in team sports. In Section 7.2, the data utilised in this analysis are described. The next section, Section 7.3 describes the methodology used to transform the in-play betting data to normalized implied probabilities and generate the real-time plot of implied probabilities against score difference. A couple of case studies were investigated in Section 7.4 for validation purposes. Furthermore, Section 7.5 examines the forecasting capabilities of the implied probabilities as the game progresses against score difference. To conclude, Section 7.6 provides a brief discussion on the limitations of using the implied probabilities to forecast match results. Material from this chapter has been published in Ryall and Bedford (2009).

7.1 Introduction

Often in sporting events, the in-play betting odds and score are misaligned. This indicates that there is a clear difference in the market opinion of victory and the current score. This difference could be attributed to the in-play betting odds incorporating other information (i.e. team quality, home advantage, injuries, any perceived momentum, time remaining etc.) in addition to current score. Using in-play betting odds as a statistical benchmark of expectation of the two competing teams is becoming extremely popular in team sports. More recently at AFL matches, the in-play odds are displayed on the big screen at the conclusion of each quarter. This gives spectators an indication of whether their team is *expected* to win.

For example, in round 20 season 2009, Essendon were hosting St Kilda at Docklands stadium on a Sunday afternoon. St Kilda went into the game as strong favourites as they were undefeated all season, meanwhile Essendon were clinging onto a top eight spot. The bookmakers thought it would be a one sided affair with St Kilda paying \$1.14 for a win while Essendon were paying \$6.50. Interestingly, at three quarter time St Kilda were trailing by 29 points which was unexpected to say the least. There was a murmur in the crowd that St Kilda would surely come back in the final quarter, after all they were the best team in the competition. The in-play betting odds were then displayed on the big screen; they were a complete turnaround to the pre-match odds with Essendon the favourite (\$1.30) and St Kilda (\$3.00) the underdog. The crowd was stunned as Essendon were now expected to win the match with a high level of certainty. Incidently, Essendon went on to win the match but not without some controversy. Nick Riewoldt (St Kilda) had a set shot from 45 metres out after the final siren to win the match, the shot sailed wide and Essendon won by a meagre two points.

Additionally at AFL matches, a real-time plot of the in-play betting odds for both competing teams are displayed on the big screen during the half time break. This is somewhat similar to a “score worm” (graph of time elapsed against score difference) as it tells a story of what has happened and when it happened. However, there are several inherent risks of a graphical representation of in-play betting odds of both teams. First and foremost, the betting odds of one team is approximately the inverse of the betting odds of the opposing team. Therefore, a real-time plot of the in-play betting odds using a linear scale for the y-axis (betting odds) will be inadequate when the match is one-sided. For example, in round 1 season 2009, Port Adelaide hosted Essendon at Football Park. Port Adelaide dominated the game leading by 20 points at quarter time, 28 points at half time, 22 points at three quarter time eventually winning the match by 41 points. The betting odds reflected this dominance and Figure 7.1 shows a real-time plot of the in-play betting odds for both teams. It is virtually impossible to track the subtle changes in Port Adelaide's odds during the match due largely to the scale of the y-axis (back price), although a transformation could tease out this detail.

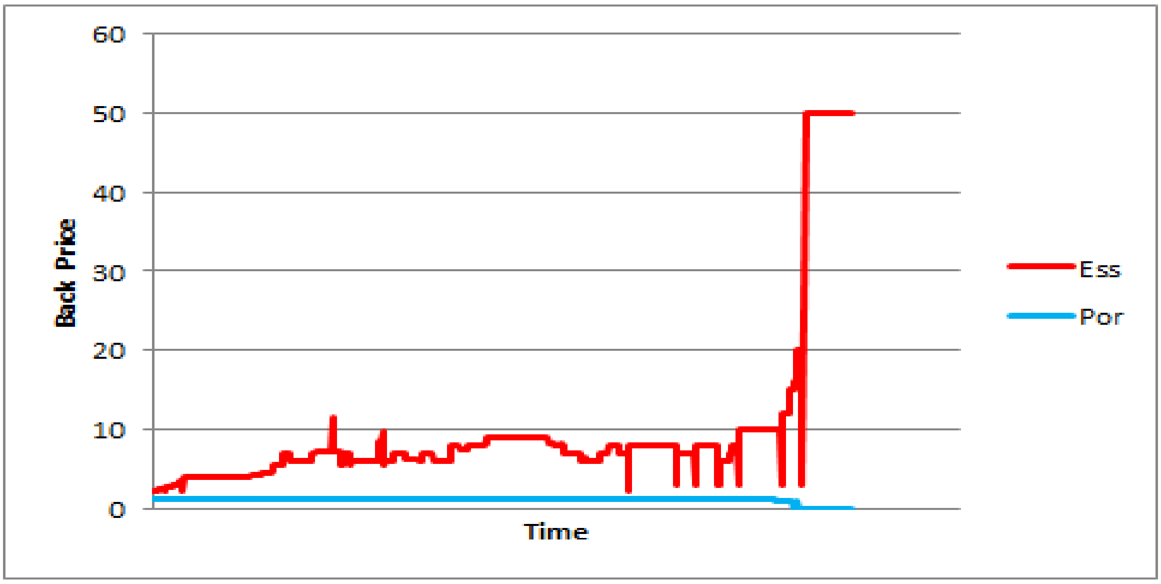


Figure 7.1: In-play betting odds of Port Adelaide vs. Essendon, round 1 season 2009

Therefore, an alternative benchmark of expectation is required based on the in-play betting data. This new measure should incorporate all available in-play betting data that was gathered in Chapter 6 (back price, back volume, lay price and lay volume) and be representative of who is going to win the match. Therefore, a probability assessment was deemed suitable for several reasons. First and foremost, it has a finite range which is important when the match is extremely one sided. Furthermore, a single probability is simple enough to be widely interpretable.

A previous study by Debnath et al. (2003) investigated information incorporation for in-play betting markets in Association Football and the National Basketball Association. Implied probabilities were calculated by taking the midpoint between back price and lay price and then normalising this value between 0 and 1. Two metrics were then used to measure the uncertainty of the implied probabilities, namely an Average Logarithmic Score (ALS) and Average Entropy (AE). The ALS is a standard measure of the accuracy of probability forecasts which can only be computed at game's end as it depends on the identity of the winning team. Conversely, the AE can be computed as the match progresses. The calculation of the ALS and AE are given in (7.1) and (7.2) respectively as in Debnath et al. (2003).

$$\text{ALS} = \frac{1}{N}\sum_{i=1}^{N} \text{log} p(t) \tag{7.1}$$

and

$$\text{AE} = \frac{1}{N}\sum_{i=1}^{N} -p(t)\text{log}p(t) - [1-p(t)]\log[1-p(t)] \tag{7.2}$$

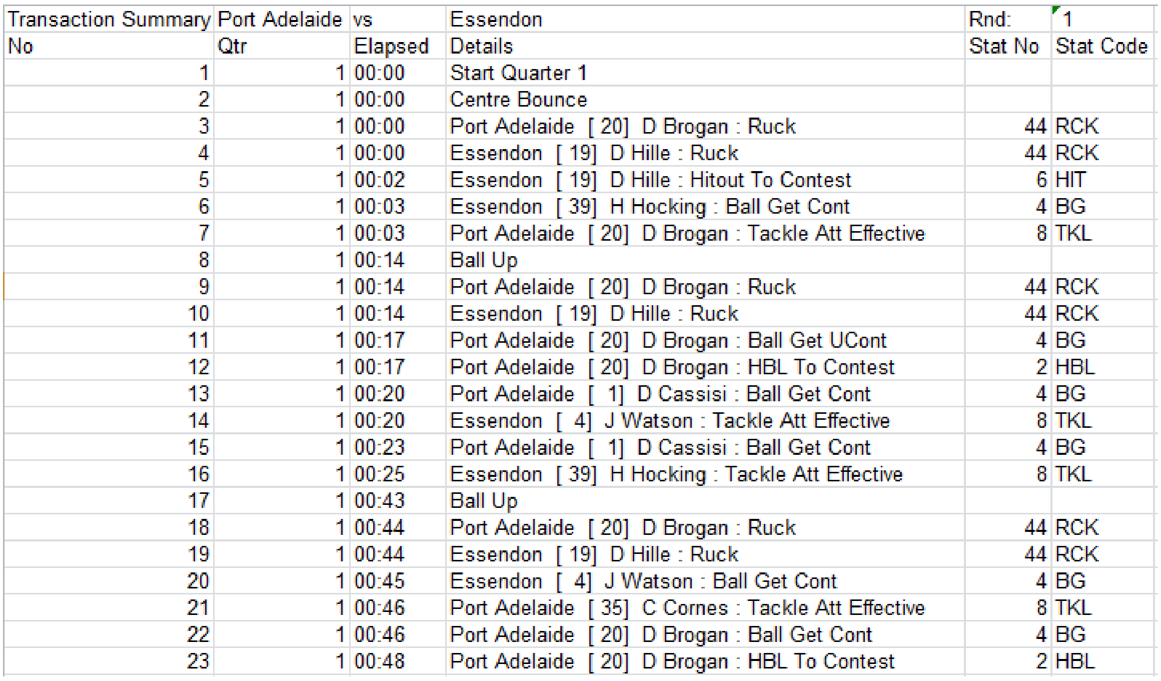
where $p(t)$ is the normalized implied probability deduced from the midpoint between back and lay prices.

However, the method for calculating the implied probabilities assumes the difference between the back price and lay price is negligible (i.e. the market is liquid), and it also does not incorporate the volume a punter is willing to risk. Therefore, in this chapter a new method is proposed to calculate implied probabilities using in-play betting data specifically for illiquid markets such as AFL.

7.2 Data

In Chapter 6, the in-play betting data for the 2009 AFL season was obtained using a fully customized program which integrates with Betfair's API. This yielded time-stamped odds for 115 out of 176 matches of the 2009 Home and Away season. Note that 61 matches were missing due to problems outside of control (e.g. internet connection cutting out). It should be noted that this sample did not contain a single draw. Recall in Section 6.3 of Chapter 6, the in-play betting data was recorded approximately every 12 seconds during a match and the data consisted of six variables for each match (timestamp, team name, back price, back volume, lay price, lay volume).

Recall one aim of this chapter was to examine the forecasting capabilities of the in-play odds against score difference as the match progresses. Furthermore, a real-time plot of implied probabilities deduced from in-play betting odds against score difference was also essential. Therefore, real-time performance data are required to extract the score difference as a function of time elapsed. The real-time performance data provided by ProWess Sports (e.g. transaction data) featured detailed timestamped performance data which meets this criteria. Figure 7.2 shows an extract of real-time performance data.



Transaction Summary	Port Adelaide	vs	Essendon	Rnd:	1
No	Qtr	Elapsed	Details	Stat No	Stat Code
1	1	00:00	Start Quarter 1		
2	1	00:00	Centre Bounce		
3	1	00:00	Port Adelaide [20] D Brogan : Ruck	44	RCK
4	1	00:00	Essendon [19] D Hille : Ruck	44	RCK
5	1	00:02	Essendon [19] D Hille : Hitout To Contest	6	HIT
6	1	00:03	Essendon [39] H Hocking : Ball Get Cont	4	BG
7	1	00:03	Port Adelaide [20] D Brogan : Tackle Att Effective	8	TKL
8	1	00:14	Ball Up		
9	1	00:14	Port Adelaide [20] D Brogan : Ruck	44	RCK
10	1	00:14	Essendon [19] D Hille : Ruck	44	RCK
11	1	00:17	Port Adelaide [20] D Brogan : Ball Get UCont	4	BG
12	1	00:17	Port Adelaide [20] D Brogan : HBL To Contest	2	HBL
13	1	00:20	Port Adelaide [1] D Cassisi : Ball Get Cont	4	BG
14	1	00:20	Essendon [4] J Watson : Tackle Att Effective	8	TKL
15	1	00:23	Port Adelaide [1] D Cassisi : Ball Get Cont	4	BG
16	1	00:25	Essendon [39] H Hocking : Tackle Att Effective	8	TKL
17	1	00:43	Ball Up		
18	1	00:44	Port Adelaide [20] D Brogan : Ruck	44	RCK
19	1	00:44	Essendon [19] D Hille : Ruck	44	RCK
20	1	00:45	Essendon [4] J Watson : Ball Get Cont	4	BG
21	1	00:46	Port Adelaide [35] C Cornes : Tackle Att Effective	8	TKL
22	1	00:46	Port Adelaide [20] D Brogan : Ball Get Cont	4	BG
23	1	00:48	Port Adelaide [20] D Brogan : HBL To Contest	2	HBL

Figure 7.2: Sample of AFL transaction data

7.3 Methods

Traditionally, implied probabilities of betting odds are calculated by taking the inverse of the betting odds (1/price) and normalizing the probabilities such that they sum to one. For example, in round 22 season 2009 Hawthorn hosted Essendon at the MCG. Hawthorn went into the game slight favourites paying \$1.87 for a win, while Essendon were paying \$2.05. Therefore, the implied probabilities of Hawthorn and Essendon are 53.48% (1/1.87) and 48.78% (1/2.05) respectively. Normalizing these probabilities gives Hawthorn and Essendon a 52.30% [53.48/(53.48+48.78)] and 47.70% [48.78/(53.48+48.78)] chance of winning respectively. However, betting exchanges allow punters to back (bet on) and lay (bet against) certain teams and there is often a discrepancy between these prices, particularly in illiquid markets such as in-play betting markets in AFL. Therefore, the back and lay prices need to be transformed into a unique value such that implied probabilities can be computed and then normalised. Debnath et al. (2003) suggests taking the midpoint between the back and lay price given by:

$$\text{Odds}_i(t) = \frac{BaPr_i(t) + LaPr_i(t)}{2} \tag{7.3}$$

where $BaPr_i(t)$ and $LaPr_i(t)$ are the back and lay price of team t at time i respectively.

However, this method for calculating the implied probabilities assumes the difference between the back price and lay price is negligible (i.e. the market is liquid) and it also does not incorporate the volume a punter is willing to risk. For example, in round 1 season 2009, Hawthorn were hosting Geelong in a grand final rematch of 2008 at the MCG. At quarter time punters could bet up to \$1,260 to back (bet on) Geelong to win at \$1.10, conversely punters could only lay a meagre \$5 (bet against) on Geelong at \$1.55. Due to the significant differences in the prices and volume bet, it is unreasonable to assume that the midpoint of the back and lay price in this example is a "fair price". Therefore, since the volume associated with each price is a known quantity, the back and lay price can be weighted against their respective volumes to give a "fairer" price. Now the odds of team t winning at time i is denoted:

$$\text{Odds}_i(t) = \frac{BaPr_i(t)BaVo_i(t) + LaPr_i(t)LaVo_i(t)}{BaVo_i(t) + LaVo_i(t)} \tag{7.4}$$

where $BaPr_i(t)$ and $LaPr_i(t)$ are the back and lay price of team t at time i respectively, similarly $BaVo_i(t)$ and $LaVo_i(t)$ are the volumes associated with the back and lay price of team t at time i respectively.

Now a unique price has been quantified for each team t at time i, the implied probabilities can be deduced by taking the inverse of the price given by

$$\text{Prob}_i(t) = \frac{1}{\text{Odds}_i(t)} \tag{7.5}$$

where $\text{Odds}_i(t)$ are the fair odds of team t at time i given in (7.4)

Akin to implied probabilities deduced from bookmakers odds, the implied probabilities deduced in (7.5) need to be normalized such that they sum to one. The normalized probabilities are given by

$$PROB = \text{Relative}_i(\text{Home}) = \frac{\text{Prob}_i(Home)}{\text{Prob}_i(Home) + \text{Prob}_i(Away)} \tag{7.6}$$

where $\text{Prob}_i(Home)$ and $\text{Prob}_i(Away)$ are the implied probabilities at time i of the nominated home and away team respectively given in (7.5).

7.4 Real-Time Plots

Recall the purpose of this chapter was to graphically display normalized implied probabilities deduced from betting odds against score difference as the match progresses. Therefore, the real time performance data (i.e. transaction data) need to be matched up against score difference. To match the real-time performance data against the in-play betting data both data sets need to have the same timestamp. Recall the transaction data provides the time elapsed for each quarter (mm:ss) while the timestamp for the in-play betting data is a 24 hour clock (hh:mm:ss). Therefore, the time between each quarter is an unknown quantity which can vary from match to match. However, according to the AFL, a maximum allocation of six minutes is allowed between the first and second, and third and fourth quarters; and 20 minutes between the second and third quarters. Therefore, given the transaction data contains the duration of each quarter, and assuming the match starts on time and that all matches have the previously mentioned quarter breaks, it's possible to approximate real-time elapsed from the performance data.

The units for time elapsed for the transaction data and the in-play betting data are both transformed to match seconds for simplicity. Therefore, since time elapsed for the in-play betting data is a 24 hour clock (hh:mm:ss), Excel functions can be easily implemented to calculate match seconds. Firstly, the time the match started needs to be transformed to seconds:

$$\text{start time} = (hh \times 24 \times 60) + (mm \times 60) + ss \tag{7.7}$$

Then the match seconds are computed after every transaction by calculating the current time elapsed (seconds) and subtracting the time the match started (seconds). For example, a Friday night match typically starts at 7:40pm. Therefore, the start time (seconds) is $(19 \times 60 \times 60) + (40 \times 60) = 70,800$, and the match seconds at time 8:23pm is given by $(20 \times 60 \times 60) + (23 \times 60) - 70,800 = 2,580$. However, the unit measurement for time elapsed for the transaction data is (mm:ss) which starts at 00:00 at the beginning of each quarter. Therefore, this can be transformed to match seconds by firstly computing the seconds elapsed for each quarter. Then match seconds is simply the sum of all previous quarter seconds plus the approximate quarter time breaks discussed earlier (quarter time = 360 seconds, half time = 1200 seconds, three quarter time = 360 seconds). For example, suppose the first quarter was 31:27 and the second quarter is midway through (09:15). The match seconds is simply $(31 \times 60 + 27) + (9 \times 60 + 15) + 360 = 2802$.

The next step is to extract the required information from the performance data and the in-play betting data. For example, from the performance data any changes in the score are required (i.e. goal or behind) and which team was responsible for the score. This sounds relatively straightforward, however the team name is embedded in a column with additional information (i.e. "Essendon [19] D Hille : Hitout To Contest") thus making it more difficult to extract. Furthermore, the number of characters for each team name varies, so careful consideration must be given as to which Excel function (or combination of functions) should be implemented. Additionally, the previously mentioned variables for the in-play betting data (back price, back volume, lay price and lay volume) need to be transformed to the normalized implied probabilities given in (7.6)

The final step is to generate a graphical display of the normalized implied probabilities deduced from betting odds against score difference as the match progresses. To further enhance the plot, vertical lines representing the beginning (or end) of quarter time breaks are superimposed on the graph. Therefore, the length of each quarter needs to be computed from the transaction data. Furthermore, a secondary axis is required for the score difference for easier readability. A macro is written in Excel to automate this procedure. Due to the sheer volume of VBA code this program was omitted from the Appendix. A couple of case studies are investigated to see what effect changes in score had on the normalized implied probabilities deduced from betting odds.

The first example is from round 1 season 2009 where Collingwood hosted Adelaide at the MCG. As it was the opening round of the season, the relative quality of the two teams could only be assessed on previous seasons results and the pre-season competition. In season 2008, Adelaide finished 5th and Collingwood finished 8th, therefore Adelaide hosted Collingwood at Football Park in an elimination final in the first week of the finals. Incidently, Collingwood won that game by 31 points only to be eliminated by St Kilda in the following week by 34 points. Furthermore, in the 2009 pre-season knockout competition, Adelaide were eliminated in the opening round albeit to the eventual winner Geelong. However, Collingwood made the pre-season final and was eventually defeated by Geelong by 76 points. It is important to note that the purpose of the pre-season competition is to prepare teams for the first round of the season proper, therefore not a great deal of importance is placed on winning matches. These previous results in addition to home advantage (Adelaide have to travel approximately 650kms) resulted in Collingwood ($1.33 to win) going into round 1 as favourite against Adelaide ($3.60 to win). Figure 7.3 shows the normalized implied probability of Collingwood winning the match against score difference.

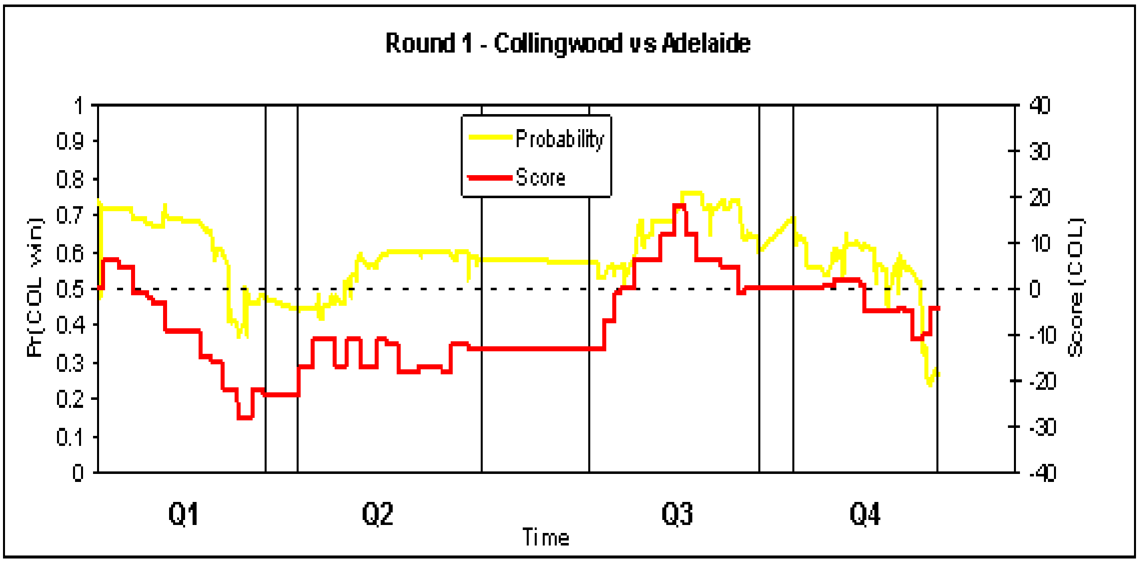


Figure 7.3: Real-time expectations deduced from betting odds: Collingwood vs. Adelaide, round 1 season 2009

At the end of the first quarter, Collingwood were behind 23 points, yet they were deemed almost an even chance of winning the match. In this instance, this expectation was later justified as Collingwood went down by a measly four points. Midway through the third quarter Adelaide were behind by almost 20 points and were deemed approximately a 25% chance of victory. Therefore, from this point in time, they far exceeded their expectations.

The second example is also in round 1 season 2009 where Hawthorn hosted Geelong at the MCG in a rematch of the 2008 Grand Final. Although Hawthorn won the grand final in season 2008, Geelong had by far the better season finishing top of the ladder (winning 21 out of 22 matches) and were four wins clear of the 2nd placed Hawthorn. Geelong also won the premiership in season 2007 and dominated the pre-season competition in season 2009, defeating Collingwood by 76 points in the final. Hawthorn also had several key players missing through injury. Therefore, Geelong went into the game as favourites ($1.42 for a win) against Hawthorn ($3.14 for a win). Figure 7.4 shows the normalized implied probability of

Hawthorn winning the match against score difference.

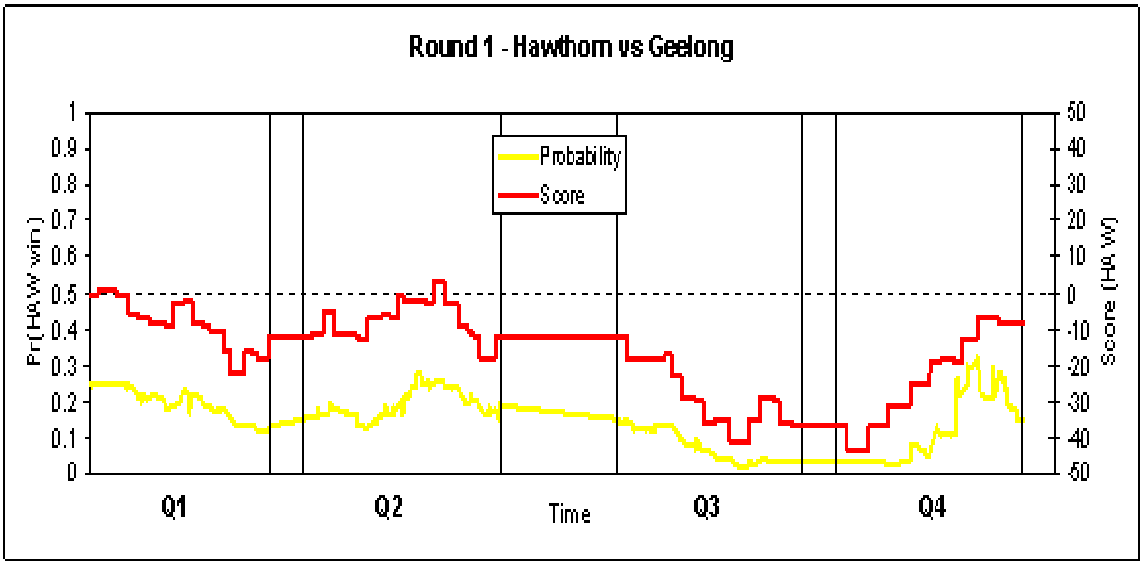


Figure 7.4: Real-time expectations deduced from betting odds: Hawthorn vs. Geelong, round 1 season 2009

At the end of the first quarter Geelong were leading by a meagre 12 points yet had approximately an 85% chance of winning the match. The margin stayed the same at half time as Hawthorn scored a late goal, Geelong's probability of winning decreased to approximately 80% since Hawthorn led at one point during the 2nd quarter. Geelong accumulated a significant lead by three quarter time leading by 37 points, making Hawthorn's task virtually impossible. Hawthorn were now rated about a 3% chance of winning the match, or $30 for a win. Interestingly, Hawthorn dominated the final quarter and it wasn't until they had kicked several goals in a row that the market rated Hawthorn a reasonable chance to win the match. Incidently, Geelong won the game by eight points after this late scare.

In summary, these figures show that early goals in a match seem to have little influence on the normalized implied probabilities, however a succession of goals by either team forces the market to update their expectations. Furthermore, the lead becomes more critical as the match progresses as teams have less opportunity to make up the deficit. Although these features may seem obvious they clearly become apparent when isolating individual matches.

7.5 Results

To measure the accuracy of the implied probability forecasts over time, comparisons are made between the percentage of games correctly classified by the probability forecasts and score difference at each of the quarter time breaks. The additional information the probability forecasts incorporated should be of greater importance at the earlier stages of the match since the outcome is largely unknown. However, as the game progresses the score difference should have greater influence as teams have less opportunity to make up a deficit. Obviously score difference is constant throughout the quarter time breaks, however slight changes in the probability forecasts can be expected (see Figure 7.3 and 7.4 for example). Therefore, the probability forecasts are taken at approximately the midpoint of each of the quarter time breaks. Table 7.1 displays the percentage of games correctly classified by the probability forecasts defined in (7.6) and score difference at each of the quarter time intervals. For example, if a team is leading on the scoreboard at quarter time then they are predicted to win according to score difference. Therefore, quarters whereby the score difference was equal to zero are removed from the analysis. Conversely, the team with a probability forecast of greater than 50% are predicted to win. Note that the probability forecasts are continuous therefore no quarters need to be remove from the analysis (i.e. prob$\neq$0.50).

Quarter	Score Difference	$PROB$
1	73.9	73.9
2	79.6	82.6
3	91.7	90.0

*reduced sample defined earlier

Table 7.1: Percentage of games correctly classified by $PROB$ and score difference, 2009*

Interestingly, both the probability forecasts and score difference predict the same number of games in the 1st quarter, probability forecasts outperform score difference in the 2nd quarter and in the 3rd quarter score difference outperforms the probability forecasts. The results are somewhat counterintuitive since the probability forecasts should outperform score difference at each of the quarter time breaks, since the probability forecasts incorporate additional information besides score difference. Furthermore, the difference between the performance of the probability forecasts and score difference in favour of the probability forecasts should decrease as the match progresses.

It is assumed that there is no significant year effect; that is, the likelihood of teams winning that are ahead on the scoreboard at the end of each quarter is the same as previous seasons. Table 7.2 shows the percentage of games correctly classified by score difference for seasons 2000 to 2009. Interestingly, the percentage of games correctly classified by score difference in the second quarter for the reduced sample (2009*) is very similar to the long term trend (2000 to 2008). However, in the first and third quarters the percentage of games correctly classified by score difference for the reduced sample (2009*) is significantly greater than the long term trend (2000 to 2008). This suggests that season 2009 was an aberration in terms of the likelihood of a team winning the match when they are ahead of the scoreboard. Therefore, all other things being equal, the probability forecasts should outperform score difference in subsequent seasons.

Season	Quarter 1	Quarter 2	Quarter 3
2000	72.0	78.6	86.8
2001	68.0	80.3	90.2
2002	75.7	80.1	86.9
2003	67.3	77.9	88.4
2004	66.8	83.0	90.6
2005	68.8	76.3	84.6
2006	66.5	81.7	88.5
2007	69.3	81.0	90.0
2008	67.5	75.1	82.8
AVE	69.1	79.3	87.6
2009	72.1	79.6	90.9
2009*	73.9	79.6	91.7

Table 7.2: Percentage of games correctly classified by score difference, 2000 to 2009

7.6 Discussion

It is important to note that the normalized implied probabilities deduced from the in-play betting odds does not represent the true probability of either team winning, since the true probability will always be unknown. However, over time the in-play betting markets will *always* correct towards the true probability. That is, although there are likely to be some inefficiencies for in-play betting markets, these inefficiencies will dissipate as the match progresses. For example, it is possible that in-play betting markets over (or under) inflate the importance of a goal relative to the likelihood of a team winning. However, as the match progresses this artificial inflation will evaporate. This self correction will occur due to the "wisdom of crowds", that is, the collective ideas (i.e. probability of winning) of a large crowd (i.e. betting exchange).

Chapter 8

The Efficiency of In-Play Betting Markets

This chapter examines the efficiency of in-play fixed odds betting markets in AFL at quarter time, half time and three quarter time. To begin, Section 8.1 provides a brief introduction to the Efficient Market Hypothesis (EMH). The data utilised in the analysis are detailed in Section 8.2. The following section, Section 8.3 showcases the methodology of the statistical tests of market efficiency used in the analysis. Furthermore, Section 8.4 shows the practical importance of the specific biases found using simple betting strategies. The next section, Section 8.5 details the assumptions of the data utilised in the analysis. To conclude, Section 8.6 provides possible reasons why specific biases were shown to be present. Material from this chapter has been published in Ryall and Bedford (2010a).

8.1 Introduction

The efficiency of both financial and betting markets has received great attention in academic literature. The fundamental question in both markets is whether the price incorporates all publically available information. A direct test of market efficiency in financial markets is complicated since the real value of a share in a company and the expected payoff is always unknown. Betting markets on the other hand, provide the perfect opportunity to test for market efficiency. The expected payoff (betting odds) for each wager is fixed and the outcome of each wager is settled at the conclusion of an event.

The much acclaimed paper on Efficient Market Hypothesis (EMH) by Fama (1970) defined market efficiency into three subsets: *Weak Form Efficiency*, whereby future prices can not be predicted by past prices; *Semi-Strong Efficiency*, whereby future prices can not be predicted by publically available information; and *Strong Form Efficiency*, whereby prices reflect all information, both public and private. Betting markets that fail these econometric tests of efficiency are only of practical importance if the bias is significant enough to be exploited via a profitable betting strategy in excess of commissions. Therefore, the efficiency of in-play fixed odds betting markets in AFL can be tested using the EFH of *Semi-Strong Efficiency.*

8.2 Data

In Chapter 6, the in-play betting data for the 2009 AFL season was obtained utilising Betfair's Application Programming Interface (API). This yielded time-stamped odds for 115 out of 176 matches for the 2009 Home and Away season. Note that 61 matches were missing. It should be noted that this sample did not contain a single draw. Recall in Section 6.3 of Chapter 6, the in-play betting data was recorded approximately every 10 seconds during a match and the data also consisted of six variables for each match (timestamp, team name, back price, back volume, lay price, lay volume). Since the purpose of this chapter is to determine the efficiency of in-play fixed odds betting markets for AFL matches at each of the quarter time breaks, the betting odds at each of the quarter time breaks needs to be determined. However, there are several inherent risks involved in extracting the quarter by quarter odds from the in-play betting data obtained in Chapter 6. Firstly, the betting odds can fluctuate significantly during quarter time breaks, therefore the point at which the odds are extracted is subjective. Furthermore, the timestamp for the in-play betting data is the real time, therefore the stage of the match (quarter) is uncertain. If real-time performance data are matched up against in-play betting data it is possible to estimate when the quarter time breaks occurred in the betting data (hence extract the in-play quarter by quarter odds), since performance data (i.e. a goal) should be reflected in the betting odds. This assumption is justified in Section 7.4 of Chapter 7.

To match the real-time performance data against the in-play betting data, both data sets need to have the same timestamp. Recall this was accomplished in Section 7.4 of Chapter 7 using a macro in Excel. This program was then further modified to extract the back price, back volume, lay price and lay volume for both teams at approximately the midpoint of the quarter time breaks. For matches that had large fluctuations of betting odds during quarter time breaks, a subjective decision was made as to what time point the odds should be extracted. Again due to the sheer volume of VBA code this program was omitted from the Appendix.

8.3 Methods

A unique feature of AFL is the fact the nominal home team does not always have an *priori* home ground advantage, the match can be played on a neutral ground and in some rare cases the opposition's home ground. Schnytzer and Weinberg (2008) overcame the problem of the nominated home team in AFL not necessarily having a perceived home advantage by defining a *priori* home team as follows:

> "*HOME* equals 1 if this home team is from a different city than the away team and the ground is the home team's ground (i.e. home advantage); *HOME* equals 0 if both teams are from the ground's city or from two cities other than the ground's city (i.e., no home advantage) or if this away team is from a different city than the home team and the ground is the home team's ground (i.e. away disadvantage)" (Schnytzer and Weinberg, 2008, p. 179).

Home teams could then be further split depending on whether the priori home team is Victorian or Non-Victorian, since it is widely assumed that Non-Victorian teams have a greater home advantage (Clarke, 2005); or whether the priori home team is Geelong since they are the only Victorian team to have a unique home ground. This chapter will use the same definition of a *priori* home team to test whether home or away bias exists for in-play fixed odds betting markets in AFL.

Line and fixed odds markets make up the majority of betting markets in game sports. In line-betting markets, the quality of the two teams is adjusted such that, in theory, both teams have an equal chance of winning. For example, if team A is deemed the lesser of the two teams they would receive a +l point advantage. Similarly, team B would receive a -l point disadvantage. In fixed odds betting markets, the objective is to predict the eventual winner regardless of the final margin.

If line-betting markets are efficient, the line is an unbiased predictor of the actual result and incorporates all publically available information. That is, the line should not be systematically higher or lower than the actual result. It has been suggested by Levitt (2004) that bookmakers in the NFL set prices (or lines) to maximise profits rather than balance the book since they are more skilled at predicting the outcome of games than bettors. Studies testing this balanced book approach include Paul and Weinbach (2007) and Paul and Weinbach (2008). These studies find mixed evidence of pricing as a means to exploit bettor biases and maximise profits. However, this chapter utilises data from a betting exchange thus removing the possibility of prices being set with the intent of exploiting better biases. Earlier studies test for market efficiency by running a simple linear regression model on the point spreads given by:

$$Y = \beta_0 + \beta_1 X_1 \tag{8.1}$$

where Y and X_1 are $N \times 1$ vectors of actual margins and point spreads respectively.

A statistical test of the joint null hypothesis, $\beta_0 = 0$ and $\beta_1 = 1$ is a test of betting market efficiency. A previous study by Golec and Tamarkin (1991) showed that the statistical test of efficiency in (8.1) is of low statistical power for testing the null hypothesis of unbiasedness when compared to a model that tests for specific biases, since β_0 measures the average of the biases. For example, consider a bias against favourites in AFL, a random sample of teams will result in approximately half favourites and half longshots leading to $\beta_0 = 0$.

Gray and Gray (1997) replace the dependent variable Y with the outcome of a bet, that is, whether a particular team beat the spread, since the margin a team beats the spread is irrelevant. Dare and MacDonald (1996) point out that a team that is favourite/underdog and home/away are characteristics that are interdependent. However, their specification does not account for home teams which are more likely than visiting teams to be the betting favourite. This leads authors such as Gray and Gray (1997) to biased coefficients. Dare and Holland (2004) consolidate research methods by both Gray and Gray (1997) and Dare and MacDonald (1996) in testing the efficiency of NFL betting markets. The specification proposed by Dare and Holland (2004) is used in this research with some slight modifications to account for (i) matches played at neutral grounds and (ii) betting markets that are in-play fixed odds. In this chapter a logistic regression model is applied using the binary variable Y defined as follows:

$$Y_i = \begin{cases} 1 \text{ win} \\ 0 \text{ loss} \end{cases} \tag{8.2}$$

For convenience, let team i be the favourite since every match has a favourite and thus an underdog. Hence the betting data needs to be transformed into a single probability assessment at quarter time, half time and three quarter time. Recall in Section 7.3 of Chapter 7, a relative probability was deduced in (7.6) from the in-play betting data by weighting the back price and lay price for both teams relative to their respective volume. Let $PROB_i$ denote the relative probability of the pre-game favourite winning at quarter i. Then the home-favourite (HF), neutral-favourite (NF), away-favourite (AF) and scoreboard bias ($AHEAD$) are introduced to the regression model defined in (8.1) akin to Dare and Holland (2004). Now the model for each quarter i can be written as:

$$Y_i = \beta_{0i} + \beta_{1i}PROB_i + \beta_{1i}HF_i + \beta_{1i}NF_i + \beta_{1i}AF_i + \beta_{1i}AHEAD_i \tag{8.3}$$

where $PROB$ is the relative probability deduced from the betting odds defined in (7.6) and

$$HF = \begin{cases} +1 \text{ \textit{priori} home team} \\ 0 \text{ otherwise} \end{cases}$$

$$NF = \begin{cases} +1 \text{ neutral home team} \\ 0 \text{ otherwise} \end{cases}$$

$$AF = \begin{cases} +1 \text{ \textit{priori} away team} \\ 0 \text{ otherwise} \end{cases}$$

$$AHEAD = \begin{cases} -1 \text{ behind on scoreboard} \\ 0 \text{ scores level} \\ +1 \text{ ahead on scoreboard} \end{cases}$$

Note that $PROB$ and $AHEAD$ are *in-game* measures and hence recalculated each quarter; whereas HF, NF and AF are *pre-game* measures and hence constant throughout the match. Logistic regression is then applied to the sample data which is split into three subsets: quarter time, half time and three quarter time. A statistical test of the joint null hypothesis $\beta_0 = 0, \beta_1 > 1, \beta_2 = 0, \beta_3 = 0$, and $\beta_4 = 0$ is a test of market efficiency for fixed odds in-play betting markets in AFL. Table 8.1 shows the results of the logistic regression analysis at quarter time, half time and three quarter time.

	Coefficient (p-value)		
Parameter	Quarter 1	Quarter 2	Quarter 3
PROB	3.21 (0.030)	9.88 (<0.001)	6.74 (0.015)
HF	-1.09 (0.269)	-4.96 (0.001)	-3.81 (0.035)
NF	-0.84 (0.416)	-4.72 (0.002)	-1.73 (0.314)
AF	-1.30 (0.203)	-5.53 (0.001)	-1.92 (0.281)
AHEAD	0.79 (0.007)	0.14 (0.710)	2.00 (0.011)

Table 8.1: Logistic regression results: Semi-strong efficiency estimates of (8.3)

The results of the logistic regression model defined in (8.3) at quarter time suggest *PROB* gives a good indication of the eventual winner (p = 0.030), however there is strong evidence that teams which are leading are underbet (p = 0.007). At half time however, *PROB* gives a strong indication of the eventual winner (p < 0.001) and strong evidence that *all* favourites (*HF*, *NF* and *AF*) are overbet with p-values 0.001, 0.002 and 0.001 respectively. At three quarter time *PROB* gives a good indication of the eventual winner (p = 0.015), there is some evidence that *HF* are overbet (p = 0.035) and strong evidence that teams which are leading are underbet (p = 0.011).

Akin to Dare and Holland (2004), another model is proposed whereby *PROB* and *AHEAD* are multiplied by the dummy variables *HF*, *NF* and *AF* to account for the possibility that the effects of *PROB* and *AHEAD* differ by the types of teams playing the game. Therefore, the alternative logistic regression model is given by:

$$\begin{aligned} Y_i &= \beta_{0i} + \beta_{1i} HF + \beta_{2i} NF + \beta_{3i} AF \\ &+ \beta_{4i} HF_{PROB} + \beta_{5i} NF_{PROB} + \beta_{6i} AF_{PROB} \\ &+ \beta_{7i} HF_{AHEAD} + \beta_{8i} NF_{AHEAD} + \beta_{9i} AF_{AHEAD} \end{aligned} \tag{8.4}$$

Table 8.2 shows the results of the logistic regression analysis at quarter time, half time and three quarter time. Note that some coefficients and their corresponding p-values are missing due to the outcome variable being completely determined by the corresponding predictor variables. Therefore, some caution should be taken with these results due to small sample sizes.

	Coefficient (p-value)		
Parameter	Quarter 1	Quarter 2	Quarter 3
HF	-0.87 (0.643)	-7.37 (0.014)	-3.34 (0.243)
NF	8.61 (< 0.001)	-66.9 (0.992)	
AF	-2.08 (0.134)	-3.67 (0.042)	6.20 (< 0.001)
HF_PROB	2.82 (0.355)	14.14 (0.007)	6.11 (0.189)
NF_PROB	4.44 (0.033)	6.96 (0.011)	5.35 (0.098)
AF_PROB	1.40 (0.665)		
HF_AHEAD	0.49 (0.370)	-0.79 (0.208)	1.76 (0.054)
NF_AHEAD	0.74 (0.071)	0.52 (0.294)	
AF_AHEAD			

Table 8.2: Logistic regression results: Semi-strong efficiency estimates of (8.4)

The results suggest that in the first quarter AF are significantly underbet ($p < 0.001$), however this could be attributed to all AF ($n = 8$) who are leading at quarter time going on to win the match in this sample. At half time HF and NF are significantly overbet ($p =$

0.014 and $p = 0.042$ respectively). $PROB$ gives a strong indication of the eventual winner for HF, NF ($p = 0.007$ and $p = 0.011$ respectively) and in the case of AF predicts the eventual winner perfectly ($n = 12$). At three quarter time NF are significantly underbet ($p < 0.001$) and there is some evidence to suggest that HF which are $AHEAD$ are underbet ($p = 0.054$).

8.4 Betting Strategies

Although it has been shown that in-play fixed odds betting markets in AFL for season 2009 do not incorporate all publically available information, this is only of practical importance if a betting strategy exists that results in positive profits after the 5% commission for winning bets deduced from standard users. To test this hypothesis two simple betting strategies are implemented. The first strategy is the "back" approach whereby \$5 (minimum bet on betfair) is bet on team A to win, the second strategy is the "lay" approach whereby \$5 is bet against team A to win. An upper limit of \$30 was placed on the lay price due to some unrealistic in-play lay prices in the vicinity of \$1000 thus having a significant influence on the overall return on investment (ROI). For example, in round 13, season 2009, the Kangaroos (ranked 3rd) hosted the Western Bulldogs (ranked 13th) at the MCG. At three quarter time the Western Bulldogs were leading by five points and one punter (or combination of punters) wanted to back the Kangaroos at odds of 1000 to 1 to win! Incidently, Western Bulldogs went on to win by 22 points so a \$5 bet against the Kangaroos to win would have netted \$5, however the punter would have risked an astonishing \$5000 for this bet. On the other hand, the upper limit of \$30 is not necessary for the back approach since total liability is simply volume bet, however for the lay approach total liability is volume bet × lay price. Betting strategies include home favourite (HF), neutral favourite (NF), away favourite (AF), home underdog (HU), neutral underdog (NU) and away underdog (AU) depending on whether they are ahead or behind on the scoreboard. Table 8.3 shows the results for the first quarter.

	Back				Lay			
Strategy	# Bet	Liability	Profit/Loss	ROI	# Bet	Liability	Profit/Loss	ROI
HF is ahead on scoreboard	26	$130.00	-$2.00	-1.60%	26	$48.40	- $16.50	-34.00%
NF is ahead on scoreboard	31	$155.00	$1.10	0.70%	31	$52.00	-$25.50	-49.00%
AF is ahead on scoreboard	8	$40.00	$14.70	36.70%	8	$24.20	-$24.20	-100.00%
HU is ahead on scoreboard	12	$60.00	-$4.60	-7.60%	9	$76.50	-$12.20	-16.00%
NU is ahead on scoreboard	21	$105.00	$20.60	19.60%	19	$211.50	-$60.80	-28.70%
AU is ahead on scoreboard	14	$70.00	$2.10	2.90%	14	$128.60	-$37.40	-29.10%
HF is behind on scoreboard	14	$70.00	$4.40	6.30%	11	$99.00	-$25.10	-25.40%
NF is behind on scoreboard	21	$105.00	-$47.70	-45.40%	18	$285.50	$19.20	6.70%
AF is behind on scoreboard	12	$60.00	-$19.80	-33.00%	9	$91.90	-$33.30	-36.30%
HU is behind on scoreboard	8	$40.00	-$40.00	-100.00%	7	$127.00	$33.30	26.20%
NU is behind on scoreboard	32	$160.00	-$77.50	-48.50%	13	$494.40	$12.80	2.60%
AU is behind on scoreboard	26	$130.00	-$81.40	-62.60%	12	$479.50	$42.30	8.80%

Table 8.3: Betting simulations for AFL in-play betting markets at *quarter time*

Firstly, note the discrepancies between the number of bets for the back and lay approach using the same betting strategy, particularly when betting on a team which is behind. This discrepancy can be attributed to a team's chance of victory being extremely high and no one is willing to lay (not win) the bet. For example, in round 8 season 2009 Geelong (ranked 2nd) were leading Kangaroos (ranked 14th) 33 to 8 at quarter time. At that point no-one was willing back the Kangaroos to win hence you couldn't accept odds to lay (bet against) the Kangaroos. This discrepancy is likely to increase as the match progresses since teams have more time to manufacture a bigger differential between scores.

Two clear betting strategies exist at quarter time: backing the AF when they are ahead on the scoreboard (ROI = 36.7%, $n = 8$) or conversely laying the HU when they are behind on the scoreboard (ROI = 26.2%, $n = 7$); and backing the NU when they are ahead on the scoreboard (ROI = 19.6%, n=21) or conversely laying the NF when they are behind on the scoreboard (ROI = 6.7%, $n = 18$). However, from Table 8.2 AF that were ahead on the scoreboard went on to win every match which is surely the exception, not the rule, since $n = 8$. Table 8.4 shows the results at half time.

	Back				Lay			
Strategy	# Bet	Liability	Profit/Loss	ROI	# Bet	Liability	Profit/Loss	ROI
HF is ahead on scoreboard	23	$115.00	-$10.90	-9.50%	26	$36.9	$0.20	0.40%
NF is ahead on scoreboard	31	$155.00	-$3.10	-2.00%	34	$41.45	-$21.90	-52.80%
AF is ahead on scoreboard	12	$60.00	$7.40	12.30%	12	$26.15	-$26.20	-100.00%
HU is ahead on scoreboard	8	$40.00	$3.40	8.40%	6	$50.1	-$0.40	-0.70%
NU is ahead on scoreboard	19	$95.00	-$3.60	-3.80%	20	$211.35	-$61.90	-29.30%
AU is ahead on scoreboard	13	$65.00	-$13.20	-20.30%	12	$75.05	$8.60	11.40%
HF is behind on scoreboard	13	$65.00	-$15.60	-23.90%	9	$89.55	-$18.30	-20.40%
NF is behind on scoreboard	20	$100.00	-$48.30	-48.30%	14	$155.3	$10.50	6.80%
AF is behind on scoreboard	8	$40.00	-$33.00	-82.40%	4	$169.45	$9.80	5.80%
HU is behind on scoreboard	12	$60.00	-$60.00	-100.00%	7	$172.75	$6.00	3.50%
NU is behind on scoreboard	34	$170.00	-$118.70	-69.80%	16	$732.75	$30.30	4.10%
AU is behind on scoreboard	27	$135.00	-$93.00	-68.90%	10	$5300.9	$8.90	0.20%

Table 8.4: Betting simulations for AFL in-play betting markets at *half time*

Again two clear betting strategies exist at half time: backing the AF when they are ahead on the scoreboard (ROI = 12.3%, $n = 12$) or conversely laying the HU when they are behind on the scoreboard (ROI = 3.5%, $n = 7$); and backing the HU when they are ahead on the scoreboard (ROI = 8.4%, $n = 8$) or conversely laying the AF when they are behind on the scoreboard (ROI = 5.8%, $n = 4$). However, again care should be taken with these betting strategies due to extremely small sample sizes. Table 8.5 shows the results at three quarter time.

	Back				Lay			
Strategy	# Bet	Liability	Profit/Loss	ROI	# Bet	Liability	Profit/Loss	ROI
HF is ahead on scoreboard	20	$100.00	$5.60	5.60%	28	$26.70	-$17.50	-65.50%
NF is ahead on scoreboard	22	$110.00	$6.10	5.60%	34	$30.40	-$30.40	-100.00%
AF is ahead on scoreboard	9	$45.00	$8.10	17.90%	10	$15.60	-$15.60	-100.00%
HU is ahead on scoreboard	7	$35.00	$0.20	0.60%	6	$10.50	-$10.50	-100.00%
NU is ahead on scoreboard	17	$85.00	$7.40	8.70%	20	$68.40	-$40.50	-59.20%
AU is ahead on scoreboard	9	$45.00	$16.60	36.90%	8	$50.10	-$20.40	-40.60%
HF is behind on scoreboard	10	$50.00	-$40.30	-80.50%	6	$99.50	$13.80	13.80%
NF is behind on scoreboard	21	$105.00	-$68.00	-64.80%	6	$86.70	-$15.80	-18.20%
AF is behind on scoreboard	8	$40.00	-$40.00	-100%	2	$24.80	-$5.30	-21.20%
HU is behind on scoreboard	10	$50.00	-$50.00	-100%	4	$152.00	$19.00	12.50%
NU is behind on scoreboard	34	$170.00	-$170.00	-100%	7	$497.00	$33.30	6.70%
AU is behind on scoreboard	29	$145.00	-$133.70	-92.20%	2	$102.50	$9.50	9.30%

Table 8.5: Betting simulations for AFL in-play betting markets at *three quarter time*

In the third quarter two clear betting strategies exist: backing the AF when they are ahead on the scoreboard (ROI = 17.9%, n = 9) or conversely laying the HU when they are behind on the scoreboard (ROI = 12.5%, n = 4); and backing the AU when they are ahead on the scoreboard (ROI = 36.9%, n = 8) or conversely laying the HF when they are behind on the scoreboard (ROI = 13.8%, n = 6). Again care should be taken with these betting strategies due to extremely small sample sizes.

8.5 Assumptions

In this chapter, testing the efficiency of fixed odds in-play betting markets in AFL was conducted on the 2009 Home and Away season. It is assumed that there is no significant year effect; that is, the likelihood of teams wining that are HF, NF, AF, HU, NU, AU and ahead or behind on the scoreboard at the end of each quarter is similar to previous years. Table 8.6 displays the proportion of games won by the HF, NF, AF, HU, NU and AU which is leading at the end of each quarter for seasons 2000 to 2009.

	HF			*NF*			*AF*			*HU*			*NU*			*AU*		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
2000	80.5	86.7	95.5	90.2	89.4	93.6	77.8	78.6	82.6	60.0	73.1	83.3	53.8	54.2	66.7	51.9	69.6	83.3
2001	84.8	90.0	95.8	69.0	81.0	90.5	75.0	84.2	100.0	38.5	50.0	69.2	51.6	65.6	83.3	66.7	90.0	91.3
2002	89.8	94.0	94.5	81.0	83.8	89.7	72.0	76.2	94.4	94.4	80.8	90.6	65.2	67.6	71.6	33.3	53.8	70.0
2003	87.8	89.6	96.9	75.0	85.7	97.4	75.0	91.7	94.1	50.0	63.9	82.1	60.6	75.8	81.6	35.4	42.9	62.5
2004	81.6	92.5	91.2	76.5	88.9	94.2	54.5	78.6	100.0	53.8	90.0	85.7	54.0	64.1	86.5	41.7	75.0	82.4
2005	85.7	90.7	95.7	78.0	76.0	84.0	81.8	75.0	76.9	73.7	72.2	76.5	55.9	63.0	75.0	45.5	69.6	85.0
2006	84.6	95.0	95.1	79.7	88.9	95.9	76.5	88.2	100.0	50.0	66.7	66.7	41.3	61.8	82.1	60.7	77.8	77.8
2007	84.4	85.4	94.1	78.6	85.6	92.6	77.8	95.8	96.2	57.7	100.0	100.0	50.0	63.3	75.9	56.9	70.8	86.0
2008	86.8	92.7	90.6	81.0	83.0	91.1	69.2	73.3	81.3	41.2	46.7	53.8	44.0	54.5	70.5	51.9	64.6	73.7
AVE	85.1	90.7	94.4	78.8	84.7	92.1	73.3	82.4	91.7	57.7	71.5	78.7	52.9	63.3	77.0	49.3	68.2	79.1
2009	84.6	85.0	93.5	84.1	90.2	96.7	84.6	96.7	93.3	65.8	82.4	90.0	60.0	70.4	82.1	43.5	45.5	78.6
2009*	84.6	85.2	96.6	85.9	89.7	98.5	100.0	95.8	100.0	62.5	87.5	93.8	61.9	65.0	81.0	50.0	53.8	90.0

Table 8.6: Percentage of games won by the Home Favourite (*HF*), Neutral Favourite (*NF*), Away Favourite (*AF*), Home Underdog (*HU*), Neutral Underdog (*NU*) and Away Underdog (*AU*) which is leading at the end of each quarter, 2000 to 2009

The results suggests there is a pronounced year effect particularly in the first and third quarter. However, the mere fact that certain teams which are ahead on the scoreboard have won more games in 2009 than previous years is no cause for concern, the average lead over the years could also be different which in turn would affect the in-play odds.

A few anomalies include the *AF* leading on the scoreboard winning 100% of the time at quarter time and three quarter which leaves serious doubt about the ROI of this strategy shown in Table 8.3 and Table 8.5. Secondly, *AU* leading on the scoreboard at three quarter time won 90.0% in 2009* compared to 79.1% in seasons 2000 to 2008, however even if the most profitable win was removed (odds $3.40) the ROI would still be 13.6% based on the *AU* winning 80% of matches at three quarter time when they are leading.

8.6 Discussion

In Australia, the Interactive Gambling Act 2001 makes it an offence to place a bet on a sporting event via the internet once the event has started, bets in-play are only permitted via telephone. This is likely to result in bettors missing potential opportunities due to time constraints. In AFL, the quarter time and three quarter time breaks are approximately six minutes, however at half time the break is approximately 20 minutes. This gives bettors more time to evaluate the likelihood of a team winning during the half time break. This suggests that the short quarter time and three quarter time breaks combined with the time constraints of betting via telephone could explain the inefficiencies of the in-play betting markets in AFL at these stages of the game. Similarly, the long half time break could explain the efficiency of the in-play betting markets in AFL at half time.

There is also the possibility that in-play betting takes the form of a hedge against original wagers which were placed in the pre-game betting market. It is difficult to justify under standard utility theory that a person willing to gamble at an expected loss in a pre-game betting market would turn around and hedge in-play betting markets. However, if the expected value of the pre-game wager is positive (due to favourite-longshot bias or its reverse in the pre-game market), then rational bettors may choose to hedge during the game via in-play betting markets. If the market is dominated by bettors who wish to hedge, it could explain excess returns generated within the in-play betting market since the market is not sufficiently liquid.

Chapter 9

Intra-Match Home advantage

Home advantage typically refers to the net advantage of several factors which, generally speaking, have a positive effect on the home team and a negative effect on the away team. However, this practice excludes the in-course dynamics of home advantage throughout the match, including the interrelationship between pre-game and in-game team characteristics. In this chapter, the aim is to calculate the intra-match home advantage for each quarter in AFL by incorporating the interaction between team quality and current score. Section 9.1 provides a brief introduction of home advantage in sport with a specific focus on intra-match home advantage. Section 9.2 details the methodology used in this Chapter. To conclude, Section 9.3 discusses the results. Material from this chapter has been published in Ryall and Bedford (2011b).

9.1 Introduction

In predicting the outcome of AFL matches it has been shown that both home advantage and the quality of the two teams play an important role in predicting success as outlined in Chapter 4 and Chapter 5. Home advantage typically refers to the net advantage of several factors which, generally speaking, have a positive effect on the home team and a negative effect on the away team (Harville and Smith, 1994). The much acclaimed paper (Schwartz and Barsky, 1977) on home advantage in American team sports (major league baseball, college and professional football, professional ice hockey, and college basketball) showed its existence and how it varied from one sport to another. They attributed home advantage to a combination of learning/familiarity (tactical) factors, travel (physiological) factors and crowd (psychological) factors. Courneya and Carron (1992) build on this and suggest referee bias as another factor to consider. Although these factors are usually cited as the cause of home advantage in team sports, the precise contribution of each factor still remains relatively unknown (Pollard, 2008).

Several studies support the argument that sport performance consists of a complex series of interrelationships between performance variables, and simple frequency data can't fully explain this interaction process (Borrie et al., 2002). If this argument holds true, one could contend that post-match point differentials between home and away teams does not fully explain home advantage. Therefore, although it was shown home advantage was attributed to a combination of travel and familiarity factors in Chapter 4, this does not account for the possibility of in-game home advantage attributes. More relevant studies on intra-match home advantage in team sports include Jones (2007) and Marcelino et al. (2009) on basketball (NBA) and volleyball respectively. Both studies investigated how home advantage is accumulated during the course of a match. It should be noted that the work of Jones (2007) forms the foundation of this chapter and is therefore quoted extensively throughout. The goal of this chapter is to build upon the work of Jones (2007) and quantify home advantage in AFL as an *intra*-match measure by incorporating complex interrelationships between team quality and score difference.

9.2 Methods

This chapter's analysis is based on AFL seasons 2000 to 2009. AFL data was gathered from ProEdge a statistical package developed by ProWess Sports (http://www.prowess.com.au). Data consisted of year, round, quarter, (nominal) home team, away team, home team score and away team score. Furthermore, information regarding the stadium of each match was gathered to establish whether the match involved a *priori* home team. Akin to Section 8.3 of Chapter 8, the *priori* home team defined in Schnytzer and Weinberg (2008) was utilised. Of the 1760 matches in seasons 2000 to 2009 there were 989 matches which involved a *priori* home team.

In a balanced schedule, where each team plays each other team as many times with one team at home as the other, home advantage is typically expressed as the average difference between the home and away team score (Stefani and Clarke, 1992). This balance allows home advantage to be obtained which is not confounded with team quality. For example, in Association football there are currently 20 teams with 38 matches in a regular season, such that each team plays every other team once at home and once away. However, in AFL the competition is unbalanced with respect to team quality and home advantage. Therefore, it is important to adjust the margin of victory for team ratings when quantifying home advantage in AFL (Clarke, 2005).

However, when investigating the intra-match home advantage in team sports, it is important to adjust the results for team quality for balanced *and* unbalanced competitions during the game. Jones (2007) concludes that:

> "Before the game starts the home team can expect to win the game roughly 62.0% of the time. If the home team is behind at the end of the first quarter, that percentage drops to 44.4% in 2002-03 and 43.8% in 2003-04. The home advantage is not something that the home team retains regardless of how it performs during the game. If the home team lets itself be outscored in the first

quarter, then the advantage it had when the game started is lost." (Jones, 2007, p. 11).

This concluding remark contradicts the finding that home advantage is greatest when the home team is behind on the scoreboard. It can be argued that the decrease in home win percentage from 62% pre-game to 44% at the end of the first quarter if the home team is behind, is most likely going to be *caused* by the difference in team quality. For example, in round 14 season 2010 West Coast (16th) hosted Collingwood (1st) at Subiaco Oval. Collingwood led at the end of every quarter going on to win by 81 points. This suggests that when home teams are behind during a match this is arguably more indicative of a superior opponent than any home advantage being negated. Therefore, it is important to obtain quarter by quarter team ratings to adjust margin of victory when quantifying intra-match home advantage. In Chapter 4 the Average Winning Margin (AWM) for each team split by season were used team as ratings to deduce the home advantage. That is,

$$h_{ij} = a_{ij} - (r_i - r_j) \tag{9.1}$$

where, r_i is the rating of team i, r_j is the rating of team j, a_{ij} is the actual margin of victory of team i against team j and h_{ij} is the home advantage which is aggregated and averaged out.

Since the purpose of this research is to quantify home advantage as an intra-match measure, ratings for each team for each quarter are required. This can be achieved by calculating teams' AWM for each quarter or using teams' AWM at game's end and dividing it by four. Since teams' AWM for subsequent quarters after the first quarter are not necessarily independent, teams' AWM at game's end divided by four are used as team ratings. Now home advantage is defined as

$$h_{ij}^k = a_{ij}^k - \left(r_i^k - r_j^k\right) \tag{9.2}$$

where the previously mentioned ratings, actual margin of victory and home advantage are now within quarter measures for given quarter $k = 1, \ldots, 4$, such that

$$h_{ij} = \sum_{k=1}^{4} h_{ij}^{k} \tag{9.3}$$

For example, in round 2 season 2009 Essendon defeated Fremantle by +38 points (a_{ij}), Essendon's AWM in 2009 was -2.5 points (r_i) similarly Fremantle's AWM in 2009 was -18.2 points (r_j). Table 9.1 tabulates the values of the parameters in (9.2) for this example.

Quarter	r_i^k	r_j^k	a_{ij}^k	h_{ij}^k
1	-0.6	-4.6	+22	+18.0
2	-0.6	-4.6	-5	-9.0
3	-0.6	-4.6	+6	+2.0
4	-0.6	-4.6	+15	+11.0

Table 9.1: Intra-match home advantage parameter values in (9.2) for Essendon vs. Fremantle example

The next stage was to incorporate pre-game and in-game characteristics of home and away teams to determine their influence on home advantage during the course of the game. These characteristics included the ratings of the two teams (pre-game) and score difference (in-game). Therefore, the ratings of the two teams are subtracted to ascertain whether the home team is the favourite ($r_{home} - r_{away} > 0$) or the underdog ($r_{home} - r_{away} < 0$). Similarly, the current score of the two teams are subtracted to ascertain whether the home team is ahead or behind on the scoreboard at the end of each quarter. This results in four unique categories of the home team in quarter two, three and four, namely Home Favourite Ahead (*HFA*), Home Favourite Behind (*HFB*), Home Underdog Ahead (*HUA*), Home Underdog Behind (*HUB*). Akin to Jones (2007), a small percentage of matches are excluded from the analysis when a quarter was neither won nor lost by the home team. Although it is possible to break differences in ratings and score into further subsets, this was not undertaken primarily because the choice of ranges is subjective. This makes the interpretation of results more challenging due to a reduction of the sample size and therefore the results are weaker.

9.3 Results

The overarching aim of this chapter was to determine the dynamic interaction the difference in team ratings and score difference have on home advantage throughout the match. Before this can be ascertained, it was important to investigate the descriptive statistics of home and away teams throughout the match as a point of reference. Table 9.2 displays the mean difference between the home and away team score ($\bar{\Delta}$) which is adjusted for team quality and split by quarters. There is some objective disagreement with Jones (2007) that home advantage is frontloaded (greatest at the beginning of the match) since home advantage in AFL is greatest in the third quarter. However, a paired t-test showed the decrease in $\bar{\Delta}$ from the third to the fourth quarter was significant at the 5% significance level ($p = 0.04$).

Quarter	Home	Away	$\bar{\Delta}$
1st	24.34	20.74	3.59
2nd	24.31	21.56	2.75
3rd	25.74	22.01	3.73
4th	24.76	22.24	2.52

Note. Results are adjusted for team ratings in (9.2)

Table 9.2: Mean difference between the home and away team score ($\bar{\Delta}$), 2000 to 2009.

In the first quarter, it is possible to distinguish whether home advantage is greater for Home Underdogs (HU) or Home Favourites (HF) since the results are adjusted for team quality. Table 9.3 displays $\bar{\Delta}$ in the first quarter which is split by HU and HF. Although there is a slight increase in $\bar{\Delta}$ when the home team is also the underdog, an independent unpaired

t-test showed that this difference in $\bar{\Delta}$ was not statistically significant at 5% significance level ($p = 0.80$).

Team	Home	Away	$\bar{\Delta}$
HF	22.45	19.23	3.22
HU	26.97	22.87	4.09

Note. Results are adjusted for team ratings in (9.2)

HF = Home Favourite, HU = Home Underdog

Table 9.3: Mean difference between the home and away team score ($\bar{\Delta}$) in the first quarter split by underdog/favourite, 2000 to 2009

Many studies based on archival research provide evidence that supportive audiences can actually affect players to perform poorly in Championship matches (see, for example, Baumeister and Steinhilber (1984) in baseball and basketball; and Wright et al. (1995) in ice hockey). Even though the results are somewhat counterintuitive, this has been well supported by subsequent laboratory experiments (Butler and Baumeister, 1998). In their study, performers believed that supportive audiences were more helpful and less stressful. However, the results indicated that when respondents were required to perform a difficult task in front of supportive audiences they elicited cautious behaviour, that is, speed decreased without improving accuracy. Another study by Wolfson et al. (2005) showed that 11% of supporters believed home advantage could be detrimental to the home team due to players feeling more pressure at home.

Championships are generally determined over a best of N matches. Championships which are thus determined by the final match of the series are indicative of two teams of a similar standing where the outcome is highly uncertain. The same theory can be applied to AFL during the match, that is, do home teams perform poorly when the match is there to be won? Therefore, the next stage was to determine the impact, if any, the difference in team ratings and score difference have on home advantage throughout the match. Table 9.4

displays the mean difference between home and away team score ($\bar{\Delta}$) and standard deviation (S_{Δ}) in the 2nd, 3rd and 4th quarters as a function of pre-game and in-game characteristics of the home team at the end of the previous quarter. In this table, $\overline{SD}$=mean score difference and $\overline{RD}$=mean rating difference.

						Next Quarter			
Quarter	Team	N	$\overline{SD}$	$\overline{RD}$	Win %	Home	Away	$\bar{\Delta}$	$S_{\bar{\Delta}}$
1	*HFA*	342	16.82	24.06	89.60%	23.55	21.97	1.58	16.47
	HFB	155	-10.54	17.89	63.90%	24.39	21.62	2.77	16.48
	HUA	229	13.04	-18.65	60.30%	24.83	21.48	3.35	17.73
	HUB	246	-13.62	-22.89	27.80%	24.26	22.10	2.16	16.29
2	*HFA*	373	25.74	24.95	93.70%	25.81	22.19	3.62	16.54
	HFB	121	-14.47	13.78	45.50%	23.99	22.52	1.47	18.82
	HUA	216	19.57	-17.11	72.20%	25.94	21.06	4.88	16.81
	HUB	258	-21.13	-23.99	18.60%	26.17	22.06	4.11	15.93
3	*HFA*	398	36.12	24.45	96.00%	25.41	23.05	2.36	16.71
	HFB	105	-20.32	14.08	30.00%	24.94	19.80	5.14	16.37
	HUA	224	24.00	-15.43	81.50%	24.66	21.49	3.17	17.02
	HUB	255	-27.36	-25.49	10.20%	23.81	22.68	1.13	16.77

Note. Lead, win %, home, away and Δ are adjusted for team ratings

HFA = Home Favourite Ahead, *HFB* = Home Favourite Behind, *HUA* = Home Underdog Ahead, *HUB* = Home Underdog Behind.

Table 9.4: Mean difference between home and away team score ($\bar{\Delta}$) in the 2nd, 3rd and 4th quarters as a function pre-game and in-game characteristics of the home team at the end of the previous quarter, 2000 to 2009

An example might help to clarify how to interpret the results. If home favourites are ahead on the scoreboard (*HFA*) at the end of the first quarter then they can expect, on average, to outscore their opponents by +1.58 points in the second quarter after adjusting for team quality. When interpreting the results careful consideration must be given to the change in characteristics of each of the four categories of the home team (*HFA*, *HFB*, *HUA* and *HUB*) as the match progresses. For example, *HUA* in the first quarter had an average lead ($\overline{SD}$) of +13.04 and an average rating difference ($\overline{RD}$) of -18.65, however as the match progress the *HUA* lead increases and the rating difference decreases. This indicates that the *HU* that is leading in subsequent quarters are likely to be opposed to weaker favourites. Another example is the *HFB*. Note the decrease in N and the average rating difference as the match progresses, this indicates that the *HF* that is behind is likely to regain the lead as the match progresses, and those home teams that don't regain the lead are likely to be weak favourites.

Also note the standard deviation in each quarter was greatest when there is a high level of uncertainty (win% $\approx$ 50%), which is indicative of the home team being the *HFB* or *HUA*. Interestingly, home advantage is greatest (+5.14) in the final quarter when the home team is the HFB. Indeed, this advantage (+5.14) by the *HFB* in the final quarter is obtained defensively limiting the away side to +19.80 points compared to +22.24 points (Table 9.2) whilst maintaining a similar offensive output. This provides objective agreement with Marcelino et al. (2009) that home teams should manage risk in the latter stages of the match. To test the significance of pre-game and in-game characteristics of the home team at the end of the previous quarter, a two way factorial analysis of variance (ANOVA) was conducted. Table 9.5 shows the results.

Quarter	Source	Partial SS	Df	MS	F
1	Model	481	3	161	0.57
	AHEAD	8	1	8	0.03
	FAVOURITE	79	1	79	0.28
	AHEAD*FAVOURITE	320	1	320	1.14
	RESIDUAL	276005	985	280	
2	Model	1174	3	391	1.39
	AHEAD	439	1	440	1.56
	FAVOURITE	997	1	997	3.54*
	AHEAD*FAVOURITE	170	1	170	0.6
	RESIDUAL	277642	985	282	
3	Model	1040	3	346	1.23
	AHEAD	4	1	4	0.01
	FAVOURITE	343	1	343	1.22
	AHEAD*FAVOURITE	895	1	895	3.18*
	RESIDUAL	278251	985	281	

*significant at the .10 level

Table 9.5: Analysis of variance summary: Mean difference between the home and away team score ($\bar{\Delta}$) in the 2nd, 3rd and 4th quarters as a function pre-game and in-game characteristics of the home team at the end of the previous quarter, 2000 to 2009

Firstly, the results provide some evidence ($p < 0.10$) that *HU* in the third quarter, receive a greater advantage than *HF*. Secondly, in the final quarter there is some evidence ($p < 0.10$) that when there is a high level of uncertainty (i.e. *HFB* and *HUA*) home teams receive a greater advantage. This provides objective disagreement with previous research that suggest home teams “choke” when they are under a high level of pressure such as sports championships (Baumeister and Steinhilber, 1984; Wright et al., 1995).

Chapter 10

In-Play Predictions

In this chapter, a generalised Logistic Model (GLM) is used to model outcomes of AFL matches in real-time. To begin, Section 10.1 provides a brief introduction on the challenges of real-time predictions in sport with a specific focus towards AFL. Section 10.2 discusses the data utilised in this Chapter. Section 10.3 details a Brownian Motion Model (BMM) of which comparisons are made throughout against the GLM. Section 10.4 illustrates how slight changes in each parameter of the GLM skew the overall distribution, and the optimisation process of the GLM which is a function of team quality and score difference for each quarter is also discussed. Section 10.5 evaluates the results of the GLM against the BMM based on various measures of performance including betting simulations. Material from this chapter has been published in Ryall and Bedford (2010b).

10.1 Introduction

Sports commentators in game sports constantly talk about the likelihood of either team winning at any point in time rarely with any empirical evidence to support their suggestions. Comments such as “Boston Celtics rarely loses the match if they are leading at three quarter time” are common. It has been shown by Cooper et al. (1992) that either team leading after three quarters of the game in Basketball (NBA), Football (NFL) and Hockey (NHL) won approximately 90% of the time. This is of course without making any adjustments for quality of the two competing teams. However, AFL is known for its high level of uncertainty during the match with the team leading at three quarter time winning approximately 85% of matches. The colloquial saying “the match is not over until the final siren is blown” has never been more appropriate. It is this uncertainty that draws spectators to matches and entices academics to try and explain it.

In predicting the outcome of AFL matches it has been shown that both home advantage and the quality of the two competing teams play an important role (Stefani and Clarke, 1992). Furthermore, Bailey and Clarke (2004) showed that by constructing a model for AFL prediction at a player based level improved the forecasting capabilities. For example, the number of player changes for each team can vary considerably from week to week due to injuries or suspensions, this in turn can have a significant impact on the likelihood of a team winning depending on the importance of the players in question. There are also a plethora of other factors which are likely to influence match outcomes which are yet to be investigated, including the importance of the match. For example, in round 22 (final round) season 2010 Hawthorn (7th) hosted Collingwood (1st) at the MCG. Regardless of the outcome of the match Collingwood were going to finish top of the ladder because they were one and a half wins ahead of the 2nd place Geelong. Therefore, Collingwood had no incentive to win and incidently lost the match by three points.

In predicting the outcome of sporting events during the game, careful consideration must be given to the relative importance of pre-game factors as the match progresses. For example, it is reasonable to assume that team quality is of more importance earlier in the match when the result is still unknown. However, you could argue the opposite, that any difference in team quality is critical late in the match provided the score difference is marginal. Furthermore, in-game factors must also be incorporated into the model and their relative importance must be weighted as the match progresses. For example, a lead of $+x$ points is more valuable as the match progresses since the opposition has less time to regain the lead. Most research in real-time match prediction to date incorporates an adjustment for team/player quality, score difference and the proportion of the match which has been completed. For example, Stern (1994) and Glasson (2006) used a Brownian motion Model (BMM) for modelling high scoring sports using time elapsed, a pre-game point estimate and score difference. Similarly, Klaassen and Magnus (2003) developed *TENNISPROB* a computer algorithm which instantaneously calculates the in-game probability of either player winning based on the current score in the match (game score, set score and match score) and the probability of player A or player B winning a point on service.

It is important to note additional in-game factors which might influence the outcome of an AFL match during the game. For example, injuries can also occur during the match, and although they can be interchanged, the calibre of the player replaced could be substantially different. Furthermore, it is not uncommon for several injuries to occur in a given match which limits the number rotations a team can make. For example, in round 20 season 2009 Essendon (8th) hosted St Kilda (1st) at Docklands Stadium. Although Essendon led by 29 points at three quarter time they had lost three players through injury by this stage. This resulted in a several tired Essendon players, as most of them had to play without rest for the remainder of the match. Incidently, Essendon went on to win by a meagre two points. Schembri and Bedford (2010) calculated the impact of injuries during an AFL match based on scoring patterns, interchange rotations, and the likelihood of winning the match.

Also, due to the discrete nature of scoring in AFL, a five point deficit can be restored by kicking a single goal (worth six points), therefore the magnitude of a small lead (< 6 points) is virtually redundant at the death of the match. Incidently, when a team is ahead by six points in the final quarter and they score a behind this is commonly referred to as a “handy point”, since the opposition now need several scoring attempts to draw level or regain the lead. Furthermore, when the lead is small ($<$ 6 points) it is important to note which team is in possession of the ball. For example, in round 16 season 2008 Richmond hosted Essendon at the MCG. With a couple of minutes remaining in the match, Richmond led by seven points after Essendon kicked a behind. Richmond then controversially rushed numerous behinds whenever they were under pressure in order to maintain possession of the ball and prevent Essendon from scoring. Incidently, Richmond went on to win by four points but they received much scrutiny for there tactics in the media. Clarke and Norman (1998) identified when to rush a behind in AFL using a dynamic programming approach. Their preliminary results suggest that it is often to a team's advantage to concede a point through a rushed behind either to avert the possibility of an imminent goal or to increase the likelihood of scoring a goal themselves.

In Chapter 9, it was shown that home advantage may depend on the in course dynamics of the match. Home advantage for *priori* home teams was found to be greatest in the latter stages of the match when there is a high level of uncertainty. Therefore, there is an argument that assigning a constant home advantage prior to the start of the match is inappropriate for real-time match prediction.

Additionally, although previous research on real-time predictions incorporate score difference (Stern, 1994; Glasson, 2006), the types of score that comprise the score difference (i.e. goals and behinds in AFL) is yet to be investigated. As a result two teams could be level on the scoreboard but one team could have considerably more (or less) scoring opportunities. For example, in round 22 season 2010 Geelong (2nd) hosted West Coast (16th) at Kardinia Park. At quarter time West Coast (2 goals, 2 behinds) 14 points led Geelong (1 goal, 7 behinds) 13 points, although Geelong had considerably more scoring opportunities they were behind. However, in the subsequent quarters Geelong kicked truly, eventually winning by 44 points. On the contrary, in the 2008 Grand Final between Hawthorn and Geelong, Geelong (6 goals, 12 behinds) had considerably more scoring opportunities in the first half than Hawthorn (8 goals, 3 behinds), however they ended up losing by 26 points.

Therefore, more scoring opportunities with a poor conversion rate could be either a sign of things to come (i.e. more scoring opportunities) or simply indicate wasted opportunities.

Many supporters also believe in the idea of team momentum during a match. However, to the author's knowledge this is yet to be empirically tested. The primary problem in quantifying momentum during a game is that it is confounded with team quality. For example, if a team kicks a succession of goals then that team is more often than not going to be of greater quality than the opposition.

With so many factors to incorporate for real-time match prediction it is important to identify what data are available during the game and the ease with which to access this information. For example, the transaction data supplied by Prowess Sports is only available post-match. Furthermore, there is no readily available database which contains information on injuries during the game, and even if there were, it becomes extremely subjective to measure the quality of the player(s) lost. Similarly, in order to determine the relative importance of momentum, a model must incorporate all the score changes as they occur, not just the current score difference. It is also important to determine whether probability forecasts are required throughout the entire match or at specific intervals (i.e. quarter time breaks). Therefore, after due consideration, it was decided to develop a model with minimal inputs that focussed on the interaction, if any, between team quality and score difference as the match progresses. This path was taken primarily because previous research assumed the effect of team quality was independent of score difference during the match (Stern, 1994; Glasson, 2006).

10.2 Data

This chapters analysis is based on AFL seasons 2000 to 2009. AFL data was gathered from ProEdge, a statistical package developed by ProWess Sports. Data consisted of year, round, quarter, (nominal) home team, away team and home team margin. A pre-game point estimate (or *LINE*) was calculated for each match for seasons 2000 to 2009 using the ratings model developed in Chapter 5. For example, in round 9 season 2010 Essendon (home) played Richmond (away) the *LINE* was +28 points in favour of Essendon. That is, prior to the start of the match Essendon are expected to win by 28 points.

10.3 Brownian Motion

Stern (1994) applied a Brownian Motion Model (BMM) to forecast the outcome of basketball (NBA) and Baseball (MLB) matches in real time. The model incorporates time remaining, home advantage and score difference yielding a probability forecast. He stated that out of the major American sports, basketball is best suited to the BMM due to the almost continuous nature of the game and score. Glasson (2006) builds on this by replacing home advantage with the bookmakers line to forecast AFL matches in real time. Since the bookmakers line should incorporate home advantage *and* team quality this should be more representative of who is going to win prior to the start of the match, which in turn should increase the forecasting capabilities of the BMM during the game.

In this section the BMM defined in Glasson (2006) is replicated by replacing the bookmakers lines with the pre-game point estimates developed in Chapter 5. Throughout the remainder of this chapter comparisons are made between the forecasting capabilities of the GLM and BMM.

Firstly, time elapsed during a match needs to be transformed to the unit interval $t \in [0, 1]$, where t describes the proportion of the match completed. Let $X(t)$ represent the lead, l, by the home team relative to the away team at time t. If $X(t) > 0$ this indicates an $+l$ point lead to the home team at time t. Similarly, if $X(t) < 0$ this indicates an $|-l|$ point lead to the away team at time t and if $X(t) = 0$ scores are level at time t. Assuming that $X(t)$ can be modelled as a Brownian motion process with drift μ and variance σ^2, per unit in time, where μ denotes the difference in quality inclusive of home advantage between the two teams prior to the start of the match in terms of points. If $\mu > 0$ this indicates a $+\mu$ point advantage to the home team, similarly if $\mu < 0$ this indicates a $|-\mu|$ point advantage to the away team, and if $\mu = 0$ there is no distinct difference between the two teams. Under the BMM, $X(t)$ can be described as

$$X(t) \sim N\left(\mu t, \sigma^2\right) \tag{10.1}$$

Therefore, prior to the start of the match, the probability that the home team wins given difference in quality inclusive of home advantage μ, and variance σ^2 is

$$P\left(X(1) > 0\right) = \Phi\left(\frac{\mu}{\sigma}\right) \tag{10.2}$$

Now once the match is underway the probability that the home team wins at time t, given they have an l point advantage (or deficit) can be estimated by the BMM:

$$\begin{aligned} P_{\mu,\sigma}(l,t) &= Pr\left(X(1) > 0 | X(t) = l\right) \\ &= Pr\left(X(1) - X(t) > -l\right) \\ &= \Phi\left(\frac{l + (1-t)\mu}{\sqrt{(1-t)}\sigma}\right) \end{aligned} \tag{10.3}$$

For simplicity, the probability of a draw has been ignored but can be incorporated using the continuity correction (Stern, 1994, p. 1129) which is given by:

$$P_{\mu,\sigma}(l,t) = 0.5\Phi\left(\frac{l - 0.5 + (1-t)\mu}{\sqrt{(1-t)}\sigma}\right) + 0.5\Phi\left(\frac{l + 0.5 + (1-t)\mu}{\sqrt{(1-t)}\sigma}\right) \tag{10.4}$$

The BMM can now be implemented to forecast AFL matches in real time subject to difference in quality inclusive of home advantage (μ), the variance (σ^2), time elapsed ($t \in [0,1]$) and score difference (l). As previously stated μ is replaced by the pre-game point estimates developed in Chapter 5. Glasson (2006) used the bookmakers lines μ and odds $Pr(X(1) > 0)$ to estimate σ by rearranging (10.3) yielding $\hat{\sigma} = 38$, he notes that varying σ by a few points either way has little influence on the probability forecasts. Applying the same principles to the Elo ratings $\hat{\sigma} = 40$ provides an adequate fit between the pre-game point estimates and probability forecasts. Since this chapter focuses on real-time probability forecasts at each of the quarter time breaks, time elapsed is replaced by $t = 0.25$ (quarter time), $t = 0.50$ (half time) and $t = 0.75$ (three quarter time). Finally, the score difference l at each of the quarter time breaks was gathered from ProEdge a statistical package developed by ProWess Sports (www.prowess.com.au).

10.4 Generalised Logistic Model

Akin to regression analysis, curve fitting is the procedure of fitting a probability distribution which gives the best fit to a series of data points. Typical probability distributions used in curve fitting include Beta, Exponential, Gamma, generalised Logistic, Gompertz, Linear, Lognormal and Weibull. Kuper and Sterken (2006) applied the inverted S-shaped Gompertz function to model the development of world records in running. Due to the asymptotic behaviour of the Gompertz function, implied limits of world records could be deduced.

There are several prerequisites that the probability distribution must satisfy for real-time match prediction in AFL. First and foremost, the probability distribution must have a lower asymptote of zero and an upper asymptote of one in order to satisfy basic probability theory. Furthermore, as the score difference approaches $-\infty$ the probability of winning should approach zero, similarly as the score difference approaches $+\infty$ the probability of winning should approach one. Also, for teams of relatively equal ability, the point of inflection should occur when the score difference equals zero. Therefore, the four-parameter generalised Logistic function seemed suitable which is given by

$$Pr_{i,SD}(t) = \frac{1}{[1 + Q_i e^{-B_i(SD-M_i)}]^{1/v_i}} \tag{10.5}$$

where $Pr_{i,SD}(t)$ denotes the probability of team t winning at quarter i, for given score difference SD and B_i, M_i, Q_i and ν_i are unknown parameters for each quarter i.

The four parameters of the GLM (B, M, Q and ν) skew the overall distribution in different ways: B controls the rate of growth, M shifts the time of maximum growth, Q depends on the value $Pr_{i,0}(t)$ and ν affects which asymptote maximum growth occurs. Figure 10.1 illustrates the effect each of the parameters (excluding Q for reasons defined later) has on $Pr_{i,SD}(\text{home})$ keeping all the other parameters constant.

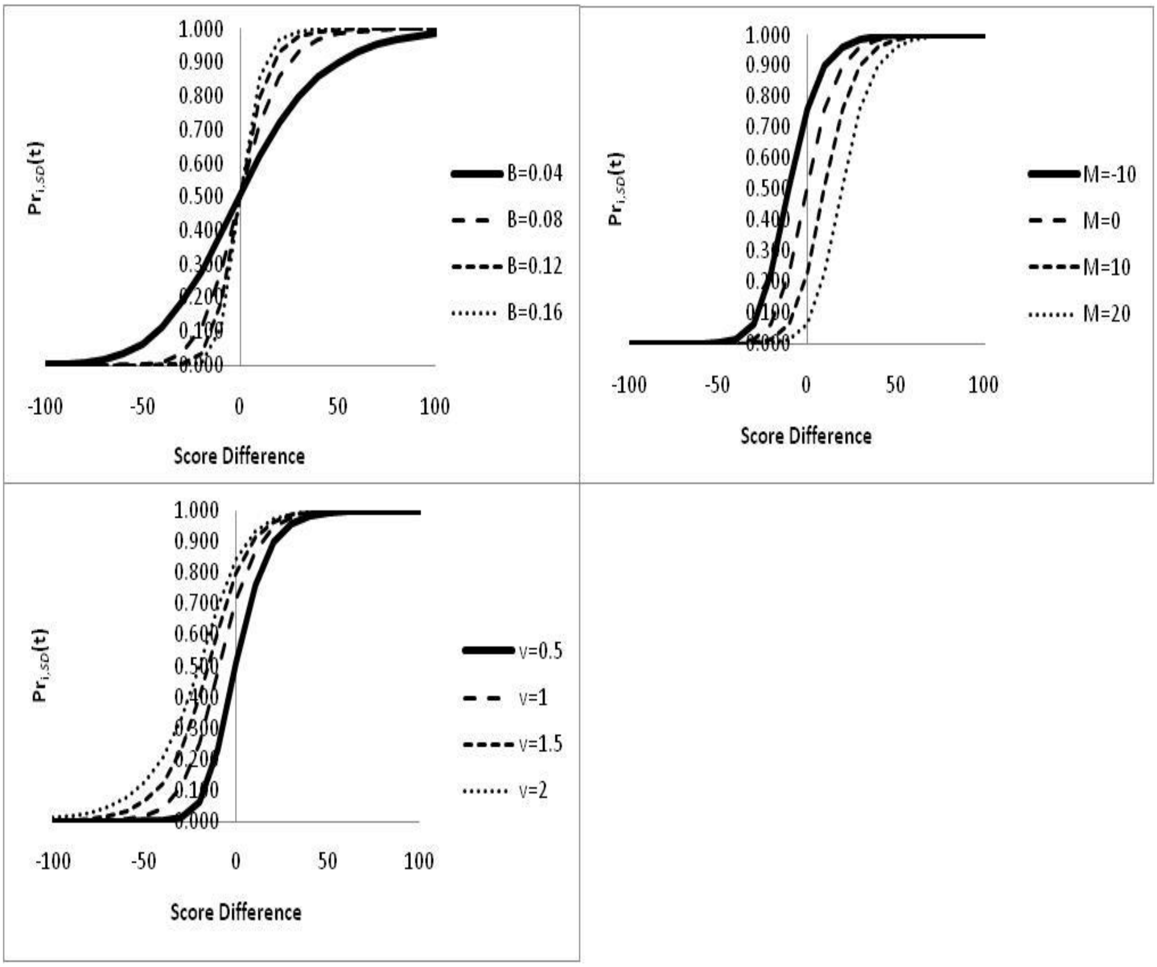


Figure 10.1: The generalised Logistic Function for varying parameter values keeping other parameters constant

The four parameters B_i, M_i, Q_i and ν_i of the GLM need to be optimised for each quarter i. However, the model given in (10.5) does not allow for any difference in team ability. Therefore, after due consideration, each of these parameters was replaced by a simple linear equation which was a function of difference in team quality. It is important to note that a linear model was selected purely for simplicity. The contribution of team quality towards the probability of winning for each quarter i is now clearly *dependent* on the score difference and vice versa. However, is this assumed interaction between team quality and score difference during the match sufficiently justified? To gauge the effect score difference has on team quality during the game, the error term defined in (10.6) is calculated at the end of each quarter depending on whether the pre-game favourite was ahead (or behind) on the scoreboard at the end of the previous quarter. Table 10.1 shows the results.

$$\epsilon_t = X(t) - X(t-1) - \frac{\mu}{4} \approx 0 \tag{10.6}$$

now let

$$\epsilon_n = \begin{cases} \epsilon_{t+1}, & \text{if quarter=1} \\ \epsilon_{t+1} + \epsilon_t, & \text{if quarter=2} \\ \epsilon_{t+1} + \epsilon_t + \epsilon_{t-1}, & \text{if quarter=3} \end{cases} \tag{10.7}$$

Quarter	Score Difference	$\mu/4$	$X(t)$	$X(t+1) - X(t)$	ϵ_{t+1}	ϵ_n
1	Ahead	5.94	15.69	5.04	-0.90	-0.90
1	Behind	4.75	-12.15	3.80	-0.95	-0.95
2	Ahead	6.12	23.49	6.23	0.11	-0.79
2	Behind	4.29	-17.08	2.42	-1.87	-2.82
3	Ahead	6.24	31.88	6.81	0.57	-0.22
3	Behind	3.97	-21.55	3.39	-0.57	-3.39

Table 10.1: Quarter by quarter observed minus expected results, 2000 to 2009

An example might help to clarify how to interpret the results. At the end of the 2nd quarter, if the pre-game favourite was ahead ($\mu/4 = +6.12$), then on average they outscored opponents by +6.23 points in the following quarter, slightly exceeding expectations by +0.11 points. However, if the pre-game favourite was behind at the end of the 2nd quarter ($\mu/4 = +4.29$) then on average they outscored opponents by +2.42 points in the following quarter, falling well short of expectations by -1.87 points. It is clearly evident that in-game scoring expectations of the pre-game favourite become more dependent on score difference as the match progresses. Therefore, the error term defined in (10.6) is clearly biased as it measures the average error between pre-game favourites that are ahead and behind on the scoreboard.

Since the parameter Q depends solely on the value $Pr_{i,0}(t)$, and $Pr_{i,0}(t) = 0.5$ when $LINE = 0$ (i.e. probability of winning equals 0.5 when scores are level and quality of both teams is the same), M must equal zero when this occurs. Therefore, Q becomes a function of ν given by:

$$Q_i = \sqrt[1/v_i]{2} - 1 \tag{10.8}$$

Therefore, there are now five variables to be optimised for each quarter i which are given by

$$\begin{aligned} B_i &= B1_i + B2_i|LINE| \\ M_i &= M2_i|LINE| \\ \nu_i &= \nu1_i + \nu2_i|LINE| \end{aligned}$$

Since every match has a nominated home team and a nominated away team, the sum of these two probabilities must equal one for quarter i and given score difference SD. That is, for every match

$$Pr_{i,SD}(home) + Pr_{i,SD}(away) = 1 \tag{10.9}$$

Therefore,

$$Pr_{i,SD} = \begin{cases} \frac{1}{\left[1+Q_i e^{-B_i(SD-M_i)}\right]^{1/v_i}}, & \text{if } t = \text{home} \\ 1 - \frac{1}{\left[1+Q_i e^{-B_i(-SD-M_i)}\right]^{1/v_i}}, & \text{if } t = \text{away} \end{cases} \tag{10.10}$$

Similarly to Chapter 5, the Brier Score is used as the objective function which is to be minimised. AFL Seasons 2000 to 2004 were used as a training set in the forward prediction of AFL seasons 2005 to 2009. Simulations were carried out utilising the Monte Carlo algorithm using Riskoptimiser, an add-in for Excel.

10.5 Results

Figure 10.2 displays the empirical probability of winning as a function of score difference (SD) at each of the quarter time breaks for varying levels of difference in team quality. Akin to Stern (1994), as the match progresses score difference (SD) has more influence while difference in team quality has less.

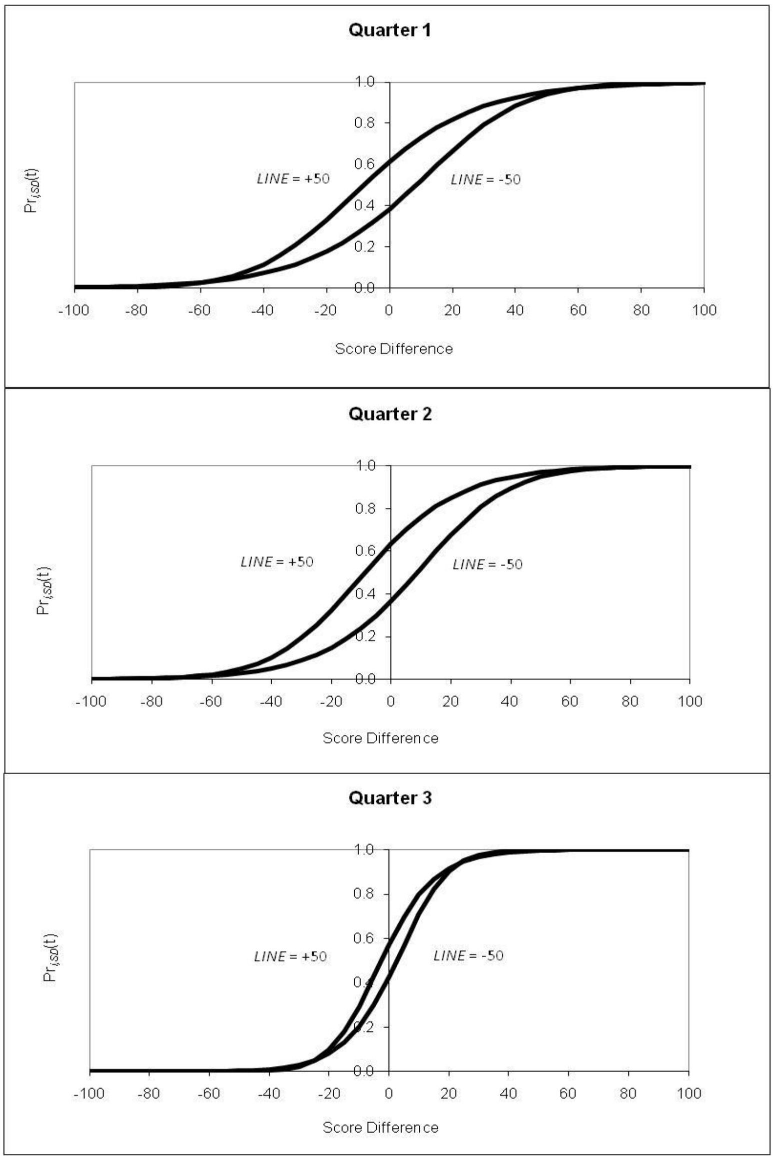


Figure 10.2: Smooth curves showing the probability of winning an AFL match at quarter time, half time and three quarter time for given score difference (SD)

Various measures can be used to evaluate the performance of prediction models in game sports. Some commonly used measures in the literature include Average Absolute margin of Error (AAE), number of predicted winners and Return on Investment (Bailey and Clarke, 2004). Since the number of predicted winners will tend towards one as the match progresses, an alternative measure is needed to evaluate the performance of the GLM. Akin to Stefani and Clarke (1992), the reliability of the probability forecasts are investigated by comparing the predicted and actual probabilities of winning. Firstly, the predicted probability of the in-game favourite winning is banded into five subgroups. The number of games and the actual probability of winning for each subgroup of predicted probabilities are shown in Table 10.2.

	Quarter 1		Quarter 2		Quarter 3	
Predicted Probability	BMM	GLM	BMM	GLM	BMM	GLM
0.50 0.59	0.561(0.256)	0.468 (0.213)	0.602 (0.192)	0.574 (0.153)	0.500 (0.117)	0.588 (0.076)
0.60 0.69	0.615 (0.225)	0.634 (0.206)	0.697 (0.171)	0.699 (0.127)	0.684 (0.107)	0.603 (0.071)
0.70 0.79	0.756 (0.230)	0.713 (0.235)	0.815 (0.170)	0.740 (0.169)	0.802 (0.142)	0.720 (0.092)
0.80 0.89	0.859 (0.167)	0.823 (0.228)	0.889 (0.212)	0.843 (0.222)	0.945 (0.163)	0.822 (0.152)
0.90 1.00	0.959 (0.110)	0.958 (0.107)	0.968 (0.243)	0.951 (0.318)	0.980 (0.461)	0.970 (0.598)
% correct	71.36	69.77	80.62	80.00	86.02	86.82

Table 10.2: Predicted and actual probabilities of winning and corresponding proportions (in parenthesis) of the Brownian Motion Model (BMM) and the generalised Logistic Model (GLM) at quarter time, half time and three quarter time, 2005 to 2009

An example might help to clarify how to interpret the results. In the first quarter, the BMM had 25.6% of all matches as a 50-59% favourite, teams that fell in this category won on average 56.1% of the time. Although the number of winners predicted by the two different models is approximately equal, the distribution of predicted probabilities for the GLM is heavily skewed towards one (win). In the 1st quarter the BMM clearly outperforms the GLM in terms of the total number of predicted winners (+1.59%), and the predicted probabilities also provide a more reliable indication of the chance of victory. This can be attributed to the BMM incorporating the contribution of team quality independently of score difference, whereas the GLM (incorrectly) assumes team quality is dependent on score difference at the end of the 1st quarter. This independence was verified in Table 10.1 which showed there was no significant difference between scoring behaviour of the pre-game favourite in the 2nd quarter based on whether they were ahead or behind at the end of the 1st quarter [$\epsilon_n(ahead) = -0.90, \epsilon_n(behind) = -0.95$]. However, in the 2nd quarter, although the BMM outperforms the GLM in terms of total number of predicted winners (+0.62%), the reliability of the predicted probabilities of the BMM should be questioned, since the actual probability of winning consistently falls outside the range of predicted probabilities for each subgroup. Conversely, the predicted probabilities of the GLM in the 2nd quarter are reliable since the actual probability of winning is approximately the midpoint of each subgroup of predicted probability ranges. Although Table 10.1 verifies the dependence of team quality and score difference at the end of the 2nd quarter [$\epsilon_n(ahead) = -0.79, \epsilon_n(behind) = -2.82$] for some reason this does not increase the number of predicted winners of the GLM. Finally, in the 3rd quarter, the GLM outperforms the BMM in terms of total number of predicted winners (+0.80%), and the predicted probabilities also provide a more reliable indication of the chance of victory. Table 10.1 verifies the dependence of team quality and score difference at the end of the 3rd quarter [$\epsilon_n(ahead) = -0.22, \epsilon_n(behind) = -3.39$].

10.6 Applications to Betting Markets

Another way to compare the performance of the two models is to investigate their respective return on investment using standard wagering strategies. In Chapter 6, in-play betting data was collected for 118 matches during the 2009 AFL season, and in Chapter 8 a program was written to extract the approximate in-play odds during each of the quarter time breaks. The advantage (or disadvantage) a punter has over a bookmaker is derived by comparing the probability of winning against the bookmaker odds which is given in (5.9) of Chapter 5. Akin to Section 5.5 of Chapter 5, a constant Kelly system is implemented using a constant pool of $1000. Table 10.3 and 10.4 displays the betting results of the BMM and GLM respectively. The results include the total number of bets, total bets won, percentage of bets won, total bet, profit/loss and the return on investment (ROI).

Quarter	# bets	# won	% Won	Total bet	Profit/Loss	ROI
1	37	15	40.5%	$12,477	$1408	11.3%
2	37	11	29.7%	$10,070	-$1348	-13.4%
3	58	9	15.5%	$12,846	-$4957	-38.6%
All Bets	132	35	26.5%	$35,394	-$3,256	-13.8%

Table 10.3: In-play head to head betting using Brownian Motion Model (BMM), 2009*

Quarter	# bets	# won	% Won	Total bet	Profit/Loss	ROI
1	66	14	21.2%	$19,678	-$5064	-25.7%
2	50	10	20.0%	$11,549	-$3516	-30.4%
3	30	18	60.0%	$14,122	$2356	16.7%
All Bets	146	42	28.8%	$45,350	-$6,224	-13.7%

Table 10.4: In-play head to head betting using generalised Logistic Model (GLM), 2009*

It is immediately evident that the ROI of the BMM *decreases* as the match progresses whereas the performance of the GLM *increases* as the match progresses. These results are consistent with the performance of the GLM and BMM in terms of total predicted winners and the reliability of the probability forecasts given in Table 10.2. However, it is important to note that both models show negative returns across all quarters. This can be attributed to AFL season 2009 being an aberration in terms of the likelihood of teams winning when they are ahead on the scoreboard. Section 8.5 of Chapter 8 showed that teams that were ahead on the scoreboard in season 2009 (across all quarters) won considerably more games than the long term average for seasons 2000 to 2008. Therefore, it is reasonable to assume that the ROI of the GLM (and BMM) would increase substantially in subsequent seasons.

Chapter 11

Phases of Play

Presenting statistical predictions that are simultaneously representative of a team's likelihood of winning, and graphically simple enough to be widely interpretable, remains a constant challenge for the sport statistician. This chapter focuses on the process involved in transforming a mass of performance variables from "live-streaming" data into a single web-based phases of play plot. Section 11.1 provides a brief introduction on phases of play in sport. Section 11.2 details the data used throughout this Chapter. Section 11.3 explains how the real-time performance data is transformed into a single probability assessment using logistic regression. Section 11.4 discusses how the phases of play plot is generated automatically post-match using macros in Excel. Graphically the plot is enhanced by adding images of a player's guernsey when a goal is scored. Additionally, with some minor modifications the plot becomes interactive such that the match can be "played out" in pseudo real-time. By integrating interchange data, team performance can be deduced relative to a players Time on Ground (TOG), this provides a novel evaluation of individual player performance. Section 11.5 evaluates the performance of the model by investigating the predictive power against score difference at each of the quarter time breaks. Furthermore, this section also examined the residuals to see if there are any specific biases in the model which are accounted for.

Material from this chapter has been published in Ryall and Bedford (2008).

11.1 Introduction

Phases of play posits that two teams or players interact in a dynamic system, that is, an active-reactive nature (McGarry et al., 2002). This concept can refer to the advantage (or disadvantage) a player has in a single point in squash in terms of their physical displacement (McGarry et al., 2002), the collective actions which lead to a goal in soccer (Grehaigne et al., 1997) and a measure to describe the performance of teams in NHL during the match (Bedford and Baglin, 2009). Borrie et al. (2002) suggest that simple frequency data can't capture the complex series of interrelationships between a wide variety of performance variables. Bedford and Baglin (2009) noted this and proposed that the sum of all teams adaptive winning behaviours along with their maladaptive losing behaviours could explain outcomes in NHL during the game. In their example, phases of play posits that teams fluctuate between periods of "high (in) phase" and "low (out of) phase", where high phase is a characteristic of winning teams and low phase is a characteristic of losing teams, with both teams being able to be in either state at any point in time. However, the authors noted that teams were typically "anti-phase stable", that is, if one team was in high phase the other team would be in low phase and vice versa. Here "relative phase" describes the difference between the team phases.

Franks and Miller (1986) found that coaches have the same level of difficulty in remembering critical events as eyewitnesses have in recalling criminal events. Furthermore, Franks and Miller (1991) showed that coaches can't accurately recall pertinent sequential information prior to a critical event occurring. This led them to develop a new method to train coaches to observe and remember. They proposed the idea to train the observational skills of coaches using a video training method. The results suggested that although coaches were incapable of remembering more than 40% of pertinent sequential information, coaches could be trained to observe and remember sequential information prior to a critical event occurring. This finding suggests that a simple reflective measure is needed to assist coaches in event recall. Therefore, the purpose of this Chapter is to provide a visual representation of team performance which is easy to interpret and emphasizes critical points during the match.

11.2 Data

This chapter's analysis is based on the 2007 AFL season. Real-time performance data was gathered from ProEdge, a statistical package developed by ProWess Sports. The data, herein referred to as transaction data, provides a list of comprehensive event details and the time at which the event occurred for a single match. Each match consists of approximately 2,500 unique transactions, with each transaction consisting of up to three actions, or unique statistics, (e.g. kick long; kicking to a contest; inside 50) attributed to one of the 44 players contesting a game. It is important to note that this transaction data was collected post-match. Therefore, in order to implement the phases live, it was important to only extract variables which were also generated in real-time.

Throughout season 2007, ProWess Sports updated real-time performance data on the Real Footy web site (www.realfooty.com.au/livestats) which unfortunately is no longer in existence. Nonetheless, alternative web sites such as AFL match day (http://xml.afl.com.au/sw f/live_stats.htm) showcase similar data in real-time. The operational Real Footy web site refreshed 20 live statistics approximately every 30 seconds which included kicks (*KCK*), handballs (*HBL*), marks (*MRK*), inside 50's (*I50*), tackles (*TKL*), spoils (*SPL*), hitouts (*HIT*), 1st possession from an umpire control situation (*1ST*), clearances (*CLE*), goals (*GLS*), behinds (*BHS*), rushed behinds (*RUS*), frees for (*FF*), marks inside 50 (*MI50*), turnovers (*TNS*), goals from general play (*GFG*), goals from free kicks (*GFF*), goals from marks (*GFM*), goals from kick ins (*GFK*) and goals from stoppages (*GFS*). Therefore, variables extracted from the ProEdge database were restricted to these 20 performance variables so the phases could be theoretically run live. In ProEdge, *TNS* is broken down by kicks to opposition (*KOP*), ineffective handballs (*IHBL*), kicks to contest (*KTC*) and kicks to space (*KTS*). Since the proportion of *KTC* and *KTS* that result in the opposition having the next possession is unknown, *TNS* was removed from the analysis. Additionally, ProEdge had two different definitions of a tackle, *TKL* which is defined as "a reasonable attempt by the player to tackle the opposition" and *TKE* defined as "a tackle that effectively disrupts or changes the way the opposition player disposes the ball". However, the definition of a tackle on the Real Footy web site is somewhere in between *TKL* and *TKE.* Since tackling is widely assumed as an integral part of winning an AFL game, removing it from the statistical analysis was not a feasible option. Therefore, *TKL* was included in the model since it was clearly more representative of the tackle variable from the Real Footy web site. Note that the discrepancy between these different definitions of what constitutes a tackle is only cause for concern if the model is run live using the cumulative statistics from the Real Footy web site.

11.3 Methods

Firstly, the contribution of each performance variable to a team winning a game needs to be considered. Stewart et al. (2007) set out to find which individual performance variables in AFL were important, and how much each variable contributed to a team winning a match. The objective was to identify inefficiencies in the market for recruiting professional AFL players. This was completed by regressing 51 "primary variables" to a single variable margin of victory using Ordinary Least Squares regression (OLS). Since margin of victory was used as the dependent variable, goals, behinds and rushed behinds had to be excluded from the model, as their inclusion was an exact predictor of margin. This meant that the final model would be biased against forwards, in particular full forwards.

Bedford and Baglin (2009) applied logistic regression to NHL summary game data for the season 2005-2006 for use in the forward prediction of season 2006-2007 based on 19 performance variables. Win/loss was used as the dependent variable since the research focused on what contributes to a win (or loss) rather than to scoring a goal (or not). Furthermore, score could not be ignored as an independent variable as it in itself is an outcome of a perturbation in the phases of play. Therefore, logistic regression was applied to the previously mentioned 19 performance variables (excluding turnovers) for the 2007 AFL season and retrospectively fitted. Separate logistic regression models were applied to nominated home/away teams due to overwhelming evidence of home advantage (Clarke, 2005). By including cumulative win percentage into the model, the model initializes prior to the start of the match to account for the difference in team quality. Table 11.1 shows the logistic regression equation results for teams based on home and away games. The equation takes the following form:

$$y_{j,t} = logit(p_{j,t}) = \beta_{0,j} + \beta_{1,j}x_{1,j,t} + \cdots + \beta_{19,j}x_{19,j,t} \tag{11.1}$$

where j=home/away, β_i=logistic coefficient, x_i=variable and t=time.

Now the probability of team j winning at time t regardless of opposition is given by:

$$P_{j,t} = \frac{e^{y_{j,t}}}{1 + e^{y_{j,t}}} \tag{11.2}$$

It is important that the probability of the nominated home team and away team equals one throughout the entire match. Therefore, a relative probability can be deduced by normalizing the probabilities which is given by:

$$\text{relative}_{home,t} = \frac{P_{home,t}}{P_{home,t} + P_{away,t}} \tag{11.3}$$

	Coefficient		P-Value		Odds Ratio	
Variable	Home	Away	Home	Away	Home	Away
CUM	1.26	1.45	0.13	0.08	3.51 (0.70, 17.74)	4.27 (0.83, 22.08)
KCK	0.03	0.00	0.14	0.84	1.03 (0.99, 1.07)	1.00 (0.96, 1.05)
HBL	-0.01	-0.01	0.16	0.44	0.99 (0.98, 1.00)	0.99 (0.98, 1.01)
MRK	-0.04	-0.01	0.07	0.78	0.96 (0.92, 1.00)	0.99 (0.95, 1.04)
$I50$	-0.11	-0.08	0.02*	0.06	0.89 (0.81, 0.98)	0.92 (0.85, 1.00)
TKL	-0.02	0.01	0.37	0.37	0.98 (0.95, 1.02)	1.01 (0.98, 1.04)
SPL	0.01	-0.01	0.71	0.75	1.01 (0.94, 1.10)	0.99 (0.92, 1.06)
HIT	0.05	0.01	0.07	0.78	1.05 (0.99, 1.12)	1.01 (0.96, 1.06)
$1ST$	-0.06	-0.07	0.28	0.29	0.94 (0.83, 1.05)	0.93 (0.81, 1.06)
CLE	-0.03	-0.07	0.70	0.36	0.97 (0.85, 1.12)	0.93 (0.81, 1.08)
GLS	0.30	0.48	0.09	0.01*	1.36 (0.96, 1.92)	1.62 (1.12, 2.33)
BHS	0.15	0.08	0.10	0.32	1.16 (0.97, 1.39)	1.09 (0.92, 1.28)
RUS	0.09	0.12	0.49	0.32	1.09 (0.85, 1.40)	1.13 (0.89, 1.43)
FF	-0.06	-0.02	0.19	0.56	0.94 (0.85, 1.03)	0.98 (0.90, 1.06)
$MI50$	0.09	0.01	0.28	0.90	1.09 (0.93, 1.28)	1.01 (0.88, 1.16)
GFG	0.19	0.14	0.27	0.40	1.20 (0.87, 1.67)	1.15 (0.83, 1.58)
GFF	0.14	0.04	0.59	0.86	1.15 (0.69, 1.93)	1.04 (0.64, 1.72)
GFM	0.09	0.14	0.67	0.46	1.09 (0.74, 1.61)	1.15 (0.80, 1.65)
GFK	-0.41	-0.36	0.06	0.09	0.66 (0.43, 1.02)	0.70 (0.46, 1.06)
GFS	0.15	-0.18	0.24	0.21	1.16 (0.91, 1.48)	0.84 (0.63, 1.11)

*significant at the .05 level

Table 11.1: Logistic regression results: Real-time AFL performance data, 2007

The sole performance variable that has a significant negative influence on the nominated home teams phase is $I50$. Although there are no performance variables that has a significant negative influence on the nominated away team's phase, $I50$ is bordering on significance (p=0.06). These results also dispel preconceived notions from so called media "experts" about what are the most important statistics. For example, there are many media articles which show a strong correlation between cumulative $I50$ and $I50$ differentials and on field success (afl, 2009b). However, once accounting for all other variables $I50$, which were statistically significant for home teams are actually *negatively* correlated with winning. This suggests it is not the quantity of the $I50$ but rather the quality that are correlated with winning.

There are no performance variables that have a significant positive influence on the nominated home team's phases, however GLS is bordering on significance (p=0.09); meanwhile GLS has a significant positive influence on the nominated away team's phases. These results obviously make conceptual sense, the more goals a team scores the more likely they are to win the match. The justification of using a model with such a poor fit is later revealed in its visual appeal which is explored in more detail in Section 11.4.2. Figure 11.1 shows an example of the two types of cumulative phases of play plots.

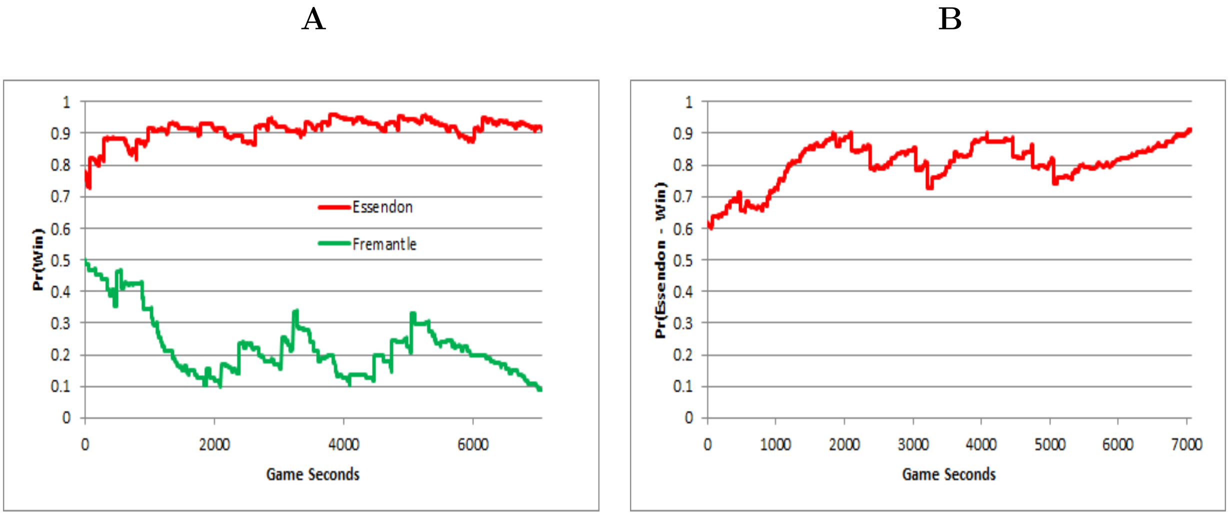


Figure 11.1: Essendon vs Fremantle, round 2 season 2009 (A) Cumulative phase plot. (B) Relative phase plot.

In the left pane of Figure 11.1 the plot shows the cumulative team phase for both teams. This allows the viewer to see if the team phases are anti-phase stable (i.e. one team is in high phase the other team is in low phase) or in-phase (i.e. both teams are in high phase or low phase). The right pane of Figure 11.1 shows the phase of the home team (Essendon) relative to the away team (Fremantle). Interestingly, the relative phase plot shows Essendon dominates the last quarter, however the cumulative team phase for both teams shows that this is attributed to Fremantle playing poorly rather than Essendon playing well.

11.4 Automation

Since each match is unique in terms of the total number of transactions, round number, names of competing teams etc., it was decided that an algorithm should be developed to automate the generation of the phases of play plot. Therefore, using VBA programming in Excel, the algorithm extracts the required performance variables, calculates the probability of either team winning by multiplying the regression coefficients to the cumulative statistics, and generates the phases of play plot which is visually enhanced by adding images of player's guernsey when a goal is scored. The remainder of this section describes the intricacies of how the algorithm works.

11.4.1 Extraction

The first port of call was to extract the previously mentioned 20 performance variables and assign each statistic to either the nominated home or away team. Figure 7.2 in Chapter 7 shows an extract of the transaction data. The first row in the spreadsheet summarizes the game and includes the name of the two teams, round number, match number, date and venue. Column headings occur in the second row and the data starts from the third row.

It is relatively straightforward to use IF statements in Excel to generate binary variables which correspond to KCK, HBL, MRK etc. using the statistic code in the last column. However, to extract $1ST$ and the type of goal scored, the sequence of transactions needed to be investigated. For example, to extract a GFM, the previous transaction before a GLS needed to have been a MRK. This process becomes more complicated for GFS as the team must have an uninterrupted chain of possessions immediately after a stoppage which resulted in a GLS. Therefore, comprehensive code is written to extract the required performance variables.

The next stage is to extract the name of the team which the statistic should be attributed to. For a typical transaction, this can be achieved by using the SEARCH command to search for the left square bracket ("[") and grab everything to the left of that square bracket using the LEFT command. However, several complications arise from using this code for all transactions. For example, RUS are a team variable hence the variable is not attributed to a single player, which means no player number and more importantly no square brackets for Excel to search for. Therefore, additional code is needed using the same two commands SEARCH and LEFT, however this time Excel searches for a semi-colon (:) if and only if a rushed behind occurs. Furthermore, there are transactions to denote the beginning and end of quarters, and umpire control situations (centre bounce, throw-in and ball up), therefore comprehensive code is written to account for these differences.

To calculate the probability of winning for the nominated home and away team, cumulative statistics are required for both teams at any point in time. Therefore, an additional sheet is created to generate the cumulative statistics for both teams using simple IF statements using the newly extracted team names and binary performance data. Then the regression coefficients in Table 11.1 can be multiplied by the cumulative statistics for both teams and a relative probability can be deduced throughout the entire match.

Since time elapsed in the transaction data is measured in minutes and seconds (mm:ss) from the beginning of each quarter, it was important to generate a variable which measured the total time elapsed since the beginning of the game for the phases of play plot. Therefore, the commands LEFT and RIGHT were used to extract the total minutes and seconds. This can then be easily modified to quarter seconds, that is, seconds elapsed since the beginning of the current quarter. Then the total match seconds can be calculated by enumerating each of the total quarter seconds variables depending on the current quarter. For example, during the third quarter, *match seconds = total quarter 1 seconds + total quarter 2 seconds + quarter 3 seconds elapsed.*

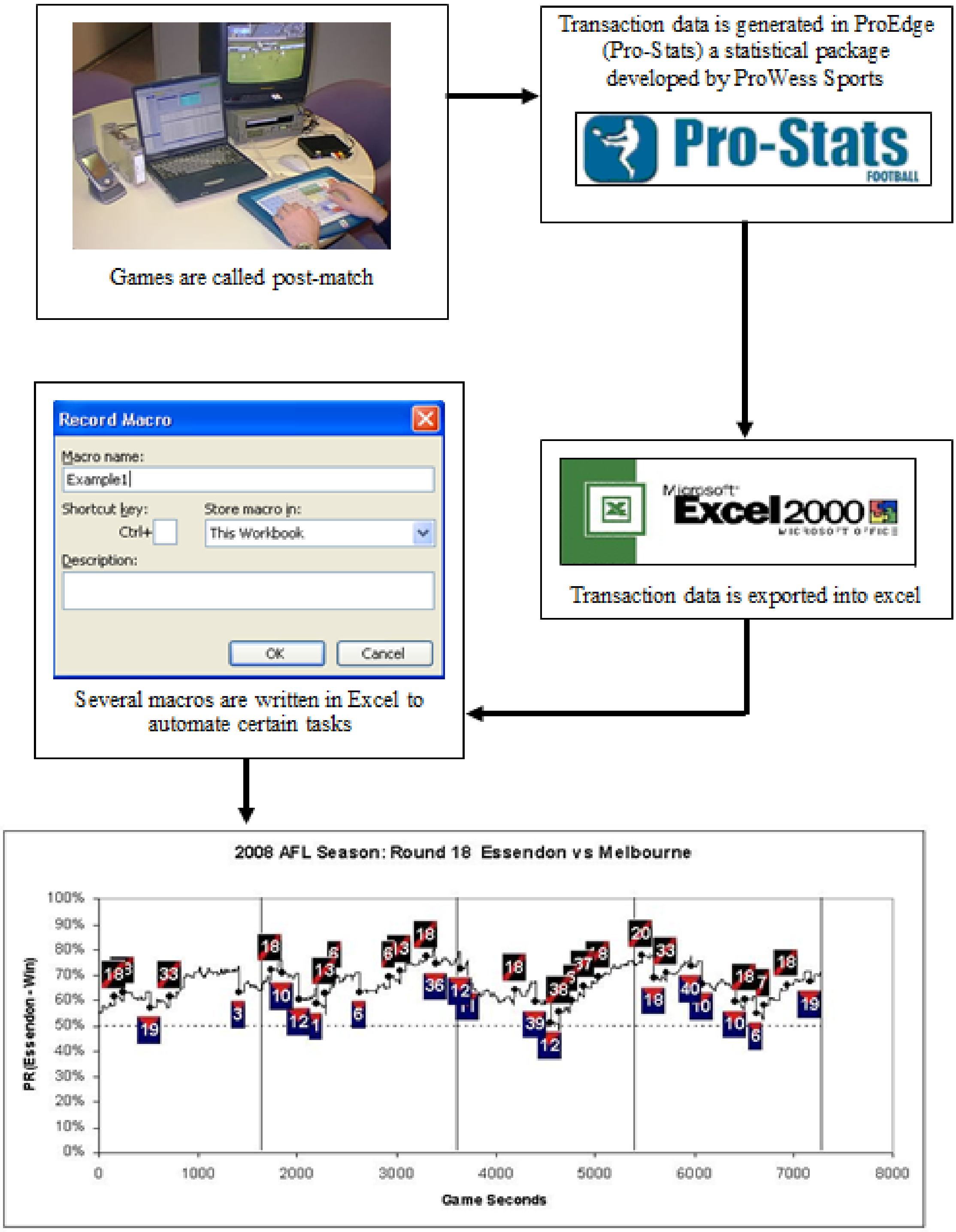
11.4.2 Generating the Plot

In order to generate the relative phase plot, the variable CUM needs to be quantified prior to the start of the match. Therefore, a separate sheet was created which contained all the previous match results such that CUM for any team for the preceding round could be accessed by using the VLOOKUP function in Excel. Recall the team names and round number was provided in the first row of the spreadsheet.

To generate the phases of play plot, total match seconds is plotted against the relative probability. When plotting a graph in Excel, the length of both axis (x and y) must be selected in terms of an array. However, each match is unique in that the total number of transactions can vary significantly from one match to the next. The most obvious solution would be to make the array adequately large (i.e. 4,000), since the average number of transactions in each match is approximately 2,500. Therefore, for each match, code was written to change all cells to missing (i.e. “”) for both arrays (time and relative probability) when the transactions for each match stopped (i.e. for a match with 2,500 transaction cells 2,501 to 4,000 would be empty). However, Excel does not treat empty cells with code (=“”) as truly being empty, and thus the plot looks fairly unattractive. Therefore, code is written to delete the cells after the final transaction of each match to account for this potential blemish. The title of the plot can also be automated to include the round number and the names of the two competing teams. Recall this information was provided in the first row of the transaction data.

To further enhance the plot, images of a player’s guernsey when a goal is scored along with the goal scorers number was superimposed on top. This was achieved by using an additional series (Z) with the same XY-coordinates consisting solely of a player’s number when a goal is scored. However, in Excel, data labels can only contain the X-value, Y-value or the series name. Since the player number corresponds to neither of these, this additional feature can not currently be achieved in Excel. However, XY data labels (an add-in for Excel) allows each data point to be labelled using a separate series (Z). Furthermore, to also include the goal scorer's respective team guernsey, the data labels can be formatted under the fill effects option and a picture from a file can be selected. It is important to note that this entire process has been automated such that the team guernseys will match the two competing teams.

Line breaks for each quarter were also included to easily differentiate between quarters. Additionally, a line was also generated across the X-axis with relative$_{home,t}$ $= 0.5$ to easily differentiate whether the home team is predicted to win (or lose) at any point in time. Figure 11.2 shows a flowchart of how the algorithm works.


Games are called post-match
Transaction data is generated in ProEdge (Pro-Stats) a statistical package developed by ProWess Sports
Pro-Stats
FOOTBALL
Microsoft
Excel 2000
MICROSOFT OFFICE
Transaction data is exported into excel
Record Macro
OK
Cancel
Several macros are written in Excel to automate certain tasks
2008 AFL Season: Round 18 Essendon vs Melbourne
PR(Essendon - Win)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
1000
2000
3000
4000
5000
6000
7000
8000
Game Seconds

Macro name: Example1

Shortcut key: Ctrl+

Store macro in: This Workbook

Description:

Figure 11.2: Flowchart of how the phases of play plot is automatically generated

Firstly, the transaction data is generated post-match by ProWess "callers" who provide detailed commentary which ultimately forms the foundation of the transaction data. At the conclusion of each round, the transaction data for each match of the current round can be exported as a separate CSV file for manipulation in Excel. To generate the phases of play plot for a single match, the transaction data is copied from the CSV file into an Excel spreadsheet where the macros have been pre-recorded. The macros, which have been pre-assigned a keyboard shortcut (ctrl+r), can then be run with the click of a button. Figure 11.3 showcases a relative phases of play plot for the 2008 Grand Final between Geelong and Hawthorn.

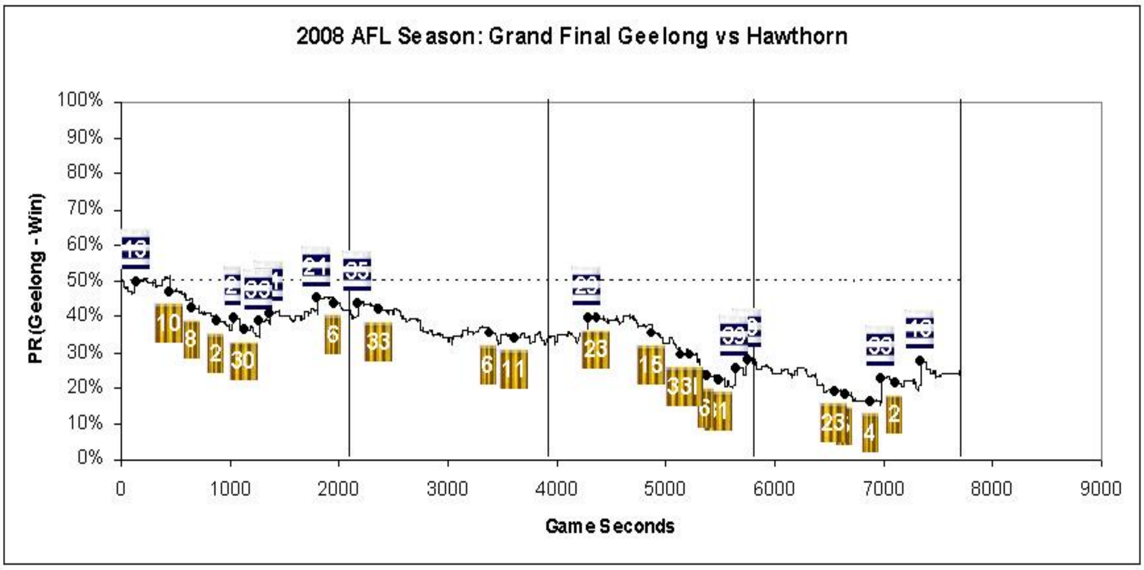


Figure 11.3: Phases of play Hawthorn vs. Geelong, 2008 Grand Final

Geelong went into the game as strong favourites only being defeated once throughout the entire 2008 season, whereas Hawthorn had been defeated on five occasions. Interestingly, at quarter time Geelong were leading by one point but the phases showed Hawthorn were playing a style of football that was more correlated with winning tendencies when compared to Geelong. Incidently, Hawthorn went on to win by 26 points but that is not the point. The point is that the score difference does not tell the entire story and its possible to model and isolate low and high phases of play where there is no change in score. For example, in Figure 11.3 midway through the 2nd quarter, it is clear Hawthorn is dominating play but this is not translated on the scoreboard. However, throughout the 3rd quarter, this dominance in play continues which ultimately led to more scoring opportunities. So the phases, in this instance, have in essence preempted future behaviour. This is an incredibly powerful concept which could be utilised directly as a coaching tool. For example, if a team is behind on the scoreboard yet the phases show they are outplaying their opponents, this could be used as a motivational tool (i.e. if the team in question keeps doing what they are doing this should eventually translate to the scoreboard). Neither the players or the coach need to understand the mathematics behind the model, all they need to comprehend is what the model is actually conveying, which is made relatively easy with the aide of the visuals. However, it is important to note that the phases will not always predict the eventually winner due to the nature of the model. Figure 11.4 shows an example of this scenario between Richmond and Essendon in round 16 season 2008.

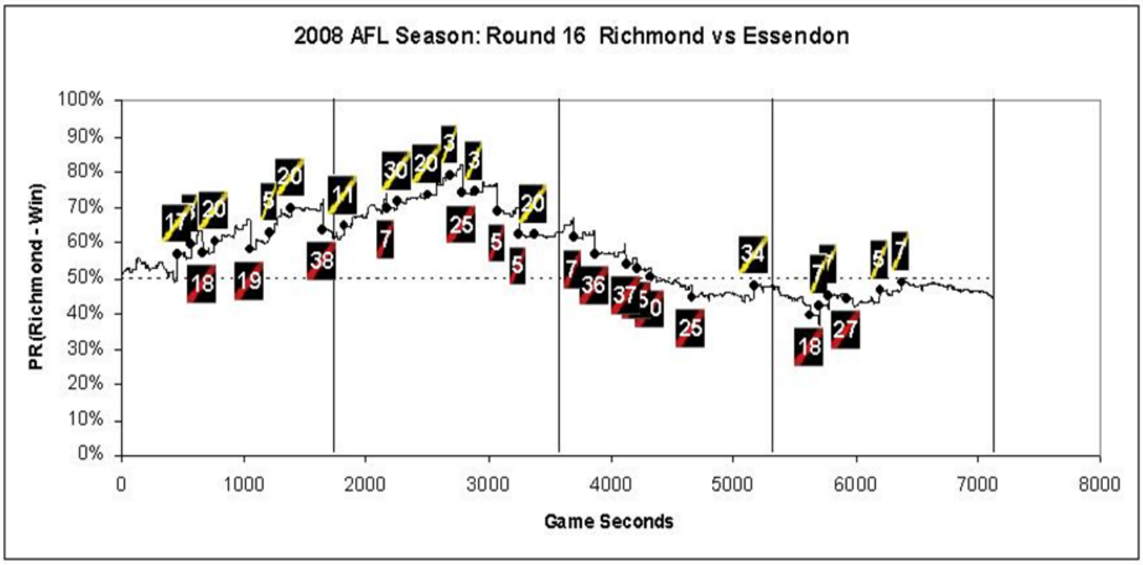


Figure 11.4: Phases of play Richmond vs. Essendon, round 16 season 2008

This match is infamous predominately due to the strategies employed by the Richmond players during the dying stages of the final quarter. Richmond, leading by seven points, knew that Essendon needed to score at least one goal to win the match. realising there was no more than a couple of minutes remaining in the match, the Richmond players maintained possession of the ball by rushing numerous behinds whenever they were under pressure. Although these tactics were later scrutinized, they worked as Richmond went on to win by four points. However, the phases suggest that Essendon were the “better” team on the day, playing a style of football that was more correlated with winning compared to Richmond. This begs further investigation as to why the losing team did not win.

11.4.3 Phases of Play in Real Time

The main aim of this chapter was to construct live statistical predictions that are both representative of a team’s likelihood of winning, and graphically simple enough to be widely interpretable for coaches and the general football public alike. Recall the variables extracted from the post-match performance data were also generated in real time on the Real Footy web site. Originally it was thought that this research would be turned into a web application with the phases updated on the Real Footy web site as the match progresses. Therefore, a feasibility study was conducted to see if a program could be developed to update the phases in real-time. An additional algorithm was developed using VBA programming which generates an interactive plot such that the match can be "played out" in pseudo real-time. Furthermore, since interchange data (time series of when players are rotated on and off the bench) was also available this was integrated with the transaction data such that the bench was also updated in pseudo real-time. This provides coaches with an object assessment of their teams performance and which players are contributing towards that performance.

A macro was written in Excel which cycles though each transaction and updates the phases plot as the match progresses. Since the difference in match seconds from one transaction to the next can vary from zero seconds to 30 seconds (i.e. when a goal is scored), the phases updates according to the frequency of transactions. However, if a football club was interested in the software, the program could be adapted so that it updates in real time so it could run parallel to video footage post match. To update the interchange bench, it is important to recognize that up to four interchanges for each team can occur simultaneously. For example, more recently in AFL matches, it is common for numerous interchanges to occur when there is a significant break in play (i.e. when a goal is scored). Therefore, a loop is written in Excel to allow for up to four interchanges to occur after each transaction. The program was developed for Geelong's matches in season 2008, as they showed interest in this research at the early stages of devlopment. Furthermore, if interchange was made available for subsequent and previous seasons and additional interest was shown in the results, the model could be adapted for all matches.

Previous analysis in team sports with the trait of measurable player independence, quantifies a player's individual performance on individual quantifiable statistics. For example, in baseball an offensive statistic is batting average (hits divided by number at bats). However, in AFL football, if we accept the idea that a player's individual performance (specifically key position players such as full forwards) is dependent on their teammates, then we arrive at an inadequate method for rating a player's impact on the a game as a whole. For example, it is widely thought that many "good" players in AFL are made to look good by their teammates and if they were in a lesser team their output would be arguably much smaller. An additional feature of running the phases in pseudo real-time and integrating the interchange data is that it is possible to calculate the Time on Ground (TOG) for each player. Furthermore, the influence each player has on team performance either directly or indirectly, can be measured by the average probability of winning relative to TOG. This would be a novel contribution to existing player ratings systems in dynamic team sports.

From (11.3) the average relative phase of the team can be calculated relative to a player's TOG. However, for players that are on the field for the entire game, such as midfielders, their measurable impact on the game relative to TOG does not give a true representation on their impact to the game, given they have not left the field. Furthermore, some players receive an inordinately high average relative probability due to the fact they happen to be on the field when the team is in high phase without contributing towards this performance. This is seen as a limitation and worth further investigation. It is important to note that these measures are not enough on their own to adequately measure a player's impact on the match and should be used in addition to a player rating system.

For example, Table 11.2 showcases an example from the 2007 AFL season in round 3 between Essendon and Carlton and ranks the player's average team phase relative to their respective TOG. Notably, the two ruckmen for Essendon appear at opposite ends of the table, David Hille and Jason Laycock. Hille finished top of the table for Essendon while the much maligned Laycock finished bottom and it is widely thought by Essendon supporters that Hille is a far superior player.

Carlton				Essendon			
Player no.	Player name	TOG	Rank	Player no.	Player name	TOG	Rank
6	Simpson	105:32	1	19	Hille	66:23	1
2	Russell	108:06	2	30	Ryder	107:46	2
8	Whitnall	82:09	3	25	Lucas	117:20	3
4	Gibbs	99:11	4	4	Watson	67:47	4
28	Cloke	65:46	5	13	Lovett	111:36	5
44	Carrazzo	114:58	6	1	Johnson	126:31	6
34	Wiggins	91:16	7	10	McVeigh	126:31	7
17	O'hAilpin	113:41	8	18	Lloyd	126:31	8
24	Stevens	122:15	9	22	Michael	126:31	9
25	Fevola	126:31	10	29	Davey	126:31	10
29	Scotland	126:31	11	31	Fletcher	126:31	11
30	Waite	126:31	12	33	McPhee	126:31	12
32	Thornton	126:31	13	11	Peverill	121:22	13
33	Houlihan	126:31	14	5	Hird	104:05	14
7	Bentick	84:48	15	26	Heffernan	110:06	15
14	Fisher	99:22	16	7	Jetta	86:13	16
19	Betts	108:18	17	24	Stanton	107:44	17
3	Murphy	108:30	18	8	Winderlich	112:26	18
12	Lappin	97:45	19	2	Dyson	53:21	19
1	Walker	100:27	20	20	Slattery	111:30	20
5	Kennedy	95:46	21	6	Monfries	75:22	21
11	Ackland	73:18	22	27	Laycock	63:54	22

Table 11.2: Alternative player rating system: Round 3 season 2007 Carlton vs. Essendon

11.5 Results

Akin to Section 7.5 of Chapter 7, to measure the reliability of the relative probability forecasts over time, comparisons are made between the percentage of games correctly classified by the probability forecasts against score difference at each of the quarter time breaks. The additional information the probability forecasts incorporate should be of greater importance at the earlier stages of the match since the outcome is largely unknown. However, as the match progresses, the score difference should have greater influence as teams have less opportunity to make up a deficit. If a team is leading on the scoreboard at quarter i then they are predicted to win according to score difference. Therefore, quarters whereby the scores were equal were removed from the analysis. Conversely, the team with a probability forecast of greater than 50% are predicted to win. Note that the probability forecasts are a decimal therefore no quarters need to be remove from the analysis (i.e. $\neq$0.50). Table 11.3 displays the percentage of games correctly classified by probability forecasts defined in (11.3) and score difference at each of the quarter time breaks.

Quarter	Score	Phases
1	69.71	73.53
2	78.53	80.59
3	90.88	87.65
4	100.00	91.17

NB. Data excludes draws and matches where data was not available (n=12)

Table 11.3: Percentage of games correctly classified by score difference and the phases at the quarter time breaks, 2007

Interestingly, the phases outperforms score difference during the first half but the score difference outperforms the phases in the second half. Furthermore, the phases incorrectly classifies 8.83% of matches. It is important to remember the target audience in this situation which is of course coaches and players. Therefore, less emphasis is placed on the forecasting capabilities (provided of course they are reasonably reliable) and more emphasis is placed on the applications of the model (i.e. motivational tool for players).

The next stage was to detect potential observations that may have a significant influence on the regression coefficients. These data points could be attributed to data entry errors, however they may be of interest to study on their own. For example, does the model consistently incorrectly classify teams with certain characteristics? In OLS, the difference between the observed value and the fitted value (i.e. residual) can be plotted against several metrics to check whether the assumptions of the linear regression model are valid. In logistic regression, there are several residuals including the Pearson residual and the Deviance residual. The Pearson residual measures the relative deviations between the observed and fitted values by standardizing the difference between the observed frequency and the predicted frequency. The Deviance residual measures the discrepancy between the maxima of the observed and the fitted log likelihood functions. Since there is a separate logistic regression model for nominated home and away teams, it is important to split the Pearson and Deviance residuals by home and away teams. Furthermore, there are several metrics which can be used to plot the residuals against, including case number and predicted probabilities. However, it is reasonable to assume that the residuals are independent of the case number due to the unordered nature of the data set. Generally speaking, if the absolute value of the Pearson or Deviance residual exceeds two it is worth further investigation. Figure 11.5 shows the Pearson residual for home and away teams respectively.

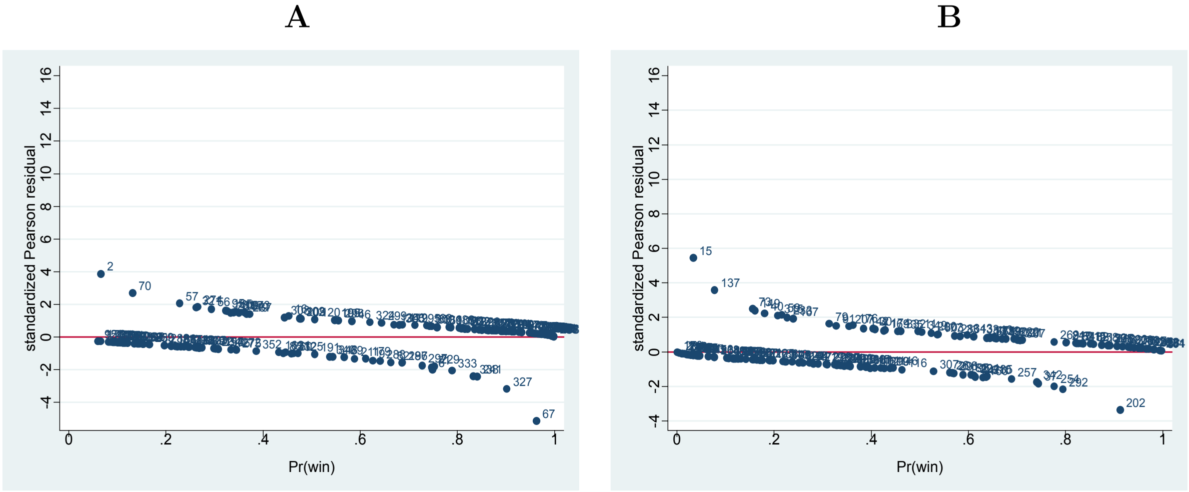


Figure 11.5: Pearson residuals (A) Home teams. (B) Away teams.

Clearly case ID 2, 70, 327 and 67 could have a significant influence on the regression model for nominated home teams, similarly case ID 15, 137 and 202 could have a significant influence on the regression model for nominated away teams. These cases will be investigated individually later in this section. Figure 11.6 showcases the Deviance residuals for nominated home and away teams respectively.

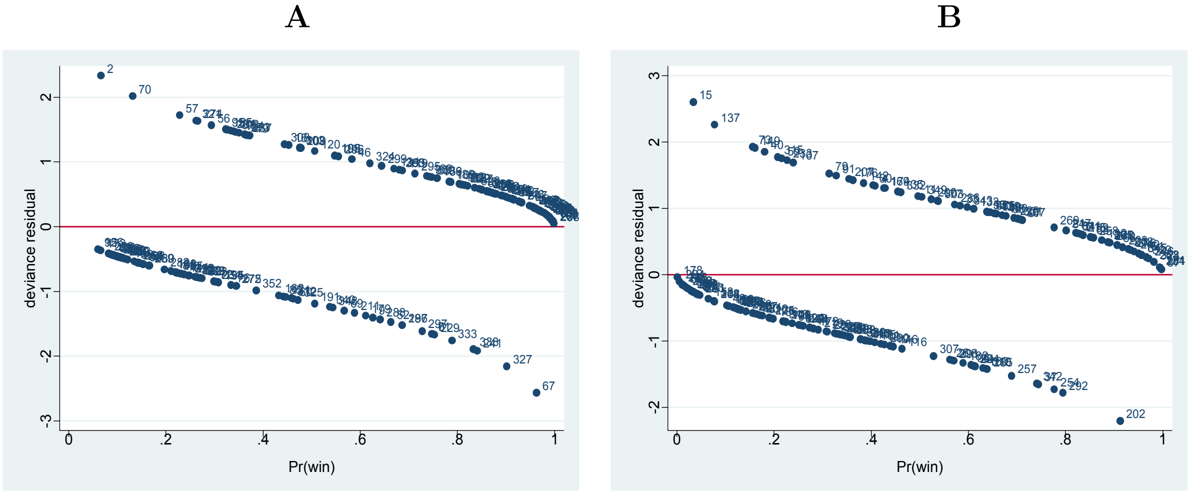


Figure 11.6: Deviance residuals (A) Home teams. (B) Away teams.

It is no coincidence that the same case ID's have been flagged for nominated home and away teams according to the Deviance residual when compared to to the Pearson residual. It is important to analyse these these cases to gain an insight as to *why* there is such a great discrepancy between observed and predicted results. Table 11.4 and 11.5 list the parameter values for each of the previously mentioned influential case ID's for nominated home and away teams respectively against the average.

Team	Brisbane	Carlton	Fremantle	Geelong	Home Average
Case ID	2	67	70	327	
CUM	0	0.5	0.25	0.85	0.47
KCK	205	187	223	202	205.35
HBL	175	123	145	162	147.57
MRK	115	84	82	106	104.22
$I50$	54	55	49	48	53.15
TKL	63	66	90	75	68.92
SPL	27	16	31	26	20.73
HIT	19	32	33	48	31.05
$1ST$	26	35	46	39	35.86
CLE	26	35	44	38	33.20
GLS	9	18	7	15	14.26
BHS	10	11	11	10	9.83
RUS	5	5	5	1	2.95
FF	28	26	26	25	21.46
$MI50$	12	13	9	15	14.25
GFG	4	7	2	7	4.58
GFF	0	2	1	3	1.55
GFM	3	6	2	2	6.07
GMK	0	0	0	1	1.06
GFS	4	6	4	5	4.82

Table 11.4: Parameter values for home teams which may influence regression coefficients

Team	Western Bulldogs	Kangaroos	Melbourne	Away Average
Case ID	15	137	202	
CUM	0	0.625	0.166	0.48
KCK	224	190	205	201.53
HBL	194	135	109	142.53
MRK	103	63	93	103.45
$I50$	52	56	58	51.38
TKL	62	104	58	68.85
SPL	33	31	17	20.85
HIT	28	47	31	30.56
$1ST$	41	46	32	33.77
CLE	33	36	35	31.05
GLS	17	10	18	13.42
BHS	8	15	11	9.58
RUS	3	4	4	2.74
FF	31	30	21	20.39
$MI50$	7	8	18	13.59
GFG	5	5	10	4.35
GFF	4	2	0	1.40
GFM	6	1	6	5.53
GFK	1	0	1	1.09
GFS	4	4	2	4.20

Table 11.5: Parameter values for away teams which may influence regression coefficients

Note that the same team does not appear more than once for both the nominated home and away model, this suggests that models are somewhat team independent. So what is it about these specific matches that may influence the coefficients of the regression models? It is immediately evident that these cases for home teams were flagged as being influential due to a multitude of variables being significantly different than the average including CUM and GFM. Similarly for away teams, the variables CUM, GFM and also HBL are significantly different from the average for these specific cases. The next stage is to test whether removing these cases alters the results of the logistic regression models. Table 11.6 shows the results.

	Coefficient		P-Value		Odds Ratio	
Variable	Home	Away	Home	Away	Home	Away
CUM	2.28	2.24	0.02*	0.02*	9.75 (1.48, 64.28)	9.44 (1.50, 59.38)
KCK	0.03	0.01	0.26	0.78	1.03 (0.98, 1.07)	1.01 (0.96, 1.05)
HBL	-0.01	-0.01	0.08	0.23	0.99 (0.97, 1.00)	0.99 (0.97, 1.01)
MRK	-0.04	-0.01	0.13	0.81	0.96 (0.92, 1.01)	0.99 (0.95, 1.04)
$I50$	-0.13	-0.08	0.02*	0.11	0.88 (0.79, 0.98)	0.93 (0.85, 1.02)
TKL	-0.02	0.01	0.35	0.47	0.98 (0.94, 1.02)	1.01 (0.98, 1.04)
SPL	0.01	-0.04	0.90	0.32	1.01 (0.92, 1.10)	0.96 (0.89, 1.04)
HIT	0.08	0.01	0.02*	0.68	1.08 (1.01, 1.16)	1.01 (0.96, 1.07)
$1ST$	-0.07	-0.17	0.26	0.04*	0.93 (0.82, 1.06)	0.84 (0.72, 0.99)
CLE	-0.03	0.00	0.68	0.98	0.97 (0.83, 1.13)	1.00 (0.85, 1.18)
GLS	0.46	0.60	0.03*	0.01*	1.59 (1.06, 2.38)	1.83 (1.19, 2.80)
BHS	0.18	0.05	0.08	0.55	1.20 (0.98, 1.46)	1.06 (0.88, 1.26)
RUS	0.05	0.14	0.73	0.28	1.05 (0.79, 1.39)	1.15 (0.89, 1.50)
FF	-0.09	-0.04	0.13	0.33	0.92 (0.82, 1.02)	0.96 (0.87, 1.05)
$MI50$	0.10	0.03	0.27	0.71	1.11 (0.93, 1.32)	1.03 (0.89, 1.20)
GFG	0.28	0.19	0.15	0.29	1.33 (0.91, 1.94)	1.21 (0.85, 1.72)
GFF	0.30	0.03	0.31	0.93	1.36 (0.75, 2.44)	1.03 (0.60, 1.76)
GFM	0.03	0.14	0.88	0.51	1.03 (0.67, 1.60)	1.14 (0.77, 1.71)
GFK	-0.57	-0.38	0.03*	0.10	0.57 (0.34, 0.94)	0.68 (0.44, 1.07)
GFS	0.08	-0.29	0.56	0.07	1.08 (0.83, 1.42)	0.75 (0.54, 1.02)

Table 11.6: Logistic regression results excluding influential cases: Real-time performance data, 2007

By removing these influential observations the statistical significance of several coefficients in the logistic regression model changed from significant to not significant and vice versa. For example, CUM is statistically significant (< 0.05) and has positive influence on the phase for home and away teams; HIT is statistically significant (< 0.05) and has positive influence on the phase for home teams; GLS is now also a statistically significant (< 0.05) and has positive influence on the away team phase; GFK is statistically significant (< 0.05) and has negative influence on the phase for home teams. These changes all seem to be conceptually correct except for GFK. This could be attributed to the infrequency of GFK for both home and away teams (approx one per game per team). Furthermore, it could be argued that all goals are worth exactly six points, therefore the contribution towards winning a match should be independent of goal type. Table 11.7 shows the percentage of games correctly classified by the two different logistic regression models and score difference.

Quarter	Score	Phases (Table 11.1)	Phases (Table 11.6)
1	69.71%	73.53%	69.36%
2	78.53%	80.59%	78.61%
3	90.88%	87.65%	85.55%
4	100.00%	91.17%	90.17%

NB. Data excludes draws and matches where data was not available (n=12)

Table 11.7: Percentage of games correctly classified by score difference and two different phases at the quarter time breaks, 2007

Interestingly, the predictive power of the phases decreases based on the logistic regression model in Table 11.6 which *excludes* the influential cases. However, it is reasonable to assume that this model would perform better in subsequent seasons when compared to the logistic regression model in (11.1) which *includes* the influential cases.

11.6 Discussion

All too often football clubs isolate individual statistics for which to draw conclusions from during a game and post match. However, this research shows that some metrics are meaningless on their own and thus need to be taken in context of the state of the match. For example, all teams are actually penalized when the ball goes into their forward 50 and they do not score. This suggests it is the quality of Inside 50's that are correlated with winning not the quantity. Therefore, teams should *never* adopt the strategy of getting the ball inside 50 as frequently as possible regardless of the consequences.

Chapter 12

Conclusions and Further Research

This dissertation has utilised mathematical models and computer programming techniques to provide further insight in relation to predicting outcomes in AFL. Furthermore, there are numerous direct applications of this research including betting markets and performance analysis which have been explored. In broad terms, the early chapters of this dissertation concentrated on home advantage and pre-game team ratings, while the later chapters had a central theme of measuring real-time outcomes. Each of the eight chapters which form the foundation of this dissertation have previously been peer-reviewed in either a journal or conference proceedings. The remainder of this chapter summarizes each of the previous chapters and the potential for future research.

12.1 Home Advantage

In Chapter 4, a new paradigm was proposed to quantify the precise cause of home advantage in AFL. It was thought that if travel, familiarity and crowd factors could be quantified through objective definitions, then their contribution towards home advantage could be deduced independently of all other factors. The factor defined for ground familiarity (GF) consisted of looking at the number of matches teams played at specific venues for each season relative to their direct opposition. It was reasonable to assume that the more games teams played at specific venues, the more familiar the team would become with the surroundings. Three separate factors were defined to measure the effect of travel fatigue. The first factor was a binary variable to distinguish when the away team was from a different state than the home team ($TRAV$); the second factor differentiated between Victorian and non-Victorian teams traveling interstate (VIC), since non-Victorian teams travel more frequently they might become accustomed to traveling and the resulting disadvantage might not be as significant; and the third factor measured the distance the away team travelled ($DIST$), if any, since it was widely assumed that home advantage was greater when the away team travelled large distances. It is important to reiterate that crowd numbers are unknown prior to the start of the match, and arguably of more importance, the breakdown of home, away and neutral supporters is always unknown. Therefore, the number of home and away supporters was estimated by incorporating the number of club members, team performance and interstate travel. This model explained an astonishing 93.41% of the variation in crowd numbers. From this, two separate factors to model the influence of crowd support were deduced. These factors were the difference between the estimated number of home and away supporters ($CROWD$) and the difference between the estimated number of home and away supporters divided by the capacity of the ground ($DENS$). The contribution of each of these factors towards home advantage was calculated by utilising a multiple linear regression model using margin of victory adjusted for team quality as the outcome variable. The results suggested that ground familiarity and distance travelled by the visiting team were the major determinants of home advantage in AFL. Furthermore, the amount of variation explained in margin of victory by this paradigm was a significant improvement over benchmark home advantage models (Clarke, 2005).

Although formal definitions provided an objective assessment of the precise cause of home advantage in AFL, the characteristics of ground familiarity and distance travelled that would yield this effect remain unclear. For example, is it the familiarity of the playing surface; familiarity of wind and weather conditions/climates due to different locations and types of stadiums; familiarity of playing fields; or a combination of these factors? Similarly for distance travelled, is the effect attributed to fatigue, time difference or something else? Therefore, future research in this area should further define these previously mentioned factors (familiarity, travel and crowd) into additional subsets. For example, the time difference between Australian states can easily be quantified and thus the effect, if any, can be deduced. If home advantage can be more accurately quantified by explaining the precise cause, then this should increase the predictive power of the model when integrated with a ratings system.

12.2 Ratings

In Chapter 5, Elo ratings (originally used to rate chess players) was adapted to forecast AFL matches. The model incorporated any difference in team quality such that a team receives a greater ratings increase if they defeat a stronger opponent compared to a weaker opponent. Furthermore, a change in ratings multiplier which is a function of margin of victory, weighted large wins (and losses) more heavily than small wins (or losses). By integrating the home advantage paradigm defined in Chapter 4 it was reasonable to assume that the predictive power of the model would also increase. The parameter values of the ratings model including the home advantage factors, were optimised using Riskoptimiser, an add-in for Excel. Additionally, by adjusting the initial ratings as the season progresses, the Elo ratings were arguably more representative of a team's true ability since less emphasis is placed on the previous season's results. Several methods were used to evaluate the performance of the model, including the reliability of the probability forecasts, number of predicted winners, average absolute margin of error and return on investment using standard wagering strategies to identify value bets. To compare the results against previous ratings models across different eras, a new metric was developed to gauge the evenness of the competition based on the standard deviation of premiership points at seasons end. This helps to isolate seasons where prediction models are expected to perform at a very low or high level.

The ratings system developed in Chapter 5 focused purely on match results and home advantage. Therefore, future research in this area should incorporate additional factors which are likely to influence match outcomes. For example, injuries and/or suspensions to key players, the importance of the match, or the departure of senior coaches mid-season. The nature of these factors are more important as the season progresses. For example, the departure of a coach mid-season is usually attributed to poor results during that season. Clearly, the subjective input of those knowledgable in AFL would enable other factors to be taken into consideration, thus increasing the return on investment. It is also important to note that although additional factors are likely to increase the predictive power of the model, there is a cost involved (time and resources). Therefore, the question needs to be asked whether the time taken to gather the additional information is worth the small increase in the predictive power of the model. The answer to this question is dependent upon the amount of resources available to the user. For example, for a single punter with a small bank size, it would not make sense to do the additional work for a small increase in ROI, however a betting syndicate with a virtually unlimited bank size would go to great lengths for a negligible increase in ROI.

12.3 Collecting In-Play Betting Data

In Chapter 6, a fully customized program was developed in Perl that integrates seamlessly with Betfair's API in order to record in-play betting data for AFL matches with minimal human intervention. The program was set to run at the beginning of each round and returned the back price, back volume, lay price and lay volume alongside a timestamp and the team name for all matches of the current round. This information alongside the current time (24 clock hh:mm:ss) and team name, was printed on the screen and recorded in a MySQL database. For convenience, the in-play betting data are recorded in a separate MySQL table for each match. At the conclusion of each round, the tables were exported as a CSV file for easy manipulation in Excel. This data are an extremely valuable commodity which has several practical applications.

Future research in this area would develop a program which is not AFL or in-play specific. The program should be adaptable for *any* betting market listed on the Betfair exchange and the user can state whether they would like pre-game odds or in-play odds as well as the frequency of collection. This would be an extremely valuable program which would have significant appeal especially to those with strong quantitative skills but a lack of computer programming. For example, throughout my PhD candidature, I have met several academics who wanted specific betting data (i.e. in-play tennis data) but did not have the skill set to write code to collect this information. A program such as this would enable several other conjectures to be formally tested across all sports. For example, matching pre-game odds for AFL matches against critical events to calculate the relative value of such events occurring. An illustration of this is when key players are in doubt for a match and an announcement is made that the player is not playing. There can often be a significant change in the betting odds when this occurs.

12.4 In-Play Betting Data as a Measure of Expectation

In Chapter 7, a new method for transforming in-play betting data to normalized implied probabilities was developed. This method weighted the back and lay price with their respective volumes to generate a unique price for each team at any point in time. These prices were then transformed into probabilities by taking the inverse of the price. A relative probability could then be deduced by normalizing the probabilities such that the probabilities of the two teams winning at any point in time sum to one. Furthermore, a graphical representation of the dissonance between score difference and the probability forecasts was obtained by matching performance data against in-play betting data. This process was also automated via an Excel macro developed exclusively for this dissertation. Several case studies show that there is often a clear difference in market opinion of victory and score difference. Interestingly, the results of the current thesis suggest that the forecasting capabilities of the implied probabilities is no different to score difference. However, a pronounced year effect was shown to be present for season 2009, with teams with certain characteristics winning considerably less/more frequently compared to previous seasons. Therefore, based on historical data, it is reasonable to assume that the implied probabilities would outperform score difference in subsequent seasons.

Future research in this area is needed to generate a live plot of the real-time expectations deduced from the in-play betting odds and score difference. This would require an additional program to scrape the score difference straight from the web in real-time into a workable format. Furthermore, the in-play betting odds would need to be utilised throughout the entire match as opposed to just at the conclusion of the match. Currently at AFL matches, the in-play betting odds are displayed graphically at each of the quarter time breaks which suggests what has been proposed is entirely plausible.

12.5 The Efficiency of In-Play Betting Markets

In Chapter 8, the efficiency of in-play AFL fixed odds betting markets at quarter time, half time and three quarter time were examined. Tests of semi-strong efficiency under the Efficient Market Hypothesis (*EMH*) were performed on the 2009 AFL season using logistic regression analysis. The results demonstrate that the team which is ahead during the match is underbet to win, particular late in the game. This bias was shown to be significant enough to yield profits utilising standard wagering strategies. Since the betting simulations conducted in this research were "in-sample" using a small sample size, and a strong year effect was shown to be present, clearly "out-of-sample" betting simulations would have to be conducted before one could conclude that in-play fixed odds betting markets in AFL are truly inefficient.

Future research in this area would be to investigate how in-play betting markets react to critical events. For example, when a goal is scored, does the market over/under inflate the importance relative to the true value? If a bias such as this exists, it provides a potential opportunity to trade frequently throughout the entire match. For example, if there is value in a team immediately after the opposition scores a goal, it would make sense to bet on this team, and when the value dissipates, bet against the same team and pocket the difference. However, there are many inherent problems with this strategy, particularly the low volume which results in large discrepancies between the back and lay price.

12.6 Intra-Match Home Advantage

In Chapter 9, the importance of considering the magnitude of team quality and score difference when quantifying home advantage as an intra-match measure in AFL was investigated. Home advantage in AFL was found to be a dominant factor in the first quarter irrespective of whether the home team was the favourite or the underdog. Furthermore, in the third quarter Home Underdogs (HU) had a distinct advantage over Home Favourites (HF) irrespective of the score difference at half time. However, one could argue the advantage the HU received in the third quarter is in vain due to only having an 18.9% chance of victory. Interestingly, in the final quarter home advantage was greatest when there was a high level of uncertainty [i.e. Home Underdog Ahead (HUA) or Home Favourite Behind (HFB)]. This suggests home crowds are most involved and vociferous late in the game when the outcome is largely unknown. Stefani (2008) suggests that the large playing field in Australian Rules football reduces the crowd's psychological influences on the match. If this argument holds true, intra-match home advantage is likely to be greater in sports such as basketball where spectators are in close proximity to the players.

Future research in this area should model home advantage as a game long process. For example, given X time has elapsed for given score difference Y and difference in team quality Z, how many points should be attributed to home advantage for the remainder of the match? Furthermore, if home advantage can be more accurately quantified during the match, then a real-time prediction model that accounts for this should outperform a similar model which utilises a constant home advantage. The changes in expectation from one quarter to the next could also be explored more directly using a simple Markov analysis.

12.7 In-Play Predictions

In Chapter 10, the importance of considering the interaction between team quality and score difference when modelling the outcome of an AFL match during the game was investigated. This interaction was found to be non-existent in the first quarter but increased significantly as the match progressed. The results suggest by taking into account the intricacies of team quality and score difference, the forecast probabilities provide a more accurate reflection of likelihood of victory, particularly late in the match. Furthermore, the Brownian Motion Model (BMM) and the generalised Logistic Model (GLM) yielded favourable ROI in the first and third quarters respectively for season 2009. Therefore, it is reasonable to assume that the ROI would increase in subsequent seasons since season 2009 was unprecedented and worked against these prediction models.

This finding will assist future research in this area by providing useful insight into the scoring behaviour of teams with certain characteristics. Future research should model the interaction between team quality and score difference as a game-long process. For example, given X time has elapsed for given score difference Y and difference in team quality Z, what is the probability of winning? I believe a Bayesian model might be appropriate to model this data. Furthermore, if home advantage can be more accurately quantified during the match, then a real-time prediction model that accounts for this should outperform a similar model which utilises a constant home advantage. There are also other additional factors which are likely to influence the outcome of the match during the game. For example, if a team is leading by $+l$ in standard weather conditions, then it starts raining heavily, it is reasonable to assume that the $+l$ point lead is now more valuable since the frequency of scoring is likely to decrease in wet conditions. Furthermore, it is not uncommon for several game-ending injuries to occur during a match. Again, it is important to reiterate that although additional factors are likely to increase the predictive power of the model there is a cost involved (time). Therefore, the cost needs to be weighed up against the magnitude of the predictive power increase.

12.8 Phases of Play

In Chapter 11, a new paradigm was developed which provided an objective measure to evaluate a team's performance during the game using logistic regression. A graphical representation of this objective measure was also enhanced by integrating a player's guernsey when a goal is scored and superimposing the goal scorer's number on top. This combination provides a real-time narrative of the ebbs and flows during a match and allows coaches to easily identify critical points during the game. An algorithm was established using macros in Excel which automatically transformed the "live-streaming" data into a single web-based phases of play plot for any given match. The results suggest that the probability assessment deduced from the real-time performance data is a better indication of the actual result compared to score difference during the first half of the match. However, score difference outperforms the probability assessment in the second half of the match. After investigating the residuals, there were several observations with similar characteristics that were deemed influential and were thus removed, following which the regression model was re-run. This resulted in the statistical significance of several coefficients in the logistic regression model changing from significant to not significant and vice versa. Furthermore, by analyzing team performance relative to a player's Time on Ground (TOG), it is possible to deduce the relative impact a player has on the overall performance of the team. However, there are several inherent problems relying solely on this metric as a player rating system, thus it should be used in addition to previous player rating systems.

Future research in this area should implement a similar model as a web application which updates in real time. To implement the phase live, a web scraper needs to be developed to extract the required performance variables in real-time. This would then be able to be utilised as an instantaneous coaching tool to isolate critical stages during a match. Further model development should also be investigated to improve the forecasting capabilities of the model. Since the model is based on data which are no longer available during the game, other avenues of real-time data should be explored in order to implement the phases live. Furthermore, the integration of real-time spatial data and performance data for phases of play in AFL should be investigated.

12.9 Summary

Drawing together all the research problems on AFL in this dissertation, a clearer picture exists as to *why* home advantage exists; *who* is the better football team; *when* critical events occur in a match; *who* represents good (betting) value during a game; *when* is home advantage greatest during a match; *who* is going to win during the match; and *why* did the losing team not win? These questions and many others can now be resolved objectively through statistical models developed in this dissertation. I look forward to continuing and building upon this work over the coming years and I honestly believe we are only limited by our own imagination. Anything is possible as data are the new oil!

Bibliography

(2006). What the papers say... AFL Website.

(2007). NFL media rights deals for ‘07 season. Sports Business Daily.

(2007). Show us the money. The Age newspaper.

(2008). Betfair announces 2,000,000th customer. Online Casino News.

(2009a). The AFL explained. AFL Website.

(2009b). The stats that matter! (and what you should be keeping). Coach AFL Website.

Bailey, M. J. (2000). Identifying arbitrage opportunities in AFL betting markets through mathematical modelling. In Cohen, G. and Langtry, T., editors, *Fifth Australian conference on Mathematics and Computers in Sport*, pages 37–42.

Bailey, M. J. and Clarke, S. R. (2002). Predicting the Brownlow medal winner. In Cohen, G. and Langtry, T., editors, *Sixth Australaian Conference on Mathematics and Computers in Sport*, pages 56–62.

Bailey, M. J. and Clarke, S. R. (2004). Deriving a profit from Australian Rules football: A statistical approach. In Morton, H. and Ganesalingam, S., editors, *Seventh Australasian Conference on Mathematics and Computers in Sport*, pages 48–56.

Bailey, M. J. and Clarke, S. R. (2006). Predicting the match outcome in one day international cricket matches, while the game is in progress. *Journal of Sports Science and Medicine*, 5(4):480–487.

Bailey, M. J. and Clarke, S. R. (2008). Assessing goodness of fit and optimal data size for a Brownlow prediction model. In Hammond, J., editor, *Ninth Australasian Conference on Mathematics and Computers in Sport*, pages 100–107.

Bailey, M. J., Clarke, S. R., and Forbes, D. (2010). Home continent advantage on the American and European professional golf tours. In Bedford, A. and Ovens, M., editors, *Tenth Australasian Conference on Mathematics and Computers in Sport*, pages 173–180.

Baumeister, R. F. and Steinhilber, A. (1984). Paradoxical effects of supportive audiences on performance under pressure: The home field disadvantage in sports championships. *Journal of Personality and Social Psychology*, 47(1):85–93.

Bedford, A. and Baglin, J. (2009). Evaluating the performance of an ice hockey team using interactive phases of play. *IMA Journal of Management Mathematics*, 20(2):159–166.

Biddle, S. J. H. (1993). *Handbook of research on sport psychology*, chapter Attribution research and sport psychology, pages 437–464. McMillan, New York.

Borland, J. and Lye, J. N. (1992). Attendance at Australian Rules football: A panel study. *Applied Economics*, 24(9):1053–1058.

Borrie, A., Jonsson, G. K., and Magnusson, M. S. (2002). Temporal pattern analysis and its applicability in sport: An explanation and exemplar data. *Journal of Sports Sciences*, 20(10):845–852.

Brailsford, T. J., Gray, P. K., Easton, S. A., and Gray, S. F. (1995). The efficiency of Australian football betting markets. *Australian Journal of Management*, 20(2):167–195.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Butler, J. L. and Baumeister, R. F. (1998). The trouble with friendly faces: Skilled performance with a supportive audience. *Journal of Personality and Social Psychology*, 75(5):1213–1230.

Clarke, S. R. (1988). Tinhead the tipster. *OR Insight*, 1(1):18–20.

Clarke, S. R. (1993). Computer forecasting of Australian Rules football for a daily newspaper. *Journal of the Operational Research Society*, 44(8):753–759.

Clarke, S. R. (1996). Calculating premiership odds by computer: An analysis of the afl final eight playoff system. *Asia-Pacific Journal of Operational Research*, 13(1):89–104.

Clarke, S. R. (1998). Evaluating the fairness and efficiency of the Australian Football League home and away schedule. *ASOR bulletin*, 17(3):2–11.

Clarke, S. R. (2005). Home advantage in the Australian Football League. *Journal of Sports Sciences*, 23(4):375–385.

Clarke, S. R., Bailey, M. J., and Yelas, S. (2008). Successful applications of statistical modeling to betting markets. *Mathematics TODAY*, 44(1):38–44.

Clarke, S. R. and Norman, J. M. (1995). Home ground advantage of individual clubs in English soccer. *Statistician*, 44(4):509–521.

Clarke, S. R. and Norman, J. M. (1998). When to rush a 'behind' in Australian Rules football: A dynamic programming approach. *Journal of the Operational Research Society*, 49(5):530–536.

Cooper, H., DeNeve, K. M., and Mosteller, F. (1992). Predicting professional sports game outcomes from intermediate game scores. *New directions for statistics and computing*, 5(3-4):18–22.

Courneya, K. S. and Carron, A. V. (1991). Effects of travel and length of home stand/road trip on the home advantage. *Journal of Sport & Exercise Psychology*, 13(1):42–49.

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.

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data.* Chapman & Hall, London, second edition.

Dare, W. H. and Holland, S. A. (2004). Efficiency in the NFL betting market: Modifying and consolidating research methods. *Applied Economics*, 36(1):9–15.

Dare, W. H. and MacDonald, S. (1996). A generalized model for testing the home and favorite team advantage in point spread markets. *Journal of Financial Economics*, 40(2):295–318.

Debnath, S., Pennock, D. M., Giles, C. L., and Lawrence, S. (2003). Information incorporation in online in-game sports betting markets. In *Fourth ACM Conference on Electronic Commerce.*

Dixon, M. J. and Pope, P. F. (2004). The value of statistical forecasts in the UK association football betting market. *International Journal of Forecasting*, 20(4):697–711.

Dobson, A. (1990). *An Introduction to Generalized Linear Models.* Chapman and Hall, London.

Dohmen, T. J. (2008). The influence of social forces: Evidence from the behavior of football referees. *Economic Inquiry*, 46(3):411–424.

Dowie, J. (1982). Why Spain should win the World Cup. *New Scientist*, 94:693–695.

Easton, S. and Uylangco, K. (2007). An examination of in-play sports betting using one-day cricket matches. *The Journal of Prediction Markets*, 1(2):93–109.

Easton, S. and Uylangco, K. (2010). Forecasting outcomes in tennis matches using within-match betting markets. *International Journal of Forecasting*, 26(3):564–575.

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present.* Arco, New York, second edition.

Falter, J. M. and Perignon, C. (2000). Demand for football and intramatch winning probability: An essay on the glorious uncertainty of sports. *Applied Economics*, 32(13):1757–1765.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

Flitman, A. M. (2006). Towards probabilistic footy tipping: A hybrid approach utilising genetically defined neural networks and linear programming. *Computers & Operations Research*, 33(7):2003–2022.

Franks, I. M. and Miller, G. (1986). Eye witness testimony in sport. *Journal of Sport Behavior*, 9(1):38–45.

Franks, I. M. and Miller, G. (1991). Training coaches to observe and remember. *Journal of Sports Sciences*, 9(3):285–297.

Gandar, J. M., Zuber, R. A., Johnson, R. S., and Dare, W. (2002). Re-examining the betting market on Major League Baseball games: Is there a reverse favourite-longshot bias? *Applied Economics*, 34(10):1309–1317.

Gandar, J. M., Zuber, R. A., O'Brien, T., and Russo, B. (1988). Testing rationality in the point spread betting market. *The Journal of Finance*, 43(4):995–1008.

Gil, R. and Levitt, S. D. (2007). Testing the efficiency of markets in the 2002 World Cup. *The Journal of Prediction Markets*, 1(3):255–270.

Glasson, S. (2006). A Brownian motion model for the progress of Australian Rules football scores. In *Eighth Australasian Conference on Mathematics and Computers in Sport.*

Golec, J. and Tamarkin, M. (1991). The degree of inefficiency in the football betting market: Statistical tests. *Journal of Financial Economics*, 30(2):311–323.

Gray, P. K. and Gray, S. F. (1997). Testing market efficiency: Evidence from the NFL sports betting market. *The Journal of Finance*, 52(4):1725–1737.

Greene, W. H. (2002). *Econometric analysis*. Prentice Hall, New Jersey, fifth edition.

Grehaigne, J. F., Bouthier, D., and David, B. (1997). Dynamic-system analysis of opponent relationships in collective actions in soccer. *Journal of Sports Sciences*, 15(2):137–149.

Harville, D. A. (1980). Predictions for National Football League games via linear-model methodology. *Journal of the American Statistical Association*, 75(371):516–524.

Harville, D. A. and Smith, M. H. (1994). The home-court advantage: How large is it, and does it vary from team to team? *The American Statistician*, 48(1):22–28.

Hess, R., Nicholson, M., Stewart, B., and De Moore, G. (2008). *A national game: The history of Australian Rules football*. Penguin Viking, Camberwell.

Holder, R. L. and Nevill, A. M. (1997). Modelling performance at international tennis and golf tournaments: Is there home advantage? *The Statistician*, 46(4):551–559.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley, New York, second edition.

Hvattum, L. M. and Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, 26(3):460–470.

IBISWorld (2009). Horse and sports betting in Australia. Australian Industry Report.

James, W. and Stein, C. M. (1961). Estimation with quadratic loss. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379.

Jones, M. B. (2007). Home advantage in the NBA as a game-long process. *Journal of Quantitative Analysis in Sports*, 3(4):Article 2.

Jones, M. V., Bray, S. R., and Olivier, S. (2005). Game location and aggression in Rugby League. *Journal of Sports Sciences*, 23(4):387–393.

Kelly, J. (1956). A new interpretation of information rate. *Bell System Technical Journal*, 35:917–926.

Klaassen, F. J. G. M. and Magnus, J. R. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research*, 148(2):257–267.

Kuper, G. H. and Sterken, E. (2006). *Statistical thinking in sports*, chapter Modelling the development of world records in running., pages 7–31. Chapman and Hall/CRC, Boca Raton.

Lames, M. (2006). Modelling the interaction in game sports - relative phase and moving correlations. In Hammond, J. and de Mestre, N., editors, *Eighth Australasian Conference on Mathematics and Computers in Sport*, pages 29–34.

Leitch, G. and Tanner, J. E. (1991). Economic forecast evaluation: Profits versus the conventional error measures. *The American Economic Review*, 81(3):580–590.

Leitner, C., Zeileis, A., and Hornik, K. (2009). Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008. *International Journal of Forecasting*, 26(3):471–481.

Levitt, S. D. (2004). Why are gambling markets organized so differently from financial markets? *Economic Journal*, 114(495):223–246.

MacLean, L. C., Ziemba, W. T., and Blazenko, G. (1992). Growth versus security in dynamic investment analysis. *Management Science*, 38(11):1562–1585.

Magee, C. (2008). *Automatic Exchange Betting: Automating the Betting Process from Strategy to Execution.* High Stakes, London.

Marcelino, R., Mesquita, I., Palao, J. M., and Sampio, J. (2009). Home advantage in high-level volleyball varies according to set number. *Journal of Sports Science and Medicine*, 8(3):352–356.

McGarry, T., Anderson, D. I., Wallace, S. A., Hughes, M. D., and Franks, I. M. (2002). Sport competition as a dynamical self-organizing system. *Journal of Sports Sciences*, 20(10):771–781.

McGarry, T., Khan, M. A., and Franks, I. M. (1999). On the presence and absence of behavioural traits in sports: An example from championship squash match-play. *Journal of Sports Sciences*, 17:297–311.

Neave, N. and Wolfson, S. (2003). Testosterone, territoriality and the 'home advantage'. *Physiology Behavior*, 78(2):269–275.

Nevill, A. M. and Holder, R. L. (1999). Home advantage in sport: An overview of studies on the advantage of playing at home. *Sports Medicine*, 28(4):221–236.

Nevill, A. M., Newell, S. M., and Gale, S. (1996). Factors associated with home advantage in English and Scottish soccer matches. *Journal of Sports Sciences*, 14(2):181–186.

Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization.* Springer Series in Operations Research. Springer, New York.

Pace, A. and Carron, A. V. (1992). Travel and the home advantage. *Canadian Journal of Sports Science*, 17(1):60–64.

Palut, Y. and Zanone, P. G. (2005). A dynamical analysis of tennis: Concepts and data. *Journal of Sports Sciences*, 23:1021–1032.

Paul, R. J. and Weinbach, A. P. (2007). Does sportsbook.com set pointspreads to maximize profits? Tests of the Levitt model of sportsbook behavior. *Journal of Prediction Markets*, 1(3):209–218.

Paul, R. J. and Weinbach, A. P. (2008). Price setting in the NBA gambling market: Tests of the Levitt model of sportsbook behavior. *International Journal of Sport Finance*, 3(3):137–145.

Pollard, R. (1986). Home advantage in soccer: A retrospective analysis. *Journal of Sports Sciences*, 4(3):237–248.

Pollard, R. (2008). Home advantage in football: A current review of an unsolved puzzle. *The Open Sports Sciences Journal*, 1:12–14.

Ryall, R. and Bedford, A. (2009). An automated approach to compare in-the-run markets with score in evaluation of team performance. In Lyons, K., Baca, A., and Lebedew, A., editors, *Seventh International Symposium on Computer Science in Sport*, pages 155–162.

Ryall, R. and Bedford, A. (2010a). The efficiency of "in-play" betting markets in Australian Rules football. *International Journal of Sports Finance*, 5(3):193–207.

Ryall, R. and Bedford, A. (2010b). Fitting probability distributions to real-time AFL data for match prediction. In Bedford, A. and Ovens, M., editors, *Tenth Australasian Conference on Mathematics and Computers in Sport*, pages 121–128.

Ryall, R. and Bedford, A. (2010c). An optimized ratings-based model to forecast Australian Rules football. *International Journal of Forecasting*, 26(3):511–517.

Ryall, R. and Bedford, A. (2011a). Independent effects that augment home ground advantage. *Journal of Sports Sciences*. Manuscript in review.

Ryall, R. and Bedford, B. (2008). An algorithm to plot an AFL teams performance in real-time using interactive phases of play. In Hammond, J., editor, *Ninth Australasian Conference on Mathematics and Computers in Sport*, pages 108–114.

Ryall, R. and Bedford, B. (2011b). The intra-match home advantage in Australian Rules football. *Journal of Quantitative Analysis in Sports.* Manuscript accepted for publication 27th January 2011.

Schembri, A. J. and Bedford, A. (2010). Longtitudinal impact of in-game player injuries on scoring. Manuscript in preparation.

Schnytzer, A. and Weinberg, G. (2008). Testing for home team and favorite biases in the Australian Rules football fixed-odds and point spread betting markets. *Journal of Sports Economics*, 9(2):173–190.

Schwartz, B. and Barsky, S. F. (1977). The home advantage. *Social Forces*, 55(3):641–661.

Stefani, R. T. (1983). Observed betting tendencies and suggested betting strategies for European football pools. *The Statistician*, 32(3):319–329.

Stefani, R. T. (1987). Applications of statistical methods to American football. *Journal of Applied Statistics*, 14(1):61–73.

Stefani, R. T. (1998). *Statistics in Sport*, chapter A taxonomy and survey of sports rating systems, pages 253–258. Arnold, New Jersey.

Stefani, R. T. (2008). *Statistical Thinking in Sports.*, chapter Measurement and interpretation of home advantage., pages 203–216. Chapman and Hall/CRC, Boca Raton.

Stefani, R. T. (2010). A world of sports and rating systems. In Bedford, A. and Ovens, M., editors, *Tenth Australasian Conference on Mathematics and Computers in Sport*, pages 1–12.

Stefani, R. T. and Clarke, S. R. (1992). Predictions and home advantage for Australian Rules football. *Journal of Applied Statistics*, 19:251–261.

Stern, H. S. (1994). A Brownian motion model for the progress of sport scores. *Journal of the American Statistical Association*, 89(427):1128–1134.

Stewart, M. F., Mitchell, H., and Stavros, C. (2007). 'Moneyball' applied: Econometrics and the identification and recruitment of elite Australian footballers. *International Journal of Sport Finance*, 2(4):231–248.

Sumner, J. and Mobley, M. (1981). Are cricket umpires biased? *New Scientist*, 91(1260):29–31.

Thaler, R. H. and Ziemba, W. T. (1988). Parimutuel betting markets: Racetracks and lotteries. *Journal of Economic Perspectives*, 2(2):161–174.

Thirer, J. and Rampey, M. (1979). Effects of abusive spectator behaviour on the performance of home and visiting intercollegiate basketball teams. *Perceptual and Motor Skills*, 48:1047–1053.

Westfall, P. H. (1990). Graphical presentation of a basketball game. *The American Statistician*, 4(4):305–307.

Witten, I. H. and Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Elsevier, San Francisco, second edition.

Wolfson, S., Wakelin, D., and Lewis, M. (2005). Football supporters perceptions of their role in the home advantage. *Journal of Sports Sciences*, 23(4):365–374.

Woodland, L. M. and Woodland, B. M. (1994). Market efficiency and the favorite-longshot bias: The baseball betting market. *The Journal of Finance*, 49(1):269–279.

Woodland, L. M. and Woodland, B. M. (2001). Market efficiency and profitable wagering in the National Hockey League: Can bettors score on longshots? *Southern Economic Journal*, 67(4):983–995.

Wright, E. F., Jackson, W., Christie, S. D., McGuire, G. R., and Wright, R. D. (1991). The home-course disadvantage in golf championships: Further evidence for the undermining effect of supportive audiences on performance under pressure. *Journal of Sport Behavior*, 14(1):51–60.

Wright, E. F., Voyer, D., Wright, R. D., and Roney, C. (1995). Supporting audiences and performance under pressure: The home-ice disadvantage in hockey championships. *Journal of Sport Behavior*, 18(1):21–28.

Zuber, R. A., Gandar, J. M., and Bowers, B. D. (1985). Beating the spread: Testing the efficiency of the gambling market for National Football League games. *Journal of Political Economy*, 93(4):800–806.

Chapter 13

Appendix

13.1 Perl

This sections covers the Perl code used in Chapter 6 to automate the collection of in-play betting data for AFL matches. Firstly, Section 13.1.1 provides the perl code to generate eight tables in a specified database within MySQL. Each table represents a single match of a given round with each table consisting of six columns (timestamp, team name, back price, back volume, lay price and lay volume). Section 13.1.2 details the Perl program (developed *exclusively* for this dissertation) to automate the collection of in-play betting data for AFL matches. The final section, Section 13.1.3 details the Perl library for accessing Betfair API services. Note that the subroutines listed in the Perl library were extracted in full (or slightly adapted) from Magee (2008) excluding the subroutine `getActiveEventTypes`.

13.1.1 Define Tables in MySQL

```
#CREATE table for each Betfair event in the automatic betting example database

DROP TABLE IF EXISTS AFL_1;
#@ _CREATE_TABLE_
CREATE TABLE AFL_1
{
        timestamp       time NOT NULL,
        team            VARCHAR(25) NOT NULL,
        back price      DECIMAL(6,2),
        back volume     DECIMAL(10,0),
        lay price       DECIMAL(6,2),
        lay volume      DECIMAL(10,0),
};
#@ _CREATE_TABLE_

DROP TABLE IF EXISTS AFL_2;
#@ _CREATE_TABLE_
CREATE TABLE AFL_2
{
        timestamp       time NOT NULL,
        team            VARCHAR(25) NOT NULL,
        back price      DECIMAL(6,2),
        back volume     DECIMAL(10,0),
        lay price       DECIMAL(6,2),
        lay volume      DECIMAL(10,0),
};
#@ _CREATE_TABLE_

DROP TABLE IF EXISTS AFL_3;
#@ _CREATE_TABLE_
CREATE TABLE AFL_3
{
        timestamp       time NOT NULL,
        team            VARCHAR(25) NOT NULL,
        back price      DECIMAL(6,2),
        back volume     DECIMAL(10,0),
        lay price       DECIMAL(6,2),
        lay volume      DECIMAL(10,0),
};
```

```
#@ _CREATE_TABLE_

DROP TABLE IF EXISTS AFL_4;
#@ _CREATE_TABLE_
CREATE TABLE AFL_4
{
        timestamp       time NOT NULL,
        team            VARCHAR(25) NOT NULL,
        back price      DECIMAL(6,2),
        back volume     DECIMAL(10,0),
        lay price       DECIMAL(6,2),
        lay volume      DECIMAL(10,0),
};
#@ _CREATE_TABLE_

DROP TABLE IF EXISTS AFL_5;
#@ _CREATE_TABLE_
CREATE TABLE AFL_5
{
        timestamp       time NOT NULL,
        team            VARCHAR(25) NOT NULL,
        back price      DECIMAL(6,2),
        back volume     DECIMAL(10,0),
        lay price       DECIMAL(6,2),
        lay volume      DECIMAL(10,0),
};
#@ _CREATE_TABLE_

DROP TABLE IF EXISTS AFL_6;
#@ _CREATE_TABLE_
CREATE TABLE AFL_6
{
        timestamp       time NOT NULL,
        team            VARCHAR(25) NOT NULL,
        back price      DECIMAL(6,2),
        back volume     DECIMAL(10,0),
        lay price       DECIMAL(6,2),
        lay volume      DECIMAL(10,0),
};
#@ _CREATE_TABLE_

DROP TABLE IF EXISTS AFL_7;
```

```
#@ _CREATE_TABLE_
CREATE TABLE AFL_7
{
        timestamp       time NOT NULL,
        team            VARCHAR(25) NOT NULL,
        back price      DECIMAL(6,2),
        back volume     DECIMAL(10,0),
        lay price       DECIMAL(6,2),
        lay volume      DECIMAL(10,0),
};
#@ _CREATE_TABLE_

DROP TABLE IF EXISTS AFL_8;
#@ _CREATE_TABLE_
CREATE TABLE AFL_8
{
        timestamp       time NOT NULL,
        team            VARCHAR(25) NOT NULL,
        back price      DECIMAL(6,2),
        back volume     DECIMAL(10,0),
        lay price       DECIMAL(6,2),
        lay volume      DECIMAL(10,0),
};
#@ _CREATE_TABLE_
```

13.1.2 Collect in-play betting data

```
#!/usr/bin/perl -w

#       prerequisite modules to run this script
use lib "/home/rryall/lib";
use BetfairAPI6Examples;
use LWP::UserAgent;
use LWP::Debug; # qw(+trace +debug +conns);
use HTTP::Request;
use HTTP::Cookies;
use SOAP::Lite +trace => "all";
use Data::Dumper;
use XML::Simple;
use XML::XPath;
use DBI;
use strict;
use warnings;
use Time::Local;


#       login variables
my $username = "username";
my $password = "password";
my $productId = "82";           #Free Access API access code


#       other program variables not declared in line
my $back_price;
my $lay_price;
my $lay_vol;
my $back_vol;
my $timestamp;
my $date;
my $discard;
my $event_delay;
my $marketStatus;
my $current_time;
my $current_minute;
my $current_hour;
my $finishTime;
my %prices_hash;
my @market_intherun;
my %index;
my $index;
my @market_array;
my $time1;
my $time2;
my $time3;
my %static_runner_data;
my @names;
my $sql;
my $query;
my $search;
my $count;
my $sec;
my $min;
my $hour;
my $mday;
my $month;
my $year;

my $event_menu = "25762578"; # eventId for "AFL 2010"
my @match_days = ("matchday1", "matchday2", "matchday3");
# enter round number and date (e.g. "Round 6 - May 01") for all matches
my $match;
my @event_id2 = ();
my @event_id3 = ();
my @event_id4 = ();
my @market = ();
my $event_name1;
my $event_name2;
my $market_name;

my $finish_min;
my $finish_hour;
my $finish_mday;
my $finish_month;
my $finish_year;
my $finish_time;
my %alive;

#       open our database handle for a permanent record of the prices
my $dbh = DBI->connect("DBI:mysql:database", "username") or die ("Error:  $DBI::errstr");
#substitute your database and user credentials

#       login to the Betfair API
my %login = login($username, $password, $productId);
my $token = $login{sessionToken};
my $login_error = $login{errorCode};

if ( !($login_error =~ /OK/) )
{
  print "Failed login:\n";
  print "$login_error";
}
else
{
      print "Login Successful!\n";
}


my %events_hash1 = get_events1($token, $event_menu);
my $event_ref1 = $events_hash1{events};
my @event_id1 = keys ( %{$event_ref1} );
foreach my $event_id1 (@event_id1) {
        $event_name1 = $events_hash1{events}->{$event_id1}->{eventName};
        my $match=grep $_ eq $event_name1, @match_days;
        if ($match==1) {
                push(@event_id2, $event_id1);
        }
}

foreach my $event_id2 (@event_id2) {
        my %events_hash2 = get_events1($token, $event_id2);
        my $event_ref2 = $events_hash2{events};
        my @event_id3 = keys ( %{$event_ref2} );
        foreach my $event_id3 (@event_id3) {
                $event_name2 = $events_hash2{events}->{$event_id3}->{eventName};
                push(@event_id4, $event_id3);
        }
}

foreach my $event_id4 (@event_id4) {
        my %markets_hash = get_events2($token, $event_id4);
        my $market_ref = $markets_hash{markets};
        my @market_id = keys ( %{$market_ref} );
        foreach my $market_id (@market_id) {
                $market_name = $markets_hash{markets}->{$market_id}->{marketName};
                if ($market_name eq "Match Odds") {
                        push(@market, $market_id);
                }
        }
}

($sec, $min, $hour, $mday, $month, $year)=localtime(time);
$month=$month+1;

if ($month==2) {
        $month=$month*28*24;
}
```

```
if ($month==1 || $month==3 || $month==5 || $month==7 || $month==8 || $month==10 || $month==12) {
        $month=$month*31*24;
}
if ($month==4 || $month==6 || $month==9 || $month==11) {
        $month=$month*30*24;
}
$mday=$mday*24;
$current_hour=$month+$mday+$hour;
$current_time = "$current_hour.$min";

$finish_month=month; #Enter month number (1 Janurary 2 February ... 12 December) program breaks for current round
$finish_mday=mday; #Enter day of month (1 to 31) program breaks for current round
$finish_hour=hour; #Enter hour (0 to 23) program breaks for current round
$finish_min=min; #Enter minute (0 to 59) program breaks for current round

if ($finish_month==2) {
        $finish_month=$finish_month*28*24;
}
if ($finish_month==1 || $finish_month==3 || $finish_month==5 || $finish_month==7 || $finish_month==8 ||
        $finish_month==10 || $finish_month==12) {
        $finish_month=$finish_month*31*24;
}
if ($finish_month==4 || $finish_month==6 || $finish_month==9 || $finish_month==11) {
        $finish_month=$finish_month*30*24;
}
$finish_mday=$finish_mday*24;
$finish_hour=$finish_month+$finish_mday+$finish_hour;
$finish_time = "$finish_hour.$finish_min";

while ($current_time<=$finish_time) {
        $dbh = DBI->connect("DBI:mysql:autodb", "rryall") or die ("Error:  $DBI::errstr");
        print "current time ($current_time) is still before finish time ($finish_time)\n";
        ($sec, $min, $hour, $mday, $month, $year)=localtime(time);
        $month=$month+1;
        if ($month==2) {
                $month=$month*28*24;
        }
        if ($month==1 || $month==3 || $month==5 || $month==7 || $month==8 || $month==10 || $month==12) {
                $month=$month*31*24;
        }
        if ($month==4 || $month==6 || $month==9 || $month==11) {
                $month=$month*30*24;
```

```
}
$mday=$mday*24;
$current_hour=$month+$mday+$hour;
$current_time = "$current_hour.$min";
$count=0;
@market_intherun=();
foreach my $marketid (@market) {
        %prices_hash = get_market_prices_compressed($token, $marketid);
        $event_delay = $prices_hash{'delay'};
        $marketStatus = $prices_hash{'marketStatus'};
        if ($event_delay>0 && $marketStatus eq "ACTIVE") {
                push(@market_intherun, $marketid);
                $count=$count+1;
                if ($count==1) {

                }
        }
}
if ($count==0) {
        print "no current in-the-run market\n";
}
sleep 11.5;
foreach my $marketid_intherun (@market_intherun) {
        $search="$marketid_intherun";
        @index{@market}= (0..$#market);
        $index=$index{$search};
        $index=$index+1;
        %prices_hash = get_market_prices_compressed($token, $marketid_intherun);
        $marketStatus = $prices_hash{'marketStatus'};
        $timestamp = $prices_hash{timeStamp};
        ($date, $timestamp) = split (/T/, $timestamp);
        ($timestamp, $discard) = split (/Z/, $timestamp);
        ($time1, $time2, $time3) = split (/:/, $timestamp);
        $time1=$time1+10;
        $timestamp = "$time1:$time2:$time3";
        @market_array = get_markets($token, $marketid_intherun);
        %static_runner_data = %{$market_array[0]};
        @names = keys(%static_runner_data);
        foreach my $runner (@names) {
                my $runnerId = $static_runner_data{$runner};
                $back_price = $prices_hash{prices}->{$runnerId}->{back}->{1}->{price};
                $back_vol = $prices_hash{prices}->{$runnerId}->{back}->{1}->{amountAvailable};
```

```
                        $lay_price = $prices_hash{prices}->{$runnerId}->{lay}->{1}->{price};
                        $lay_vol = $prices_hash{prices}->{$runnerId}->{lay}->{1}->{amountAvailable};
                        print "$date, $timestamp, $runner, $back_price, $back_vol, $lay_price, $lay_vol\n";
                        $sql = qq(INSERT INTO afl_$index VALUES
                        (’$timestamp’, ’$runner’, ’$back_price’, ’$back_vol’, ’$lay_price’, ’$lay_vol’) );
                        $query = $dbh->prepare($sql);
                        $query->execute;
                }
        }
}
```

13.1.3 Perl Library for Accessing Betfair API Services

```
package BetfairAPI6Examples;
use LWP::UserAgent;
use LWP::Debug; # qw(+trace +debug +conns);
use HTTP::Request;
use HTTP::Cookies;
use Data::Dumper;
use XML::Simple;
use XML::XPath;
use strict;
use warnings;
require Exporter;
our @ISA = qw(Exporter);
our @EXPORT=
qw(login get_active_event_types get_events1 get_events2 get_markets get_market_prices_compressed);
our $VERSION= 1.0;
our $cookie_jar = HTTP::Cookies->new(hide_cookie2 => 1);


sub login
        {
  my ($username,$password,$productId)=@_;
  my $xml=’<SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
        xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xmlns:xsd="http://www.w3.org/2001/XMLSchema">
        <SOAP-ENV:Body>
                <m:login xmlns:m="http://www.betfair.com/publicapi/v3/BFGlobalService/">
```

```
                        <m:request>
                                <password>'.$password.'</password>
                                <productId>'.$productId.'</productId>
                                <username>'.$username.'</username>
                                <vendorSoftwareId>0</vendorSoftwareId>
                                <locationId>0</locationId>
                        </m:request>
                </m:login>
        </SOAP-ENV:Body>
</SOAP-ENV:Envelope>';
my $userAgent = LWP::UserAgent->new();
my $request = HTTP::Request->new(POST => 'https://api.betfair.com/global/v3/BFGlobalService');
$request->header(SOAPAction => '"
https://api.betfair.com/global/v3/BFGlobalService
"');
$request->content($xml);
$request->content_type("text/xml; charset=utf-8");

my $resp = $userAgent->request($request);
my $content=$resp->content;
#print Dumper(\$content);

my $ref;

eval { $ref = XMLin($content) };
if ($@) {print Dumper ($content); print "$@\n"; die "login failed to retrieve valid XML"};

my %login_hash = ();
$login_hash{sessionToken}       =$ref->{'soap:Body'}{'n:loginResponse'}{'n:Result'}
{'header'}{'sessionToken'}{'content'};
$login_hash{headererrorCode}    =$ref->{'soap:Body'}{'n:loginResponse'}{'n:Result'}
{'header'}{'errorCode'}{'content'};
$login_hash{errorCode}          =$ref->{'soap:Body'}{'n:loginResponse'}{'n:Result'}
{'errorCode'}{'content'};
        return %login_hash;
        }



sub get_active_event_types
  {
  my ($sessionToken)=@_;
  my %active_events;
```

```
 my $xml=’<SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
       xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
       <SOAP-ENV:Body>
               <m:getActiveEventTypes xmlns:m="http://www.betfair.com/publicapi/v3/BFGlobalService/">
                       <m:request>
                               <header>
                                       <clientStamp>0</clientStamp>
                                       <sessionToken>’.$sessionToken.’</sessionToken>
                               </header>
                       </m:request>
               </m:getActiveEventTypes>
       </SOAP-ENV:Body>
</SOAP-ENV:Envelope>’;

 my $userAgent = LWP::UserAgent->new();
 my $request = HTTP::Request->new(POST => ’https://api.betfair.com/global/v3/BFGlobalService’);
 $request->header(SOAPAction => ’"https://api.betfair.com/global/v3/BFGlobalService"’);
 $request->content($xml);
 $request->content_type("text/xml; charset=utf-8");

 my $resp = $userAgent->request($request);
 #print Dumper(\$resp);
 my $content=$resp->content;
 #print Dumper(\$content);

 my $result;
 eval { $result = XMLin($content) };
 if ($@) {print Dumper ($content); print "$@\n"; die "get_active_event_types failed to retrieve valid XML"};

 my $response=\%{$result->{’soap:Body’}{’n:getActiveEventTypesResponse’}{’n:Result’}};

 $active_events{’sessionToken’}        =$response->{’header’}->{’sessionToken’}{’content’};
 $active_events{’timeStamp’}           =$response->{’header’}->{’timestamp’}{’content’};
 $active_events{’errorCode’}           =$response->{’errorCode’}{’content’};
 foreach my $key (@{$response->{’eventTypeItems’}{’n2:EventType’}})
 {
    $active_events{eventType}{$key->{id}{’content’}}{name}           =   $key->{name}{’content’};
    $active_events{eventType}{$key->{id}{’content’}}{nextMarketId}   =   $key->{nextMarketId}{’content’};
    $active_events{eventType}{$key->{id}{’content’}}{exchangeId}     =   $key->{exchangeId}{’content’};

 }
```

```
  return %active_events;
}

sub get_events1
        {
  my ($sessionToken,$eventParentId)=@_;
        my %events_hash;
  my $xml=’<SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
        xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
        <SOAP-ENV:Body>
                <m:getEvents xmlns:m="http://www.betfair.com/publicapi/v3/BFGlobalService/">
                        <m:request>
                                <header>
                                        <clientStamp>0</clientStamp>
                                        <sessionToken>’.$sessionToken.’</sessionToken>
                                </header>
                                <eventParentId>’.$eventParentId.’</eventParentId>
                        </m:request>
                </m:getEvents>
        </SOAP-ENV:Body>
</SOAP-ENV:Envelope>’;

my $userAgent = LWP::UserAgent->new();
my $request = HTTP::Request->new(POST => ’https://api.betfair.com/global/v3/BFGlobalService’);
$request->header(SOAPAction => ’"https://api.betfair.com/global/v3/BFGlobalService"’);
$request->content($xml);
$request->content_type("text/xml; charset=utf-8");

my $resp = $userAgent->request($request);
#print Dumper(\$resp);
my $content=$resp->content;
#print Dumper(\$content);
my $result;
eval { $result = XMLin($content) };
if ($@) {print Dumper ($content); print "$@\n"; die "get_events call failed"};
my $response=\%{$result->{’soap:Body’}{’n:getEventsResponse’}{’n:Result’}};
$events_hash{’errorCode’}               =$response->{’errorCode’}{’content’};
$events_hash{’timeStamp’}               =$response->{’header’}->{’timestamp’}{’content’};
$events_hash{’sessionToken’}            =$response->{’header’}->{’sessionToken’}{’content’};

        my $type=substr($response->{’eventItems’}{’n2:BFEvent’},0,4);   if ( $type eq "HASH" )
        {
        my $eventId=$response->{’eventItems’}{’n2:BFEvent’}{’eventId’}{’content’};

        $events_hash{events}{$eventId}{orderIndex}=$response->{’eventItems’}{’n2:BFEvent’}{orderIndex}{’content’};
        $events_hash{events}{$eventId}{eventName}=$response->{’eventItems’}{’n2:BFEvent’}{eventName}{’content’};
        $events_hash{events}{$eventId}{timezone}=$response->{’eventItems’}{’n2:BFEvent’}{timezone}{’content’};
        $events_hash{events}{$eventId}{startTime}=$response->{’eventItems’}{’n2:BFEvent’}{startTime}{’content’};
        $events_hash{events}{$eventId}{menuLevel}=$response->{’eventItems’}{’n2:BFEvent’}{menuLevel}{’content’};
        $events_hash{events}{$eventId}{eventTypeId}=$response->{’eventItems’}{’n2:BFEvent’}{eventTypeId}{’content’};
        }
        if ( $type eq "ARRA" )
        {
        foreach my $s (@{$response->{’eventItems’}{’n2:BFEvent’}})
                {
                $events_hash{events}{$s->{eventId}{’content’}}{orderIndex}=$s->{orderIndex}{’content’};
                $events_hash{events}{$s->{eventId}{’content’}}{eventName}=$s->{eventName}{’content’};
                $events_hash{events}{$s->{eventId}{’content’}}{timezone}=$s->{timezone}{’content’};
                $events_hash{events}{$s->{eventId}{’content’}}{startTime}=$s->{startTime}{’content’};
                $events_hash{events}{$s->{eventId}{’content’}}{menuLevel}=$s->{menuLevel}{’content’};
                $events_hash{events}{$s->{eventId}{’content’}}{eventTypeId}=$s->{eventTypeId}{’content’};
                }
        }

        return %events_hash;
}


sub get_events2
        {
  my ($sessionToken,$eventParentId)=@_;
        my %events_hash;
  my $xml=’<SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
        xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
        <SOAP-ENV:Body>
                <m:getEvents xmlns:m="http://www.betfair.com/publicapi/v3/BFGlobalService/">
                        <m:request>
                                <header>
                                        <clientStamp>0</clientStamp>
                                        <sessionToken>’.$sessionToken.’</sessionToken>
                                </header>
                                <eventParentId>’.$eventParentId.’</eventParentId>
                        </m:request>
                </m:getEvents>
        </SOAP-ENV:Body>
</SOAP-ENV:Envelope>’;

my $userAgent = LWP::UserAgent->new();
my $request = HTTP::Request->new(POST => ’https://api.betfair.com/global/v3/BFGlobalService’);
$request->header(SOAPAction => ’"https://api.betfair.com/global/v3/BFGlobalService"’);
$request->content($xml);
$request->content_type("text/xml; charset=utf-8");

my $resp = $userAgent->request($request);
#print Dumper(\$resp);
my $content=$resp->content;
my $result;
eval { $result = XMLin($content) };
if ($@) {print Dumper ($content); print "$@\n"; die "get_events call failed"};
my $response=\%{$result->{’soap:Body’}{’n:getEventsResponse’}{’n:Result’}};
$events_hash{’errorCode’}               =$response->{’errorCode’}{’content’};
$events_hash{’timeStamp’}               =$response->{’header’}->{’timestamp’}{’content’};
$events_hash{’sessionToken’}            =$response->{’header’}->{’sessionToken’}{’content’};

my $type=substr($response->{’marketItems’}{’n2:MarketSummary’},0,4);
        if ( $type eq "HASH" )
        {
        my $marketId=$response->{’marketItems’}{’n2:MarketSummary’}{’marketId’}{’content’};
        $events_hash{markets}{$marketId}{timezone}=$response->{’marketItems’}{’n2:MarketSummary’}{timezone}
        {’content’};
        $events_hash{markets}{$marketId}{menuLevel}=$response->{’marketItems’}{’n2:MarketSummary’}{menuLevel}
        {’content’};
        $events_hash{markets}{$marketId}{marketName}=$response->{’marketItems’}{’n2:MarketSummary’}{marketName}
        {’content’};

        $events_hash{markets}{$marketId}{orderIndex}=$response->{’marketItems’}{’n2:MarketSummary’}{orderIndex}
        {’content’};
        $events_hash{markets}{$marketId}{marketType}=$response->{’marketItems’}{’n2:MarketSummary’}{marketType}
        {’content’};
        $events_hash{markets}{$marketId}{startTime}=$response->{’marketItems’}{’n2:MarketSummary’}{startTime}
        {’content’};
        $events_hash{markets}{$marketId}{eventTypeId}=$response->{’marketItems’}{’n2:MarketSummary’}{eventTypeId}
        {’content’};
        }
        if ( $type eq "ARRA" )
        {
        foreach my $s (@{$response->{’marketItems’}{’n2:MarketSummary’}})
                {
                $events_hash{markets}{$s->{marketId}{’content’}}{timezone}=$s->{timezone}{’content’};
                $events_hash{markets}{$s->{marketId}{’content’}}{menuLevel}=$s->{menuLevel}{’content’};
                $events_hash{markets}{$s->{marketId}{’content’}}{marketName}=$s->{marketName}{’content’};
                $events_hash{markets}{$s->{marketId}{’content’}}{orderIndex}=$s->{orderIndex}{’content’};
                $events_hash{markets}{$s->{marketId}{’content’}}{marketType}=$s->{marketType}{’content’};

                $events_hash{markets}{$s->{marketId}{’content’}}{startTime}=$s->{startTime}{’content’};
                $events_hash{markets}{$s->{marketId}{’content’}}{eventTypeId}=$s->{eventTypeId}{’content’};
                }
        }

        return %events_hash;
}




sub get_markets
        {
  my ($sessionToken,$marketId)=@_;
        my %market_hash= ();
                my %names_hash = ();
                my $runnerId = ();
                my $runner_name = ();
  my $xml=’<SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
        xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
        <SOAP-ENV:Body>
                <m:getMarket xmlns:m="http://www.betfair.com/publicapi/v5/BFExchangeService/">
                        <m:request>
                                <header>
                                        <clientStamp>0</clientStamp>
                                        <sessionToken>’.$sessionToken.’</sessionToken>
                                </header>
                                <marketId>’.$marketId.’</marketId>
                        </m:request>
                </m:getMarket>
        </SOAP-ENV:Body>
```

```
</SOAP-ENV:Envelope>';
my $userAgent = LWP::UserAgent->new();
my $request = HTTP::Request->new(POST => 'https://api-au.betfair.com/exchange/v5/BFExchangeService');
$request->header(SOAPAction => '"https://api-au.betfair.com/exchange/v5/BFExchangeService"');
$request->content($xml);
$request->content_type("text/xml; charset=utf-8");
my $resp = $userAgent->request($request);
my $content=$resp->content;
my $result;
eval { $result = XMLin($content) };
if ($@) {print Dumper ($content); print "$@\n"; die "get_markets call failed"};
  my $response=\%{$result->{'soap:Body'}{'n:getMarketResponse'}{'n:Result'}};
  $market_hash{'errorCode'}            =$response->{'errorCode'}{'content'};
        $market_hash{'timeStamp'}               =$response->{'header'}->{'timestamp'}{'content'};
        $market_hash{'sessionToken'}            =$response->{'header'}->{'sessionToken'}{'content'};
                $market_hash{marketTime}        =$response->{'market'}{'marketTime'}{'content'};
                $market_hash{BSP}               =$response->{'market'}{'bspMarket'}{'content'};
                $market_hash{canTurnInplay}     =$response->{'market'}{'canTurnInplay'}{'content'};
                $market_hash{marketType}        =$response->{'market'}{'marketType'}{'content'};
                $market_hash{marketSuspendTime} =$response->{'market'}{'marketSuspendTime'}{'content'};
                $market_hash{numberOfWinners}   =$response->{'market'}{'numberOfWinners'}{'content'};
                $market_hash{eventTypeId}       =$response->{'market'}{'eventTypeId'}{'content'};
                $market_hash{countryISO3}       =$response->{'market'}{'countryISO3'}{'content'};
                $market_hash{timezone}          =$response->{'market'}{'timezone'}{'content'};
                $market_hash{discountAllowed}   =$response->{'market'}{'discountAllowed'}{'content'};
                $market_hash{menuPath}          =$response->{'market'}{'menuPath'}{'content'};
                $market_hash{name}              =$response->{'market'}{'name'}{'content'};
                $market_hash{marketDisplayTime} =$response->{'market'}{'marketDisplayTime'}{'content'};
                $market_hash{marketStatus}      =$response->{'market'}{'marketStatus'}{'content'};
                $market_hash{marketBaseRate}    =$response->{'market'}{'marketBaseRate'}{'content'};
                $market_hash{parentEventId}     =$response->{'market'}{'parentEventId'}{'content'};
                $market_hash{runnersMayBeAdded} =$response->{'market'}{'runnersMayBeAdded'}{'content'};
                $market_hash{marketDescription} =$response->{'market'}{'marketDescription'}{'content'};
                $market_hash{lastRefresh}       =$response->{'market'}{'lastRefresh'}{'content'};
        foreach my $s (@{$response->{'market'}->{'runners'}{'n2:Runner'}})
                {
                $market_hash{runners}{$s->{selectionId}{'content'}}{asianLineId}=$s->{asianLineId}{'content'};
                $market_hash{runners}{$s->{selectionId}{'content'}}{name}=$s->{name}{'content'};
                $market_hash{runners}{$s->{selectionId}{'content'}}{handicap}=$s->{handicap}{'content'};
        #To extract most frequently used hashes, create 2 arrays and take a reference to each

                $runnerId = $s->{selectionId}{'content'};
                $runner_name = $s->{name}{’content’};
                $names_hash{$runner_name} = $runnerId;
                }
        my @hashes = (\%names_hash, \%market_hash);
        return @hashes;

        #return %names_hash;
        }

sub get_market_prices_compressed
        {
  my ($sessionToken,$marketId)=@_;
  my $xml=’<SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
        xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
        <SOAP-ENV:Body>
                <m:getMarketPricesCompressed xmlns:m="http://www.betfair.com/publicapi/v5/BFExchangeService/">
                        <m:request>
                                <header>
                                        <clientStamp>0</clientStamp>
                                        <sessionToken>’.$sessionToken.’</sessionToken>
                                </header>
                                <marketId>’.$marketId.’</marketId>
                        </m:request>
                </m:getMarketPricesCompressed>
        </SOAP-ENV:Body>
</SOAP-ENV:Envelope>’;
my $userAgent = LWP::UserAgent->new();
my $request = HTTP::Request->new(POST => ’https://api-au.betfair.com/exchange/v5/BFExchangeService’);
$request->header(SOAPAction => ’"https://api-au.betfair.com/exchange/v5/BFExchangeService"’);
$request->content($xml);
$request->content_type("text/xml; charset=utf-8");
my $resp = $userAgent->request($request);
my $content=$resp->content;
#print Dumper(\$content);
my $result;
eval { $result = XMLin($content) };
if ($@) {print Dumper ($content); print "$@\n"; die "get_market_prices_compressed failed to retrieve valid XML"};
  my %prices_hash;
          my $response=\%{$result->{’soap:Body’}{’n:getMarketPricesCompressedResponse’}{’n:Result’}};
        $prices_hash{’errorCode’}               =$response->{’errorCode’}{’content’};
        $prices_hash{’timeStamp’}               =$response->{’header’}->{’timestamp’}{’content’};
```

```
$prices_hash{’sessionToken’}            =$response->{’header’}->{’sessionToken’}{’content’};
my $runner_price=$result->{’soap:Body’}{’n:getMarketPricesCompressedResponse’}{’n:Result’}{’marketPrices’}
{’content’};
$runner_price=~ s/\\:/colon/g;
#print $runner_price,"\n";
$prices_hash{’noRunners’}=0;
my @price_split=split(/:/,$runner_price);
my $size=@price_split;
#print $size,"\n";
my @market_attributes=split(/\~/,$price_split[0]);
$prices_hash{’marketId’}                =$market_attributes[0];
$prices_hash{’currencyCode’}    =$market_attributes[1];
$prices_hash{’delay’}           =$market_attributes[3];
$prices_hash{’marketStatus’}    =$market_attributes[2];
$prices_hash{’marketInfo’}      =$market_attributes[5];
$prices_hash{’numberOfWinners’} =$market_attributes[4];
$prices_hash{’lastRefresh’}     =$market_attributes[8];
$prices_hash{’IsBSP’}           =$market_attributes[10];        #returns a value of Y or N
for ( my $t =1 ; $t < $size ; $t++)
        {
        my @prices =split(/\|/,$price_split[$t]);
        my @header =split(/\~/,$prices[0]);
        my @back   =split(/\~/,$prices[1]);
        my @lay    =split(/\~/,$prices[2]);
        $prices_hash{’noRunners’}++;
        my $runnerId = $header[0];
        $prices_hash{’prices’}{$runnerId}{’orderIndex’}         =$header[1];
        $prices_hash{’prices’}{$runnerId}{’totalAmountMatched’} =$header[2];
        $prices_hash{’prices’}{$runnerId}{’lastPriceMatched’}   =$header[3];
        $prices_hash{’prices’}{$runnerId}{’asianHandicap’}      =$header[4];
        $prices_hash{’prices’}{$runnerId}{’reductionFactor’}    =$header[5];
        $prices_hash{’prices’}{$runnerId}{’vacantTrap’}         =$header[6];
        $prices_hash{’prices’}{$runnerId}{’farBSP’}             =$header[7];
        $prices_hash{’prices’}{$runnerId}{’nearBSP’}            =$header[8];
        $prices_hash{’prices’}{$runnerId}{’actualBSP’}          =$header[9];
        $prices_hash{’prices’}{$runnerId}{’backDepth’}=0;
        my $back_size=@back;
        my $lay_size =@lay;
        if ( $back_size == 0 )
                {
                $prices_hash{’prices’}{$runnerId}{’backDepth’}=0;
                }
```

```
if ( $back_size == 4 )
        {
        $prices_hash{'prices'}{$runnerId}{'backDepth'}=1;
        $prices_hash{'prices'}{$runnerId}{'back'}{'1'}{'price'}         =$back[0];
        $prices_hash{'prices'}{$runnerId}{'back'}{'1'}{'amountAvailable'}=$back[1];
        }
if ( $back_size == 8 )
        {
        $prices_hash{'prices'}{$runnerId}{'backDepth'}=2;
        $prices_hash{'prices'}{$runnerId}{'back'}{'1'}{'price'}         =$back[0];
        $prices_hash{'prices'}{$runnerId}{'back'}{'1'}{'amountAvailable'}=$back[1];
        $prices_hash{'prices'}{$runnerId}{'back'}{'2'}{'price'}         =$back[4];
        $prices_hash{'prices'}{$runnerId}{'back'}{'2'}{'amountAvailable'}=$back[5];
        }
if ( $back_size == 12 )
        {
        $prices_hash{'prices'}{$runnerId}{'backDepth'}=3;
        $prices_hash{'prices'}{$runnerId}{'back'}{'1'}{'price'}         =$back[0];
        $prices_hash{'prices'}{$runnerId}{'back'}{'1'}{'amountAvailable'}=$back[1];
        $prices_hash{'prices'}{$runnerId}{'back'}{'2'}{'price'}         =$back[4];
        $prices_hash{'prices'}{$runnerId}{'back'}{'2'}{'amountAvailable'}=$back[5];
        $prices_hash{'prices'}{$runnerId}{'back'}{'3'}{'price'}         =$back[8];
        $prices_hash{'prices'}{$runnerId}{'back'}{'3'}{'amountAvailable'}=$back[9];
        }
if ( $lay_size == 0 )
        {
        $prices_hash{'prices'}{$runnerId}{'layDepth'}=0;
        }
if ( $lay_size == 4 )
        {
        $prices_hash{'prices'}{$runnerId}{'layDepth'}=1;
        $prices_hash{'prices'}{$runnerId}{'lay'}{'1'}{'price'}          =$lay[0];
        $prices_hash{'prices'}{$runnerId}{'lay'}{'1'}{'amountAvailable'}=$lay[1];
        }
if ( $lay_size == 8 )
        {
        $prices_hash{'prices'}{$runnerId}{'layDepth'}=2;
        $prices_hash{'prices'}{$runnerId}{'lay'}{'1'}{'price'}          =$lay[0];
        $prices_hash{'prices'}{$runnerId}{'lay'}{'1'}{'amountAvailable'}=$lay[1];
        $prices_hash{'prices'}{$runnerId}{'lay'}{'2'}{'price'}          =$lay[4];
        $prices_hash{'prices'}{$runnerId}{'lay'}{'2'}{'amountAvailable'}=$lay[5];
        }
```

```
                if ( $lay_size == 12 )
                        {
                        $prices_hash{’prices’}{$runnerId}{’layDepth’}=3;
                        $prices_hash{’prices’}{$runnerId}{’lay’}{’1’}{’price’}          =$lay[0];
                        $prices_hash{’prices’}{$runnerId}{’lay’}{’1’}{’amountAvailable’}=$lay[1];
                        $prices_hash{’prices’}{$runnerId}{’lay’}{’2’}{’price’}          =$lay[4];
                        $prices_hash{’prices’}{$runnerId}{’lay’}{’2’}{’amountAvailable’}=$lay[5];
                        $prices_hash{’prices’}{$runnerId}{’lay’}{’3’}{’price’}          =$lay[8];
                        $prices_hash{’prices’}{$runnerId}{’lay’}{’3’}{’amountAvailable’}=$lay[9];
                        }
                }
        my $prices = ’prices’;
        return %prices_hash;
}
```

Glossary

50 Metre Arc A line drawn in the shape of an arc to signify 50 metres from the “Goal Line”.

50 Metre Penalty A distance penalty usually awarded in addition to a “Free Kick” or after a “Mark” with “The Mark” being advanced 50 metres towards the centre of the “Goal Line”.

Ball-Up A situation where a “Field Umpire” restarts play by bouncing the ball into the ground or propelling the ball into the air for a “Ruck” contest between the two opposing Ruckmen.

Behind A “Behind” (one point) is awarded when the football passes over the behind line; or the football strikes any part of the goal post; or prior to the football passing over the behind or goal line it is touched by another player; or the defending team deliberately plays the ball over the behind or goal line (“Rushed Behind”).

Behind Line A white line marked between each “Goal Post” and “Behind Post”.

Behind Post The posts either side of the “Goal Posts” at either end of the ground. These posts are significantly shorter than the “Goal Posts”.

Blood Rule If a player is found to be actively bleeding by a “Field Umpire” they must head immediately to the “Interchange Bench” and are not permitted back on the “Playing Field” until the cause of such bleeding has been abated and any blood-stained clothing has been removed.

Bounce Any player moving whilst in possession of the ball must bounce or touch the football on the ground at least every 15 metres regardless of the direction they are running. Failure to do so results in a “Free Kick” to the opposition.

Boundary Line The white line drawn on the playing surface to identify the “Playing Surface”.

Boundary Umpires They have the responsibility of determining when the ball is “Out of Bounds” or “Out of Bounds on the Full”.

Centre Bounce See “Ball-Up”. This occurs at the beginning of each quarter and after a goal has been scored.

Disposal(s) A player releasing the ball from their "Possession". Also used as a common measure to describe the number "Kicks" and "handballs" of a team or player.

Draft A structured drafting system for the recruitment of players to AFL teams. The current drafting system consists of four distinctive phases: The National Draft, Rookie Draft, Pre-Season Draft and the Trading Period. The order of the selection process of each draft is based on the reverse order of the "Ladder"; the team which finishes last receives the first pick, the team which finishes second last receives the second pick and so on.

Field Umpires A game usually consists of three "Field Umpires". The responsibilities of a field umpire include the "Ball-Up" and "Centre Bounce"; awarding penalties "Free Kick" and "50-metre penalty"; reporting players ("Report").

Free Kick "Possession" is awarded to player for breaking a rule. This player then receives an unimpeded "Kick" over "The Mark", or if they choose they can "Play-On".

Goal Six points is awarded to the attacking team when the football is kicked completely over the goal line regardless of whether or not it bounces, provided it has not touched an opposition player in any way.

Goal Line A white line marked between the two “Goal Posts”.

Goal Posts The middle two posts at either end of the ground. These posts are significantly larger than the “Behind Posts”.

Goal Umpires A game consists of two Goal Umpires at either end of the ground. They are the official score keepers and they award “Behinds” and “Goals” and work in tandem with the boundary umpires when the ball goes “Out of Bounds” or “Out of Bounds on the Full” near the “Behind Post”.

Grand Final The Grand Final is the ultimate match in AFL, the winner of which is declared the “Premiership” team.

Handball The act of holding the ball in one hand and punching it with a clenched fist of the other.

Hit-Out A statistical measure generally performed by the “Ruckmen” player which involves striking the ball with one hand after the ball has been brought back into play after a “Ball-Up”, “Centre-Bounce” or “Throw-In”.

Home and Away Season The regular season which currently consists of 22 matches.

Inside 50 A common statistical measure which is recorded to the attacking team when the ball moves inside their attacking “50 Metre Arc”.

Interchange Bench Designated area which is marked on the “Boundary Line” where players may enter and depart the “Playing Surface”.

Kick or Kicking A type of “Disposal” defined as when a players leg (below the knee) comes into contact with the football.

Kick-In A “Kick” which must occur from the “Goal Square” by the opposition immediately after a behind has been scored.

Knock-On The act of knocking the ball while a player is not in “Possession” of the football. This typically occurs in congested situations or when a player is under pressure to dispose of the ball.

Ladder A ranking system used by the AFL to determine which teams qualify for the finals series. Teams are ranked based on their “Premiership Points” and any teams with an equal number of “Premiership Points” are ranked based on their “Percentage”.

Mark A player who catches the ball immediately after it has been kicked by another player which is deemed to have traveled at least 15 metres; and not touched the ground or been touched by another player during its journey. This player then receives an unimpeded “Kick” over “The Mark”, or if they choose they can “Play-On”.

Out of Bounds When the ball completely passes over the boundary line or strikes the “Behind Post” after touching the ground or another player. When this occurs a “Boundary Umpire” performs a “Throw-In”.

Out of Bounds on the Full When the ball completely passes over the boundary line or strikes the “Behind Post” without touching the ground or another player. When this occurs the nearest opposition player receives an unimpeded free “Kick”.

Percentage A statistical measure which expresses the total number of points a team has scored as a percentage of total points scored against that team during the entire "Home and Away Season".

Play-On A verbal and visual instruction used by a "Field Umpire" to signal the football is in-play. For example, a player has attempted to dispose of the football other than the direct line over "The Mark" after a "Mark" or "Free Kick" has been awarded.

Playing Surface The area enclosed by the "Boundary Line" with which the match is played.

Possession A literal interpretation meaning the player physically holds the football. Also used to refer to "Disposals".

Premiership The team that wins the "Grand Final" is known as the "Premiership" team.

Premiership Points Teams are awarded four Premierhip Points for a win, two points for a draw and zero points for a loss during the "Home and Away Season".

Rebound 50 A common statistical measure which is recorded to the defensive team when the ball moves outside their defensive "50 metre arc".

Ruck Contest A contest usually comprising of two “Ruckmen” from a opposing teams whose objective is to win the “Hitout” to the advantage of a team mate or gain a significant amount of ground by knocking the ball as far as they can to their respective goals.

Ruckmen A position on a team whose primary objective is to win the “Ruck Contest”. These positions are typically filled by players which are tall, agile and and have a good vertical leap.

Shepherd A statistical measure used to describe a player that uses their body or arm to hinder the movement of an opposition player when that player is within 5 metres of the ball.

Suspension Players who are forced to miss matches due to disciplinary action.

Tackle The act of holding the player in “Possession” of the football. A legal tackle must be performed above the knees and below the waist. A player being tackled must immediately release the football via “Kick” or “Handball” or risk giving away a “Free Kick” to the opposition.

The Mark The position on the playing field where a player on the field stands immediately after a “Mark” or “Free Kick” has been awarded to the opposition.

Throw-In A situation where a “Boundary Umpire” restarts play by throwing the ball from the “Boundary Line” into the air for a “Ruck” contest between the two opposing Ruckmen.

Umpires Each game consists of three “Field Umpires”, four “boundary Umpires” and two “Goal Umpires”.